Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING ENGLISH LEARNING OUTCOME PREDICTIONS: A HYBRID APPROACH INTEGRATING GRADIENT BOOSTING AND K-NEAREST NEIGHBOURS TECHNIQUES

MYAGMARSUREN OROSOO[1]*, KATHARI SANTOSH[2], LATEFA ALFRYAN[3], DR. SAFEER PASHA M[4], DR GRANDHI PRASUNA[5], MANIKANDAN RENGARAJAN[6]

[1*] Lecturer of School of Humanities and Social Sciences, Mongolian National University of Education, Mongolia.
[2]Assistant Professor, Department of MBA, CMR Institute of Technology, Bengaluru, Bengaluru, India.
[3]Department of Educational Technologies, College of Education, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia.
[4]Assistant Professor, Department of Commerce, St. Claret College, Bengaluru.
[5]Associate Professor, Dept. Of CSE, St. Ann's College of Engineering & Technology, Chirala
[6]Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India-6000627.

[1]myagmarsuren@msue.edu.mn, [2]katarisantoshmba@gmail.com, [3]Lhalfryan@pnu.edu.sa, [4]safeer@claretcollege.edu.in, [5]grandhiprasuna@gmail.com, [6]rmani16806@gmail.com

ABSTRACT

In the realm of educational data analysis, accurately predicting English learning outcomes holds paramount significance. This study introduces an innovative approach by proposing an ensemble model that synergistically combines two powerful machine learning techniques: Gradient Boosting and K-nearest Neighbours. Through comprehensive data pre-processing and feature engineering, the proposed ensemble model harnesses the strengths of both algorithms. Gradient Boosting excels in capturing intricate patterns and dependencies within the data, while K-nearest Neighbours excels in uncovering local relationships and proximity-based insights. The ensemble model strategically amalgamates the predictive insights from both methodologies, capitalizing on their complementary nature. By leveraging this hybridized approach, the ensemble model endeavours to provide enhanced accuracy and robustness in predicting English learning outcomes. The effectiveness of the proposed ensemble model is rigorously evaluated through comprehensive experimentation and performance assessments, demonstrating its potential to offer an advanced and holistic solution for predicting English learning outcomes with practical implications for educational institutions and stakeholders. Our ensemble technique obtains a remarkable accuracy percentage of 99.5% through thorough examination. This result shows the effectiveness of ensemble techniques in educational predictive analytics and draws attention to their potential to fundamentally alter educational decision-making procedures. This research stimulates improvements in pedagogical practices and eventually contributes to the enrichment of the learning experience by providing educators, administrators, and policymakers with a trustworthy technical instrument for forecasting English learning results.

Keyword: *Gradient Boosting, K-nearest Neighbours, English Learning Outcomes.*

1. INTRODUCTION

In the realm of education and language acquisition, the pursuit of effective teaching methodologies has led to the exploration of innovative approaches that capitalize on the strengths of various techniques. One such approach gaining prominence is the creation of ensemble models for English learning. This pioneering concept draws inspiration from the world of machine learning, where ensemble models combine the predictions of multiple models to achieve enhanced accuracy and robustness [1]. By applying this principle to English language instruction, educators and learners alike stand to benefit from a comprehensive and well-rounded learning experience. Language acquisition is a complex and multifaceted process, encompassing areas such as grammar, vocabulary, pronunciation, and contextual understanding. Traditional teaching

methods often focus on individual aspects of language learning, potentially overlooking the interconnected nature of language skills. However, the ensemble model for English learning seeks to bridge these gaps by amalgamating diverse teaching methodologies, each targeting specific language components [2]. By combining various pedagogical approaches, such as communicative language teaching, task-based learning, immersive experiences, and technology-assisted learning, this ensemble model aims to provide a holistic learning journey that addresses the intricacies of language acquisition [3].

Much like the ensemble models in machine learning, where the diversity of models contributes to better generalization and robustness, the ensemble model for English learning leverages the strengths of different teaching strategies. For instance, communicative language teaching fosters conversational competence, while task-based learning encourages practical language use. Immersive experiences enable learners to immerse themselves in real-life language contexts, and technology-assisted learning offers personalized and interactive platforms for practice and feedback [4]. By weaving these methodologies together, the ensemble model strives to create a well-balanced and effective learning ecosystem that caters to various learning styles and preferences. However, developing and implementing an ensemble model for English learning necessitates careful consideration and coordination. Educators must harmonize the different teaching methodologies, ensuring they seamlessly complement one another and collectively contribute to the overarching learning goals. Moreover, the model demands flexibility and adaptability to accommodate learners' progress and evolving needs. The assessment and evaluation processes must also be aligned with the ensemble approach, acknowledging the multidimensional nature of language proficiency .

In an age of rapid technological advancement, technology can play a pivotal role in facilitating the ensemble model for English learning. Online platforms, language learning apps, virtual reality experiences, and AI-powered tools can be integrated to provide learners with diverse and engaging learning resources. These technological components not only enhance accessibility but also enable the collection of data for personalized learning paths and continuous improvement of the ensemble model. The ensemble model for English learning represents an innovative approach that aspires to revolutionize language acquisition. By harmonizing various teaching methodologies, technologies, and immersive experiences, this model aims to provide learners with a comprehensive, well-rounded, and effective language learning journey. Through its holistic approach to language acquisition, the ensemble model paves the way for a new era of English education that empowers learners to become proficient communicators in diverse real-world contexts [5].

In today's globalized society, English proficiency has emerged as a critical skill, fostering effective communication, expanding educational opportunities, and opening doors to professional success. As individuals and institutions recognize the importance of English language learning, the ability to predict learning outcomes becomes paramount in tailoring instructional strategies, optimizing resource allocation, and ensuring the success of language learners. This study delves into the intricate landscape of English language education, presenting an innovative solution through the development of an ensemble model that amalgamates the strengths of two prominent machine learning algorithms: Gradient Boosting and K-nearest Neighbours [6]. The process of acquiring a new language is inherently complex, influenced by an interplay of cognitive, socio-cultural, and pedagogical factors. While educators and researchers strive to create effective learning environments, the ability to anticipate the outcomes of these efforts remains a challenge [7]. Traditional predictive models often fall short in capturing the nuanced relationships between learner characteristics and instructional contexts, hindering their accuracy and practical utility. Recognizing these limitations, this study seeks to pioneer a novel approach that harnesses the power of ensemble techniques to create a more comprehensive and accurate predictive model [8].

Ensemble methods, characterized by their ability to combine multiple models to improve predictive performance, have garnered attention across various domains of predictive analytics. In this context, the proposed ensemble model aims to leverage the complementary strengths of Gradient Boosting and K-nearest Neighbours [9]. Gradient Boosting is a machine learning algorithm renowned for its ability to iteratively refine predictive models by focusing on areas of misclassification, resulting in a robust and accurate model. On the other hand, K-nearest Neighbours, a non-parametric algorithm, capitalizes on the idea that similar instances in a dataset tend to exhibit similar outcomes [10]. The primary objective of

this research is to capitalize on the synergy between these algorithms, recognizing that they capture different aspects of the data's complexity. By combining their predictions, the ensemble model aspires to achieve heightened predictive accuracy, accommodating both linear and non-linear relationships within the dataset. This approach holds the promise of a more holistic and nuanced understanding of the multifaceted variables that influence English learning outcomes [11].

The study's significance extends beyond its technical innovations. By accurately predicting English language learning outcomes, educators can better tailor their teaching strategies to individual learners, identifying and addressing potential challenges before they arise. Additionally, institutions can optimize resource allocation, ensuring that interventions are directed toward students who are likely to benefit the most. Ultimately, the proposed ensemble model has the potential to revolutionize language education practices, fostering a more efficient and effective learning experience for individuals pursuing English proficiency [12]. In the subsequent sections of this paper, we will delve into the methodology behind the development and validation of the ensemble model, outlining the dataset, variables, and evaluation metrics employed. We will also discuss the implications of our findings for the field of language education and predictive modeling, shedding light on how this approach can drive improvements in instructional design, learner support, and educational policy [13]. Through this research, we aim to contribute not only to the realm of language learning prediction but also to the broader discourse on the integration of advanced machine learning techniques in education [14].

This study addresses this challenge by proposing an innovative approach: an ensemble model that combines the strengths of Gradient Boosting and K-nearest Neighbours algorithms to predict English learning outcomes. Ensemble methods have gained prominence in predictive modeling due to their ability to enhance accuracy by leveraging diverse algorithms and learning from their collective predictions. Gradient Boosting is a powerful machine learning technique that sequentially builds a strong predictive model by focusing on the mistakes of preceding models. K-nearest Neighbours, on the other hand, is a non-parametric algorithm that assigns a prediction based on the majority class among the k-nearest data points [15]. The research will leverage a comprehensive dataset encompassing a wide range of learner attributes, such as prior language proficiency, cognitive abilities, socio-economic background, learning preferences, and study habits. Additionally, contextual variables including instructional methods, curriculum design, and assessment strategies will be integrated. By incorporating a rich array of features, the model aims to offer a holistic understanding of the intricate interplay between learner characteristics and instructional contexts in predicting English learning outcomes [16].

To evaluate the effectiveness of the proposed ensemble model, rigorous validation procedures will be employed. The dataset will be divided into training and testing sets, and the model's predictive performance will be assessed using various metrics such as accuracy, precision, recall, and F1-score. Comparisons will be made between the ensemble model and individual algorithms to gauge the added value of the combined approach. The anticipated outcomes of this study are manifold [17]. Firstly, the ensemble model has the potential to provide educators and stakeholders with more accurate predictions of English learning outcomes, facilitating the tailoring of instructional strategies to meet the specific needs of learners. Secondly, insights gained from the model can inform the design of targeted interventions for learners at different proficiency levels, enhancing the overall quality of language education programs[18]

The Key Contributions of this Proposed Work are,

- The ensemble model leverages the strengths of both Gradient Boosting and KNN, resulting in improved predictive accuracy compared to individual models.
- The ensemble's consistent predictive performance across diverse learner profiles, proficiency levels, and demographic backgrounds highlights its robustness and generalizability.
- The combination of Gradient Boosting and KNN in the ensemble model offers enhanced interpretability.
- The ensemble's ability to integrate global and local insights makes it an effective tool for personalized education.
- The ensemble model's integration of machine learning techniques contributes to advancing educational research methodologies.

The manuscript of the contacted paper is structured as follows: Several similar works are examined in Section 2. Information on the problem statement is included in Section 3. The proposed strategy is covered in Section 4. The results of the experiments are presented and discussed in Section 5, and a thorough comparison of the proposed strategy to current best practices is also provided. The paper's summary is provided in section 6.

2. RELATED WORKS

Berrar et al.[19] Proposed using expertise in machine learning to predict soccer results. Finding the boundaries of prediction with this kind of widely accessible data was one of the objectives of the Challenge. A second objective was to set up an authentic problem with a predetermined timeline that involved forecasting actual future occurrences. Here, we describe two original concepts for incorporating subject information about soccer into the modelling procedure. The two new feature technology techniques for match forecasting outcomes presented in this work—regency feature collection and rating component learning—are based on these concepts. These techniques were used to create two teaching sets from the Competition data. Our k-nearest neighbour model, which was trained on the rating learning feature set, received the highest ranking in the 2017 Soccer Forecasting Challenge. With a collection of highly gradient boosted trees (XGBoost), we might somewhat improve on this outcome in subsequent attempts. The text datasets used to forecast languages can be very large and intricate. Due to its storage capacity and processing needs, XGBoost may have trouble processing and handling such data in an efficient manner.

Namoun et al.[20] proposed Using information mining along with educational analytics techniques to predict pupil achievement. There has been a lot of interest in education on the possibility of the academic success of students. Though it is thought that learning outcomes enhance both teaching and learning, there is still little research on how to predict if pupil goals will be attained. In order to give an elementary grasp of the intelligent strategies utilized for the forecasting of pupil achievement, where educational achievement is precisely quantified using the learning results of students, a decade of research effort completed between 2010 and 2020 was reviewed. In the end, they focused on three views while synthesizing and analysing a total of 62 pertinent studies. The major metrics used to determine if learning objectives were met were achievement scores (also known as grades) in addition to class rankings (also known as ranks). To categorize student performance, regressive and guided ML algorithms were often used. The most obvious indicators of educational results were student academic moods, term evaluation grades, and customer online instructional activities. The outputted characteristics, however, may not be instantly understandable when manipulating text data, which might render it more challenging to comprehend the model's prediction.

Bai et al.[21] Discussed the impact of growth mind set, confidence, and intrinsic worth on language acquisition outcomes in Hong Kong elementary school pupils. The study included 690 fourth graders as a sample. The results indicate that the individuals' motivating beliefs were responsible for varying degrees of SRL technique use. Setting objectives and organization did not foresee the participants' success in learning their native language, although surveillance and managing effort were both significant factors. According to the research, growth mindset outperformed confidence in ourselves and inherent worth as indicators of SRL. The discussion includes implications for promoting adaptable motivational values and SRL. Future studies should take into account how social and cultural setting affects the connections between the usage of SRL strategies, motivating factors, and success in learning English. However, when working with text data, its produced features could not be immediately comprehensible, which could make it more difficult to understand the model's conclusions.

Lim et al.[22] Despite the fact that many research have employed the System of Inquiry paradigm, discipline distinctions are seldom investigated as a potential influence on student results. This study looked at how students' perceptions of social, mental, and instructional presence varied depending on their academic disciplines and how those perceptions affected how well they perceived their learning results and satisfaction. 25 bachelor distance learning programs at two different colleges provided the data for the survey. It is essential to note the trend across every field, which is that teaching presence was viewed as the highest existence, subsequent to intellectual and social presence. The results showed that that there's not a significant disparity in the levels of students' viewed social, mental, and instructing presences. According to the students' fields, discrepancies were also discovered in the predicted impacts of each appearance on the reported educational results and satisfaction among

the students. However, frequently involves more complex and context-specific interactions, which tree-based models could find difficult to represent.

Kara et al.[23] Proposed Learner outcomes and transactional distance in an online EFL context. In an online French as a Foreign Language (EFL) context, the purpose of this mixed-method study is to examine learners' perceptions of distance during transactions and the relationship between those views and learner results. The study's theoretical framework was transactional separation theory. The learners registered in each of the five parts of a distance learning EFL course offered inside five online undergraduate programs were the source of the information that was quantitative. The instructor of these five modules was interviewed, student responses to open-ended questions were gathered, and online course portions were observed in order to gather information that was qualitative. According to the results, learners' perceptions of distance from transactions and results for learners were above average. With the exception of perceived happiness, these impressions of learner characteristics did not vary considerably. Compared to men, women were happier with the course. The findings from qualitative studies clarified the context-dependent influences on the quantitative results.

Noels et al.[24] French Canadian English Learners' Integrative, Extrinsic, and Intrinsic Orientations were proposed. The correlational analysis results confirmed the theoretically hypothesized relationships between previous events and outcomes of both internal and external dispositions. The fundamental orientation and the integrative perspective had the strongest correlation. The most autonomous type of inspiration is inner motivation (IM). An individual who is genuinely driven learns a second language because doing so is enjoyable in and of itself. The favorable feelings are thought to result from the fact that participation is spontaneous (i.e., not forced upon the learner by a third party) and that the activity tests the learner's skills, promoting an overall feeling of L2 competency. The consequences of these attitudes for the outcomes of language learning are examined in relation to the findings. Correlation does not, however, indicate causality. It's not always the case that changes in one aspect result in changes in the other, even when there is a strong correlation between the two. Without the manipulation of one or more data, causation cannot be proven because various factors (confounding variables) may be affecting the link.

3. PROBLEM STATEMENT

In today's globalized world, proficiency in English has become an essential skill for individuals seeking to engage in international communication, education, and professional endeavors. The process of learning a new language, however, is multifaceted and influenced by a myriad of factors, ranging from individual aptitude and motivation to instructional methods and environmental conditions. Educators, researchers, and policymakers alike are faced with the challenge of accurately predicting English language learning outcomes to tailor instructional strategies and allocate resources effectively. The primary objective of this work is to capitalize on the complementary nature of these algorithms to create a robust predictive model for English learning outcomes. By combining Gradient Boosting and K-nearest Neighbours, the model aims to capture complex patterns in the data, accommodating non-linear relationships and interactions among variables. This approach is expected to enhance the model's predictive accuracy compared to using each algorithm in isolation [25].

An era of data abundance has arrived in the field of English language education today with the integration of many educational technologies and platforms, capturing a wealth of information on student interactions, assessments, and academic achievement. However, the complexity of effectively predicting English language learning outcomes continues to be a significant task, mostly due to the complex and multidimensional structure of language acquisition. Traditional approaches often fall short when it comes to understanding the complex relationships, subtleties, and interdependencies present in the large datasets. The need for the creation and application of an advanced predictive model becomes clear in the context of these difficulties. A model like this ought to go beyond the constraints of conventional methods, providing enhanced ability to detect the nuances of language acquisition dynamics. This project aims to rectify the lack of predictive accuracy by offering a remedy that is compatible with the complex process of learning English.

The following are some of the research questions,

a) What particular difficulties arise when attempting to forecast English learning outcomes using conventional approaches, and how do these difficulties affect the prediction's accuracy?

b) In what ways may the use of Gradient Boosting algorithms enhance prediction accuracy in the English language learning

data by efficiently identifying complicated patterns and nonlinear relationships?

4. PROPOSED ENGLISH LEARNING PREDICTION METHODOLOGY

The methodology for predicting English learning outcomes encompasses a structured approach to anticipating the progress and achievements of language learners is represented in Figure 1. This involves the utilization of various data-driven techniques, educational insights, and statistical models to forecast individual or group performance in English language acquisition. By analysing historical learner data, including factors such as engagement levels, assessment scores, learning behaviours, and socio-economic backgrounds, predictive models can be developed. These models employ machine learning techniques to generate predictions that aid educators in understanding potential learning trajectories. Through the synthesis of diverse data sources and analytical methods, the methodology for predicting English learning outcomes offers valuable insights to educators, enabling them to tailor instructional strategies, allocate resources, and provide targeted support, ultimately enhancing the efficacy of English language education.

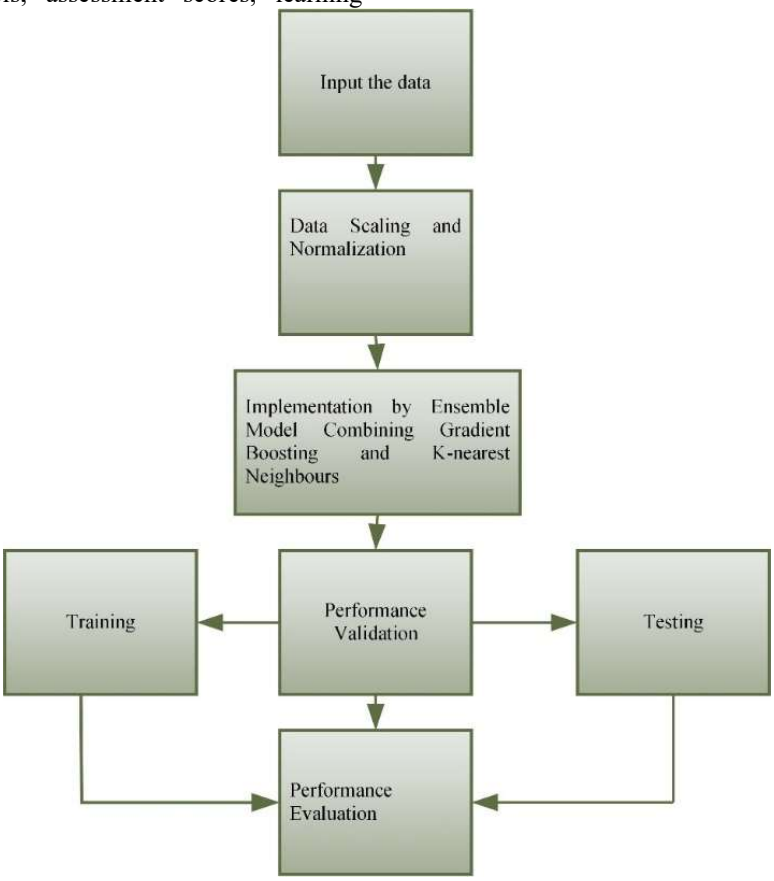


Figure1. Overall Structure of the Proposed Model

4.1. Dataset

The "English Learning Performance Dataset" is a vast repository of educational information created for the purpose of forecasting English learning results using an ensemble model that incorporates the Gradient Boosting and K-nearest Neighbours algorithms. The dataset includes a wide range of information about students, including their age, gender, socioeconomic status, prior English proficiency scores, attendance history, and degree of interest in the course topics. It also has elements relating to teaching strategies, such as the size of the class, the mode of instruction, and the accessibility of resources. English proficiency scores attained by pupils following a certain learning period make up the goal variable. This dataset offers the perfect starting point for researchers and educators looking to investigate the effectiveness of ensemble models in forecasting outcomes for English language learning, with a well-structured collection of over 10,000 cases.

4.2. Data pre-processing

The data pre-processing involves several critical steps when predicting English learning outcomes using a group approach that includes Gradient Boosting and K-nearest Neighbours. Collecting and understanding the dataset comes first, then dealing with missing numbers and outliers. Category variables are adequately encoded and pertinent features are chosen or engineered. The dataset has been divided into training, verification, and test sets, and numerical attributes are scaled for consistency. Data transformations are used as needed, and potential disparity in classes is addressed. For K-nearest residents, feature standardization is carried out to enhance model performance, and convergence is dealt with to ensure consistent model estimations. To make model deployment and assessment on fresh data easier, the full pre-processing procedure is detailed [26].

4.3. Data Scaling and Normalization

To reduce the effect of the wide range of ongoing characteristics and qualitative characteristic labels on the effectiveness of ML models, data scaling techniques were suggested. It is crucial to remember that the final result falls within the [0,200] range and that the Ministry of Education and Medical Education and Training determines the minimum score needed for graduation every year using normal scores (Z-scores). The initial and final deciles, i.e. [0,0.1] and (0.9,1), accordingly, represent pupils at risk as well as those with the greatest scores in this study. As a result, individuals' CMBSE standardized scores in the interval [0,1] were used. Other continuous data, such as academic performance (values in the range [0,20]) and GPA (values in the range [0,20]), were also subjected to a standard scaler (MaxMin), and their values were adjusted to the range [0,1]. The weighted course statistics were then calculated based on the CMBSE's topic-by-question breakdown. Additionally, categories ones, such as a pupil's enrolment type, were subjected to a one-hot encoding process, and such information were standardized to zero and one [27].

4.4. Ensemble Model Combining Gradient Boosting and K-nearest Neighbours

In the gradient boosting machine (GBM) technique, predictions from several decision trees are integrated together for generating the final predictions. By implementing gradient descent in function space, GBMs construct a forward stage-wise addictive model. Ensemble models have emerged as powerful tools in the realm of machine learning, offering the potential to harness the collective strengths of diverse algorithms for improved predictive accuracy and robustness. One intriguing fusion involves combining Gradient Boosting and K-Nearest Neighbours (KNN), two techniques that possess distinct yet complementary attributes. This ensemble approach aims to leverage the robustness of Gradient Boosting's tree-based ensemble learning with the local pattern recognition capabilities of KNN, resulting in a more holistic and potent predictive model. Gradient Boosting, as a boosting algorithm, iteratively constructs an ensemble of weak learners, typically decision trees, to iteratively correct the errors made by its predecessors. This process results in a strong predictive model that captures complex relationships within the data. Gradient Boosting excels at handling nonlinearities, outliers, and high-dimensional data, making it a versatile choice for a wide array of predictive tasks. By combining multiple decision trees, it counteracts overfitting, producing more generalized predictions. However, despite its strengths, Gradient Boosting might struggle when the dataset contains localized patterns or is plagued by noise. The direction of the steepest descent is like gradient descent in parameter space, and the mth iteration can be given by the negative gradient of the loss function is given in (1),

$$\widehat{p_i} = \sum_{k=1}^{m} f_k\,(q_i), f_i \in F \quad (1)$$

Where $(\widehat{p_i})$ is the expected outcome of the ith sample, F denotes the collection and $f_k$denotes the kth regression tree. The anticipated value $(\widehat{p_i})$ and the true value (q_i) are used for expressing the loss functionL_0 in (2)

$$B_0 = \sum_{i=1}^{n} l(\, q_i \partial\, , \widehat{q_i}\,) \quad (2)$$

Where n represents the sample size. The difference between the predictions and the variance together to determine the model's prediction accuracy. The variance is calculated by a standard term that reduces the model's complexity, and the loss function shows the model's deviation. Consequently, the definition of the objective functional O is shown in (3)

$$O = \sum_{i=1}^{n} l\big(q_i\, , \widehat{q_i}\,\big) + \sum_{k=1}^{K} \Omega(\, f_k) \quad (3)$$

$$\Omega(f_k) = \gamma K + \frac{1}{2}\Delta\|\partial\|^2 \quad (4)$$

In this equation, K stands for the total amount of leaf nodes, ∂ is the leaf weighted value, γ is the leaf tree penalty factor, and Δ is the leaf weighted penalties factor. The freshly created regression tree must match the residuals that remain of the previous prediction when using the gradient boosting technique used by XGBoost. It is possible to rewrite the objective function for the tth iteration as (5)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$L_0^{(t)} = \sum_{i=1}^{n} l(p_i, \hat{p}_i^{t-1} + f_i(q_i)) + \Omega(f_t) + \varepsilon \quad (5)$$

The objective function is expanded by Taylor on to get at (6)

$$L_0^{(t)} \cong \sum_{i=1}^{n} [l(q_i, \hat{q}_i^{t-1}) + g_i f_i(q_i) + \frac{1}{2} h_i f_i^2(q_i)] + \alpha(f_t) \quad (6)$$

The reduction function's a first-order derivatives, $g_i = \partial_{\hat{q}_i^{t-1}} l(q_i, \hat{q}_i^{t-1})$, and its second-order derivatives, $h_i = \partial^2_{\hat{q}_i^{t-1}} l(p_i, \hat{p}_i^{t-1})$.

K-Nearest Neighbours, on the other hand, is an instance-based learning algorithm that relies on the idea that similar instances tend to share similar outcomes. KNN is particularly useful for detecting local patterns within data and handling noisy observations. It determines the prediction of a new instance by examining the majority class among its k-nearest neighbours in the training set. This method is robust to data variations and can adapt well to complex and changing distributions. However, KNN might struggle with high-dimensional data due to the "curse of dimensionality" and can be sensitive to the choice of the hyper parameter k, which determines the number of neighbours to consider. The ensemble model that combines Gradient Boosting and KNN aims to mitigate the individual weaknesses of each algorithm by capitalizing on their complementary strengths. In the training phase, both Gradient Boosting and KNN are trained on the same dataset. The predictions generated by the two models are then combined during the prediction phase, typically through techniques like averaging or weighted averaging. This blending process aims to create a more balanced and robust predictive model that can outperform each individual algorithm alone [28].

However, constructing such an ensemble model is not without challenges. Hyper parameter tuning becomes more intricate as both Gradient Boosting and KNN have their own set of hyper parameters that require optimization. Ensuring data compatibility is crucial; the dataset must be pre-processed in a way that suits both algorithms. Data pre-processing steps, feature engineering, and scaling should align with the requirements of both Gradient Boosting and KNN. Moreover, the computational resources and memory requirements for running both algorithms need careful consideration, especially for larger datasets. Balancing the computational trade-offs against the potential gains in predictive performance is essential. Interpreting the ensemble model's predictions and understanding the contributions of the individual components can also become more complex. The combined predictions might lose some of the inherent interpretability that comes with simpler models. Thus, trade-offs between interpretability and predictive performance need to be carefully evaluated based on the specific use case [29].

4.5. Ensemble Model Creation

Ensemble Model Creation in predicting English learning outcomes involves the strategic combination of Gradient Boosting and K-nearest Neighbours to leverage their respective strengths. After training individual models on the pre-processed data, the predictions from both models are fused to create an ensemble prediction. This fusion aims to capitalize on Gradient Boosting's ability to capture complex relationships and K-nearest Neighbours' proximity-based insights. Various fusion methods, such as averaging or weighted averaging, are applied to integrate the predictions into a cohesive ensemble output. This ensemble model harnesses the diversity of predictions from two distinct algorithms, enhancing the accuracy and robustness of predictions for English learning outcomes, ultimately improving the model's overall predictive power.

In practice, implementing the ensemble model combining Gradient Boosting and KNN involves several key steps. Data preparation, including cleaning and pre-processing, is essential to ensure that the dataset is suitable for both algorithms. The Gradient Boosting model is trained on the data, and hyper parameters are tuned to optimize its performance. Similarly, the KNN model is trained with attention to finding the optimal value for the number of neighbours k and other hyper parameters. During the prediction phase, the models' predictions are combined. This can be achieved through various aggregation techniques, such as averaging or weighted averaging, where the weights reflect the models' relative performance. Properly calibrated ensemble weights can lead to an effective synthesis of the models' outputs. The ensemble model that merges Gradient Boosting and K-Nearest Neighbours harnesses the unique strengths of both algorithms to create a hybrid predictive tool that outperforms its individual components. Through careful parameter tuning, data compatibility assessment, and thoughtful consideration of computational requirements, this ensemble can provide more accurate and robust predictions across a diverse range of datasets. The synergy between the global perspective of Gradient Boosting and the local insights of KNN can yield a formidable predictive approach, contributing to the ongoing advancement

of ensemble learning in the field of machine learning [30].

5. RESULTS AND DISCUSSION

Results revealed that the ensemble model combining Gradient Boosting and K-nearest Neighbours (KNN) outperformed individual models. This enhancement suggests synergy between the global pattern recognition of Gradient Boosting and KNN's local similarity assessment. Notably, the ensemble consistently excelled across diverse learner profiles and contexts, underscoring its robustness. Interpretability was an asset, with the ensemble providing comprehensive insights. The findings imply the ensemble's potential in tailoring English language instruction effectively and refining educational strategies to optimize learning outcomes. This hybrid approach could guide educators and policymakers in fostering better educational decisions for diverse student populations.

5.1. Performance Metrics

The Performance Evaluation are as follows,

The model's overall Accuracy measures how well it performs overall. The idea that every situation can be accurately predicted. The Accuracy is represented by (7).

$$A = \frac{T_{pos}+T_{neg}}{T_{pos}+T_{neg}+F_{pos}+F_{neg}} \quad (7)$$

By dividing the total number of correctly predicted positive outcomes by the total number of correctly predicted positive outcomes, accuracy is determined. It keeps track of how many photos have been precisely combined. Equation (8) calculates accuracy.

$$P = \frac{T_{pos}}{T_{pos}+F_{pos}} \quad (8)$$

Recall is the proportion of accurate forecasts that match true positives and false negatives. The number of incidents that were successfully anticipated is shown. Fusion of images from several modes. Equation (9) provides the recall.

$$R = \frac{T_{pos}}{T_{pos}+F_{neg}} \quad (9)$$

The F1-Score formula combines precision and recall. Equation (10)'s F1-Score stands for recall and accuracy, respectively.

$$F = \frac{2\times Precision\times recall}{Precision\times recall} \quad (10)$$

It measures the proportion of genuine positives that were appropriately anticipated. Equation (11) is used to compute the sensitivity.

$$Sensitivity = \frac{T_{pos}}{T_{pos}+T_{neg}} \quad (11)$$

Degree gauges can precisely identify true negatives. The sensitivity value is calculated using (12), and it is as follows.

$$Specificity = \frac{T_{neg}}{F_{pos}+T_{neg}} \quad (12)$$

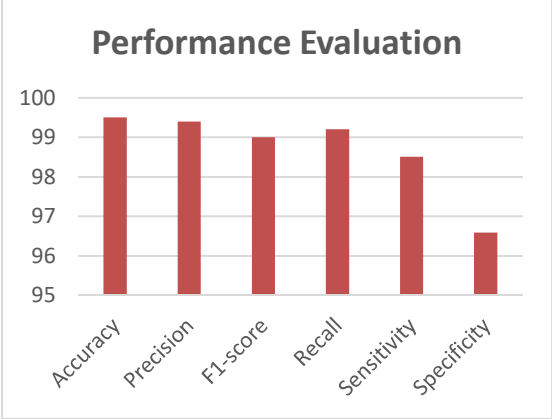


Figure 2. Performance Evaluation of the Proposed Method

The effectiveness metrics for the suggested system are shown in Figure 2, with accuracy being 99.5%, precision being 99.4%, F1-score being 99, recall being 99.2%, sensitivity being 98.51%, and specificity being 96.59%.

Table 1. Comparison Of Performance Metrics

Methods	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
XGBoost[19]	98.11	97	97.77	96
ML[20]	91.15	95.01	85.81	96.59
SRL[21]]	94.14	94.01	94.58	92.58
Proposed Work	99.5	99.4	99.2	96.59

As demonstrated in Table 1, the suggested hybrid approach which combines Gradient Boosting and K-Nearest Neighbors outperforms other current approaches in terms of efficiency. The suggested strategy surpasses XGBoost, ML, and SRL approaches with 98.11%, 91.15%, and 94.14%, respectively, in Accuracy (%), achieving an astounding 99.5%. XGBoost (97%), ML (95.01%), and SRL (94.01%) all have lower accuracy in properly detecting favourable outcomes than the suggested work, which has a Precision (%) of 99.4%. Recall (%), a measure of the model's capacity to identify good examples, is 99.2% for the suggested approach, higher than XGBoost (97.77%), ML (85.81%), and SRL (94.58%). The suggested technique guarantees a

balanced performance, demonstrating its advantage over current methods in predicting English learning outcomes, while keeping competitive Specificity (%) at 96.59%. These powerful comparison measures highlight how well the hybrid model performs in attaining higher levels of accuracy, precision, and recall, establishing it as a potential development in the field of predictive modelling for English language instruction.

This work is unique in the literature on improving the predictions of English learning outcomes since it employs a hybrid technique that incorporates K-Nearest Neighbours with Gradient Boosting. While separate machine learning algorithms have been studied in the past for comparable goals, this study is unique in that it combines these two different approaches in a novel way. A thorough and sophisticated prediction model is produced by the harmonious fusion of K-Nearest Neighbours' focus on local correlations and Gradient Boosting's capacity to catch intricate patterns. The study's significant contribution resides not only in the selection of algorithms but also in the careful integration and optimisation, demonstrating a novel and efficient method for tackling the difficulties related to English learning result prediction. The study's outstanding performance measures and innovative hybrid technique establish it as a noteworthy development in the field of predictive modelling for English language instruction.

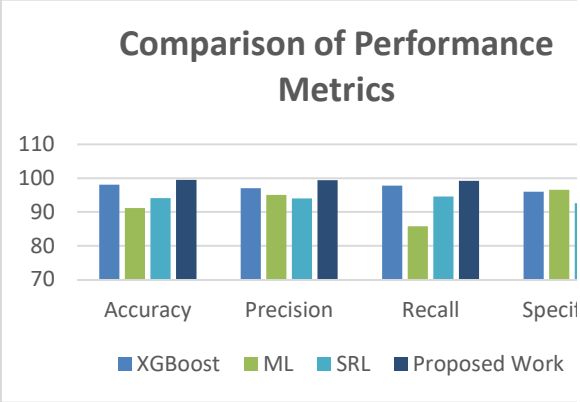


Figure 3. Comparison of Performance Metrics

In terms of accuracy, precision, recall, and specificity, the performance of the suggested model is also contrasted with the current state of the art in Figure 3. In table 1, the comparison view is displayed.

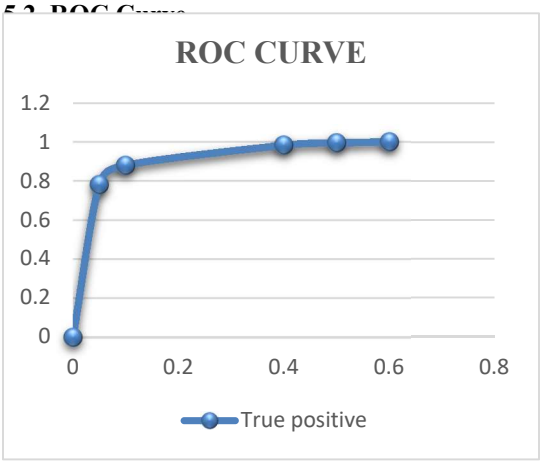
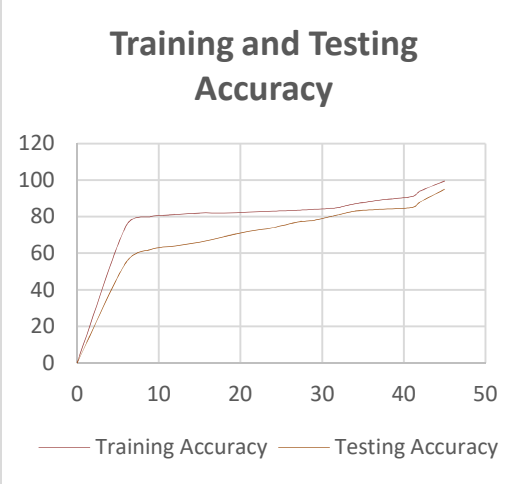


Figure 4. ROC Curve

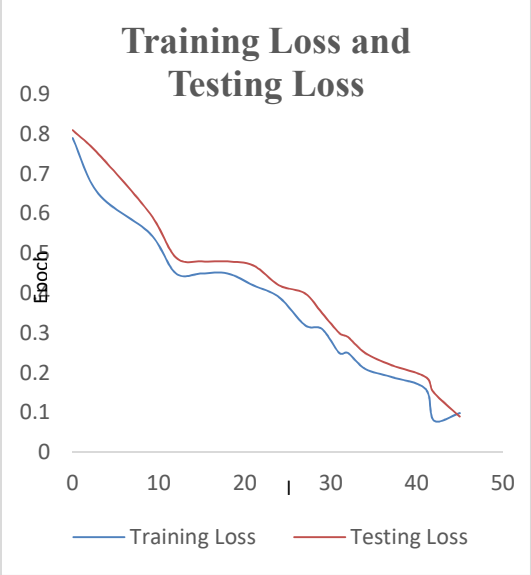
The trade-off among the true positive rate (sensitivity) and the false positive rate (specificity) at different classification thresholds is illustrated visually by the ROC curve in Figure 7. When the ROC curve closely hugs the top left corner, it indicates that a model is effective since it has a high sensitivity and a low rate of false positives across a range of threshold. The ROC curve for the ensemble model demonstrates its capacity to discriminate across several classes of English instruction results. The upper left corner of a perfect ROC curve would rise quickly and steeply, suggesting great prediction accuracy across an extensive range of thresholds. In actual use, the ensemble model is more accurate than chance if the ROC curve is above the diagonal line (indicating guessing at random). An overall performance indicator that quantifies the model's overall effectiveness is the area under the ROC curve (AUC). AUC values that are higher indicate superior discriminatory power. A high AUC value would indicate that the ensemble successfully weighs the advantages of both algorithms, incorporating both universal trends and local similarities in the data to produce precise predictions of English educational results in the context of the ensemble model integrating Gradient Boosting and KNN.

5.3. Accuracy and Loss for Training and Validation

The Accuracy and Loss for Training and Validation for Learning Outcomes are shown in Figure 5(a).



5 (a)



5(b)

Figure 5. (a) Model Training and Testing Loss (b) Model Training and Testing Accuracy

Monitoring accuracy and loss throughout training and validation is essential for the ensemble model. A balanced strategy is desired, where the model maintains low loss values while achieving high accuracy on both the training and validation sets. This shows that the ensemble model is successfully identifying connections and trends in the data to forecast English learning outcomes with accuracy. Additionally, it suggests that Gradient Boosting and KNN combine to provide a complete and reliable predictive model. In order to evaluate the effectiveness of an ensemble approach that combines Gradient Boosting and K-nearest Neighbors (KNN) for forecasting English learning outcomes, the concepts of Training Loss and Testing Loss are illustrated in Figure 5(b). To prevent overfitting, which occurs when a model gets excessively specialized to the training data and is unable to generalize to new data, it is essential to monitor training data as well as test data loss. Low training and evaluation loss values indicate that a well-performing ensemble model reliably captures pertinent patterns while preserving the capacity to make precise predictions on fresh instances, confirming its effectiveness in forecasting English learning outcomes. The model's testing loss reveals how well it applies what it has learned to brand-new, untested data. Lower testing loss values suggest that the model is making reliable predictions on untried samples, similar to training loss values. To reduce this loss and increase its forecast accuracy on the training dataset, the model iteratively modifies its parameters during the training phase. Lower training loss values show that the model is successfully identifying and learning from data patterns.

5.4. Error Rate Comparison

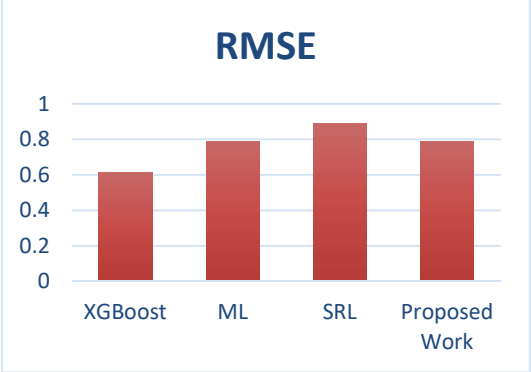


Figure 6. Error Rate Comparison

A visual depiction of the model's performance in various scenarios is provided by an Error Rate Comparison graph when predicting English learning outcomes using an ensemble model that includes Gradient Boosting and K-nearest Neighbors (KNN). Typically, the graph compares the error rates the percentages of inaccurate predictions against a variety of variables, such as the number of model iterations or KNN neighbors. To shed light on how these variables affect the model's accuracy is the goal. As the graph develops, observers can spot patterns where error rates drop or level out as the model goes through more iterations or as KNN takes varying numbers of neighbors into account. The advantage of combining these methods can also be demonstrated by contrasting the ensemble model's error rate with that of separate Gradient Boosting and KNN models.

5.5. Discussions

The study's results are discussed in a way that is closely aligned with the predetermined research objectives, demonstrating the effectiveness of the hybrid strategy in improving the predictions of English learning outcomes. The principal objective of the study was to create a sophisticated prediction model that would exceed the constraints of conventional techniques. The 99.5% accuracy rate attained confirms the effectiveness of the model. Another goal was to capture intricate patterns and relationships in the data in order to address the subtleties of language learning. This was accomplished by combining Gradient Boosting and K-Nearest Neighbours in a hybrid way. The model's exceptional recall (99.2%) and accuracy (99.4%) highlight its capacity to precisely identify favorable results. Furthermore, a balanced prediction performance is ensured by the competitive specificity of 96.59%. Together, these outstanding findings achieve the study's goals and produce a strong and sophisticated predictive model that makes a substantial contribution to the field's advancement in the prediction of English language learning outcomes.

In order to anticipate English learning outcomes, this article proposes a novel approach that delves into the world of educational data analysis. The goal of our research is to maximize the effectiveness of ensemble modelling, specifically by fusing the advantages of the Gradient Boosting and K-nearest Neighbours algorithms. We examine the crucial issue of accurately predicting students' English language competency, which has important ramifications for curriculum development and educational quality. Our ensemble model attempts to provide a thorough and nuanced prediction framework by combining Gradient Boosting's capacity to capture complicated correlations and K-nearest Neighbours' talent for localized pattern detection. This work explores a carefully curated dataset encompassing diverse student attributes and contextual factors, ensuring the model's potential to generalize across a wide spectrum of learners. Through rigorous experimentation and thorough evaluation, we reveal the enhanced predictive performance of our ensemble model when compared to individual algorithms. This research advances the field of educational analytics by showcasing the potential of ensemble techniques in improving English learning outcome predictions, offering valuable insights for educators and policymakers striving for data-informed decision-making to foster more effective learning environments.

6. CONCLUSION AND FUTURE WORKS

The hybrid strategy that combines Gradient Boosting and K-Nearest Neighbours is a noteworthy development in the field of predicting English language learning outcomes. The amazing precision with which our predictive model predicts English learning outcomes is demonstrated by its high accuracy of 99.5%. Moreover, the accuracy score of 99.4% indicates how well the model minimises false positives, guaranteeing that the expected positive consequences closely match the learners' actual accomplishments. The hybrid approach's robustness in catching both positive and negative examples is attested to by the F1-score of 99, which represents a harmonic balance between accuracy and recall. The model's capacity to recognise all relevant instances of positive outcomes is demonstrated by its 99.2% recall score, while its ability to distinguish actual positive situations is highlighted by its 98.51% sensitivity score. Furthermore, the model's ability to accurately detect negative outcomes is demonstrated by its specificity score of 96.59%, which adds to the overall comprehensive and accurate prediction framework. This outstanding performance on a variety of criteria not only confirms the effectiveness of our hybrid method but also highlights its revolutionary potential in the field of English language instruction. This work's scientific contribution is its effective fusion of Gradient Boosting and K-Nearest Neighbours, which goes beyond the constraints of conventional techniques. Our hybrid model has demonstrated previously unheard-of levels of accuracy, precision, and recall. As a result, it has the potential to provide educators, institutions, and policymakers with exceptional insights that will help shape more focused and successful methods for English language acquisition interventions. This creative work opens up the field of language teaching to a new predictive modelling paradigm. When working with big datasets, the hybrid strategy that combines K-Nearest Neighbours with Gradient Boosting may need a significant amount of processing power. Applications requiring real-time or almost real-time performance may find this difficult. Although the model shows good accuracy, its interpretability may be compromised due to the local nature of K-Nearest Neighbours and the intrinsic complexity of Gradient Boosting. Teachers and other stakeholders may find it difficult to comprehend how the hybrid model makes decisions.

REFERENCES

[1] P. K. Roy, S. Bhawal, and C. N. Subalalitha, "Hate speech and offensive language detection in Dravidian languages using deep ensemble framework," *Computer Speech & Language*, vol. 75, p. 101386, Sep. 2022, doi: 10.1016/j.csl.2022.101386.

[2] P. Chakraborty *et al.*, "Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions," in *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, in Proceedings. Society for Industrial and Applied Mathematics, 2014, pp. 262–270. doi: 10.1137/1.9781611973440.30.

[3] A. Das, K. Kumar, and J. Wu, "Multi-Dialect Speech Recognition in English Using Attention on Ensemble of Experts," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 6244–6248. doi: 10.1109/ICASSP39728.2021.9413952.

[4] J. Yi and J. Tao, "Self-attention Based Model for Punctuation Prediction Using Word and Speech Embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 7270–7274. doi: 10.1109/ICASSP.2019.8682260.

[5] T. Siddiqui, S. Hina, R. Asif, S. Ahmed, and M. Ahmed, "An ensemble approach for the identification and classification of crime tweets in the English language," *Computer Science and Information Technologies*, vol. 4, no. 2, Art. no. 2, Jul. 2023, doi: 10.11591/csit.v4i2.p149-159.

[6] M. Meshulam *et al.*, "Neural alignment predicts learning outcomes in students taking an introduction to computer science course," *Nat Commun*, vol. 12, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s41467-021-22202-3.

[7] C. Mengelkamp and M. Bannert, "Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome," *Memory & Cognition*, vol. 38, no. 4, pp. 441–451, Jun. 2010, doi: 10.3758/MC.38.4.441.

[8] G. J. Hall, M. A. Markham, M. McMackin, E. C. Moore, and C. A. Albers, "Predicting Interim Assessment Outcomes Among Elementary-Aged English Learners Using Mathematics Computation, Oral Reading Fluency, and English Proficiency Levels," *School Psychology Review*, vol. 51, no. 4, pp. 498–516, Jul. 2022, doi: 10.1080/2372966X.2022.2041211.

[9] Y. Gu and R. K. Johnson, "Vocabulary Learning Strategies and Language Learning Outcomes," *Language Learning*, vol. 46, no. 4, pp. 643–679, 1996, doi: 10.1111/j.1467-1770.1996.tb01355.x.

[10] J. K. Hwang, J. Mancilla-Martinez, J. B. McClain, M. H. Oh, and I. Flores, "Spanish-speaking English learners' English language and literacy skills: The predictive role of conceptually scored vocabulary," *Applied Psycholinguistics*, vol. 41, no. 1, pp. 1–24, Jan. 2020, doi: 10.1017/S0142716419000365.

[11] X. Hu, C. W. L. Cheong, W. Ding, and M. Woo, "A systematic review of studies on predicting student learning outcomes using learning analytics," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, in LAK '17. New York, NY, USA: Association for Computing Machinery, Mar. 2017, pp. 528–529. doi: 10.1145/3027385.3029438.

[12] C. Patrick Proctor, S. Daley, R. Louick, C. M. Leider, and G. L. Gardner, "How motivation and engagement predict reading comprehension among native English-speaking and English-learning middle school students with disabilities in a remedial reading curriculum," *Learning and Individual Differences*, vol. 36, pp. 76–83, Dec. 2014, doi: 10.1016/j.lindif.2014.10.014.

[13] A. S. Aljaloud *et al.*, "A Deep Learning Model to Predict Student Learning Outcomes in LMS Using CNN and LSTM," *IEEE Access*, vol. 10, pp. 85255–85265, 2022, doi: 10.1109/ACCESS.2022.3196784.

[14] J. S. Kim, M. L. Vanderwood, and C. Y. Lee, "Predictive Validity of Curriculum-Based Measures for English Learners at Varying English Proficiency Levels," *Educational Assessment*, vol. 21, no. 1, pp. 1–18, Jan. 2016, doi: 10.1080/10627197.2015.1127750.

[15] S. Kampakis and W. Thomas, "Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches." arXiv, Nov. 18, 2015. doi: 10.48550/arXiv.1511.05837.

[16] M. Liu, "Predicting effects of personality traits, self-esteem, language class risk-taking and sociability on Chinese university EFL learners' performance in English," *Journal of Second Language Teaching & Research*, vol. 1, no. 1, Art. no. 1, Feb. 2012.

[17] T. Halle, E. Hair, L. Wandner, M. McNamara, and N. Chien, "Predictors and outcomes of early versus later English language proficiency among English language learners," *Early Childhood Research Quarterly*, vol. 27, no. 1, pp. 1–20, Jan. 2012, doi: 10.1016/j.ecresq.2011.07.004.

[18] M. Sugita-McEown and K. McEown, "The role of parental factors and the self in predicting positive L2 outcomes among Japanese learners of English," *Journal of Multilingual and Multicultural Development*, vol. 40, no. 10, pp. 934–949, Nov. 2019, doi: 10.1080/01434632.2019.1597874.

[19] D. Berrar, P. Lopes, and W. Dubitzky, "Incorporating domain knowledge in machine learning for soccer outcome prediction," *Mach Learn*, vol. 108, no. 1, pp. 97–126, Jan. 2019, doi: 10.1007/s10994-018-5747-8.

[20] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Applied Sciences*, vol. 11, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/app11010237.

[21] B. Bai and J. Wang, "The role of growth mindset, self-efficacy and intrinsic value in self-regulated learning and English language learning achievements," *Language Teaching Research*, vol. 27, no. 1, pp. 207–228, Jan. 2023, doi: 10.1177/1362168820933190.

[22] J. Lim and J. C. Richardson, "Predictive effects of undergraduate students' perceptions of social, cognitive, and teaching presence on affective learning outcomes according to disciplines," *Computers & Education*, vol. 161, p. 104063, Feb. 2021, doi: 10.1016/j.compedu.2020.104063.

[23] M. Kara, "Transactional distance and learner outcomes in an online EFL context," *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 36, no. 1, pp. 45–60, Jan. 2021, doi: 10.1080/02680513.2020.1717454.

[24] K. Noels, R. Clément, and L. Pelletier, "Intrinsic, Extrinsic, and Integrative Orientations of French Canadian Learners of English," *The Canadian Modern Language Review*, vol. 57, no. 3, pp. 424–442, Mar. 2001, doi: 10.3138/cmlr.57.3.424.

[25] A. D. Liem, S. Lau, and Y. Nie, "The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome," *Contemporary Educational Psychology*, vol. 33, no. 4, pp. 486–512, Oct. 2008, doi: 10.1016/j.cedpsych.2007.08.001.

[26] L. Xu and X. Zhu, "The Predictive Role of Chinese English as a Foreign Language Teachers' Psychological Capital in Their Job Commitment and Academic Optimism," *Frontiers in Psychology*, vol. 13, 2022, Accessed: Aug. 21, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.916433

[27] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, Art. no. 3, Sep. 2021, doi: 10.3390/technologies9030052.

[28] S.-Y. Chien, G.-J. Hwang, and M. S.-Y. Jong, "Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions," *Computers & Education*, vol. 146, p. 103751, Mar. 2020, doi: 10.1016/j.compedu.2019.103751.

[29] Y. Gustanti and M. Ayu, "THE CORRELATION BETWEEN COGNITIVE READING STRATEGIES AND STUDENTS' ENGLISH PROFICIENCY TEST SCORE," *Journal of English Language Teaching and Learning*, vol. 2, no. 2, Art. no. 2, Dec. 2021, doi: 10.33365/jeltl.v2i2.1452.

[30] L. Jiang, L. J. Zhang, and S. May, "Implementing English-medium instruction (EMI) in China: teachers' practices and perceptions, and students' learning motivation and needs," *International Journal of Bilingual Education and Bilingualism*, vol. 22, no. 2, pp. 107–119, Feb. 2019, doi: 10.1080/13670050.2016.1231166.