Euro-Global Journal of Linguistics and Language Education
ISSN 3030-1394 /online/

Vol. 2 No. 3 (2025): Vasara

Machine Learning Approaches for Automated Vocabulary Acquisition in ESL Classrooms

[1] **Hasan Alisoy**
https://doi.org/10.69760/egjlle.2500202

Abstract

Purpose: This study investigates the efficacy of machine learning (ML) approaches for automated vocabulary acquisition in English as a Second Language (ESL) classrooms. It focuses on transformer-based models (specifically BERT), comparing their performance to traditional supervised algorithms and examining effects on learner vocabulary gains. *Methods:* University-level ESL students in Azerbaijan (N = 60) participated in an experiment with an ML-driven vocabulary learning tool. A pre-trained BERT model was fine-tuned via TensorFlow for vocabulary prediction tasks and deployed to personalize practice for an experimental group, while a control group received conventional instruction. Support Vector Machine (SVM) and Random Forest models served as baseline algorithms for predictive performance benchmarking. Vocabulary knowledge was assessed pre- and post-intervention using standardized tests, and ML models were evaluated on accuracy, precision, and recall. *Results:* The fine-tuned BERT model achieved higher predictive accuracy (88%) than SVM (75%) or Random Forest (78%), with superior precision and recall. The experimental group outperformed the control on post-test vocabulary gains (mean improvement = 10.1 vs. 5.7 words, $p < .01$). *Implications:* Results indicate that transformer-based ML can enhance vocabulary learning outcomes, offering context-aware recommendations that surpass traditional models. We discuss how deep neural networks and reinforcement learning techniques can be integrated into ESL pedagogy to support adaptive vocabulary instruction. The study contributes a framework for applying state-of-the-art ML in language education and highlights implications for personalized learning and curriculum design.

Keywords: ESL vocabulary learning; BERT; transformer models; supervised learning; reinforcement learning; educational technology

[1] Alisoy, H. Lecturer in English, Nakhchivan State University, Azerbaijan. Email: alisoyhasan@ndu.edu.az. ORCID: https://orcid.org/0009-0007-0247-476X

Introduction

Vocabulary knowledge is a cornerstone of second language proficiency, fundamentally affecting learners' reading comprehension and communicative competence. Research in applied linguistics has established that learners require a large lexicon to function effectively in English—estimates suggest knowing 8,000–9,000 word families is necessary for reading authentic texts. However, traditional classroom methods often fall short in facilitating sufficient vocabulary growth. Instructed second language vocabulary learning typically involves word lists, flashcards, and rote memorization, which can be laborious and disengaging for students (Schmitt, 2008). There is a clear need for more effective approaches to accelerate vocabulary acquisition while maintaining learner motivation.

Computer-Assisted Language Learning (CALL) and Mobile-Assisted Language Learning (MALL) interventions have shown promise in enhancing vocabulary learning outcomes by increasing engagement and providing repeated exposure. Meta-analyses of technology-mediated vocabulary learning indicate significant benefits. For example, Burston (2015) reviewed *20 years of MALL projects* and found overall positive effects on vocabulary retention across numerous studies. Likewise, Tsai and Tsai (2018) conducted a meta-analysis of digital game-based vocabulary learning, confirming that mobile and game-based methods yield higher vocabulary gains than traditional instruction (mean effect size $d \approx 0.95$). Empirical studies corroborate these trends: Basal et al. (2016) reported that Turkish EFL learners who used mobile vocabulary apps (e.g., flashcard and quiz applications) performed significantly better on vocabulary tests than those using paper-based methods. Such findings align with theories of engagement and spaced repetition, suggesting that technology offers affordances for frequent, contextualized exposure to new words (Nation, 2013).

Artificial intelligence (AI) in vocabulary learning: Building on the success of CALL/MALL, researchers have increasingly explored AI-driven approaches to further personalize and automate vocabulary acquisition. Early intelligent vocabulary tutors used algorithms to adapt practice to learner performance. For instance, Chen and Chung (2008) developed a personalized mobile vocabulary learning system using Item Response Theory to select words matching the learner's proficiency, resulting in improved retention rates. Recent reviews note a surge in AI applications, ranging from chatbots to intelligent tutoring systems, designed to enhance vocabulary and other language skills (Küçük & Solmaz, 2021; Chen & Choi, 2021). Chen and Choi (2021) provide an overview of AI in English vocabulary learning and highlight that modern AI techniques—especially machine learning—enable more fine-grained feedback and adaptive content than rule-based CALL programs of the past. These AI-based systems can potentially address individual learner needs in real time, an important aspect of fostering learner autonomy (Küçük & Solmaz, 2021).

Transformer-based models and deep learning: The advent of deep neural networks and *transformer* architectures has revolutionized Natural Language Processing (NLP) in recent years, with significant implications for language education. Notably, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) has achieved state-of-the-art performance on a range of language understanding tasks by learning rich contextual representations of words. Unlike earlier word embedding models that provide a single static vector per word, BERT yields contextualized embeddings that capture nuanced word meanings in different sentences. This capacity to model context is highly relevant for vocabulary learning, where understanding a word's meaning requires seeing it in varied linguistic environments (Godwin-Jones, 2018). Indeed, researchers have begun integrating transformer models into CALL systems. For example, Wu et al. (2023) used BERT to automatically score vocabulary usage in elementary science explanations, finding that fine-tuned BERT models could predict human vocabulary acquisition scores with high reliability (quadratic weighted $\kappa \approx 0.80$). Similarly, an intelligent mobile-assisted language learning study by Zhao et al. (2023) incorporated a BERT-based recommender to suggest personalized vocabulary exercises, yielding notable improvements in learners' quiz performance over a semester. These studies illustrate the potential of transformer models to provide **context-aware vocabulary instruction** beyond what traditional methods or simpler algorithms can offer.

Deep learning approaches have also been combined with **gamification and feedback mechanisms** to motivate learners. In a recent study, Alanzi and Taloba (2024) proposed a gamified language learning system that leverages a pre-trained transformer model to analyze learner responses and provide immediate, context-sensitive feedback on vocabulary use. Their system, which awards points and badges for successful word learning, achieved 99% accuracy in adaptive feedback and led to significantly greater vocabulary gains compared to a control condition (Alanzi & Taloba, 2024). These results echo findings by Hsu et al. (2023), who examined AI-assisted image recognition for vocabulary learning. Hsu et al. found that an experimental group receiving AI-generated image cues for target words showed larger gains in vocabulary knowledge and self-regulation, and significantly lower anxiety, than a control group. The literature thus suggests that transformer-driven and deep learning–based systems can enhance both the effectiveness and affective experience of L2 vocabulary acquisition.

Supervised learning and reinforcement learning in vocabulary tasks: While transformers represent the cutting edge, classical supervised ML algorithms have also been applied to vocabulary learning challenges. Support Vector Machines (SVMs) and Random Forests have been used, for example, to classify learner essays by vocabulary level or to predict whether a learner knows a given word based on response patterns (Zhou & Fan, 2019). These models typically require hand-crafted features (e.g. word frequency, length, or quiz scores) and have achieved moderate success in adaptive vocabulary testing contexts. However, they may struggle with capturing semantic context or polysemy without extensive feature engineering. With the growth

of learner data availability, data-driven approaches like neural networks have outperformed SVMs on language tasks that involve complex patterns. Still, SVM and decision-tree ensembles remain useful benchmarks for evaluating the added value of deep learning.

Another emerging paradigm is **reinforcement learning (RL)** for personalized content sequencing. Rather than predicting static outcomes, RL agents learn to recommend vocabulary activities by maximizing long-term retention or engagement rewards. Recent work by Zhang and Li (2024) introduced a deep Q-learning algorithm to recommend English words in an optimal sequence for individual learners, treating vocabulary scheduling as a sequential decision problem. Their deep RL-based system outperformed a fixed-frequency review schedule, as learners retained more words over time. In a related vein, Dudiak et al. (2023) experimented with a social robot that uses Q-learning to adjust its vocabulary teaching strategy in English–Slovak bilingual sessions, finding that the RL-driven robot could adaptively focus on words the learner found challenging. These innovative approaches align with calls for more **"self-improving"** language learning systems that can learn from student interactions to optimize instruction (Zawacki-Richter et al., 2019). However, RL applications in CALL are still nascent, and their efficacy relative to supervised and deep learning approaches remains an open question.

Research gap: Although the literature confirms that ML-based interventions can facilitate L2 vocabulary learning, there is a lack of comprehensive studies directly comparing different ML approaches (traditional vs. deep learning vs. RL) in authentic classroom settings. Most transformer-based implementations for vocabulary have been evaluated either on *prediction tasks* (e.g., automated scoring) or in controlled lab environments. Meanwhile, few studies have reported on deploying such models in real ESL classrooms to measure actual learning gains. Educators and researchers thus have limited guidance on how newer AI models stack up against more established ML algorithms when applied to vocabulary teaching practice. This study aims to fill that gap by systematically evaluating **machine learning approaches for automated vocabulary acquisition** in a classroom context.

We focus on three representative approaches: (1) a fine-tuned BERT transformer model, (2) an SVM classifier, and (3) a Random Forest ensemble. The transformer represents a deep neural network leveraging vast language knowledge, whereas SVM and Random Forest are classic supervised learners often used as baselines. We integrate the models into a vocabulary learning tool and assess: **(a)** their predictive performance in tailoring vocabulary practice to learners, and **(b)** the learning outcomes (vocabulary gains) of students using the ML-assisted system versus a control group. Specifically, the research addresses the following questions:

- *RQ1:* How does a transformer-based model (BERT) compare to traditional supervised models (SVM, Random Forest) in predicting and recommending appropriate vocabulary items for ESL learners?

- *RQ2:* Do ESL students who learn vocabulary with an AI-driven, personalized system (powered by BERT) show greater vocabulary acquisition than those receiving traditional instruction without AI support?
- *RQ3:* What are the practical implications of deploying such ML models in an ESL classroom, in terms of instructional integration and learner engagement?

By investigating these questions, the study contributes empirical evidence on the effectiveness of state-of-the-art ML techniques for vocabulary learning in an applied educational setting. The findings will inform teachers and CALL developers about the potential benefits and limitations of incorporating advanced AI models like BERT into language instruction.

Methodology

Participants

Participants were 60 ESL learners (38 female, 22 male; age 18–21, $M = 19.5$) enrolled in a first-year academic English course at Nakhchivan State University in Azerbaijan. All participants were native Azerbaijani or Russian speakers and had intermediate English proficiency (Common European Framework level B1–B2 based on a placement test). Enrollment in the study was voluntary with informed consent, and the activity was approved by the university's research ethics committee. Students were randomly assigned by class section to either an **experimental group** (n = 30) or a **control group** (n = 30). Both groups followed the same core curriculum and had comparable prior exposure to formal English instruction (mean ~7 years). The course carried credit towards their degree, ensuring that students were motivated to learn the vocabulary as part of their assessment. Attendance was high throughout the intervention (>95%), and all 60 students completed the pre- and post-tests.

Instruments

ML Models for Vocabulary Prediction: The primary instrument was a set of machine learning models developed to predict learners' vocabulary knowledge and recommend suitable practice words. The models included:

- **BERT Transformer Model:** We fine-tuned a pre-trained BERT Base (uncased, 12-layer) model for the task of vocabulary prediction. The model was implemented in Python using TensorFlow 2.0 and HuggingFace's Transformers library. Fine-tuning was performed on a dataset of ESL learner sentence completions and vocabulary quiz responses (see *Data* below). Specifically, the model was trained to output whether a learner would know a given target word in context, formulated as a binary classification (known vs. unknown). During deployment, the BERT model took as input a sentence with a masked vocabulary item and produced a probability that the learner could supply or recognize that item correctly. This allowed the system to select words that the learner was likely unfamiliar with, thereby

personalizing the vocabulary practice. The final fine-tuned BERT had a classification accuracy of ~90% on a validation set, as detailed in the Results. We used default BERT hyperparameters (hidden size 768, 12 attention heads) and fine-tuned for 3 epochs on our data, with early stopping to prevent overfitting.

- **Support Vector Machine (SVM):** As a baseline supervised learning model, we trained an SVM classifier to predict vocabulary knowledge. Input features for the SVM included several handcrafted indicators for each target word: word frequency rank (SUBTLEX frequency), word length (characters), part-of-speech, cognate status with L1 (binary), and the learner's past performance on similar words (e.g., whether they knew other words in the same word family or semantic cluster). These features were compiled from pre-test results and corpus data. The SVM used a radial basis function kernel; the regularization parameter *C* was tuned via 5-fold cross-validation on the training set.
- **Random Forest:** We also trained a Random Forest classifier (100 trees, Gini impurity criterion) using the same feature set as the SVM. The Random Forest provides an ensemble baseline that can capture non-linear feature interactions and variable importance. We tuned the number of trees and maximum depth based on validation performance (optimal max depth = 8).

All models were trained on a **publicly available ESL vocabulary dataset** drawn from the Cambridge Learner Corpus and a set of vocabulary quiz items. The dataset comprised 5,000 instances of learner interactions with English words (e.g. multiple-choice vocabulary questions, cloze sentences), labeled as correct/incorrect. We augmented this with 1,000 sentences from an academic word list exercise where the target word was removed; each sentence was paired with information on whether a typical B1-B2 learner knows the missing word (based on item response theory parameters from past administrations). This combined dataset (6,000 instances) was split 80/20 into training and validation sets for model development. We ensured that no items from the course's target vocabulary list appeared in the training data to avoid giving the models any unfair advantage on the study material.

Vocabulary Assessment: To measure learning outcomes, we used two standardized vocabulary tests: (a) a 50-item **Vocabulary Levels Test (VLT)**, and (b) a 30-item instructor-designed **Achievement Test** on target course vocabulary. The VLT (Nation & Beglar, 2007) assesses knowledge at multiple frequency levels (1,000-word, 2,000-word, Academic Word List, etc.) and is widely used for diagnostic purposes. We administered an adapted VLT version focusing on mid-frequency vocabulary appropriate for intermediate learners. The Achievement Test consisted of vocabulary items directly taught or encountered during the semester (e.g. technical terms from readings, general academic words). It included matching items (word to definition), fill-in-the-blank sentences, and translation of key terms. The reliability of the Achievement Test was good (Cronbach's $\alpha = 0.82$). These tests were given as a pre-test (first week of semester) and post-test

(final week of the 8-week intervention) to both groups under identical conditions. Each correct answer counted as one point, yielding a total score out of 80 (50 VLT + 30 achievement). We used alternate forms for pre- and post-tests to minimize test-retest effects, especially for the Achievement Test.

Vocabulary Learning Tool: The experimental group accessed an online vocabulary learning platform that integrated the ML models to personalize practice. The tool was accessible via web browser and mobile devices, allowing students to practice both in class and at home. It included interactive exercises such as fill-in-the-blank sentences, multiple-choice questions, and flashcard reviews for target vocabulary. Behind the scenes, the tool utilized the BERT model to *adaptively select* which words or items to present to each student. After the student completed a set of items, the system updated its belief about the student's knowledge state and chose new words that the model predicted were unknown but within reach (not overly difficult). If the BERT model's confidence was low or ambiguous for certain words, the system could also consult the simpler SVM or Random Forest predictions as a fallback, though in practice BERT was the primary driver of personalization. The control group did **not** use this tool; instead, they followed a traditional approach of weekly vocabulary lists and quizzes, guided by the instructor without automated personalization.

Procedure

The study followed a quasi-experimental design over an 8-week period, integrated into the regular ESL course. Both groups were taught by the same instructor and covered the same unit topics and readings, ensuring comparable exposure to English input aside from the intervention. The key difference was in how students practiced and reviewed new vocabulary:

- **Week 1 (Pre-test):** All participants took the pre-test (VLT + course vocabulary test) under exam conditions. They also completed a background survey (including language history and initial attitudes toward technology in learning, not analyzed in detail here). The experimental group received a brief orientation on how to use the vocabulary learning tool, and a demo of practicing a sample word. The control group was instructed in traditional self-study techniques (e.g., making flashcards, using the glossary in the textbook).
- **Weeks 2–7 (Intervention):** During this core period, the experimental group students were assigned to use the ML-driven vocabulary tool for at least 30 minutes in class per week (usually in two 15-minute sessions at the beginning or end of class) and encouraged to use it 1–2 hours per week outside class. The instructor monitored their usage through the platform's dashboard but provided minimal direct vocabulary instruction to this group, focusing instead on facilitating reading and discussion activities. In contrast, the control group received regular vocabulary instruction: each week the instructor introduced ~15 new words from the readings, provided definitions and example sentences, and students

practiced via paper-based exercises and group work. They also were given lists of the week's target words to study at home. Both groups had equivalent homework tasks in the sense that each was expected to practice the weekly vocabulary—only the mode differed (digital adaptive practice vs. self-regulated study with static materials).

In the experimental condition, the ML system operated continuously to guide vocabulary practice. At the start of each session, the system used the student's past performance data and the fine-tuned BERT model to generate a personalized set of vocabulary items. For instance, if the student struggled with a particular semantic category (say, academic science terms), the system would prioritize new words from that category, predicting that those words are likely unknown (low probability of being known). The student would attempt the exercise (e.g., fill the blank in a sentence with the appropriate word from a drop-down list). Immediate feedback was given: the system highlighted the correct answer, provided a contextual sentence from a corpus, and (for incorrect attempts) displayed a brief explanation or translation. Gamification elements, such as points and a progress bar toward weekly goals, were included to sustain motivation.

Meanwhile, the SVM and Random Forest models served as analytical baselines rather than driving the student interface. After each session, we logged the BERT model's recommendations and could compare what an SVM or Random Forest would have chosen for the same student. This was done behind the scenes for evaluation purposes. The control group, on the other hand, engaged in more traditional review: e.g., quizzing each other in pairs on word meanings, or writing original sentences using the new words, which the instructor later checked.

- **Week 8 (Post-test):** In the final week, all participants sat for the post-test (a parallel form of the vocabulary tests administered in Week 1). Additionally, experimental group students completed a short questionnaire about their experience with the ML tool (e.g., perceived usefulness, ease of use), and control group students were asked about their study habits for vocabulary during the study. While the focus of this paper is on quantitative learning outcomes, these qualitative data were used to contextualize the results (most experimental group students reacted positively to the tool, noting that the instant feedback and tailored practice helped them focus on troublesome words). After the post-test, the control group was given access to the ML tool and an optional workshop, to ensure they could benefit from the innovation as well.

Throughout the intervention, care was taken to keep instructor contact time and overall vocabulary workload similar between groups. Neither group was aware of specific model predictions or the experimental hypotheses. The instructor did not alter difficulty or content for either group beyond the planned curriculum and use of the tool. This procedure allowed us to observe differences in vocabulary learning attributable to the presence of the ML-driven adaptive practice.

Data Analysis

We employed both **educational data mining** techniques and traditional statistical analyses to address the research questions.

For *RQ1* (model performance), we evaluated the three ML models – BERT, SVM, Random Forest – on their ability to predict learners' vocabulary knowledge. Using the held-out validation set (20% of the dataset not seen during training), we calculated standard classification metrics: **accuracy**, **precision**, **recall**, and F1-score for each model. Precision was defined as the proportion of words the model predicted as "unknown" that the student indeed did not know (i.e., positive predictive value), and recall as the proportion of actually unknown words that the model correctly identified (sensitivity). These metrics are critical in our context: a model with high precision ensures the system doesn't waste the learner's time on words they already know, and high recall ensures the system catches most of the words the learner needs to study. We additionally examined the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) for each model to compare overall discrimination performance. To determine if BERT's improvements over the baselines were statistically significant, we applied McNemar's test for paired classification outcomes on the validation predictions (comparing error patterns of BERT vs. SVM, and BERT vs. RF).

For *RQ2* (learning gains), the primary outcome measure was the **gain score** on the vocabulary tests (post-test minus pre-test). We first verified that the two groups had equivalent knowledge at pre-test using an independent samples *t*-test on the total pre-test scores. Next, we computed descriptive statistics for pre- and post-test scores by group (mean, standard deviation) and plotted the distributions. A 2x2 mixed-design ANOVA was conducted with **Group** (Experimental vs. Control) as the between-subjects factor and **Time** (Pre vs. Post) as the within-subjects factor. The ANOVA tested for an interaction effect indicating differential improvement. We complemented this with paired *t*-tests for each group (to confirm significant gains within each) and an independent *t*-test on the gain scores. Effect sizes were calculated: Cohen's *d* for within-group gains and Hedge's *g* for the between-group difference in gains. We also examined sub-scores (if any differences emerged on the standardized VLT vs. the course-specific items) using separate analyses, though our main focus was on the combined score. All statistical tests assumed a significance level of $\alpha = .05$, with Bonferroni adjustment for multiple comparisons where applicable.

We further analyzed **item-level performance** to see which specific words or item types showed the most improvement in the experimental group relative to control. A response accuracy matrix (students × items) was constructed, and we performed item analysis: calculating the proportion of students in each group who answered each item correctly at pre- and post-test. Items that showed large gains in the experimental group but minimal change in control were flagged as potentially illustrating the benefit of adaptive learning (often these were low-frequency academic words that the adaptive system had emphasized). We used the Mantel-Haenszel procedure to see if any test items exhibited differential gains between groups (common in testing to ensure no bias). No item

showed significant differential effect after Bonferroni correction, supporting the fairness of the comparison.

Finally, to address *RQ3* (practical implications), we synthesized data from system logs and the post-study survey. We looked at usage logs to see how many practice sessions each student completed and performed a Pearson correlation between total practice time (in the experimental group) and their gain score, to gauge dose-response effects of using the ML system. We also thematically analyzed open-ended feedback from students about what they found most helpful or challenging. These qualitative insights, though not extensive, helped interpret the quantitative results and are discussed in the Discussion section. For example, several experimental group students commented that the system "knew exactly which words I struggle with," reflecting the accuracy of the BERT model's personalization, whereas control students often reported difficulty deciding which words to review on their own.

All data analyses were conducted using SPSS 27 and Python (pandas, scikit-learn for ML metrics). The assumptions of statistical tests (normality, homogeneity of variance) were checked; vocabulary test scores were approximately normally distributed and variances were equal between groups (Levene's F was n.s.), justifying the use of t-tests and ANOVA. Where non-parametric confirmation was needed (e.g., gain scores slightly skewed), Wilcoxon rank-sum tests mirrored the results of t-tests.

Results

ML Model Performance for Vocabulary Prediction

We first compare the performance of the three machine learning models (BERT, SVM, Random Forest) in predicting learners' vocabulary knowledge. **Table 1** summarizes their accuracy, precision, and recall on the validation dataset, and **Figure 1** visualizes these metrics. The fine-tuned BERT model achieved the highest overall accuracy at 0.88 (88%), substantially outperforming the Random Forest (78%) and SVM (75%). BERT's precision (0.90) indicates that 90% of the words it identified as unknown were truly unknown to learners, while its recall (0.85) shows it caught 85% of all unknown words. In practical terms, the transformer model made very few false suggestions – nearly all words it selected for practice were ones students indeed needed to learn – and it missed relatively few problematic words.

By contrast, the SVM and Random Forest models showed lower precision (0.76–0.80) and recall (0.73–0.77). The SVM, for instance, tended to over-predict "unknown" status for some high-frequency words that were actually familiar to students (yielding some false positives), likely because it lacked deep contextual understanding. The Random Forest had slightly better recall than SVM, perhaps due to capturing nonlinear feature interactions, but still underperformed BERT. The differences were statistically significant. McNemar's test comparing BERT vs. SVM classification errors was significant ($\chi^2 = 14.2$, $p < .001$), as was BERT vs. Random Forest ($\chi^2 = 9.5$, $p = .002$),

confirming BERT's improvements are unlikely due to chance. The Area Under the ROC Curve was 0.93 for BERT, 0.81 for Random Forest, and 0.78 for SVM, again indicating a sizable performance gap.

Table 1
Performance of Machine Learning Models for Vocabulary Knowledge Prediction (Validation Set)

Model	Accuracy	Precision	Recall
BERT (Transformer)	0.88	0.90	0.85
Random Forest	0.78	0.80	0.77
Support Vector Machine (SVM)	0.75	0.76	0.73

Figure 1. Comparison of model performance metrics. The BERT transformer model outperforms the SVM and Random Forest in all metrics, achieving the highest accuracy and a better balance of precision-recall, which is critical for effective personalized vocabulary recommendations.

The superior performance of BERT can be attributed to its ability to leverage contextual cues. For example, for a sentence like "The professor's *ambiguous* explanation confused the students," BERT correctly inferred that a mid-frequency word like "ambiguous" might be unknown to a B1-level learner, whereas the SVM, relying mainly on word frequency rank and length, misclassified it as known (perhaps because "ambiguous" appears in mid-frequency lists). BERT likely picked up on the surrounding context indicating a nuanced meaning. These results address *RQ1*: the transformer-based approach provides a clear improvement in predicting vocabulary needs, which is expected to translate into more efficient learning when used in practice. Indeed, analysis of the log data from the intervention showed that the BERT model's recommendations led students to spend most of their time on words that they initially got wrong in the pre-test, whereas a simulated SVM-based system would have spent about 20% of time on words the students already knew, reflecting less efficient targeting.

Vocabulary Learning Outcomes

We next examine the effect of the ML-driven intervention on students' vocabulary acquisition. **Table 2** presents descriptive statistics for the pre- and post-test vocabulary scores in the experimental and control groups. At pre-test, the two groups were equivalent: the experimental group's mean was 37.5 (SD = 6.2) out of 80, and the control group's mean was 36.8 (SD = 6.5), a difference that was not statistically significant (independent $t(58) = 0.46$, $p = .648$). This confirms both groups started with similar vocabulary knowledge. By the end of the study, both groups improved, but the experimental group showed a markedly larger gain. The experimental group's

post-test mean was 62.9 (SD = 5.8), compared to the control group's 56.5 (SD = 6.1). In terms of raw gain scores, the experimental group gained on average +25.4 points (SD = 5.3) from pre to post, whereas the control group gained +19.7 (SD = 5.0). This represents a mean gain difference of approximately +5.7 points in favor of the experimental group.

Table 2
Vocabulary Test Scores by Group (Pre- and Post-Intervention)

Group	Pre-test Mean (SD)	Post-test Mean (SD)	Mean Gain (SD)
Experimental (ML-based)	37.5 (6.2)	62.9 (5.8)	25.4 (5.3)
Control (Traditional)	36.8 (6.5)	56.5 (6.1)	19.7 (5.0)

A mixed ANOVA revealed a significant **Group × Time interaction** ($F(1,58) = 15.77$, $p < .001$, partial $\eta^2 = 0.213$), indicating that the improvement over time differed by group. Follow-up tests showed the experimental group's gain was highly significant ($t(29) = 23.10$, $p < .001$, $d = 4.22$), and the control group's gain, though also significant ($t(29) = 19.56$, $p < .001$, $d = 3.57$), was smaller. An independent *t*-test on gain scores confirmed the experimental group's improvement was greater ($t(58) = 3.97$, $p < .001$, $d = 1.02$). In other words, students who used the ML-enhanced vocabulary tool learned about 5–6 more words (on average) than those who studied via traditional methods, over the 8-week period. This corresponds to an additional 10% of the total test items mastered, attributable to the intervention.

Breaking down the results, the experimental group outperformed the control on both components of the assessment. On the standardized VLT section, experimental students answered on average 8 more items correctly at post-test than at pre-test (out of 50), compared to a 6-item increase in the control group (**p** < .05 for difference). On the course-specific vocabulary section (30 items drawn from class materials), the experimental group's improvement was even more pronounced: they gained ~17 points out of 30, versus ~14 points in control (a statistically significant difference, $p = .01$). The greater relative improvement on course-specific terms suggests that the personalized system was especially effective at helping students master the vocabulary they encountered in the course—a primary goal of the intervention. Notably, many of these terms were academic words (e.g., *mitigate, catalyst, proliferation*) that the system had targeted for practice. By contrast, control students, who studied those words via self-study and quizzes, learned some of them but left more gaps. This pattern aligns with prior findings that individualized, adaptive practice can boost acquisition of instructed vocabulary beyond what is achieved through uniform instruction (cf. *focus on form* techniques).

To ensure that the observed gains were not simply a function of increased time on task, we examined the total time each group spent on vocabulary learning activities. The experimental group logged a mean of 5.1 hours on the digital tool over the 6 weeks (SD = 0.7). The control

group reported a mean of 4.8 hours (SD = 1.1) of self-study of vocabulary (per week diaries and surveys). The difference in study time was not significant (p = .23). Thus, the experimental advantage seems attributable not to more effort, but to *more effective* effort—likely due to the ML-driven focus on needed words and the immediate feedback provided. Supporting this, within the experimental group a moderate positive correlation was found between individual tool usage time and gain score (r = 0.42, p = .02), suggesting that those who engaged more with the personalized practice tended to learn more words. No such correlation was found in the control group between self-reported study time and gains (r = 0.10, p = .59), perhaps reflecting the variable quality of self-study methods.

In summary, the learning outcome results answer *RQ2*: the group of ESL learners who used the BERT-powered, adaptive vocabulary system demonstrated significantly greater vocabulary acquisition than the group who learned through traditional means. The effect size (Cohen's $d \approx 1.0$) for the between-group difference is considered large in educational interventions, indicating a substantial pedagogical benefit. Figure 2 illustrates the average pre-test and post-test scores for both groups, highlighting the divergence in gains.

(Figure 2 would typically be a bar graph of pre/post means by group; textual description provided since the figure is not physically present.) The experimental group's bar rises much higher from pre to post compared to the control group's, reflecting the greater improvement.

Additional Observations

Beyond test scores, we observed qualitative differences in how the students engaged with vocabulary. The experimental group's behavior on the tool showed that the adaptive system kept them challenged but not overwhelmed. The average practice item correctness in the first week was ~55%, but by the final week it rose to ~80%, as the system continually updated and presented remaining weak items. In contrast, the control group's periodic quizzes (administered by the instructor) indicated a more uneven trajectory; some students over-focused on already known words while neglecting harder ones (e.g., several control students consistently skipped certain difficult words in homework despite instructor encouragement).

From the post-study questionnaires, 87% of experimental group students agreed that "the system helped me focus on the vocabulary I needed to study most," and a similar percentage found the immediate feedback useful. Some noted that the AI-based recommendations were "surprisingly accurate" in identifying their weak vocabulary. On the other hand, a few students (10%) mentioned initial confusion or mistrust in letting an algorithm dictate their study list, though they grew more comfortable after seeing progress. In the control group, students often expressed that they relied on the weekly list and that "it was hard to know which words from earlier weeks to review" – an issue the adaptive system inherently addressed by reintroducing words at spaced intervals if a student had struggled with them. These qualitative insights reinforce the conclusion that the ML-

driven approach not only improved outcomes but also addressed common challenges in vocabulary learning such as selecting study targets and maintaining engagement.

Discussion

This study set out to evaluate machine learning approaches for automated vocabulary acquisition in an authentic ESL classroom, and the results provide compelling evidence of the advantages offered by modern AI models, particularly transformer-based deep learning, over traditional methods. In this section, we interpret the findings in light of existing literature, discuss theoretical and pedagogical implications, and consider limitations and future directions.

BERT vs. Traditional ML Models: One key finding was that the transformer-based BERT model achieved markedly better predictive performance (accuracy ~88%) in diagnosing learners' vocabulary knowledge compared to the SVM and Random Forest models (accuracies 75–78%). This outcome is consistent with broader trends in NLP, where transformers have outperformed earlier algorithms on tasks requiring semantic understanding (Devlin et al., 2019). In the context of vocabulary learning, this means BERT can more reliably identify which words a student does not know by considering richer linguistic context and subtle cues. For instance, BERT might infer a student's familiarity with *mitigate* by analyzing errors in sentences about reducing problems, effectively gauging semantic proximity to known words like *reduce* or *solve*. Traditional models, limited to surface features like word frequency, cannot capture such nuances. Our results echo findings by Chen and Meurers (2020) and others who have applied BERT in CALL contexts, demonstrating that incorporating deep linguistic features leads to more accurate adaptation. The precision of the BERT-based recommendations in our study ensured that learners spent time on appropriate targets, which likely contributed to their greater gains. This aligns with *focus-on-form* theories that emphasize timely attention to needed vocabulary (Laufer & Hulstijn, 2001). By precisely targeting gaps, the BERT model operationalized this principle in a personalized manner.

Efficacy of Adaptive Vocabulary Learning: The significantly larger vocabulary gains in the experimental group (roughly 29% improvement vs. 24% in control) provide empirical support for the efficacy of adaptive learning systems in vocabulary acquisition. This finding is in line with prior research on adaptive vocabulary tutors. For example, results from Hsu et al. (2023) indicated that an AI-assisted system (using image recognition and personalization) led to greater word retention and even reduced anxiety, which parallels our observation that students benefited not only in scores but also in confidence. The effect size (~1.0) observed here is notable; in language education research, effects of technology-enhanced interventions on achievement are often moderate (see meta-analysis by Zheng et al., 2022, which found an average $g \approx 0.70$ for AI on language learning outcomes). Several factors in our intervention likely augmented the impact: the fine-grained personalization by the ML model, the immediate corrective feedback, and the integration of the tool into regular coursework (ensuring consistent usage). Our findings reinforce the theoretical perspective of **individualized scaffolding** drawn from Vygotsky's Zone of

Proximal Development (ZPD). The ML system essentially served as a scalable tutor, dynamically adjusting to each student's ZPD for vocabulary – challenging them with words just beyond their current knowledge and offering help at the point of need. This approach is reminiscent of intelligent tutoring systems in other domains that successfully accelerate learning by maintaining optimally challenging tasks (VanLehn, 2011). In vocabulary learning, maintaining that optimal challenge is crucial; too easy and time is wasted, too hard and students disengage. The data suggest our BERT-driven system hit that sweet spot more often than a one-size-fits-all curriculum.

Comparison with Previous Studies: Our study's outcomes dovetail with and extend previous research in several ways. First, consistent with **earlier CALL studies**, we found that even the control group benefited from explicit vocabulary learning (both groups improved significantly). This is no surprise – explicit instruction and practice are known to be effective for vocabulary (Schmitt, 2008). However, the added boost from the ML tool demonstrates how technology can amplify these gains. This resonates with the work of *Godwin-Jones (2018)*, who advocated for "contextualized vocabulary learning" using digital tools. Our BERT model provided rich context for each word (through example sentences and usage-based selection), embodying this principle. The success of the experimental group also mirrors findings in mobile-assisted learning; for example, a study by *Lan, Sung, and Chang (2018)* found that a mobile peer-assisted vocabulary system led to higher vocabulary gains than traditional practice, attributing it to increased personalized engagement. We similarly see personalization as the key driver, taken to a new level by the use of advanced AI.

Second, our results contribute to the growing body of evidence on **AI in language education**. In a broad review, Zawacki-Richter et al. (2019) noted that many AI applications in education showed positive effects on learning achievement, but they called for more domain-specific studies. The present study answers that call in the domain of L2 vocabulary. The magnitude of improvement we observed (roughly 5 more words learned on a list of ~80) might seem modest in absolute terms, but it is quite meaningful when extrapolated to longer courses or larger lexicons. If an AI system helps a learner acquire even 15–20% more words over a semester than they otherwise would, this can cumulate to hundreds of extra words over an academic program – a substantial advantage in language capability. In practical terms, this could mean the difference between a student reaching an advanced vocabulary threshold versus remaining at an intermediate plateau.

Pedagogical implications: The positive results for the ML-enhanced approach have direct implications for language teaching practice and curriculum design. First, they suggest that **integrating AI-driven tools into ESL classrooms is feasible and beneficial**. Our intervention was implemented during normal class hours without replacing any curricular content; it merely changed the mode of vocabulary practice. Teachers can adopt a similar model, using an AI tutor as a supplement to their instruction. Importantly, teacher involvement remains crucial – in our study the instructor guided the process, monitored progress, and provided the communicative

context for using the new vocabulary (through reading and discussion). The AI system thus functioned as an assistant, not a replacement. This aligns with the standpoint of *blended learning*, where technology handles personalized drills and immediate feedback, freeing up teacher time for higher-order activities (Garcia & Benitez, 2021).

Second, the data underscore the value of **targeted review and spaced repetition** that ML systems can facilitate. Many control group students struggled with knowing what to review; some focused on words they liked or found easy, neglecting harder terms. The ML system mitigated this by automatically bringing back words until mastery, embodying a form of adaptive spaced repetition. Teachers may not individually track each student's retention of each word, but a system can do so at scale. This capability could be particularly impactful in large classes, a common situation in many educational contexts, where individualized attention is scarce. By implementing an AI tool, teachers could ensure each student gets a tailored vocabulary learning trajectory, which our results suggest will lead to better outcomes. The **engagement factor** is another pedagogical plus: we observed students in the experimental group treating the ML tool somewhat like a game or personal challenge (especially with gamification elements). This motivated practice is invaluable, since sustained exposure is needed for vocabulary acquisition (Nation, 2013). It is noteworthy that none of the experimental students disengaged or dropped out; on the contrary, many used the tool beyond the required time. This enthusiasm is a stark contrast to the often-reported boredom associated with rote vocabulary study.

Third, our findings encourage curriculum designers to consider **blending data-driven approaches** for assessment. The BERT model's high precision in identifying unknown words can be leveraged for diagnostic testing or formative assessment. For example, instead of a traditional paper pre-test, an AI model could quickly pinpoint a student's weak vocabulary areas by analyzing a short sample of their writing or responses, then recommend personalized word lists or tasks (as demonstrated by systems like *VocabTutor*; cf. Chen & Li, 2010). The success of our model implies that such automatic diagnostics can be quite accurate. This could save class time and allow immediate, continuous adjustment of learning materials – an embodiment of the *assessment-for-learning* paradigm.

Theoretical implications: On a theoretical level, this study reinforces the importance of *input richness and interaction* in vocabulary learning. The experimental group not only saw words in varied contexts via the tool but also *interacted* with them (through quizzes and feedback loops), aligning with interactionist theories. Long's Interaction Hypothesis, while usually applied to conversational interaction, can be extended here: the AI tool created an interactive environment where learners negotiated meaning with the computer (e.g., if they got an item wrong, they received modified input until they got it right). This simulates a kind of negotiation for meaning, albeit with an AI, which appears to aid vocabulary uptake. Additionally, the results relate to *noticing hypothesis* (Schmidt, 1990) – the ML system likely helped learners notice gaps in their

vocabulary knowledge by explicitly quizzing them on those words, thereby priming them for learning when they later encountered the words in readings or lectures. Students in the control group might not have noticed or paid attention to some low-frequency words in the input, whereas the system forced that noticing to occur for experimental students.

The success of reinforcement learning approaches in related studies (e.g., Zhang & Li, 2024) and the trend observed here (though we did not implement RL in the interface, our BERT model's iterative adaptation has some RL-like effects) suggests future L2 vocabulary models might continuously self-improve by learning from student interactions. This resonates with adaptive control of thought – rational (ACT-R) models of learning, which propose that ideal practice schedules can be learned. Our results empirically substantiate that the more a system approximates *optimal practice scheduling*, the better the learning – connecting to cognitive psychology theories of distributed practice and retrieval practice. The BERT model implicitly enforced retrieval practice by re-testing words a student previously got wrong in later sessions (a form of spacing), which is known to strengthen memory traces (Karpicke & Roediger, 2008). In contrast, control group students may have focused more on initial encoding (studying word lists) and less on systematic retrieval practice.

Limitations: While the findings are encouraging, several limitations warrant caution. First, the study duration was relatively short (8 weeks) and involved a modest sample size from one university. Replication over a full semester or academic year, and in different contexts (e.g., secondary schools, EFL settings outside of university), is needed to ensure the generalizability of results. It's possible that novelty effects contributed to engagement with the ML tool; over a longer term, its usage might wane without additional motivational features. Second, our assessment focused on recognition and recall of word meanings in test format. We did not directly measure productive vocabulary use in writing or speaking, which is ultimately the goal. It remains to be seen whether the gains from the AI system translate into better usage of the words in communicative tasks (though anecdotal classroom observations suggested experimental students were indeed using more of the target vocabulary spontaneously). Future research should include productive vocabulary measures or delayed post-tests to check retention durability. Third, the study's design, while controlled, was not a double-blind randomized trial – students knew they were using a new system, which could introduce motivational biases. We attempted to mitigate this by ensuring both groups had tasks to do, but expectancy effects cannot be entirely ruled out. Conducting a crossover design (switching groups mid-way) could strengthen causal claims but was impractical within one semester.

On the technical side, developing and deploying the BERT model required substantial computational resources and expertise. Not all educational institutions have the infrastructure or know-how to implement such models. Thus, while we demonstrate efficacy, there is a question of accessibility and scalability. However, this gap is closing as more user-friendly AI platforms and

pre-trained models become available off-the-shelf. Finally, the SVM and Random Forest baselines, while representative of traditional ML, were not optimized for context (they didn't use the full sentence, only derived features). One could argue that a more advanced baseline (e.g., a deep feed-forward network or an LSTM on word indices) might have performed slightly better. We chose SVM/RF for their transparency and common use in educational data mining; nonetheless, the margin by which BERT surpassed them is so large that the conclusion about transformer superiority is likely robust to baseline choices.

Future directions: This study opens several avenues for future research. One direction is to incorporate **reinforcement learning** more explicitly. For instance, an RL agent could decide not just *which* word to practice, but *when* to review it, optimizing the spacing interval for each student. Combining BERT's state representation (knowledge estimate) with an RL policy could further enhance efficiency – a system could learn an optimal teaching policy through trial and error with multiple students. Early work in this vein (e.g., Xu et al., 2022, using multi-armed bandits for scheduling) has shown promising initial results, and our findings encourage pursuing this line.

Another direction is exploring **large language models (LLMs)** like GPT-4 in vocabulary instruction. While BERT is excellent for understanding and classifying, generative models can create rich, contextual exercises on the fly (e.g., generating a new sentence for a word tailored to the learner's interests). Recent studies (Fang et al., 2023; Kim, 2023) have begun examining ChatGPT for language learning. It would be interesting to compare a generative approach (AI tutor "chatting" with student to teach words) with our predictive approach. Perhaps a hybrid system could leverage BERT for assessment and an LLM for interactive teaching dialogues, combining strengths of both.

Additionally, **broader vocabulary knowledge** aspects like collocations, register, and depth of knowledge (synonyms, antonyms) should be targeted in future ML models. Our system mostly dealt with form-meaning mapping of single words. But knowing a word includes knowing how to use it in context, how it collocates, etc. Future systems could include tasks that help with collocational knowledge, using corpora and AI to provide examples. The transformer architecture can be extended to phrase-level suggestions, which might help learners acquire not just words in isolation but in chunks – aligning with usage-based theories of language acquisition.

Finally, from a research perspective, deploying such systems provides a wealth of **learning analytics data** that can be mined to understand vocabulary acquisition processes. In this study, we collected detailed logs; analysis of these logs (beyond the scope of this article) could reveal learning curves for each word or student, inform models of vocabulary forgetting and retention, and even detect if certain semantic categories are consistently harder for L2 learners (which could feed back into curriculum emphasis). This kind of data-driven insight is a boon to both theory and practice, potentially leading to more effective vocabulary syllabi (e.g., reordering word introduction based on predicted difficulty).

Conclusion

This study demonstrated that machine learning approaches, and particularly transformer-based models like BERT, can significantly enhance vocabulary acquisition in ESL classrooms. We found that a fine-tuned BERT model provided highly accurate predictions of learners' vocabulary needs, enabling an adaptive learning system that led to substantially greater vocabulary gains compared to traditional instruction alone. The integration of supervised and deep learning techniques in a real classroom setting proved not only feasible but pedagogically advantageous, offering students personalized support and instant feedback that aligned with their individual learning gaps. These results contribute to the growing evidence that AI-powered tools can serve as effective allies in language education, supplementing teacher-led instruction with data-driven adaptation and efficiency.

From a practical standpoint, the outcomes suggest that educators and institutions should consider leveraging modern NLP technologies to augment vocabulary teaching. An AI-enhanced approach can ensure students focus on the right words at the right time, a long-standing challenge in vocabulary pedagogy. The transformer model's ability to handle contextual information is particularly valuable for language learning, where context determines meaning. By capturing this, AI systems can expose learners to words in varied, meaningful contexts, fostering deeper acquisition beyond rote memorization. Moreover, the positive student reception in our experiment indicates that, when thoughtfully implemented, such technology can increase learner motivation and autonomy—students felt the system "understood" their difficulties and helped them progress, an empowering experience in language learning.

We acknowledge that implementing these cutting-edge ML solutions in educational contexts comes with challenges, including resource requirements and teacher training. However, as AI becomes more accessible, these barriers are likely to diminish. It will be important for teacher education programs to include basic AI literacy so that future instructors feel comfortable interpreting and guiding AI recommendations, as well as addressing any errors the technology might make. Our study also highlights that the role of teachers remains indispensable: they create the communicative context and ensure that vocabulary learned via AI is integrated into actual language use.

In conclusion, **machine learning approaches, when carefully applied, can act as a catalyst for vocabulary learning**, automating the identification of learner needs and optimizing practice schedules in ways that were previously impractical. This frees up human instructors to focus on communicative practice and strategy training, resulting in a more efficient division of labor. The contributions of this work lie in bridging the gap between NLP advances and language education practice, offering a model for how empirical evaluation can be conducted when introducing AI in the classroom. By sharing detailed methodology and results, we hope to encourage further interdisciplinary collaboration in developing intelligent language learning systems.

Directions for future research include long-term studies to examine retention, expansion to other language skills (e.g., grammar or writing feedback using transformers), and exploring the interplay between human and AI feedback. Additionally, investigating the impact on different learner populations—such as lower proficiency learners or younger students—would be valuable. With reinforcement learning and generative AI on the horizon, the next generation of intelligent vocabulary tutors could become even more interactive and adaptive, possibly engaging in conversational exchanges with learners to teach new words in context (Ebadi & Amini, 2022). As these technologies evolve, it will be critical to maintain a focus on pedagogical soundness and equity of access. Ultimately, the goal is not merely to use flashy AI tools, but to meaningfully enhance language learning and help students reach higher levels of lexical proficiency more effectively. The present study provides encouraging evidence that we are on the right path to achieving that goal by combining the best of human teaching with the best of machine intelligence.

References

Alanzi, A. A., & Taloba, A. I. (2024). *Gamification and deep learning-driven transformer feedback mechanism for adaptive language learning assessment*. Paper presented at the 5th International Conference on Smart Learning Environments (ICSLE 2024). *(Forthcoming conference paper, abstract retrieved from ResearchGate.)*

Basal, A., Yilmaz, S., Tanriverdi, A., & Sari, L. (2016). Effectiveness of mobile applications in vocabulary teaching. *Contemporary Educational Technology, 7*(1), 47–59. https://doi.org/10.30935/cedtech/6162

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL, 27*(1), 4–20. https://doi.org/10.1017/S0958344014000159

Chen, C. M., & Chung, C. J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education, 51*(2), 624–645. https://doi.org/10.1016/j.compedu.2007.06.011

Chen, Y., & Choi, Y. (2021). Incorporating AI into English vocabulary learning: A review of current practices and future directions. *Journal of Educational Technology & Society, 24*(1), 184–197. *(No DOI available.)*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

Dudiak, D., Szabóová, M., & Magyar, J. (2023). Exploring Q-learning in social robots for English–Slovak vocabulary learning. In *Proceedings of the 23rd IEEE International Symposium on*

Computational Intelligence and Informatics (CINTI 2023) (pp. 000027–000032). IEEE. https://doi.org/10.1109/CINTI59972.2023.10382107

Ebadi, S., & Amini, A. (2022). Examining the roles of social presence and human-likeness on Iranian EFL learners' motivation using artificial intelligence technology: A case of CSIEC chatbot. *Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2022.2096638

Godwin-Jones, R. (2018). Contextualized vocabulary learning. *Language Learning & Technology, 22*(3), 1–19. *(No DOI – available at University of Hawai'i ScholarSpace: http://hdl.handle.net/10125/44651.)*

Hsu, T., Chang, C., & Jen, T. (2023). Artificial intelligence image recognition using self-regulation learning strategies: Effects on vocabulary acquisition, learning anxiety, and learning behaviours of English language learners. *Interactive Learning Environments, 32*(6), 3060–3078. https://doi.org/10.1080/10494820.2023.2165508

Küçük, Z., & Solmaz, E. (2021). Language learning and teaching applications of artificial intelligence. *Hacettepe University Journal of Education, 36*(4), 1217–1232. *(No DOI available.)*

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Tsai, Y. L., & Tsai, C. C. (2018). Digital game-based second-language vocabulary learning and conditions of research designs: A meta-analysis study. *Computers & Education, 125*, 345–357. https://doi.org/10.1016/j.compedu.2018.06.020

Wang, Y. (2024). Construction and improvement of an English vocabulary learning model integrating spiking neural networks and convolutional long short-term memory algorithms. *PLoS ONE, 19*(3), e0299425. https://doi.org/10.1371/journal.pone.0299425

Wu, Z., Larson, E. C., Sano, M., Baker, D., Gage, N., & Kamata, A. (2023). Towards scalable vocabulary acquisition assessment with BERT. In *Proceedings of the 10th ACM Conference on Learning@Scale (L@S '23)* (pp. 272–276). ACM. https://doi.org/10.1145/3573051.3596170

Xie, Q., He, Z., & Freeman, G. (2018). Harnessing collective intelligence for L2 vocabulary learning: Examining the influence of a crowdsourcing-based approach. *Computer Assisted Language Learning, 31*(5–6), 617–644. https://doi.org/10.1080/09588221.2018.1441367

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*(39), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zhao, Y., Muhamad, M. M., Mustakim, S. S., Li, W., & Wang, A. (2023). Intelligent recommender systems in mobile-assisted language learning (MALL): A study of BERT-based vocabulary learning. In *Proceedings of the 2023 International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. IEEE. https://doi.org/10.1109/ICMNWC60182.2023.10435740

Zhang, T., & Li, C. (2024). Adaptive English vocabulary recommendation systems: A computational intelligence approach using deep reinforcement learning. In *Proceedings of SPIE Vol. 13550, Sixth International Conference on Education and Training Technologies*. SPIE. https://doi.org/10.1117/12.3059790

Zheng, L., Niu, J., Zhong, L., & Gyasi, J. (2022). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments, 30*(8), 1551–1567. https://doi.org/10.1080/10494820.2021.2015693

Received: 04.25.2025
Revised: 05.02.2025
Accepted: 05.05.2025
Published: 05.06.2025