STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–21.
Published online in International Academic Press (www.IAPress.org)

A Hybrid NLP–LLM Framework for Intelligent Classification of Individual and Collaborative Tasks in Learning Environments

Amel Douar[1,*], Yacine Slimani[2], Fairouz Hadi[1], Adel Alti[1], Heythem Azzouz[3], Abdallah Marouki[3]

[1] *LRSD, Faculty of Sciences, Setif 1 University - Ferhat Abbas (UFAS1), Sétif 19000, Algeria*
[2] *Artificial Intelligence Laboratory, Faculty of Sciences, Setif 1 University - Ferhat Abbas (UFAS1), Sétif 19000, Algeria*
[3] *Department of Computer Science, Faculty of Sciences, Setif 1 University - Ferhat Abbas (UFAS1), Sétif 19000, Algeria*

Abstract Natural Language Processing (NLP) plays a crucial role in automating text classification tasks, particularly in the context of education and scientific experimentation. However, the classification of practical tasks, especially distinguishing between individual and collaborative worksheets in laboratory sessions, remains an open challenge. This work extends our previous research on collaborative virtual laboratories by introducing an intelligent classification model that automatically determines task modality from worksheet specifications, enabling adaptive pedagogical orchestration. This article investigates the performance of Recurrent Neural Networks (RNNs) and their variants, including LSTM, GRU, and bidirectional models, in addressing this issue. For contextual benchmarking, selected transformer-based models were also evaluated to compare performance and computational trade-offs. We aim to determine which architecture best balances classification accuracy and computational efficiency. RNN-based models were selected due to their efficiency in sequential text modeling and their suitability for real-time deployment in educational platforms. To enhance data diversity and improve model generalization, a data augmentation step leveraging a Large Language Model (LLM) was employed to synthetically enrich the training corpus while preserving semantic consistency. Multiple RNN architectures were trained and evaluated on a domain-specific dataset of chemistry-related worksheets, using both original and LLM-augmented data. Performance was assessed using accuracy, precision, recall, and F1-score metrics. Among the models, LSTM achieved the highest accuracy (95.02%), demonstrating superior classification capabilities. GRU models offered competitive performance with lower computational costs, while bidirectional architectures improved contextual retention but exhibited variable performance depending on the dataset features. Although LLM-based data augmentation marginally enhanced model efficiency, the dataset's inherent simplicity ensured strong baseline performance across all models. Importantly, classification errors do not compromise learning outcomes but only affect execution modality, making the approach robust for real educational deployment. Overall, the findings highlight the efficiency of deep learning models in classifying practical educational tasks and underscore the potential of LLM-assisted augmentation to enhance adaptive learning environments.

Keywords Deep Learning (DL), Natural Language Processing (NLP), Recurrent Neural Networks (RNNs), Large Language Models (LLM), Data Augmentation, Task Classification, Educational Technology.

AMS 2010 subject classifications 68T50, 68T01, 68P20.

DOI: 10.19139/soic-2310-5070-3268

1. Introduction

Deep Learning (DL) is a branch of Artificial Intelligence (AI) that enables machines to learn from large datasets and accomplish complex tasks. By combining principles from computer science, mathematics, and neuroscience, it develops models capable of understanding, interpreting, and generating information [1]. This approach effectively bridges the gap between human cognition and computational power, revolutionizing fields such as computer vision,

[*] Correspondence to: Amel Douar (Email: douaramel@gmail.com)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

speech recognition, and Natural Language Processing (NLP) [2, 3].

Among these, NLP is a cornerstone of AI, enabling computers to understand and generate human language with remarkable accuracy. Integrating computer science, linguistics, and AI, it tackles the inherent complexities of human communication through syntactic and semantic analysis, discourse processing, and text generation [4, 5].

The core of NLP lies in the development of sophisticated algorithms and DL models capable of processing and analyzing large amounts of unstructured textual data. These models leverage neural networks and transformer architectures to extract semantic meaning, detect patterns, and enhance comprehension. As a result, NLP has become an indispensable tool in various real-world applications, including machine translation, sentiment analysis, speech recognition, information retrieval, conversational AI, automated summarization, and multilingual systems [6, 7, 8].

Text classification is an important NLP task supporting applications such as sentiment analysis, spam detection, and topic categorization. Traditional machine learning methods, such as Naïve Bayes and SVM, rely on handcrafted features, limiting their effectiveness. DL has greatly improved classification accuracy using architectures such as CNNs, RNNs, and transformer-based models (e.g., BERT and GPT) that autonomously learn hierarchical representations. However, these models require large amounts of high-quality labeled data, which is often expensive and labor-intensive to collect. These drive researchers toward semi-supervised, transfer, and self-supervised learning techniques to reduce extensive labeled datasets while maintaining performance [9, 10].

NLP has brought significant advances across various domains, including education, where its integration is transforming traditional teaching and learning methodologies. By leveraging NLP-driven technologies, educational environments benefit from intelligent tools that support automated assessment, personalized learning, and interactive tutoring systems. These applications enhance student engagement, streamline content delivery, and offer instructors valuable insights into learner progress. As NLP technologies evolve, they promise to further reshape the educational landscapes [11, 12].

Neural language models have significantly evolved, revolutionizing the field of NLP. Among the most influential models are BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), both of which have redefined the capabilities of modern NLP systems. Built on the Transformer architecture, these models effectively capture long-range dependencies and process variable-length textual sequences using self-attention mechanisms and deep contextual embeddings. They have substantially improved performance in NLP tasks, including machine translation, text summarization, and sentiment analysis, paving the way for large-scale language modeling, demonstrating the transformative impact of deep learning in language understanding [13, 14].

Building on these advancements, NLP techniques can be applied beyond conventional domains to address challenges in education[15]. Practical work sessions, which play a crucial role in developing problem-solving and collaboration skills, can greatly benefit from NLP-driven automation and intelligent task management. A key challenge for instructors lies in effectively assigning students to tasks, whether individual or collaborative. Traditionally, this process is carried out manually, relying on instructor expertise to distinguish between tasks that require independent effort and those that benefit from teamwork. However, this manual approach is time-consuming, prone to inconsistencies, and difficult to scale in large classroom settings.

To address this issue, we leverage advances in NLP, DL, and Large Language Models (LLMs) to automatically classify laboratory tasks extracted from textual descriptions of practical work worksheets. By accurately identifying whether a task is individual or collaborative, the system assists instructors in assigning students more effectively, enhancing classroom management, collaboration, and improving overall learning experience. While existing studies highlight the potential of NLP and Deep Learning in enhancing education, most of them focus on adaptive feedback, language support, or general e-learning applications. To the best of our knowledge, no prior work has specifically explored the automatic classification of practical learning tasks in individual and collaborative contexts. This gap motivates our work, where we propose a novel classification framework that combines NLP, Deep Learning, and RNN architectures to enhance task allocation and support adaptive, personalized learning environments.

This work extends our previous research on collaborative virtual laboratories, in which we developed a web-based immersive environment enabling students to conduct experimental practical work interactively [16]. Although the platform supported collaborative experimentation, the identification of task modality (individual versus collaborative) relied on manual instructor intervention, thereby limiting scalability and real-time adaptability. The present study addresses this limitation by introducing an AI-driven, NLP-based classification model that can automatically determine whether a practical task should be performed individually or collaboratively based on worksheet specifications. This extension represents a significant step towards adaptive virtual laboratories with autonomous pedagogical orchestration, enabling intelligent task allocation and seamless integration with Learning. Management Systems (LMS).

The main contributions of this study are:

1. Designing a domain-specific dataset of chemistry practical tasks labeled as individual or collaborative, using structured data generation, meticulous annotation, verification, and data augmentation techniques (e.g., synonym replacement, random insertion, deletion) to enhance model robustness and generalization.
2. Applying various data augmentation techniques, including synonym replacement, random insertion, and deletion, to expand the dataset, enhance model robustness, and improve generalization on unseen task descriptions.
3. Developing RNN-based classifiers, including RNN, LSTM, GRU, and their bidirectional variants, to automatically classify task descriptions.
4. Evaluating and comparing model performance using accuracy, precision, recall, and F1-score to determine the most effective deep learning model for classifying long scientific experimental texts.
5. Conducting a comparative assessment with transformer-based models to contextualize RNN performance in terms of accuracy and resource requirements.

The remainder of this paper is organized as follows: Section 2 reviews related work on NLP and DL for text classification in the educational domain. Section 3 describes the proposed workflow, including dataset creation, preprocessing, and model architecture. Section 4 details the experimental setup and performance evaluation of RNN-based models on both original and augmented datasets, and we discuss key findings, emphasizing the superior performance of the LSTM model ,the benefits of data augmentation and a comparative analysis with transformer-based model. Finally, Section 5 concludes the paper and suggests future research directions, including the integration of advanced transformer-based model and optimization strategies.

2. Related Work

The rapid advancement of Artificial Intelligence (AI) and NLP has led to transformative developments across various fields, including education and experimental sciences. This section reviews key studies and organizes related work into four main categories:

1. NLP in education,
2. Foundational architectures and models,
3. AI/ML for adaptive learning and analytics,
4. NLP-based classification and domain-specific applications.

2.1. NLP in education

Fuchs [13] examines opportunities and challenges of NLP models such as ChatGPT and Google Bard in higher education. The study highlights their potential to enhance student learning through personalized learning experiences, automation, and real-time support. It also discusses their role in facilitating academic engagement, particularly for non-English languages that lack adequate support in traditional learning management systems. However, concerns remain regarding accuracy, bias in training data, and the risk of over-reliance on AI tools, which may affect the development of critical thinking skills. The study highlights the need for ethical guidelines and structured implementation.

Similarly, Biester and Wu [17] propose a series of homework assignments linked to in-class worksheets to help NLP students understand and debug their code. These worksheets allow students to verify their comprehension of algorithms on paper before coding and later serve as test cases during programming, enabling students to compare their results with expected outcomes. This pedagogical approach reinforces conceptual understanding and improves learning and debugging efficiency. However, while pedagogically valuable, this work focuses on general NLP learning and debugging rather than the automatic classification of practical-work worksheets, which remains an underexplored challenge addressed by our proposed framework.

The study by Li et al. [18] examined the role of NLP technology in e-learning, focusing on intelligent platforms for computer-mediated education. It reviews existing methods already integrated into learning systems and emerging techniques not yet deployed in real-world software solutions. Despite certain limitations, the study highlights the significant contribution of language technology to enhancing computer-assisted teaching and learning, particularly for underrepresented languages. A case study on Czech e-learning materials illustrates these concepts in practice.

2.2. *Foundational Architectures and Models*

The Transformer model proposed by Vaswani et al. [14], represents a major shift in sequence modeling by replacing recurrence with self-attention mechanisms. Unlike RNN-based and CNN-based architectures, it processes input sequences entirely through self-attention, significantly improving parallelization and computational efficiency. Its effectiveness was demonstrated in WMT 2014 translation benchmarks (English-to-German and English-to-French) with reduced training time. By leveraging multi-head attention and positional encodings, the Transformer captures long-range dependencies more effectively, establishing the foundation for modern attention-based NLP architecture.

Authors in [19] explored the application of ChatGPT and BERT models in mechanical engineering, focusing on tasks such as technical documentation analysis, fault diagnosis, and customer service. This work provides their architecture, advantages, and limitations, while evaluating their performance in logical inference, paraphrasing, and text similarity. It also discusses challenges faced by ChatGPT in natural language understanding and suggests hybrid approaches like the Multidimensional Informational-Variable Approach (Mivar) method to enhance performance.

Recent studies have further demonstrated the effectiveness of hybrid deep learning and NLP models. Rustad et al. [20] conducted a comprehensive systematic literature review on named-entity recognition, highlighting the synergy between statistical modelling and transformer-based architectures. Similarly, Ferdous et al. [21] explored sentiment analysis in the machine-learning era, combining hybrid architectures to improve classification under limited-data conditions. These contributions illustrate how hybrid NLP–deep learning frameworks can advance classification performance across domains where data are scarce.

2.3. *AI and ML for Adaptive Learning*

Recent advancements in AI and Machine Learning (ML) are reshaping e-learning platforms and enabling adaptive learning systems. A study conducted by [22] examines the utilization, benefits, and challenges of AI/ML in this domain through a literature review of works published since 2010. The findings show how AI/ML algorithms enhance student engagement, retention, and performance by personalizing learning experiences and optimizing learning paths. Despite challenges such as data privacy concerns, AI-driven e-learning systems hold great potential for transforming education by meeting individual learner needs more effectively.

Similarly, authors in [23] present a comprehensive study of ML in online education by synthesizing findings from 1961 to 2022. It explores methodologies for developing learning analytics tools, identifies key data resources, and analyzes the diversity of available datasets. The study evaluates different ML and DL techniques, providing valuable insights and methodological guidance for advancing research and practice in data-driven education.

2.4. NLP-based Classification and Domain-Specific Applications

Hasanen et al. [24] compare the performance and behavior of CNN, LSTM, and GRU for classification and prediction tasks across several datasets, highlighting trade-offs in accuracy, training time and resource usage. Their findings guide model selection for text classification problems relevant to educational contexts.

In [25], the authors demonstrate that exam-generation and assessment tasks in education can benefit from RNN and LSTM architectures, particularly for capturing relationships among exam questions and student performance patterns. Although the study focuses on exam papers rather than practical work worksheets, it provides valuable methodological insights for designing classification systems in educational settings.

Recent surveys and benchmarks on NLP for education and text classification highlight the rapid growth of methods and application domains, while also emphasizing domain-specific challenges for scientific text. Lan et al. [25] provide a taxonomy of educational NLP tasks and identify text classification as a core task with unique constraints in educational settings. Prior work on automated assessment of collaborative engagement demonstrates the feasibility of detecting group interactions and collaboration interactions from textual and multimodal data, yet practical work classification remains underexplored. Moreover, domain-specific efforts such as ChemNLP tools Choudhary and Kelley [26] show the feasibility of extracting structured information from experimental texts, supporting the relevance of our chemistry-focused dataset.

2.5. Discussions

Despite extensive research on NLP applied to education, few studies have specifically addressed the automatic classification of practical educational tasks, such as worksheets used for scientific practical work. Most existing research focuses on general NLP applications or specific educational tasks without directly tackling the problem of classifying practical tasks based on their nature (individual vs. collaborative). This gap in literature emphasizes the originality and significance of our contribution. By developing an approach that leverages NLP and Large Language Models (LLMs) to classify practical educational tasks, we propose a novel NLP- and LLM-driven framework for classifying practical educational tasks. Furthermore, by developing a domain-specific annotated dataset for this task and implementing deep learning architectures, we provide both methodological advancement and a valuable resource to the scientific community, filling this gap and opening new research directions in the field.

3. Proposed Hybrid NLP-LLM-Based Task Classification System

3.1. System Design

The automatic classification of individual and collaborative practical works aims to distinguish the nature of students' activities according to their working mode. A practical work is considered collaborative when it involves a division of tasks among several learners working together toward a common goal, whereas it is individual when a single student can complete it independently. In this context, implementing an automatic classification system makes it possible to analyze the content of practical works, identify indicators of collaboration or autonomy, and differentiate between collective and individual tasks. This approach enables a faster and fairer evaluation, while providing teachers with a clear understanding of each student's or group's working mode. It therefore contributes to improving pedagogical management and monitoring in online or hybrid learning environments.

This section presents a structured framework for classifying chemistry practical tasks into individual or collaborative categories using Recurrent Neural Network (RNN) models. As shown in Fig. 1, the proposed approach consists of two main phases: dataset generation and task classification. In the first phase, a high-quality, domain-specific dataset was developed from real-world practical worksheets provided by the Chemistry Department of our university. This process involved analyzing task structure, generating additional entries using prompt engineering with ChatGPT-4, and performing automated consistency checks. The final dataset, containing over $1,400$ structured documents, was further augmented to enhance diversity and model robustness. In the second phase, we focused on classifying these tasks using RNN-based architectures. We carried out thorough data preprocessing, including normalization, tokenization, and lemmatization, to prepare the text for sequential learning. The dataset was then split into training, validation, and testing sets to ensure balanced evaluation. RNN models, including LSTM, GRU, and their bidirectional variants, were employed for their ability to capture temporal dependencies in text. Model performance was assessed using standard metrics: accuracy, precision, recall, and F1-score, ensuring a rigorous and interpretable evaluation of the proposed classification system.

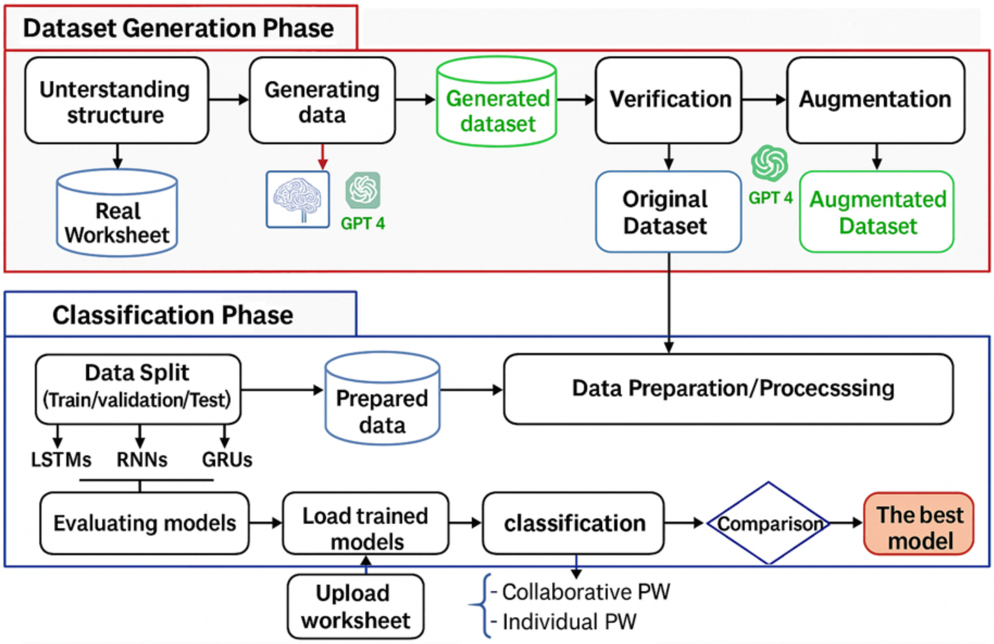


Figure 1. The Workflow of the Proposed Approach.

3.2. *Chemistry Dataset Construction*

The construction of a suitable dataset for this study presented several challenges. Initially, the absence of an existing corpus specifically designed to distinguish between individual and collaborative chemistry tasks, combined with the limited availability of domain experts for data labeling, made it difficult to ensure contextual relevance and accuracy. This limitation underlined the need for expert-informed guidelines or automated methods for dataset construction in domain-specific educational contexts. To overcome these challenges, we collaborated with the Chemistry Department of our institution to collect practical work descriptions. However, the available material was limited in scope and predominantly individual in nature, rendering it insufficient for training deep learning models.

Consequently, we developed a dedicated dataset named AMCDPT [27](Automatic Multiclass Chemistry Dataset for Practical Tasks) [†] through a structured and systematic methodology tailored to the requirements of this research. The process encompassed the collection of academic materials, data generation using advanced NLP and LLM tools, and rigorous verification to ensure label consistency and annotation quality. The resulting dataset serves as a robust foundation for training and evaluating binary text classification models aimed at distinguishing between individual and collaborative chemistry tasks.

[†] The chemistry dataset is publicly available on Kaggle at https://www.kaggle.com/datasets/pfe202419/amcdpt-dataset

3.2.1. Dataset Description and Characteristics: The AMCDPT dataset consists of structured descriptions of experimental chemistry tasks, each written in concise, domain-specific language. An in-depth analysis of the data reveals several distinctive characteristics that guided the model design and training strategy:

- **Moderate text length:** Each description contains an average of approximately 32 words, with a maximum of 104, enabling efficient model learning without information loss due to sequence truncation.
- **Structured and coherent organization:** The descriptions follow a logical procedural format describing experimental steps, facilitating the ability of sequential models to capture contextual dependencies.
- **Specialized vocabulary:** The use of precise scientific terminology enhances model generalization while reducing linguistic ambiguity and noise.
- **Low lexical diversity:** The use of controlled vocabulary simplifies the model learning process by minimizing variance and improving token consistency.
- **Minimal textual noise:** As the data originates from academic sources, grammatical and typographical errors are rare, allowing the model to focus on meaningful linguistic and structural patterns.

These characteristics collectively make the AMCDPT dataset particularly suitable for training Recurrent Neural Network (RNN) architectures such as LSTM and GRU, which excel at modeling sequential dependencies in structured textual data.

3.2.2. Structural Analysis of Practical Works: Based on 400 practical tasks worksheets provided by the Chemistry Department, the structural analysis of these materials was conducted to guide the organization of the dataset. This analysis helped us design a dataset specifically tailored for binary text classification, distinguishing between Collaborative and Individual tasks. The final dataset includes over $1,400$ detailed documents, each representing a specific chemical practical task. Fig. 2 presents key attributes of dataset, including a unique identifier (ID), task name, required laboratory glassware, involved products, and a step-by-step procedure description.

The classification label (**Collaborative** or **Individual**) serves as the target variable for training the model. These structured elements were selected to capture the complexity and collaborative dimension inherent in laboratory-based educational tasks.

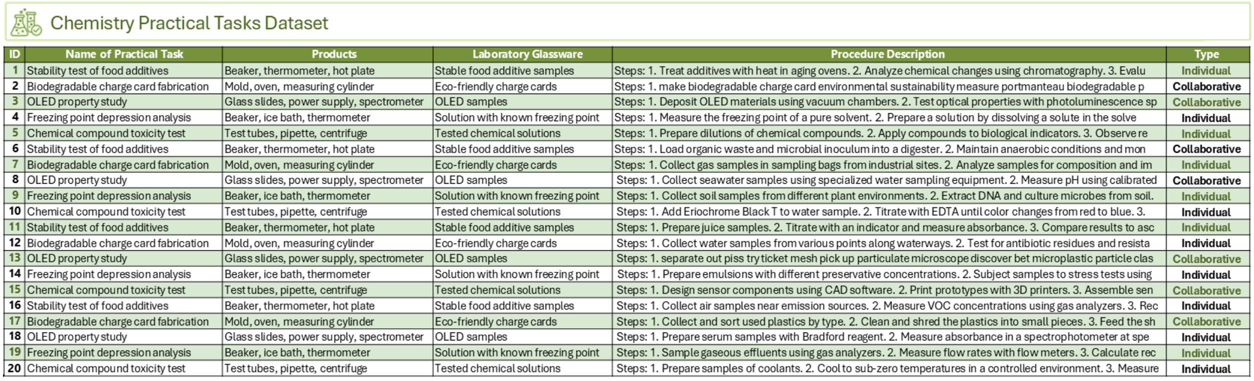
3.2.3. LLM-Based Dataset Generation: To achieve high-quality data generation, prompt engineering techniques were applied to guide ChatGPT-4 in producing structured and semantically consistent outputs. This approach involves formulating clear and detailed prompts that provide sufficient context, define the model's role (e.g., chemistry teacher), and specify the desired data structure. In our study, ChatGPT-4 was instructed to generate a customized dataset of approximately 1,700 practical chemistry tasks using predefined fields derived from Chemistry Department task sheets: task ID, task name, required laboratory glassware, products, procedures, and task type (collaborative or individual).
Besides, ChatGPT-4 has strong capability in understanding and generating domain-specific terminology related to scientific and practical laboratory tasks (represented by chemical practical activities) to enhance well-structured task descriptions and preserve semantic coherence.

3.2.4. Data Verification and Improvement: After completing the dataset generation with ChatGPT-4, we enter the rigorous verification step to refine its processing methods and enhance the subsequent performance. Aligned with its objectives, constructing a high-quality dataset, improving the accuracy of labels assigned to each task this step re-submit dataset to ChatGPT without the assigned labels, and the model is instructed to classify each task as either collaborative or individual. The newly generated classifications were systematically compared with the original labels to evaluate consistency. Tasks that retained the same classification were marked as verified, whereas those with inconsistent labels were excluded to maintain dataset integrity. This verification and sorting process was carefully managed in Google Sheets, facilitating efficient tracking and validation. After verification, the final exported CSV file contained 1405 validated practical works, ready for further analysis and model training.

3.2.5. Data augmentation: was used to expand the dataset and enhance the robustness and generalization of RNN models, especially for sequential learning tasks. Three techniques were applied:

1. synonym replacement, which substitutes words with their synonyms using WordNet to maintain meaning,
2. random insertion, which adds words at random positions to test the model's ability to handle sequence disruptions,
3. random deletion, which removes words probabilistically to encourage the model to infer meaning from limited context.

After applying these augmentation techniques, the dataset was expanded from 1405 to 7885 practical tasks. All the newly generated entries were semantically validated to ensure that the added tasks remained true to the original practical works and scientifically coherent.These augmentation strategies increase data diversity and improve the model's ability to effectively learn temporal dependencies and linguistic variability.



Chemistry Practical Tasks Dataset

ID	Name of Practical Task	Products	Laboratory Glassware	Procedure Description	Type
1	Stability test of food additives	Beaker, thermometer, hot plate	Stable food additive samples	Steps: 1. Treat additives with heat in aging ovens. 2. Analyze chemical changes using chromatography. 3. Evalu	Individual
2	Biodegradable charge card fabrication	Mold, oven, measuring cylinder	Eco-friendly charge cards	Steps: 1. make biodegradable charge card environmental sustainability measure portmanteau biodegradable p	Collaborative
3	OLED property study	Glass slides, power supply, spectrometer	OLED samples	Steps: 1. Deposit OLED materials using vacuum chambers. 2. Test optical properties with photoluminescence sp	Collaborative
4	Freezing point depression analysis	Beaker, ice bath, thermometer	Solution with known freezing point	Steps: 1. Measure the freezing point of a pure solvent. 2. Prepare a solution by dissolving a solute in the solve	Individual
5	Chemical compound toxicity test	Test tubes, pipette, centrifuge	Tested chemical solutions	Steps: 1. Prepare dilutions of chemical compounds. 2. Apply compounds to biological indicators. 3. Observe re	Individual
6	Stability test of food additives	Beaker, thermometer, hot plate	Stable food additive samples	Steps: 1. Load organic waste and microbial inoculum into a digester. 2. Maintain anaerobic conditions and mon	Collaborative
7	Biodegradable charge card fabrication	Mold, oven, measuring cylinder	Eco-friendly charge cards	Steps: 1. Collect gas samples in sampling bags from industrial sites. 2. Analyze samples for composition and im	Individual
8	OLED property study	Glass slides, power supply, spectrometer	OLED samples	Steps: 1. Collect seawater samples using specialized water sampling equipment. 2. Measure pH using calibrated	Collaborative
9	Freezing point depression analysis	Beaker, ice bath, thermometer	Solution with known freezing point	Steps: 1. Collect soil samples from different plant environments. 2. Extract DNA and culture microbes from soil.	Individual
10	Chemical compound toxicity test	Test tubes, pipette, centrifuge	Tested chemical solutions	Steps: 1. Add Eriochrome Black T to water sample. 2. Titrate with EDTA until color changes from red to blue. 3.	Individual
11	Stability test of food additives	Beaker, thermometer, hot plate	Stable food additive samples	Steps: 1. Prepare juice samples. 2. Titrate with an indicator and measure absorbance. 3. Compare results to asc	Individual
12	Biodegradable charge card fabrication	Mold, oven, measuring cylinder	Eco-friendly charge cards	Steps: 1. Collect water samples from various points along waterways. 2. Test for antibiotic residues and resista	Individual
13	OLED property study	Glass slides, power supply, spectrometer	OLED samples	Steps: 1. separate out piss try ticket mesh pick up particulate microscope discover bet microplastic particle clas	Collaborative
14	Freezing point depression analysis	Beaker, ice bath, thermometer	Solution with known freezing point	Steps: 1. Prepare emulsions with different preservative concentrations. 2. Subject samples to stress tests using	Individual
15	Chemical compound toxicity test	Test tubes, pipette, centrifuge	Tested chemical solutions	Steps: 1. Design sensor components using CAD software. 2. Print prototypes with 3D printers. 3. Assemble sen	Collaborative
16	Stability test of food additives	Beaker, thermometer, hot plate	Stable food additive samples	Steps: 1. Collect air samples near emission sources. 2. Measure VOC concentrations using gas analyzers. 3. Rec	Individual
17	Biodegradable charge card fabrication	Mold, oven, measuring cylinder	Eco-friendly charge cards	Steps: 1. Collect and sort used plastics by type. 2. Clean and shred the plastics into small pieces. 3. Feed the sh	Collaborative
18	OLED property study	Glass slides, power supply, spectrometer	OLED samples	Steps: 1. Prepare serum samples with Bradford reagent. 2. Measure absorbance in a spectrophotometer at spe	Individual
19	Freezing point depression analysis	Beaker, ice bath, thermometer	Solution with known freezing point	Steps: 1. Sample gaseous effluents using gas analyzers. 2. Measure flow rates with flow meters. 3. Calculate rec	Individual
20	Chemical compound toxicity test	Test tubes, pipette, centrifuge	Tested chemical solutions	Steps: 1. Prepare samples of coolants. 2. Cool to sub-zero temperatures in a controlled environment. 3. Measure	Individual

Figure 2. List of data properties and items.

3.3. *Practical Work Task Classification*

3.3.1. Data Preparation and Preprocessing The data preparation phase involved a careful curation of the dataset to ensure that only the most relevant and informative data records were retained for training. This filtering process ensured that the RNN model would be exposed to high-quality, meaningful input, while irrelevant or redundant data was removed. The dataset was efficiently managed and structured using Python and the panda's library, establishing a solid foundation for the subsequent stages of preprocessing and model training.

During preprocessing, the raw text data was systematically cleaned and standardized to improve model performance. The process included lowercasing, removal of extra white spaces, punctuation, numerals, and special characters, followed by tokenization. A comprehensive list of 571 stop words was used to eliminate common terms that carry minimal semantic weight. Finally, lemmatization was applied to reduce words to their root forms, ensuring lexical consistency. Together, these preprocessing steps minimize textual noise and variability, allowing the RNN model to focus on essential linguistic patterns relevant to learning task.

3.3.2. Dataset Splitting To ensure balanced training and reliable evaluation, the dataset was divided into three subsets: 70% for training, 15% for testing, and 15% for validation. This split helps reduce overfitting and ensures that models generalize well to new data. The training subset facilitated model learning, the validation subset optimized hyperparameter tuning, and the test subset provided an unbiased measure of model performance.

3.3.3. Application models: RNN and Variants Recurrent Neural Networks (RNNs) were chosen for classifying chemistry experiment descriptions into individual or collaborative tasks due to their proven effectiveness in processing sequential data. These experiment task descriptions are inherently sequential, with each step depending on the preceding and subsequent ones for full contextual understanding. RNNs, especially advanced variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, are capable of capturing both short- and long-term dependencies within such sequences. This capability makes them well-suited for handling texts averaging 32 words and extending up to 104 words. Furthermore, bidirectional RNNs enhance performance by processing text in both directions, allowing the model to leverage contextual information from both past and future words. Their proven effectiveness across a wide range of natural language processing tasks, combined with architectural flexibility in architecture, including Simple-RNN, LSTM, GRU, and their bidirectional forms, and their robustness in adapting to scientific language and moderate-length texts, make them an ideal choice for this classification task.

In addition, we apply transformer-based architecture (e.g., BERT and RoBERTa) to enhance classification performance. The main motivation for adopting transformer models lies in their strong capability to model contextual dependencies within scientific textual content, such as worksheet descriptions of practical tasks. Their attention-based mechanism enables effective semantic encoding and robust representation learning, improving embedding quality and reducing information loss. This results in more accurate and consistent classification while maintaining acceptable computational cost and processing efficiency.

3.3.4. Evaluation of classification models To assess the effectiveness of the classification models, four standard evaluation metrics were used: accuracy, precision, recall, and the F1-score.

Accuracy measures the proportion of correct predictions over the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

where TP, TN, FP, and FN true positives, true negatives, false positives, and false negatives, respectively.

Precision quantifies the accuracy of positive predictions, indicating the proportion of the predicted positives among all predicted positives.

$$Precision = \frac{TP}{TP + FN} \tag{2}$$

Recall, on the other hand, reflects the model's capability to identify all actual positive cases within the dataset.

$$Recall = \frac{TP}{TP + FP} \tag{3}$$

F1-score offers a harmonic mean between precision and recall, providing a more balanced view when there's an uneven class distribution. These metrics are defined as follows:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

4. Experimental Evaluation And Results

This section presents a comprehensive evaluation of various RNN models using both original data and augmented datasets to assess the impact of architectural optimization and data augmentation on classification performance.

4.1. Parameter settings

The models were trained for 8 epochs with batch size of 256. The CNN model utilized 32 filters with a kernel size of 7, an activation function of ReLU and a pooling size of 2. The Simple RNN, LSTM, Bi-LSTM, GRU, and Bi-GRU models were each configured with 256 units, including a 20% dropout and recurrent dropout to reduce overfitting.

The performance of the models was evaluated and compared using both the original data and augmented data to analyze the impact of data augmentation on classification accuracy. We utilized grid search and random search to optimize key hyperparameters, including learning rate, which ranged from 0.001 to 0.1, and the batch size across values 16, 32, 64, and 128. The number of epochs was evaluated between 10 and 100, while the dropout rate was

varied between 0.2 and 0.5 to ensure model regularization. Additionally, the number of units in the hidden layers was tested at 32, 64, 128, and 256 to determine the most effective configuration for sequential text processing.
All models were implemented in TensorFlow/Keras and executed on Google Colab, a cloud-based computational environment for experimental evaluation. For classification, worksheet content is converted into embedding vectors using Word2Vec to capture semantic relationships between domain-specific terms. Model optimization was performed using the Adam optimizer, with systematic hyperparameter tuning and validation-based early stopping to ensure stable convergence and prevent overfitting.

4.2. Criteria for Selecting Best Parameters

The selection of best parameters is guided by three criteria:

- **Validation accuracy:** monitored continuously to find the highest performance on unseen data,
- **Early stopping:** Implemented to prevent overfitting by halting training when validation loss does not improve,
- **Final model evaluation:** Analyzed test accuracy and classification reports, including precision, recall, and F1-score.

4.3. Impact of Optimization on RNN Performance

Figures 3, 4, and 5 illustrate the effects of optimization techniques and data augmentation on RNN models. Specifically, we examine baseline performance, improvements following architectural tuning, and gains from augmented training data. The RNN model shows better data diversity and enhanced accuracy across epochs from augmented data. However, the optimizer now follows a smoother gradient path, leading to stable convergence.

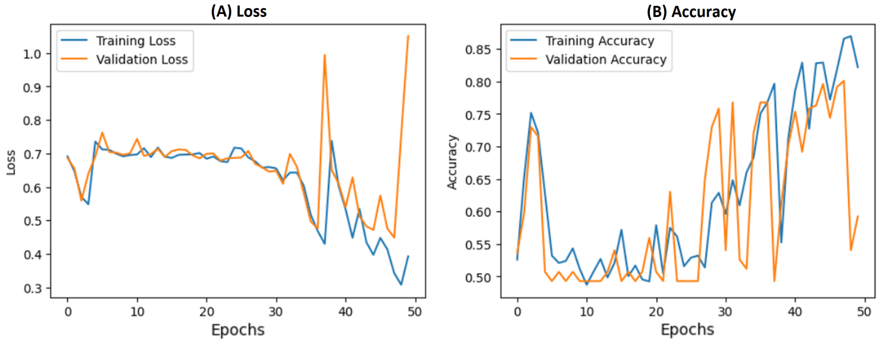


Figure 3. RNN model performance before optimization (A) Training and validation loss (B) Training and validation accuracy.

4.4. Impact of Data Augmentation on Individual RNN Variants Performance

Figures 6 to 10 illustrate results for each recurrent neural network (RNN) variant, namely Simple-RNN, long short-term memory (LSTM), and gated recurrent unit (GRU), with their bidirectional extensions. Each model is evaluated on both the original and augmented datasets to examine the improvements in contextual learning and classification accuracy. Finally, a comparative analysis is presented for Bi-Simple-RNN, LSTM, Bi-LSTM, GRU, and Bi-GRU, highlighting their performance differences between the original and augmented data.
As shown in Fig. 6, the Bi-SimpleRNN trained on original data shows slower convergence and mild overfitting, with a noticeable gap between training and validation curves. After data augmentation, the model achieves faster convergence, closer training–validation alignment, and higher accuracy, indicating improved generalization and overall robustness.

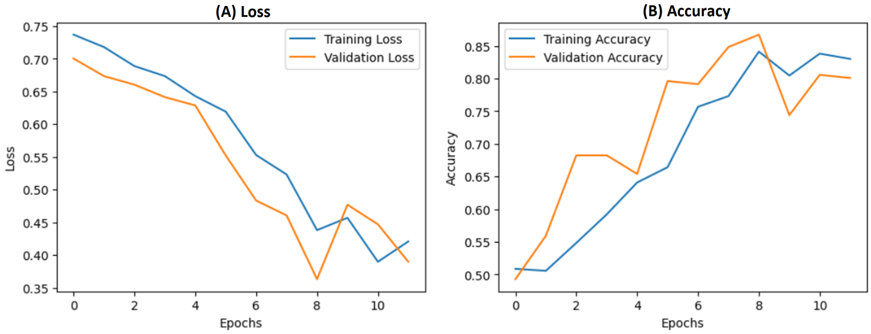


Figure 4. RNN model performance after optimization (A) Training and validation loss (B) Training and validation accuracy.

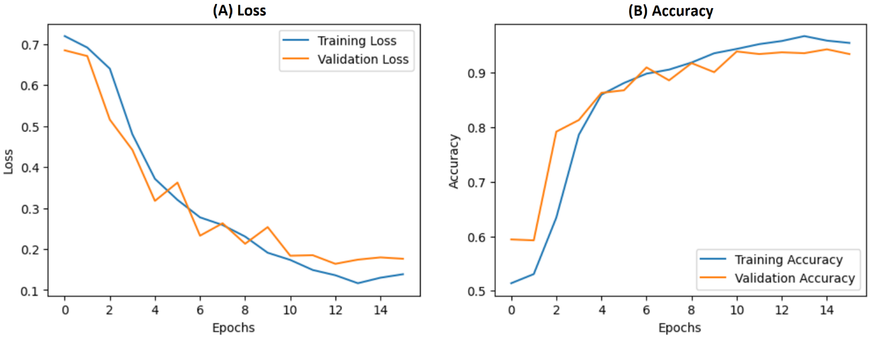


Figure 5. RNN model performance using augmented data (A) Training and validation loss (B) Training and validation accuracy.

In Fig. 7, the LSTM model trained on original data shows gradual learning with mild fluctuations in validation performance, suggesting limited generalization. After data augmentation, the model achieves faster convergence, lower loss, and higher, more stable accuracy, demonstrating improved generalization and better adaptation to diverse input patterns.

In Fig. 8, the Bi-LSTM model trained on original data exhibits steady improvement but with noticeable fluctuations in validation accuracy, indicating partial overfitting. After data augmentation, the model converges faster and more smoothly, with lower loss and closely aligned training–validation accuracy, confirming enhanced learning stability and stronger generalization capability.

In Fig. 9, the GRU model trained on original data shows gradual convergence with minor fluctuations in validation accuracy, indicating limited generalization.

After data augmentation in Fig. 10, the model achieves faster and smoother convergence, lower loss, and more stable accuracy, demonstrating improved learning efficiency, reduced overfitting, and stronger generalization to unseen data.

From above results, several key observations can be summarized as follows:

- LSTM and GRU achieved the best performances, with precision scores of 94.3% and 92.9%, respectively.
- Bidirectional models (Bi-LSTM, Bi-GRU) slightly outperformed their standard versions, benefiting from enhanced contextual understanding by processing input sequences in both forward and backward directions.

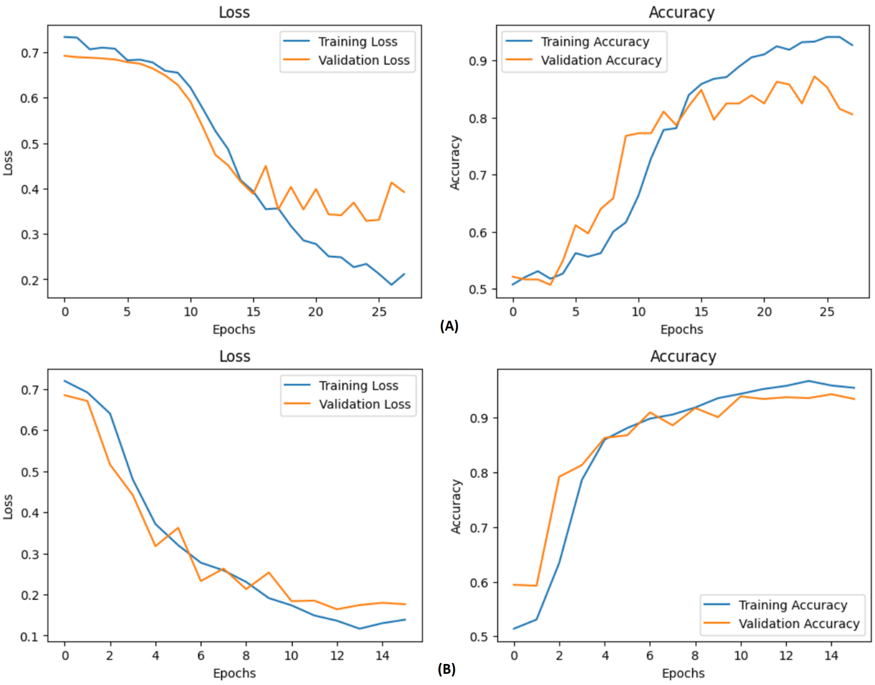


Figure 6. Bi-SimpleRNN model performance on original vs. augmented data; (A) Original data (B) Augmented data.

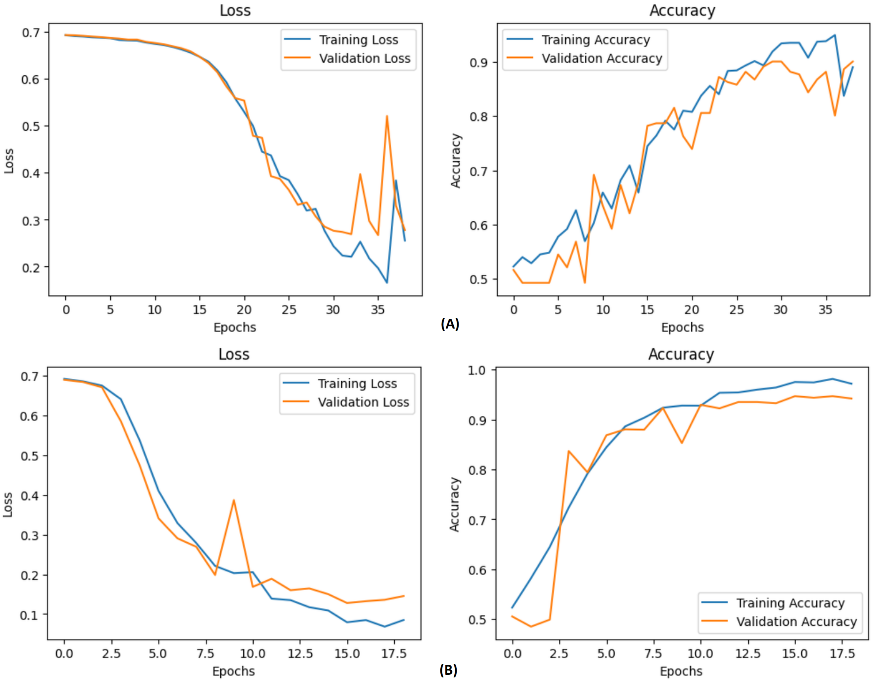


Figure 7. LSTM model performance on original vs. augmented data, (A) Original data (B) Augmented data.

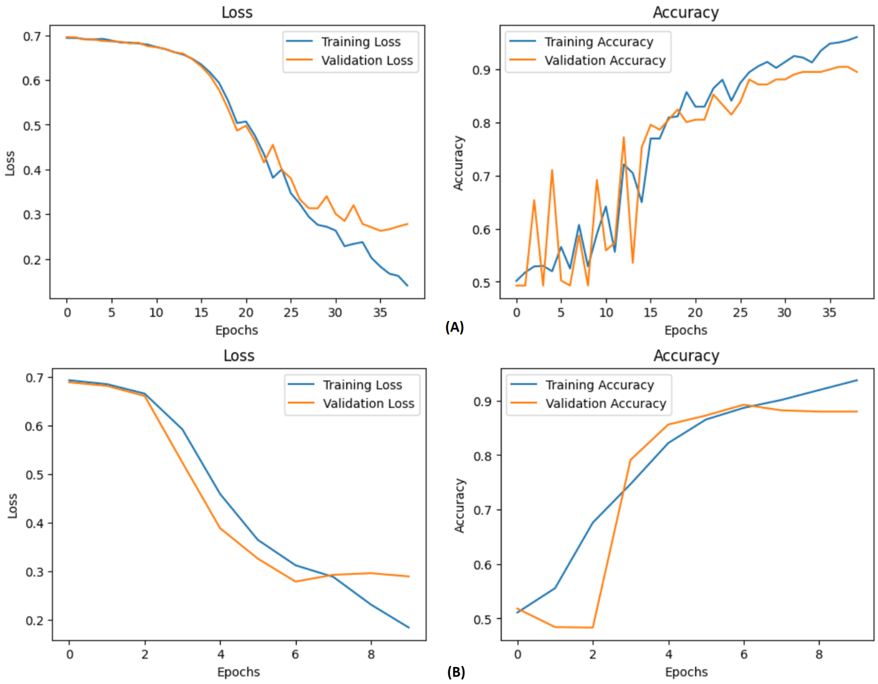


Figure 8. Bi-LSTM model performance on original vs. augmented data; (A) Original data (B) Augmented data.

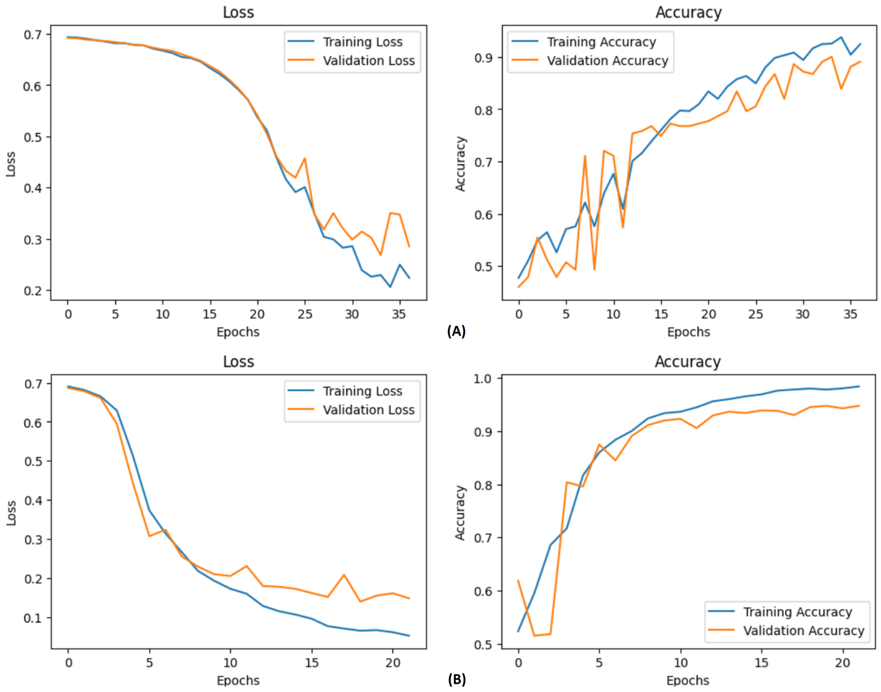


Figure 9. GRU model performance on original vs. augmented data; (A) Original data (B) Augmented data.

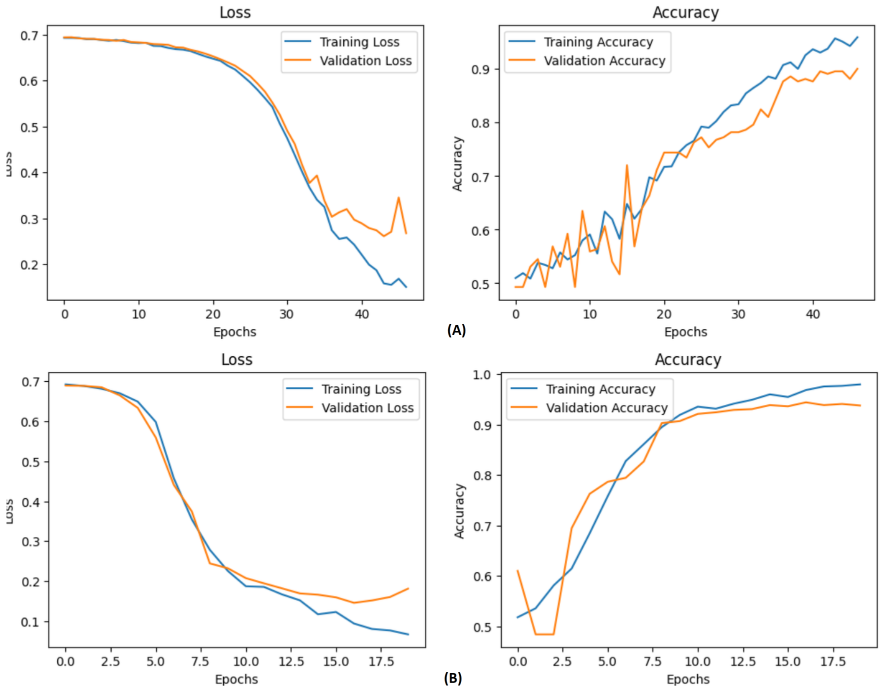


Figure 10. Bi-GRU model performance on original vs. augmented data; (A) Original data (B) Augmented data.

- The SimpleRNN achieved precision of 90.9%, indicating competitive performance despite its simpler architecture.
- With augmented data, GRU and LSTM reached 94.8% and 95.0% precision, confirming their efficiency and robustness.
- Data augmentation improved the models' generalization, particularly for Bi-GRU and Bi-LSTM.

4.5. *Model Performance Comparison Before and After Data Augmentation*

To provide a comprehensive overview, Fig. 11 and 12 summarize and compare the classification metrics of all models using original and augmented datasets, respectively. In Fig. 11, models trained on the original data show varying convergence behaviors, with LSTM and GRU achieving higher accuracy and lower loss than SimpleRNN-based models. Bidirectional variants demonstrate smoother learning curves and better overall performance, indicating stronger contextual learning.

In Fig. 12, after applying data augmentation, all models converge faster with reduced loss and improved accuracy consistency. The gap between models narrows, confirming that augmentation enhances data diversity, stabilizes training, and boosts generalization across architectures, especially for the LSTM and Bi-GRU models.

4.6. *Comparative Evaluation of RNN Models on Original Data*

To assess the classification capabilities of different RNN models, we conducted a comparative analysis that was conducted using key performance indicators: test accuracy, precision, recall, and F1-score. These metrics were used to evaluate the ability of each model to distinguish between individual and collaborative task contexts. The results are summarized in Table 1.

From Table 1, we note the following key observations:

- The SimpleRNN model achieved 90.99% accuracy, indicating good performance, despite its simplicity compared to more advanced models. It maintains balanced precision and recall for both collaborative and

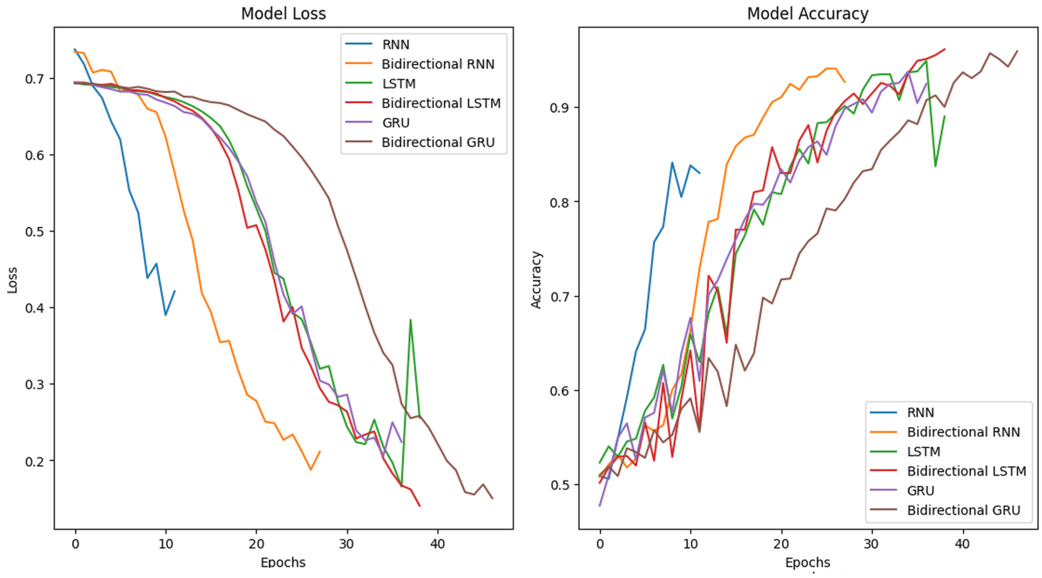


Figure 11. Comparison of Model Performance on Original Data.

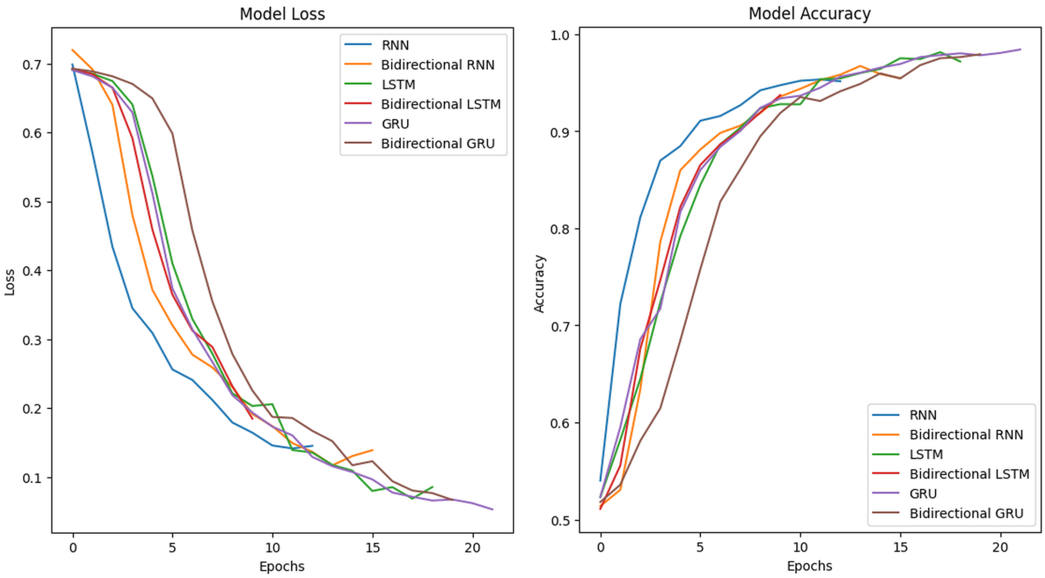


Figure 12. Comparison of Model Performance on Augmented data.

individual tasks, about 91% F1-score, showing its ability to correctly classify instances in each category effectively.

- The Bidirectional RNN slightly outperformed SimpleRNN with 91.94% accuracy, achieving higher precision for collaborative tasks (95%) and strong recall for individual tasks (95%), yielding a consistent 92% F1-score across both classes.
- The LSTM model is the best performing model, achieving 94.31% accuracy. It demonstrates exceptional recall for collaborative tasks (98%), demonstrating its superior capability to retain long-term dependencies

Models	**Epochs/ Time (s)**	**Test Accuracy %**	**Precision (Ind/Col)**	**Recall (Ind/Col)**	**F1-score (Ind/Col)**
S-RNN	25.82	90.99	93 / 89	90 / 92	91 / 91
Bi-RNN	43.51	91.94	95 / 89	89 / 95	92 / 92
LSTM	98.75	94.31	92 / 98	98 / 90	95 / 94
Bi-LSTM	177.46	92.42	93 / 92	93 / 92	93 / 92
GRU	85.54	92.89	93 / 93	94 / 92	93 / 92
Bi-GRU	189.28	91.94	93 / 91	92 / 92	92 / 92

S-RNN: Simple RNN
LSTM: Long Short-Term Memory
GRU: Gated Recurrent Unit
Ind: Individual class

Bi-RNN: Bidirectional RNN
Bi-LSTM: Bidirectional LSTM
Bi-GRU: Bidirectional GRU
Col: Collaborative class

Table 1. Performance Comparison of RNN Models on Original Data

with 94% and 95% F1-scores, reflecting its ability to achieve high accuracy while balancing precision and recall.

- The Bidirectional LSTM achieved 92.42% accuracy with well-balanced precision and recall across both categories (93% / 92%), benefiting from richer contextual understanding due to its bidirectional processing.
- The GRU model is close in performance to LSTM, reaching 92.89% accuracy with balanced precision and recall for both classes (93% / 93% and 94% / 92%, respectively, confirming its strength to be a computationally efficient alternative to LSTM with minimal performance trade-off.
- The Bidirectional GRU model matched the Bidirectional RNN in terms of accuracy (91.94%) and maintained a 92% F1-score for both classes, offering stable performance, with slightly higher precision for collaborative tasks.

Overall, all models demonstrate strong classification capabilities. However:

- The LSTM model emerged as the most effective model for classification tasks, achieving the highest test accuracy (94.31%), and superior recall, particularly for collaborative task contexts.
- The Bidirectional models (Bi-LSTM, Bi-GRU) effectively leverage contextual information from both past and future sequences, leading to well-balanced precision and recall across classes.
- The SimpleRNN model, while slightly less accurate, still performed well, making it a viable lightweight and efficient alternative for less complex scenarios.
- GRU and Bi-GRU models offered a comparable performance to LSTM, indicating their suitability as efficient alternatives with reduced computational costs.

This comparative analysis highlights the strengths and trade-offs of different RNN architectures, providing valuable insights into selecting the most suitable model based on a combination of classification accuracy, computational efficiency, and robustness (see Fig. 9).

4.7. Comparative Evaluation of RNN Models on Augmented Data

Table 2 presents a comparative analysis of different RNN-based models trained with augmented data. The evaluation metrics, including test accuracy, precision, recall, and F1-score were employed to assess the impact of data augmentation on model performance.

Models	Epochs/ Time (s)	Test Accuracy %	Precision (Ind/Col)	Recall (Ind/Col)	F1-score (Ind/Col)
S-RNN	72.58	93.36	96 / 91	91 / 96	93 / 94
Bi-RNN	164.65	94.23	96 / 93	93 / 96	94 / 94
LSTM	326.50	95.02	95 / 95	95 / 95	95 / 95
Bi-LSTM	320.47	88.93	90 / 88	87 / 91	89 / 89
GRU	306.51	94.86	96 / 94	94 / 96	95 / 95
Bi-GRU	532.32	94.47	95 / 94	94 / 95	94 / 95

S-RNN: Simple RNN
LSTM: Long Short-Term Memory
GRU: Gated Recurrent Unit
Ind: Individual class
Bi-RNN: Bidirectional RNN
Bi-LSTM: Bidirectional LSTM
Bi-GRU: Bidirectional GRU
Col: Collaborative class

Table 2. Performance Comparison of RNN Models on Augmented Data

From Table 2, we note the following key observations:

- The SimpleRNN model achieved 93.36% accuracy, showing strong performance on the augmented dataset. It maintained balanced precision and recall, with a slightly higher precision for the collaborative class (96%) and F1-scores of 93% for collaborative tasks and 94% for individual tasks, demonstrating effective performance across both task categories.
- The Bidirectional RNN slightly outperformed SimpleRNN with 94.23% accuracy, showing high recall for Individual tasks (96%), while maintaining strong precision for both classes. Its 94% F1-score demonstrates model effective management of precision-recall trade-offs.
- The LSTM recorded the highest accuracy among all models at 95.02%, exhibiting uniformly high precision and recall at and F1-score of 95%, highlighting its efficiency in maintaining high accuracy while ensuring a strong balance between precision and recall.
- The Bidirectional LSTM, despite its bidirectional architecture, achieved a lower accuracy of 88.93% suggesting possible overfitting or sensitivity to augmentation. It maintains a balance between precision 90% (individual tasks) and 88% (collaborative tasks) with corresponding recall values of recall 87% and 91% respectively. Its performance is lower compared to other models. Also, its F1-scores of 89% for both classes indicate reasonable performance, still below that of other models.
- The GRU closely matched LSTM's performance, achieving 94.86% accuracy with balanced precision and recall, with slightly higher recall for individual tasks (96%). Its 95% F1-scores across both individual and collaborative categories, confirming that GRU offers near-LSTM performance with greater computational efficiency.
- The Bidirectional GRU attained 94.47% accuracy like Bidirectional RNN. It shows well-balanced precision (95% collaborative, 94% individual) and recall (94% collaborative, 95% individual), with a slight advantage in precision for the Collaborative class. Its F1-scores of 94% (collaborative) and 95% (individual) reflect stable and effective classification performance under augmented data conditions.

The proposed approach leverages RNN-based architectures to capture contextual and temporal dependencies in worksheet description. The obtained results demonstrate that LSTM and GRU mechanisms effectively keep relevant information and reduce sensitivity to minor lexical variations introduced by augmentation. This leads to stable feature representations and improves the classification efficiency of AMCDPT textual instances.

4.8. Comparison of RNNS with Transformer-Based Models

For further contextualizing the performance of the proposed approach, we conducted a comparative analysis between the best-performing RNN model (LSTM) and two transformer-based models, BERT and RoBERTa. As shown in Table 3, transformer models achieved slightly higher accuracy on the augmented dataset, achieving

(97.39%) for BERT and (95.26%) for RoBERTa compared to (95.02)%) for LSTM. However, this improvement came at a significantly higher computational cost. In particular, BERT required approximately 3731.36 seconds for 50 epochs, while RoBERTa required approximately 9335.48 seconds for 50 epochs, compared to only 326.50 seconds for 50 epochs for the LSTM model on the augmented dataset. Despite the higher accuracy achieved by transformer models, LSTM demonstrated highly competitive performance across all evaluation metrics, especially on the original dataset, achieving a higher accuracy of (94.31%) with lower computational cost. These results indicate that for moderately sized and structurally uniform educational datasets, lightweight sequential architectures may represent a more practical solution than computationally intensive transformer models. On the other hand, transformer architectures remain promising for future research involving larger, more diverse, and multilingual datasets, where their contextual modeling capabilities may offer additional advantages. This observation is consistent with prior studies indicating that transformer superiority becomes more evident in large-scale and complex linguistic tasks rather than working on structured domain-specific datasets.

Table 3. Comparison of RNNs with Transformer-Based Models

Original Dataset

Models	**Epochs/ Time (s)**	**Test Accuracy %**	**Precision (Ind/Col)**	**Recall (Ind/Col)**	**F1-score (Ind/Col)**
LSTM	98.75	94.31	92 / 98	98 / 90	95 / 94
BERT	472.07	93.84	92 / 95	95 / 93	94 / 94
RoBERTa	1491.45	91.47	95 / 89	87 / 95	91 / 92

Augmented Dataset

Models	**Epochs/ Time (s)**	**Test Accuracy %**	**Precision (Ind/Col)**	**Recall (Ind/Col)**	**F1-score (Ind/Col)**
LSTM	326.50	95.02	95 / 95	95 / 95	95 / 95
BERT	3731.36	97.39	97 / 98	97 / 98	97 / 97
RoBERTa	9335.48	95.26	96 / 94	95 / 96	95 / 95

4.9. Comparison of LLMs for Data Augmentation

Table 4 compares the effectiveness of different LLMs for data augmentation. ChatGPT-4 achieved the highest semantic coherence (4.8) and augmentation quality (4.9), resulting in the best classification accuracy of (95.02%). GPT-3.5 showed strong performance with slightly lower accuracy, while LLaMA 2 demonstrated comparatively lower augmentation quality and classification performance. These results confirm that higher-quality semantic augmentation improves classification accuracy.

Table 4. Performance comparison of different LLMs for data augmentation

LLM Model	**Semantic Coherence** (1–5)	**Augmentation Quality** (1–5)	**Classification Accuracy** (%)
GPT-3.5	4.1	4.0	–
LLaMA-2	3.9	3.8	–
ChatGPT-4	**4.8**	**4.9**	**95.02**

4.10. Practical Integration within LMS systems

The proposed classification model will be integrated within Learning Management Systems (LMS) to automatically analyze worksheet specifications based on parameterized factors derived from learning task characteristics, whether a task should be classified as individual or collaborative. Once a worksheet is uploaded, the system automatically processes its content, predicts the appropriate task type, and configures execution parameters accordingly. This resulting task-type determination supports LMS orchestration functionalities, including automated group formation, instruction adaptation, and monitoring of experimental activity.

The integration within LMS systems enables automated group formation, instruction adaptation, and experimental activity monitoring since it operates through a structured representation of task features, learner interactions, and contextual parameters, ensuring efficient pedagogical classification. The integration of the proposed model aligns with the objectives of our previous work [16], which aimed to add intelligent task-type classification mechanisms to automate pedagogical guidance and support comprehensive learner evaluations.

4.11. Discussions

The advantage of the proposed approach is based on contextual sequence modeling through RNN-based models. The hidden states represent temporal dependencies and semantic flow within the text, while gated memory mechanisms are less sensitive to minor lexical variations that may arise from paraphrasing or synonym substitution. Because RNN variants such as LSTM and GRU selectively retain informative features and suppress irrelevant noise, the learned representations remain stable even when augmented samples are introduced. For this reason, data augmentation is preferable, as it positively influences the efficiency of classifying AMCDPT textual instances.

In general, all models exhibit strong classification capabilities, though their performances reveal distinct strengths and trade-offs. (see Fig. 10):

- The LSTM model achieved the highest accuracy (95.02%), demonstrating its superiority in classifying tasks effectively.
- The SimpleRNN model performed remarkably well, achieving competitive results across all metrics despite being a simpler model.
- The GRU models provided a balanced and robust performance, making them viable alternatives to LSTM, offering similar accuracy with reduced computational costs.
- The Bidirectional models (Bi-RNN, Bi-GRU, Bi-LSTM) leveraged contextual dependencies. However, Bidirectional LSTM showed a noticeable drop in accuracy due to increased model complexity or overfitting introduced by augmentation.

The AMCDPT dataset demonstrates the strong potential of generative AI to produce high-quality and pedagogically valid educational materials. Through structured prompts, ChatGPT-4 generated chemistry practical tasks that are both scientifically coherent and semantically accurate. The validation process confirmed a high level of consistency between generated and real laboratory tasks, ensuring data reliability. The balanced distribution between collaborative and individual tasks supports fair model training, while data augmentation to 7,885 samples enhanced linguistic diversity without compromising meaning. Overall, the dataset combines semantic precision, educational relevance, and structural balance, making it a robust foundation for automatic classification and educational NLP research.

Overall, the results obtained with both original and augmented data are closely aligned, with only slight improvements observed after augmentation. This outcome can be attributed to several factors:

- The simplicity of the dataset, allowing all models to perform effectively and achieve high classification results.
- The dataset's limited vocabulary and structured nature lead to reduced complexity, thereby facilitating more straightforward classification and minimizing the impact of additional synthetic data.
- Although data augmentation provided marginal enhancement in precision and recall, but it does not significantly change the overall performance rankings among models.

These findings indicate that, while augmentation offers small improvements, the inherent characteristics of the dataset already enable strong classification performance across all tested architectures.

Although occasional misclassifications of practical tasks may slightly affect task execution modality, their impact is organizational rather than pedagogical. For instance, a collaborative task classified as individual can still be completed effectively, whereas an individual task classified as collaborative only leads to redundant participation. Thus, misclassification mainly influences time management and coordination efficiency rather than pedagogical validity.However, we will include in future work cost-sensitive evaluation metrics to better model the differentiated impact of classification errors.

In fact, misclassification primarily affects the modality of task execution rather than the validity of the learning process, meaning that the associated costs are organizational rather than pedagogical. While this aspect does not compromise the integrity of the learning outcomes, it may impact operational efficiency and resource allocation. Future work will incorporate cost-sensitive evaluation metrics to better capture and model the differentiated impact and practical consequences of classification errors.

The evaluation of the proposed approach was achieved regarding computational time versus dataset-size sensitivity, where the obtained results are very encouraging of which accuracy is reduced with training data size decreases. However, RNN models remain relatively robust under moderate data conditions, which supports their suitability for realistic educational scenarios. The results presented in Tables 1 and 2 confirm that the proposed approach achieves a favorable balance between computational efficiency and classification accuracy across varying dataset sizes

4.12. Limitations and futur work

Although the proposed approach demonstrates strong performance and practical relevance, several limitations should be acknowledged:

- LLM-generated data may exhibit certain constraints, particularly in experimental scientific worksheets, where generated text can sometimes show reduced linguistic diversity or formulaic phrasing.
- Limited lexical variability in synthetic samples may affect expressive richness and reduce subtle semantic distinctions between task descriptions.
- Uniform language patterns across generated data may increase classification difficulty when distinguishing between individual and collaborative tasks.
- Although experiments were conducted on chemistry practical tasks, the proposed framework is domain-independent. Chemistry was selected as a prototype domain due to data availability. The underlying architecture and prompt-engineering strategy can be applied to other domains.
- The current study does not yet assess performance across multiple disciplines; therefore, extending evaluation to additional domains such as physics and biology is planned for future work.
- Future research will also focus on improving synthetic data quality by reducing artificial patterns and minimizing semantically weak or repetitive content.
- As a perspective for improving the proposed framework, attempting to create a complete approach based on recent and relevant explainable AI techniques[28][29]may likewise enhance model interpretability and instructor trust of the proposed classification approach.

5. Conclusion

This study investigated the automatic classification of practical tasks into individual and collaborative categories using several Recurrent Neural Network (RNN) models, including SimpleRNN, LSTM, GRU, and their bidirectional variants. Extensive experiments conducted on both original and augmented datasets demonstrated that LSTM achieved the best overall performance, reaching an accuracy of (95.02%), while GRU provided competitive results with lower computational cost. Bidirectional architectures further improved contextual representation, although their efficiency varied depending on dataset characteristics.

The findings indicate that data augmentation yields modest but consistent improvements, whereas the intrinsic structure and semantic coherence of the AMCDPT dataset already enable strong classification performance across models. This highlights the critical role of dataset design, balanced class distribution, and pedagogical consistency in determining model effectiveness for educational NLP tasks.

Additional comparative analysis with transformer-based architectures showed that, although transformer models can achieve slightly higher accuracy on augmented data, this gain often comes at substantially higher computational cost. These results suggest that lightweight sequential architectures such as LSTM may represent a more practical solution for structured, domain-specific educational datasets, particularly in real-time or resource-constrained learning environments.

Despite promising results, several limitations should be acknowledged. The current evaluation was conducted on a single domain and language, which restricts conclusions regarding cross-domain generalizability. Moreover, while augmentation improved performance, further investigation of augmentation strategies and statistical validation techniques would strengthen future analyses.

Future work will explore transformer-based architectures such as DeBERTa and DistilBERT, along with optimization strategies including hyperparameter tuning and model compression to improve efficiency while maintaining high predictive performance. Overall, this work demonstrates the feasibility and practical relevance of automated task-type classification in educational environments and provides a foundation for intelligent instructional support systems in virtual laboratory settings.

In addition, explainable AI techniques will be incorporated to provide interpretable predictions and enhance instructor trust in automated educational systems. A user-study evaluation will also be conducted to measure the real pedagogical impact of the proposed framework.

Acknowledgement

This work was completed with the support of DGRSDT-MESRS (Algeria).

REFERENCES

1. W. Etaiwi, D. Suleiman, and A. Awajan, *Deep Learning Based Techniques for Sentiment Analysis: A Survey*, Informatica, vol. 45, no. 7, pp. 89–95, 2021. DOI: 10.31449/inf.v45i7.3674.
2. A. Chiche, and B. Yitagesu, *Part of Speech Tagging: A Systematic Review of Deep Learning and Machine Learning Approaches*, Journal of Big Data, vol. 9, no. 1, pp. 10, 2022. DOI: 10.1186/s40537-022-00561-y.
3. A. Mohammed, and R. Kora, *A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges*, Journal of King Saud University–Computer and Information Sciences, vol. 35, no. 2, pp. 757–774, 2023. DOI: 10.1016/j.jksuci.2023.01.014.
4. Z. Feng, *Past and Present of Natural Language Processing*, in Formal Analysis for Natural Language Processing: A Handbook, Springer Nature Singapore, pp. 3–48, 2023. DOI: 10.1007/978-981-16-5172-4_1.
5. S. G. Burabod, *Leveraging Artificial Intelligence for Enhanced Linguistic Analysis: A Deep Learning Approach to Language Understanding*, in Conference on AI & Linguistics, vol. 1, 2024. DOI: 10.54878/jsw4yv84.
6. C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, *A Survey of Natural Language Generation*, ACM Computing Surveys, vol. 55, no. 8, pp. 1–38, 2022. DOI: 10.1145/3554727.
7. D. Khurana, A. Koli, K. Khatter, and S. Singh, *Natural Language Processing: State of the Art, Current Trends and Challenges*, Multimedia Tools and Applications, vol. 82, no. 3, pp. 3713–3744, 2023. DOI: 10.1007/s11042-022-3428-4.
8. B. Nethravathi, G. Amitha, A. Saruka, T. P. Bharath, and S. Suyagya, *Structuring Natural Language to Query Language: A Review*, Engineering, Technology & Applied Science Research, vol. 10, no. 6, pp. 6521–6525, 2020. DOI: 10.48084/etasr.3873.
9. M. Zhao, L. Zhang, Y. Xu, J. Ding, J. Guan, and S. Zhou, *Epida: An Easy Plug-in Data Augmentation Framework for High Performance Text Classification*, arXiv preprint arXiv:2204.11205, 2022. DOI: 10.48550/arXiv.2204.11205.
10. Z. Miao, Y. Li, and X. Wang, *Rotom: A Meta-Learned Data Augmentation Framework for Entity Matching, Data Cleaning, Text Classification, and Beyond*, in Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), 2021. DOI: 10.1145/3448016.3457258.
11. T. Shaik, X. Tao, Y. Li, C. Dann, J. McDonald, P. Redmond, and L. Galligan, *A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis*, IEEE Access, vol. 10, pp. 56720–56739, 2022. DOI: 10.1109/ACCESS.2022.3177752.
12. S. Ahriz, H. Gharbaoui, N. Benmoussa, A. Chahid, and K. Mansouri, *Enhancing Information Technology Governance in Universities: A Smart Chatbot System Based on Information Technology Infrastructure Library*, Engineering, Technology & Applied Science Research, vol. 14, no. 6, pp. 17876–17882, 2024. DOI: 10.48084/etasr.8878.
13. K. Fuchs, *Exploring the Opportunities and Challenges of NLP Models in Higher Education: Is ChatGPT a Blessing or a Curse?*, Frontiers in Education, vol. 8, pp. 1166682, 2023. DOI: 10.3389/feduc.2023.1166682.
14. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, Ll. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention Is All You Need*, in Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017. DOI: 10.48550/arXiv.1706.03762.

15. B. Wang, *A Hybrid Fuzzy Logic and Deep Learning Model for Corpus-Based German Language Learning with NLP*, Informatica, vol. 49, no. 21, 2025. DOI: 10.31449/inf.v49i21.7423.
16. A. Douar, M. Djoudi, S. Harous, and A. Adel, *3DVL@ES: A 3D web-based virtual Laboratory for Collaborative Learning in experimental science practical work*, International Journal of e-Collaboration (IJeC), vol. 19, no. 1, pp. 1–26, 2023. DOI: 10.4018/IJeC.315786.
17. L. Biester, and W. Wu, *Tightly Coupled Worksheets and Homework Assignments for NLP*, in Proceedings of the Sixth Workshop on Teaching NLP, pp. 66–68, 2024.
18. R. Li, X. Gong, B. Zhang, C. Liang, M. Li, and H. Guo, *Study in Intelligent Exam Based on RNN and LSTM*, in Artificial Intelligence in Education and Teaching Assessment, pp. 63–67, 2022. DOI: 10.1007/978-981-16-6502-8.
19. R. Kim, A. Kotsenko, A. Andreev, A. Bazanova, D. Aladin, D. Todua, A. Marushchenko, and O. Varlamov, *Evaluation of BERT and ChatGPT Models in Inference, Paraphrase and Similarity Tasks*, E3S Web of Conferences, vol. 515, pp. 03016, 2024. DOI: 10.1051/e3sconf/202451503016.
20. S. Rustad, G. F. Shidik, E. Noersasongko, and D. R. I. M. Setiadi, *Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application*, Statistics, Optimization and Information Computing, vol. 12, no. 2, pp. 907–942, 2024. DOI: 10.19139/soic-2310-5070-1631.
21. S. M. Ferdous, S. N. E. Newaz, S. B. S. Mugdha, and M. Uddin, *Sentiment Analysis in the Transformative Era of Machine Learning: A Comprehensive Review*, Statistics, Optimization and Information Computing, vol. 13, no. 1, pp. 331–346, 2025. DOI: 10.19139/soic-2310-5070-2113.
22. I. Gligorea, M. Cioca, R. Oancea, A. T. Gorski, H. Gorski, and P. Tudorache, *Adaptive Learning Using Artificial Intelligence in E-Learning: A Literature Review*, Education Sciences, vol. 13, no. 12, pp. 1216, 2023. DOI: 10.3390/educsci13121216.
23. R. Shafique, W. Aljedaani, F. Rustam, E. Lee, A. Mehmood, and G. S. Choi, *Role of Artificial Intelligence in Online Education: A Systematic Mapping Study*, IEEE Access, vol. 11, pp. 52570–52584, 2023. DOI: 10.1109/ACCESS.2023.3278590.
24. H. S. Abdullah, N. H. Ali, and N. A. Abdullah, *Evaluating the Performance and Behavior of CNN, LSTM, and GRU for Classification and Prediction Tasks*, Iraqi Journal of Science, vol. 65, no. 3, pp. 1741–1751, 2024. DOI: 10.24996/ijs.2024.65.3.43.
25. Y. Lan, X. Li, H. Du, X. Lu, M. Gao, W. Qian, and A. Zhou, *Survey of Natural Language Processing for Education: Taxonomy, Systematic Review, and Future Trends*, IEEE Transactions on Knowledge and Data Engineering, 2025. DOI: 10.1109/TKDE.2025.3621181.
26. K. Choudhary, and M. L. Kelley, *ChemNLP: A Natural Language Processing Based Library for Materials Chemistry Text Data*, The Journal of Physical Chemistry C, vol. 127 no. 35, pp. 17545–17555, 2023. DOI: 10.1021/acs.jpcc.3c03106.
27. A. Douar, *AMCDPT: Automatic Multiclass Chemistry Dataset for Practical Tasks*, Kaggle dataset, 2025. www.kaggle.com/datasets/pfe202419/amcdpt-dataset (Accessed: 2025-11-09).
28. D. A. Ali and H. T. Sadeeq, *An Interpretable Deep Learning Framework for Multi-Class Dental Disease Classification from Intraoral RGB Images*, Statistics, Optimization & Information Computing, vol. 14, no. 6, pp. 3380–3397, 2025. DOI: 10.19139/soic-2310-5070-2880.
29. D. A. Ali and H. T. Sadeeq, *A Review on Deep Learning Frameworks for Dental Anomaly and Disease Classification*, Dasinya Journal for Engineering and Informatics, vol. 1, no. 1, pp. 1–19, 2025. DOI: 10.65542/djei.v1i1.13.