

Automated Analysis of Student Handwritten Notes Based on Attention Multi-Task U-Net and Generative AI

周, 云宇

<https://hdl.handle.net/2324/7396254>

出版情報 : Kyushu University, 2025, 博士 (工学) , 課程博士
バージョン :
権利関係 :



**Automated Analysis of Student
Handwritten Notes Based on Attention
Multi-Task U-Net and Generative AI**

Zhou YunYu

2025, 8

Doctor of Philosophy

Department of Information Science and Technology

Contents

Abstract	1
1 Introduction	4
1.1 Background and Motivation	4
1.2 Research Objectives	7
1.3 Research Questions and Hypotheses	8
1.4 Thesis Structure	9
2 Literature Review	10
2.1 Introduction	10
2.2 Image Segmentation of Handwritten Notes	11
2.3 Natural Language Processing in Educational Contexts	12
2.4 AI-Driven Learning Analytics	13
2.5 Gaps and Opportunities	14
2.6 Summary	15
3 Unified Methodological Framework	16
3.1 Overview of the Framework	16
3.2 Attention Multi-task U-Net for Image Segmentation	18
3.3 OCR and Preprocessing	20

3.4	Vision-to-Language Pipeline	20
3.5	System Integration and Data Flow	21
3.6	Evaluation Design	21
3.7	Theoretical and Technical Foundations	22
3.7.1	Evolution and Principles of the Attention Multi-task U-Net	22
3.7.2	Language Understanding Capabilities of GPT-4 and Its Educational Adaptability	23
3.7.3	Advantages of Vision-Language Integration Models	24
3.8	Summary	25
4	Segmentation Enhancement: Improved Model for Note Images	26
4.1	Research Background	27
4.2	Attention Multi-task U-net Model	29
4.2.1	Attention Multi-task U-net Model	30
4.2.2	Attention R2 Multi-Task U-Net Model	31
4.3	Experiment Results	32
4.3.1	Dataset Description	33
4.3.2	Comparative experiments of U-Net model	33
4.4	Comparison with other neural networks	37
4.5	Conclusion	38
4.6	Chapter Summary	39
5	Vision–Language Fusion: Multi-Task U-Net + GPT-4	40
5.1	Research Background	41
5.2	Methodology	42
5.2.1	Attention Multi-Task U-Net	44
5.2.2	GPT-4 Builder	45
5.2.3	Handwritten Notes Processing	47

5.2.4	Question & Answer Process	50
5.3	Experiments and results	51
5.3.1	Dataset description	51
5.3.2	Evaluation Protocol	52
5.3.3	Quantitative Results	52
5.3.4	Chapter-wise Trends	54
5.3.5	Discussion and Implications	54
5.4	Conclusion	55
5.5	Chapter Summary	56
6	Advanced Unified Framework: Extracting Learning Data	57
6.1	Research Background	58
6.2	Related Work	61
6.2.1	Applications of Image Processing in Education	61
6.2.2	Applications of Natural Language Processing in Education	62
6.2.3	The Potential of Combining Image Processing and NLP in Handwritten Note Analysis	63
6.3	Methodology	64
6.3.1	Method Overview	64
6.3.2	Attention Multi-task U-Net [12]	65
6.3.3	GPT-4o	67
6.3.4	Integrated Framework	70
6.3.5	Innovation and Advantages	71
6.4	Experiments and Results	74
6.4.1	Experimental Design	74
6.4.2	Experiment Result	75
6.5	Discussion	88

6.5.1	Key Findings	88
6.5.2	Advantages of the Method	89
6.5.3	Comparison with Existing Research	91
6.5.4	Practical Applications and Educational Impact	93
6.5.5	Future Research Directions	94
6.6	Conclusion	95
6.7	Chapter Summary	97
7	Cross-Paper Synthesis and General Discussion	98
7.1	Cross-Study Findings	98
7.2	Integration of Findings with Research Questions:	100
7.3	Theoretical Contributions	101
7.4	Practical Implications	102
7.5	Limitations	104
7.6	Directions for Future Research	105
7.7	Chapter Summary	106
8	Conclusion and Future Work	107
8.1	Summary of Contributions	107
8.2	Reflections on Research Findings	108
8.3	Implications for Stakeholders	109
8.4	Future Work	109
8.5	Final Reflections	110
	Acknowledgement	112
	Reference	113

List of Figures

3.1	System Process Overview	19
4.1	Example of student handwritten notes and Page segmentation label for handwritten notes	28
4.2	Attention Multi-Task U-Net Model	31
4.3	Attention R2 Multi-Task U-Net Model	32
4.4	Multi-Task U-Net Model Page Segmentation Result and Attention Multi-Task U-Net Model Page Segmentation Result	36
4.5	Multi-Task U-Net Model Character Segmentation Result and At- tention Multi-Task U-Net Model Character Segmentation Result . .	36
5.1	Example of handwritten note segmentation	43
5.2	Attention Multi-Task U-Net	44
5.3	Building the GPT-4 Builder Process(Due to copyright issues, some images in the graph are made blurry.)	46
5.4	Examples of saving graphs and formulas in textbooks as .pdf files .	47
5.5	Handwritten Notes Processing	49
5.6	Question and Answer Process	50
5.7	Knowledge Base	51
5.8	Examples of Output Results	55

6.1	Attention Multi-task U-Net	66
6.2	Handwritten note page segmentation example	68
6.3	Process of building a knowledge base of textbooks	71
6.4	Process for student handwritten notes	72
6.5	Matching result example	72
6.6	2021 keywords Bubble Chart	89
6.7	2022 keywords Bubble Chart	90

List of Tables

4.1	Comparison of Accuracy Rates by Region	34
4.2	Overall Accuracy and F1 Score of U-Net Variants	34
4.3	Comparison with other neural networks	37
5.1	Accuracy of Chapter 1–11	53
6.1	Comparison of U-Net Accuracy with Other Neural Networks	76
6.2	Comparison of Accuracy Rates by Region	78
6.3	Matching Accuracy Comparison	80
6.4	GPT-4 Accuracy of Chapters 1–11	81
6.5	GPT-4o Accuracy of Chapters 1–11	82
6.6	Matching Accuracy Comparison	82
6.7	GPT-4 Accuracy for Chapter 1–11	83
6.8	GPT-4o Accuracy for Chapter 1–11	84
6.9	CHAPTER 1 KEYWORDS	86

Abstract

This dissertation presents a unified framework for the automated analysis of handwritten student notes by integrating state-of-the-art image segmentation and natural language processing techniques. The system combines an Attention Multi-task U-Net model to accurately detect and classify heterogeneous content areas—such as text, diagrams, and mathematical expressions—with GPT-4 to semantically interpret the extracted information and align it with textbook concepts. Through three empirical studies, the proposed framework demonstrates effectiveness in improving segmentation accuracy, extracting pedagogically relevant insights, and establishing correlations between student notes and learning behaviors. The model adapts well to various subjects, writing styles, and learning contexts, making it applicable across disciplines and institutional settings. In addition to technical advancements, the research contributes to the theory of multimodal learning analytics by redefining handwritten notes as valuable cognitive artifacts rather than supplementary records. The system enables educators to monitor student engagement, identify conceptual gaps, and support personalized instruction in both in-person and online environments. Moreover, the architecture supports scalable deployment and can be integrated into existing learning management systems for real-time feedback. This dissertation significantly advances the fields of educational data mining, AI in education, and vision-language modeling, providing

interpretable and practical tools for the next generation of evidence-based, data-informed learning systems.

Key Words:Handwritten Notes, Educational Data Mining , Learning Analytics, Attention Multi-task U-Net, GPT-4 / Generative AI, Image Segmentation

要旨

本論文では、学生の手書きノートを自動的に解析するための統一的なフレームワークを提案する。画像セグメンテーションには Attention Multi-task U-Net を採用し、テキスト、図表、数式など異なる情報を精密に領域分割し、自然言語処理には GPT-4 を活用して意味理解および教科書との概念的整合を実現する。本研究は三つの実証的な研究を通じて、セグメンテーションの精度向上、教育的に意味のある情報の抽出、学習行動との相関分析において高い有効性を示した。また、STEM 系と文系のノート形式に柔軟に適応でき、様々な教育環境や対象学生に広く応用可能である。本研究の理論的貢献として、手書きノートを単なる補助資料ではなく、認知的エビデンスとして再定義し、マルチモーダル学習アナリティクスの新たな枠組みを提示する。また、教育現場では、教師が学生の理解状況を定量的に把握し、個別対応や授業設計の改善に活かすことが可能となる。さらに、本システムは LMS との統合や大規模導入にも対応可能であり、教育データマイニング、教育 AI、視覚と言語統合モデルの分野に新たな道を切り開く実用的かつ理論的に意義深い成果を提供する。

キーワード: 手書きノート, 教育データマイニング, 学習アナリティクス, Attention Multi-task U-Net , GPT-4 / 生成 AI 画像セグメンテーション

Chapter 1

Introduction

1.1 Background and Motivation

Handwritten notes have long been a cornerstone of the learning process, representing an externalization of students' cognitive engagement during lectures, readings, and independent studies. These notes encompass diverse elements such as textual annotations, mathematical formulas, diagrams, and tables. Despite their intrinsic educational value, traditional educational analytics frequently overlooks handwritten notes due to the complexity involved in extracting meaningful insights from unstructured and heterogeneous data.

Conventionally, instructors evaluate student performance using assessments, attendance, and participation metrics. However, such metrics often lack depth in capturing the nuances of students' learning processes. In contrast, handwritten notes offer rich insights into individual cognitive behaviors, learning styles, knowledge retention, and misconceptions. For instance, the presence of visual diagrams may suggest a student's preference for spatial reasoning, whereas frequent underlining might indicate surface-level engagement. Consequently, a detailed analysis

of handwritten notes could significantly enhance educators' understanding of students' cognitive processes.

Moreover, handwritten notes serve as longitudinal records of student learning behaviors. By examining patterns in notetaking over multiple lessons or academic terms, educators can identify evolving comprehension levels, subject-specific challenges, and cognitive engagement trends. Such insights align well with competency-based educational paradigms that prioritize tracking and supporting individual student trajectories rather than applying uniform standards across diverse learners.

Recent advancements in image processing and natural language understanding technologies provide promising opportunities for automating the analysis of student-generated content. [153, 154] Specifically, deep learning models, such as convolutional neural networks (CNNs), have demonstrated remarkable effectiveness in image segmentation, object detection, and handwriting recognition tasks. Concurrently, generative AI models, notably OpenAI's GPT-4, have revolutionized natural language processing (NLP), enabling nuanced text summarization, context inference, and content alignment.

These technological advancements coincide with broader trends in educational data mining (EDM) and learning analytics (LA). As educational institutions increasingly adopt data-driven methodologies to improve student outcomes, AI-driven handwritten note analysis fills a critical gap in multimodal learning assessment. Unlike digital interaction data (e.g., clickstream data or quiz scores), handwritten notes uniquely capture thought structures, attention allocation, and content prioritization, providing richer cognitive insights.

In the context of rising interest in personalized and adaptive learning systems, the need for tools capable of accurately diagnosing individual learning trajectories is increasingly urgent. Automated handwritten note analysis represents a potential

transformative tool, bridging qualitative observation with quantitative evaluation. Additionally, the growing prevalence of remote and hybrid learning models further underscores the necessity of digitally analyzing student-generated artifacts to compensate for limited opportunities for in-person observation.

Hence, this dissertation capitalizes on these technological developments to propose a unified framework for automated handwritten note analysis, emphasizing scalability, flexibility, and multimodal integration across various educational contexts. The envisioned framework aims not only to advance technological capabilities but also to reshape pedagogical strategies by supporting evidence-based personalized learning interventions.

However, despite extensive advances in educational data analytics, current methods are insufficient for extracting actionable learning insights from handwritten notes. This dissertation addresses that gap by introducing an integrated approach that bridges computer vision and NLP to transform handwritten notes into actionable educational data.

Leveraging the processing pipeline proposed in this study, the system precisely maps students' handwritten notes to the corresponding instructional knowledge units and delivers the analysis results to instructors and learners separately, thereby facilitating both teaching refinement and learning optimization:

1)Teacher side: The system aggregates and visualizes the textbook chapters and knowledge points most frequently referenced in the notes of high-achieving students. These insights enable instructors to pinpoint critical teaching content and make targeted adjustments to course design, pedagogical strategies, and resource allocation—ultimately enhancing overall instructional effectiveness.

2)Student side: The system presents students with the focal areas and representative note-taking examples of high-GPA peers, prompting them to reflect on and improve their own note-taking practices and learning strategies, which can in turn

elevate academic performance.

This bidirectional feedback mechanism not only reinforces data-driven instructional decision-making but also offers students a peer-based pathway for self-improvement, yielding measurable gains in educational practice.

1.2 Research Objectives

The overarching objective of this dissertation is to develop and validate an integrated analytical framework that automates the extraction and interpretation of educational data from student handwritten notes. This framework relies on two primary technological innovations: the Attention Multi-task U-Net model for precise image segmentation and GPT-4 (including variants like GPT-4o) for robust natural language processing.

To address the aforementioned gap and challenges, this study is guided by the following research questions:

- 1) Improved Image Segmentation Models: Improving existing U-Net architectures through multi-task learning and attention mechanisms to effectively segment complex handwritten notes into meaningful regions, including text, formulas, and diagrams.
- 2) Integrating Vision and Language Processing: Developing a seamless pipeline bridging image segmentation and natural language processing, transforming segmented content into structured, machine-readable formats optimized for GPT models.
- 3) Extracting Pedagogically Relevant Insights: Enabling the systematic identification of core concepts, student focus areas, and learning gaps by correlating student notes with standard curricular materials.
- 4) Evaluating Model Performance: Conducting rigorous empirical evaluations to

benchmark the proposed framework against established methods, ensuring its accuracy and applicability in realistic educational environments.

5) Supporting Instructional Interventions: Equipping educators with actionable insights derived from automated note analysis to facilitate personalized instructional strategies and timely academic interventions.

By accomplishing these objectives, the dissertation aims to offer a practical, impactful tool for educational stakeholders, promoting a transformative shift towards more personalized, efficient, and scalable learning analytics.

1.3 Research Questions and Hypotheses

This dissertation is guided by the following research questions:

RQ1: How can image segmentation models be optimized to accurately handle the heterogeneous structures of handwritten notes?

RQ2: What strategies most effectively integrate OCR and NLP techniques to analyze handwritten educational content?

RQ3: To what extent can automated note analysis provide meaningful insights into student learning behaviors and content mastery?

RQ4: How does the proposed integrated framework compare to existing approaches regarding accuracy, efficiency, and educational relevance?

Correspondingly, the research advances the following hypotheses:

H1: Attention-enhanced multi-task U-Net will significantly outperform baseline U-Net and alternative segmentation models in terms of accuracy and computational efficiency.

H2: Combining OCR technologies with generative AI (GPT-4/GPT-4o) will markedly enhance the semantic interpretation of handwritten note content.

H3: Automated note analysis reliably identifies critical learning behaviors and correlates effectively with conventional academic performance indicators.

H4: The proposed integrated analytical framework will demonstrate superior scalability and user-friendliness compared to fragmented or manual methods.

These hypotheses will undergo rigorous empirical testing using real-world educational datasets, ensuring the framework’s technical robustness and practical educational relevance.

1.4 Thesis Structure

This dissertation is organised into eight chapters. Chapter 1 introduces the research background, objectives, and questions; Chapter 2 reviews pertinent literature; Chapter 3 details the unified methodological framework; Chapters 4–6 report three empirical studies that progressively refine image segmentation, integrate vision–language processing, and extract pedagogically relevant insights; Chapter 7 synthesises cross-study findings; Chapter 8 concludes with contributions and future work.

Each chapter is mapped to one of the five research objectives to guide the reader: Objective 1—Improved Image Segmentation Models is addressed in Chapter 4; Objective 2—Integrating Vision and Language Processing is the focus of Chapter 5; Objective 3—Extracting Pedagogically Relevant Insights is realised in Chapter 6; Objective 4—Evaluating Model Performance is examined across Chapters 4–6; Objective 5—Supporting Instructional Interventions is demonstrated in pedagogical applications of Chapter 6 and implications discussion in Chapter 7.

Chapter 2

Literature Review

2.1 Introduction

This chapter systematically reviews the foundational literature relevant to the development of the proposed framework for automated analysis of student handwritten notes. The review specifically explores prior research across three pivotal areas: image segmentation for handwritten documents, natural language processing (NLP) applications in educational contexts, and the integration of artificial intelligence technologies into learning analytics. Through critical synthesis of these research strands, this chapter establishes a comprehensive foundation supporting the methodological decisions detailed in subsequent chapters. Furthermore, this review highlights both achievements and limitations within prior studies, clearly motivating the integrated approach advocated in this dissertation. Special emphasis is placed on recent advancements in deep learning architectures, multimodal data analysis, and personalized educational technologies.

Recent studies have begun exploring generative AI (e.g., large language models like GPT-4) in educational contexts [57] [72] [79] [122] [134] [135] [123]. These

works demonstrate the potential of AI to evaluate or even tutor student learning, yet they largely focus on text-based inputs. In contrast, this approach uniquely combines image segmentation of handwritten notes with generative AI, thus extending the scope of AI in education to include rich, hand-written content.

2.2 Image Segmentation of Handwritten Notes

The segmentation of handwritten content has garnered considerable attention within the field of computer vision. Traditional methods typically employed rule-based heuristics and edge-detection algorithms. Despite their foundational contributions, these methods faced challenges due to the inherent variability and complexity of handwritten documents, especially when containing mixed elements like text, diagrams, and mathematical equations. However, the advent of convolutional neural networks (CNNs), particularly the U-Net architecture, has significantly transformed this research domain. [4] The encoder-decoder structure inherent in U-Net enables efficient learning and interpretation of spatial hierarchies within handwritten images. Moreover, U-Net's skip connections effectively preserve fine-grained details through multi-resolution processing, significantly enhancing its suitability for educational notes comprising small symbols, arrows, or annotations.

Recent advancements incorporating attention mechanisms and multi-task learning strategies have further improved precision in boundary detection and layout analysis. These enhancements are particularly relevant in educational contexts, where complex layouts—such as marginal annotations, layered mathematical equations, or mixed-language symbols—require precise disentanglement and interpretation. Studies, such as Ronneberger et al. (2015), originally proposed U-Net for biomedical segmentation tasks, yet demonstrated substantial cross-domain adapt-

ability. [4] Subsequent research has effectively adapted this architecture to manage more complex and heterogeneous datasets, including handwritten Japanese, Arabic scripts, and mathematical notation. Further methodological innovations have explored stacked U-Nets, attention-gated U-Nets, and dilated convolutions, [155–157] significantly improving accuracy without sacrificing computational efficiency.

More recently, transformer-based models have been introduced, providing enhanced global context modeling capabilities through sophisticated self-attention mechanisms. However, their high training cost and dependency on extensive labeled datasets pose notable challenges for educational applications, where annotated handwritten datasets are typically scarce and highly diverse. Consequently, attention-augmented U-Net architectures remain a practical and balanced choice, effectively reconciling performance requirements with feasibility constraints in implementation.

2.3 Natural Language Processing in Educational Contexts

Natural language processing (NLP) has been extensively applied in analyzing educational content such as student essays, discussion forum posts, and textbooks. Techniques including topic modeling, named entity recognition, and sentiment analysis have effectively extracted pedagogically relevant insights. For instance, latent Dirichlet allocation (LDA) has identified prevalent themes in student-generated content, [158] and dependency parsing has supported syntactic assessment in language learning scenarios.

The introduction of transformer-based models, such as BERT, RoBERTa, and the GPT series, has significantly advanced the precision and contextual sensitivity

of text analysis. [159–161] These models excel in tasks requiring nuanced interpretation, such as automated feedback generation, knowledge tracing, and question generation. Recent studies leveraging GPT models for summarization, question answering, and grading underscore their potential in educational contexts. However, accurately interpreting unstructured handwritten content remains challenging, particularly when OCR-induced noise and ambiguous diagrams are present. Integrating domain-specific fine-tuning and multimodal embeddings could address these issues effectively, enhancing accuracy in analyzing complex educational materials.

2.4 AI-Driven Learning Analytics

Learning analytics (LA) traditionally utilizes digital footprints, including learning management system logs, clickstream data, and quiz performance metrics. While useful, these data sources often overlook cognitive aspects integral to learning processes, such as problem-solving strategies and note-taking styles. AI technologies can bridge this gap by analyzing multimodal educational artifacts—written, visual, and verbal content—to gain deeper insights into student behaviors.

Advancements in computer vision facilitate the extraction of diagrams from notes, while NLP contextualizes textual reflections. Some studies have integrated speech recognition to analyze verbal interactions, enriching multimodal analytics. Emerging multimodal learning analytics (MMLA) frameworks merge diverse data streams—written notes, audio transcripts, and video captures—to form holistic learner profiles. These integrated systems leverage deep learning to identify misconceptions, engagement patterns, and dropout risks effectively.

However, current frameworks often emphasize model performance without sufficient interpretability, creating practical challenges for educators. Future devel-

opments should focus on transparent, explainable, and pedagogically actionable insights. User-centric dashboards, explainable AI algorithms, and modular system designs represent promising directions.

2.5 Gaps and Opportunities

Despite significant progress in image segmentation, NLP, and learning analytics individually, integrated systems specifically tailored to handwritten educational content remain limited. Few methods simultaneously employ advanced image segmentation with generative language understanding to deliver comprehensive educational insights. Issues of scalability, domain adaptability, and interface usability further constrain current approaches.

Standardized datasets for handwritten educational content, particularly across diverse subjects and languages, are notably scarce, complicating model generalization. Real-world applications require cross-modal fusion to connect diagrams, equations, and handwritten explanations to standardized assessments effectively.

This dissertation addresses these gaps by proposing a unified framework combining Attention Multi-task U-Net and GPT-4. The proposed method capitalizes on the strengths of each component: precise layout segmentation and robust semantic interpretation. By validating the framework across diverse datasets and educational contexts, this research aims to enhance both methodological rigor and practical utility. Importantly, the proposed design emphasizes scalability, interpretability, and ease of integration—critical features for sustainable educational technology adoption.

2.6 Summary

This chapter has systematically reviewed literature in three foundational areas—image segmentation, NLP in education, and AI-driven learning analytics—highlighting how existing technologies can be integrated and extended for robust handwritten note analysis. Building on identified limitations and innovations, the next chapter details the proposed methodology, bridging visual and semantic analysis to derive meaningful educational insights.

Chapter 3

Unified Methodological Framework

3.1 Overview of the Framework

This chapter describes the technical architecture of the proposed unified framework for automated analysis of student handwritten notes. Driven by the demand for scalable, interpretable, and pedagogically relevant solutions, this framework integrates advanced models from computer vision and natural language processing (NLP) into a cohesive pipeline comprising three interconnected modules:

1)Image Segmentation Module: Employing an enhanced Attention Multi-task U-Net, this module segments complex handwritten pages into meaningful regions, including text, mathematical expressions, diagrams, annotations, and non-verbal symbols.

2)Text Extraction and Normalization Module: Utilizing Optical Character Recognition (OCR) technologies, this module digitizes segmented regions and performs preprocessing steps such as adaptive thresholding, language identification, mathe-

mathematical symbol recognition, and structural annotation.

3) Semantic Analysis Module: Leveraging GPT-4/GPT-4o models, this module interprets textual data semantically, aligning note content with curricular standards and providing pedagogical insights.

To make it easier for readers to understand, I will explain the two versions of GPT here: GPT-4 and GPT-4O.

1) GPT-4: Publicly announced 14 March 2023 and rolled out the same day via ChatGPT Plus and the API.

2) GPT-4o (“o” for omni): Introduced at OpenAI’s Spring Update on 13 May 2024 as the new flagship model, gradually replacing GPT-4/Turbo in ChatGPT and the API.

The key improvements of GPT-4o over GPT-4:

1) Unified Multimodality –GPT-4o processes audio, vision and text within one network, eliminating the three-model voice pipeline used by GPT-4 and retaining non-verbal cues.

2) Real-time Interaction –Average voice latency falls from seconds to sub-second human-like speeds, enabling fluid dialogues and live translation.

3) Efficiency-Architectural optimisations halve cost and double throughput versus GPT-4 Turbo, making high-end reasoning affordable for broader deployment.

4) Larger Working Memory –128 k-token window (300 pages) supports long documents, code bases or meeting transcripts without manual chunking.

5) Stronger Multilingual Capability –A redesigned tokenizer compresses many non-Latin scripts (e.g., Gujarati $\times 4.4$, Telugu $\times 3.5$), improving accuracy and reducing token costs.

6) Enhanced Vision & Audio Reasoning –Outperforms GPT-4 on multimodal benchmarks, recognising scenes, interpreting charts and performing speech-to-speech tasks natively.

7) Safety by Design –Trained with cross-modal red-teaming and new guardrails; risk assessments show Medium across cybersecurity, persuasion, CBRN and autonomy categories.

Therefore, GPT-4o retains GPT-4-level reasoning while delivering real-time, low-cost, fully multimodal interaction—marking a practical leap from research-grade LLMs toward everyday conversational AI.

The modular design ensures flexibility and scalability, allowing educators or administrators to adapt individual components to different educational scenarios and technical requirements. Furthermore, the system supports seamless integration with institutional learning management systems (LMS), facilitating real-time educational feedback and analysis.

Figure 3.1 illustrates the overall framework of this system. A student’s note page is fed into the system, where **1)**the Attention Multi-Task U-Net segments it into text, formula, and diagram regions; **2)**an OCR engine then converts the segmented text into machine-readable form; and **3)**the processed text and images are input into this custom ‘GPT-4 Builder’ module—a pipeline that supplies GPT-4 with the segmented content and a domain-specific knowledge base. The GPT-4 Builder therefore serves as the mechanism for efficiently querying GPT-4 with note-specific data. Following these steps ensures that each component—vision, OCR, and NLP—contributes to a comprehensive end-to-end analysis of the notes.

3.2 Attention Multi-task U-Net for Image Segmentation

The image segmentation module serves as the foundational component for parsing handwritten notes. Originally developed for biomedical image segmentation, the

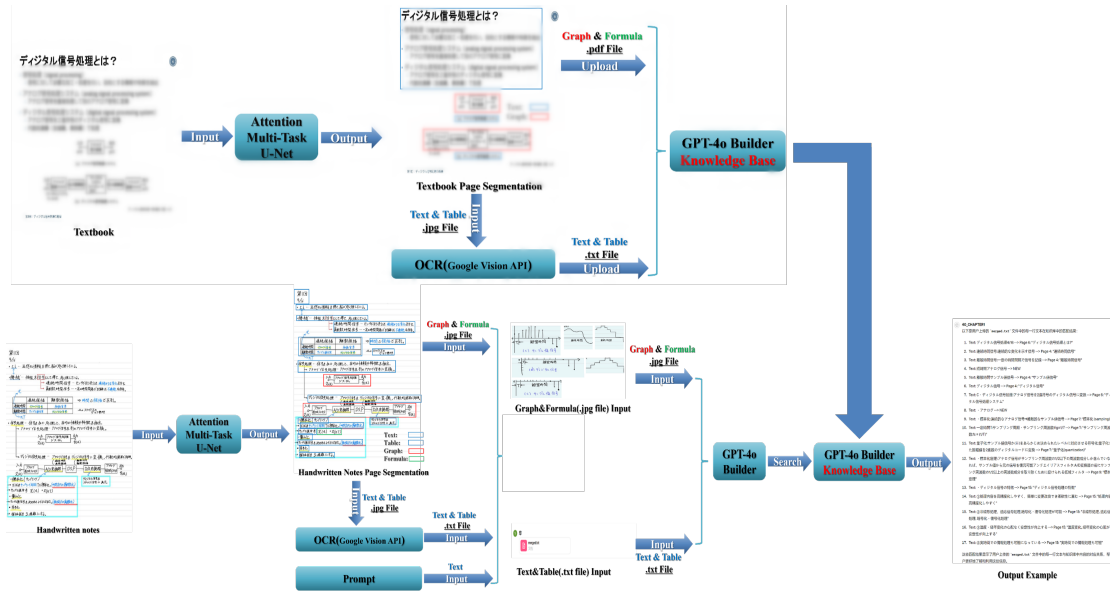


Figure 3.1: System Process Overview

U-Net architecture has been enhanced through multi-task learning and attention mechanisms to better handle educational note complexity. Key enhancements include:

- 1) Multi-task Learning: Simultaneously addresses tasks such as text segmentation, margin detection, and object classification, promoting improved generalization.
- 2) Attention Mechanisms: Attention mechanisms endow the network with the ability to “look where it matters.” Spatial attention dynamically highlights where in the page salient information resides (e.g., title banners, boxed equations, marginal annotations), while channel attention stresses what type of feature is most discriminative (text strokes vs. diagram edges vs. formula symbols). By recalibrating feature maps on both axes, the model suppresses background noise (ruled lines, shadows) and alleviates class imbalance, leading to sharper boundaries and fewer false positives. Incorporating lightweight attention blocks (e.g., SE, CBAM, Attention Gates) adds less than 2% extra parameters yet consistently boosts seg-

mentation metrics .

3) Custom Loss Functions: Combines categorical cross-entropy and Dice coefficient loss for robust boundary detection and model convergence.

Extensive training on annotated student notes from various STEM disciplines ensures robustness and transferability. Performance evaluation metrics include Intersection over Union (IoU), Dice Score, and per-class accuracy, validated through rigorous testing and cross-validation methods.

3.3 OCR and Preprocessing

Following segmentation, OCR processes are implemented to digitize handwritten content. Utilizing a hybrid approach of Tesseract and cloud-based APIs (e.g., Google Vision API), [162, 163] preprocessing involves adaptive thresholding, skew correction, and noise filtering. Each segmented region undergoes language identification, mathematical symbol recognition, and structural annotation to ensure high-quality digitization. Outputs are formatted into structured JSON documents, enriched with metadata for ease of retrieval and pedagogical analysis.

3.4 Vision-to-Language Pipeline

The digitized content is subsequently analyzed through a semantic analysis pipeline utilizing GPT-4/GPT-4o. Core analytical functionalities include:

- 1) Entity Recognition and Disambiguation:** Identification and contextual resolution of key terms.
- 2) Conceptual Alignment:** Matching note content with textbook sections and identifying coverage gaps.
- 3) Natural Language Summarization:** Condensation of dense note content into suc-

cinct summaries.

4) Learning Indicator Extraction: Detection of cognitive indicators such as emphasis and note density.

These processes produce structured XML/JSON outputs compatible with visualization tools and educational dashboards, facilitating dynamic interactions and educational interpretations tailored to instructor-defined objectives.

3.5 System Integration and Data Flow

The proposed framework is implemented as a modular, service-oriented backend using Python, Fast API, and Docker. This architecture emphasizes modularity, scalability, and interoperability, allowing easy upgrades and integration with existing LMS platforms.

Data processing steps include:

- 1)Enhanced U-Net segmentation of input images.
- 2)OCR and metadata annotation.
- 3)GPT-based semantic analysis.
- 4)Visualization of analysis results.

Validation checkpoints, error logging, and fault isolation mechanisms ensure robust performance and easy maintenance. Additionally, built-in security measures guarantee the safe handling of sensitive educational data.

3.6 Evaluation Design

Comprehensive evaluation protocols are designed across multiple dimensions:Image Segmentation: Metrics include IoU, Dice Coefficient, and precision/recall.OCR Accuracy: Character Error Rate (CER), Word Error Rate (WER), and domain-

specific symbol accuracy. Semantic Analysis: Cosine similarity evaluations, expert educator reviews, and instructional relevance assessments. System Usability: System Usability Scale (SUS) assessments conducted with educators and students.

Evaluations are conducted on extensive datasets comprising thousands of handwritten notes, employing qualitative and quantitative methodologies, including surveys, interviews, and statistical analysis.

3.7 Theoretical and Technical Foundations

To deepen the understanding of the technical framework adopted in this study, this chapter introduces the theoretical foundations and development trajectories of two core models: the Attention Multi-task U-Net used for image segmentation, and GPT-4, a representative large-scale generative pre-trained language model. By systematically outlining their structures, principles, and advantages, this chapter lays the groundwork for the technical implementation and empirical analysis in subsequent sections.

3.7.1 Evolution and Principles of the Attention Multi-task U-Net

U-Net is a classical encoder-decoder architecture that was initially applied in medical image segmentation. Its hallmark is the use of skip connections, which concatenate high-level feature maps from the encoder with lower-level decoder outputs, thereby preserving spatial information and improving segmentation precision.

In this study, to address the heterogeneity in student handwritten notes—such as text, formulas, diagrams, and highlights—I have introduced a multi-task learning mechanism. This approach enables the model to perform the primary

segmentation task while simultaneously learning auxiliary tasks (e.g., edge detection, symbol classification), enhancing the model's sensitivity to boundaries and semantic structure.

Furthermore, attention mechanisms were incorporated to strengthen the model's capacity to focus on task-relevant regions. By embedding attention gates in the decoding process, the model dynamically prioritizes content-rich areas during spatial reconstruction, improving segmentation accuracy—particularly for visually dense segments like mathematical matrices and complex schematic diagrams.

3.7.2 Language Understanding Capabilities of GPT-4 and Its Educational Adaptability

GPT-4, developed by OpenAI, is the fourth-generation large-scale language model, known for its powerful generalization and context comprehension capabilities. Based on the transformer architecture and pre-trained on massive corpora in an unsupervised fashion, GPT-4 excels in language generation, reasoning, and semantic matching.

In this study, GPT-4's main role is to process textual regions extracted through image segmentation. After OCR conversion, the handwritten notes are fed into GPT-4 prompts for:

- 1) Key concept extraction
- 2) Semantic similarity matching
- 3) Content summarization and topic inference

These capabilities make GPT-4 a central component in building functionalities such as “note-to-textbook alignment” and “student concept mastery diagnostics.”

One significant strength of GPT-4 is its support for few-shot and zero-shot learning, enabling robust generalization even in data-scarce or annotation-limited

educational contexts. Additionally, GPT-4 can identify latent cognitive patterns in notes—such as alternative expressions of the same topic, disordered logic, or incomplete conceptual chains.

3.7.3 Advantages of Vision-Language Integration Models

Compared to traditional analysis methods that rely solely on visual or textual features, vision-language integration models offer clear advantages in educational contexts. Handwritten notes represent a quintessential multimodal data form, incorporating both visual structure and semantic content.

The “Attention Multi-task U-Net + GPT-4” integrated architecture proposed in this study enables synergistic analysis through the following mechanisms:

- 1) Spatial recognition (Where): U-Net determines location and layout distribution
- 2) Semantic understanding (What): GPT-4 provides contextual interpretation and content alignment
- 3) Pedagogical linkage (Why): By referencing textbook standards, the model extracts educationally meaningful behavioral indicators

This three-tier analytical framework shifts the assessment paradigm from test score-centric metrics toward a multidimensional system focused on learning processes, content mastery, and expressive structure.

In summary, this chapter establishes the theoretical foundation of the study by grounding it in advanced image segmentation and language understanding technologies. It also paves the way for the system implementation and empirical validation that follow in the subsequent chapters.

3.8 Summary

This chapter provided a detailed description of the unified methodological framework designed for automated handwritten note analysis. By integrating advanced segmentation, OCR techniques, and generative NLP models, the proposed system translates unstructured educational artifacts into structured, interpretable learning data. Emphasizing modularity, interoperability, and educational relevance, this framework offers robust technical solutions for real-world educational analytics.

Chapter 4

Segmentation Enhancement: Improvement of Image Segmentation Model for Handwritten Notebook Analytics

Chapter 4 initiates the empirical research of this dissertation by addressing a critical technical challenge in the analysis of handwritten educational notes—accurate image segmentation. Prior to deeper semantic analyses and practical educational applications, reliably distinguishing and categorizing different content types (text, diagrams, equations, and annotations) within student notes is essential.

Therefore, lays the foundational groundwork through the development and validation of an enhanced image segmentation model, specifically tailored to the unique demands of educational analytics. By improving segmentation accuracy and robustness, this work sets the stage for subsequent integration with natural language processing models and eventual practical educational implementations

described in Chapters 5 and 6.

4.1 Research Background

The handwritten notes of students are an important source of feedback on their learning progress. They contain content such as text, formulas, graphs, tables, etc.(As shown in Figure 1.). In the label, the blue region is the text, the red region is the graph, and the green region is the formula.

By analyzing the content of students' notes, it can help teachers understand the learning situation of students. On the student's side, students can be reminded of what needs to be improved. In terms of teachers, teachers can know whether the current teaching speed and difficulty are suitable. Then the teacher can make certain improvements.

At present, the page segmentation task is to segment student notes into character regions, formula regions, and graph regions. The segmentation result allows us to understand the percentage of each region in the student's notes. On the other hand, the character segmentation task allows us to understand the number of characters in the notes taken by the students in each class. Integration of the image-segmentation results enables analysis of student learning. For normalized text content, traditional image segmentation methods can extract character content very well. However, handwritten notes are not as well formed as printed fonts. The best analysis method is to extract the content of the notes and then perform semantic analysis on the content.

In the previous research, Li [1] used U-Net and Post Processing to segment students' handwritten notes and achieved good results. However, the previous page segmentation process was complicated, and two models were used to complete page segmentation and character segmentation tasks. After U-Net completes the

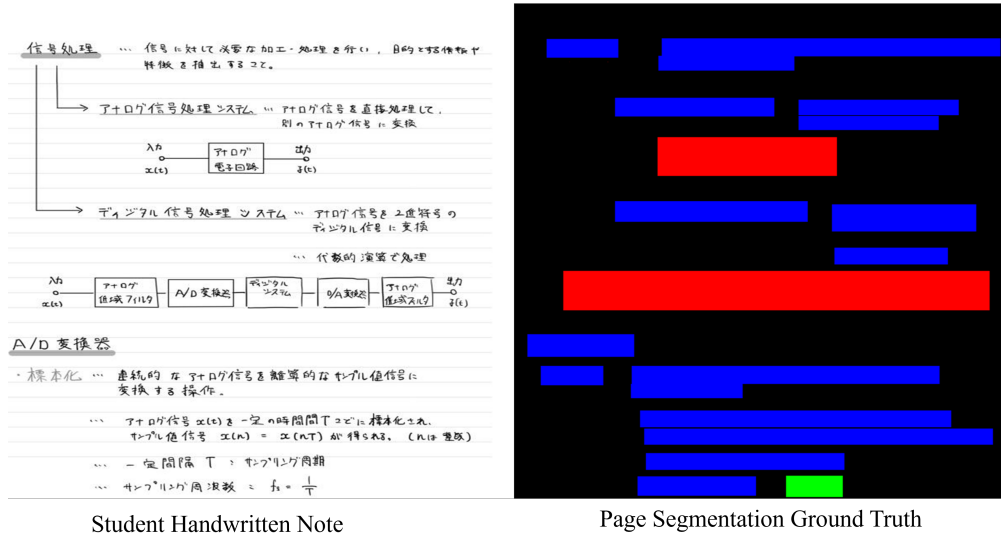


Figure 4.1: Example of student handwritten notes and Page segmentation label for handwritten notes

segmentation, post-processing of the image is required to get the result. To improve the convenience of the model, this paper tries to achieve good segmentation results without using post-processing. This paper will use one multi-task U-Net model for page segmentation and character segmentation tasks.

The performance of an image segmentation model directly affects the accuracy and reliability of the analysis results in handwritten notebook analytics. Improving the image segmentation model for handwritten notebook analytics requires a deep understanding of the challenges and limitations of existing models and the development of new model that can better address these challenges.

This work focuses on improving the image segmentation model for handwritten notebook analytics. The goal is to improve the U-Net model that can accurately and reliably segment handwritten notes in a notebook and extract relevant information from them. The model should be able to handle a wide range of handwritten notes, including different handwriting styles, sizes, and orientations.

To achieve this goal, this study explores new techniques and algorithms that can improve the accuracy and reliability of the image segmentation process, including deep learning-based approaches and traditional computer vision algorithms.

To address these issues, this study introduced attention mechanism [2] and recursive residual convolutional neural network(R2) [3] in the multi-task U-Net model.

A dataset of 2,000 annotated handwritten-note images—collected from undergraduate courses such as Digital Signal Processing—was compiled for this study. Each note image was manually annotated with region labels (text, formula, diagram, etc.) to serve as ground truth for segmentation. I split the dataset into training and validation sets (80/20 split) and evaluated performance using standard metrics: overall segmentation accuracy and F1-scores for each region type. This experimental setup ensured that the model’s improvements could be quantitatively measured on diverse, real-world handwritten notes.

4.2 Attention Multi-task U-net Model

Traditional image segmentation methods mainly include rule-based methods and feature-based methods. Rule-based methods typically use information such as color, texture, shape, etc. in the image to identify the target region in the image. The disadvantage of this method is that it is sensitive to noise and background interference in the image and is difficult to handle complex images.

Feature-based methods, on the other hand, perform image segmentation by calculating features in the image such as edge information and gray-scale histograms. The disadvantage of this method is that it requires a high degree of effectiveness in feature extraction from the image and is difficult to handle complex images.

Therefore, traditional image segmentation methods have problems such as low

accuracy and low efficiency, making it difficult to meet the requirements of handwritten note analysis.

In previous research, Li [1] used U-Net model and post-processing to solve this problem. However, his method is a bit complicated, and the accuracy is not good enough. Therefore, in this research the purpose is to achieve a better segmentation effect. I modified U-Net into multi-task U-Net. As shown in Figure 4.2, the multi-task model shown by removing the attention module from the Attention multi-task U-Net model and then the remaining part is the multi-task U-Net model. The accuracy of the model was improved without the post-processing. More importantly, character segmentation was found to enhance page-segmentation performance.

Due to the limited graph content in the handwritten notes of students, the model is unable to effectively learn the features of image regions. Accordingly, this study adopts a multi-task model for image segmentation in place of a single-task approach. The multi-task model can improve the efficiency of image segmentation and has a mutually reinforcing effect between multiple tasks. Meanwhile, an attention mechanism and a recursive residual convolutional neural network are incorporated to enable the model to learn graphical-region features quickly and accurately, even from a limited number of samples.

4.2.1 Attention Multi-task U-net Model

Attention mechanism can play a crucial role in improving the performance of deep learning models in various applications, including image classification, machine translation, and speech recognition.

Its main advantages are:

1) Improved feature representation: The attention mechanism allows the model to focus on important features in the input and weigh them differently, leading to a

better representation of the input data.

2)Reduced computational complexity: By selectively focusing on related features, the attention mechanism can reduce the computational complexity of the model and improve its computational efficiency. These advantages align precisely with

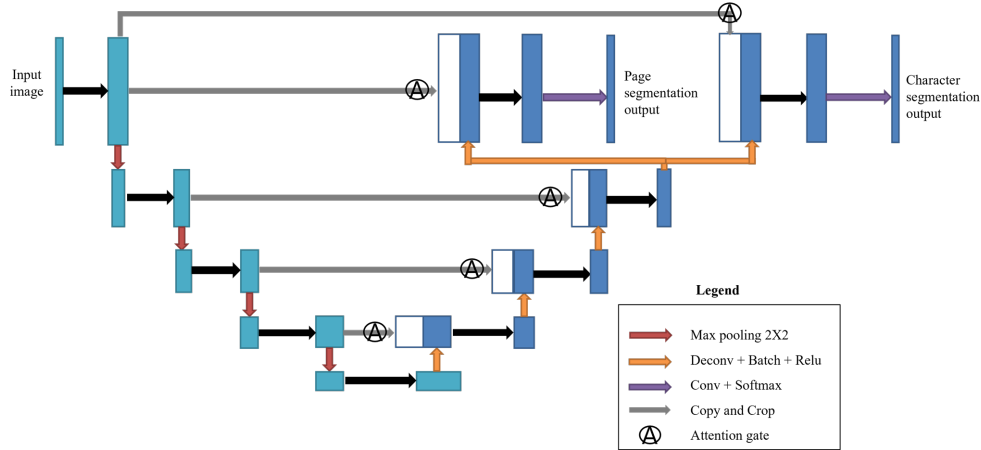


Figure 4.2: Attention Multi-Task U-Net Model

the study's objectives, enabling the model to concentrate on pertinent image regions and thus improve recognition accuracy. Figure 4.3 illustrates the resulting multi-task U-Net augmented with an attention module.

4.2.2 Attention R2 Multi-Task U-Net Model

To further improve performance on the modest student-note dataset (2,000 images), the architecture was augmented with an R2 module, as illustrated in Figure 4.3.

The R2 module has the following advantages:

1)Improved accuracy: RRCNN has a recursive structure that allows it to capture the temporal dependencies in sequential data and improve the accuracy of the model.

2) Better handling of sequential data: The recurrent structure of RRCNN enables it to handle sequential data more effectively, making it well-suited for applications such as time series analysis and speech recognition.

These advantages help the model achieve higher accuracy in small datasets. On the other hand, the R2 module greatly increases the computational complexity of the model and can reduce the training efficiency of the model.

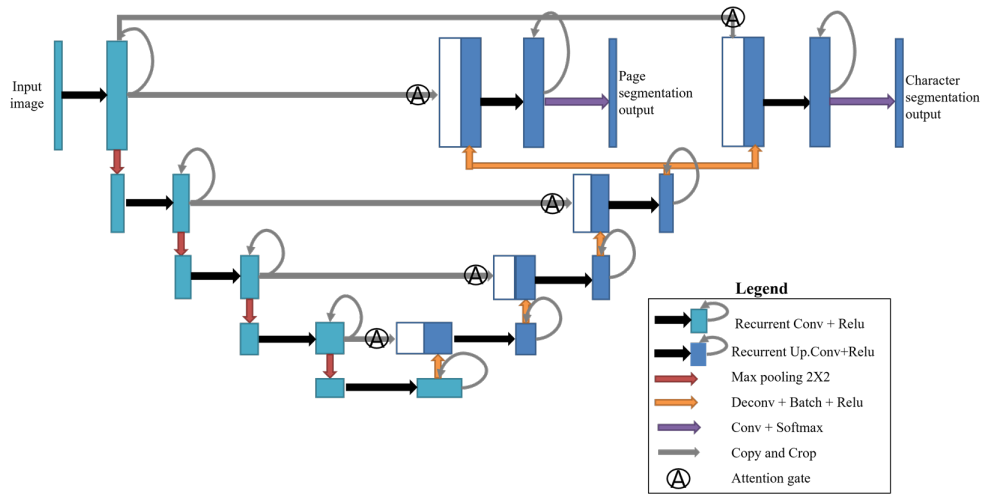


Figure 4.3: Attention R2 Multi-Task U-Net Model

4.3 Experiment Results

Section 3.1 introduces the dataset, Section 3.2 presents experimental results for the original U-Net, multi-task U-Net, attention multi-task U-Net, and attention R2 U-Net, and Section 3.3 compares the attention U-Net with other neural-network baselines.

4.3.1 Dataset Description

The dataset used in this study consists of handwritten notes produced by students in the Digital Signal Processing course between 2021 and 2022. There are a total of 1980 images. The dataset was split into training (70 %), validation (15 %), and test (15 %) subsets. All accuracy figures reported in this paper are based on the model's performance on the test set.

The accuracy of the model in each region is calculated using Equation 1. C represents the number of pixels that are correctly segmented, and L represents the number of pixels in the label.

$$Accuracy = C/L \quad (4.1)$$

4.3.2 Comparative experiments of U-Net model

To investigate the impact of the attention mechanism and the R2 module on model accuracy, comparative experiments were conducted. The experiments contained the Original U-Net [4], Multi-Task U-Net [5], Attention Multi-Task U-Net, and Attention R2 Multi-Task U-NET. As shown in Table 1. The four U-Net models were compared in terms of overall recognition accuracy and region-wise recognition accuracy.

Comparison of the multi-task U-Net model and the original U-Net. It can be clearly found that the multi-task U-Net improves the accuracy over the original single task U-Net in the text(6.74%) and formula(5.08%) region and improves 1.92% in the graph region. This indicates that in the multi-task model, the features learned by the character segmentation task help the model to correctly identify the text and formula regions in the page segmentation task.

Comparison of the Attention Multi-Task U-Net model and the Multi-Task U-Net model. It is worth noting that the accuracy of Attention Multi-Task U-Net

Table 4.1: Comparison of Accuracy Rates by Region

Model	Total Ac/F1	Back ground	Text	Formula	Graph
Original U-Net	0.8162/0.7869	0.9137	0.8253	0.7346	0.5891
Multi-Task U-Net	0.8728/0.8475	0.9406	0.8927	0.7854	0.6083
Attention Multi-Task U-Net	0.9425/0.9264	0.9733	0.9561	0.8651	0.8263
Attention R2 Multi-Task U-Net	0.9458/0.9351	0.9782	0.9608	0.8647	0.8315

Table 4.2: Overall Accuracy and F1 Score of U-Net Variants

Model	Accuracy	F1 Score
Original U-Net	0.9381	0.9172
Multi-Task U-Net	0.9423	0.9287
Attention Multi-Task U-Net	0.9637	0.9561
Attention R2 Multi-Task U-Net	0.9614	0.9596

model is improved by 21.8% in the Graph region compared to Multi-Task U-Net model. The text and formula regions are also improved, with 6.34% for the Text region and 7.97% for the Formula region. The huge improvement in accuracy shows that it is very effective to apply the attention mechanism to the Multi-task U-Net model.

As shown in figure 4.4, in the white rectangle, the model with the attention mechanism clearly identifies more image regions correctly than the multi-task model. In the yellow rectangles, the multi-task model incorrectly identifies most of them as background regions, while the multi-task model with the attention mechanism correctly identifies them as image regions. In summary, the multi-task U-NET model with the attention mechanism achieves higher accuracy in all regions, especially in the graph region.

As shown in the Table 2. It can be clearly seen that the accuracy of Attention model and Attention+R2 model is higher than that of multi-task U-Net and original U-Net . However, the accuracy of adding the R2 module is slightly lower than that of only the Attention module. This indicates that the attention module performs better in this dataset and is more suitable to perform the character segmentation task in this dataset.

On the other hand, the character segmentation results of Attention model and multi-task model are shown in figure 4.5. Attention model segmentation result is better than multi-task model. The results are more complete, with each character being segmented closer to a complete rectangle. This is well demonstrated within the yellow rectangle box. And in the red rectangle, multi-task model does not correctly identify the character, while attention model correctly identifies and segments the character.

Comparison of the Attention R2 Multi-Task U-Net model and the Attention Multi-Task U-Net model. With the addition of the R2 module, the accuracy of

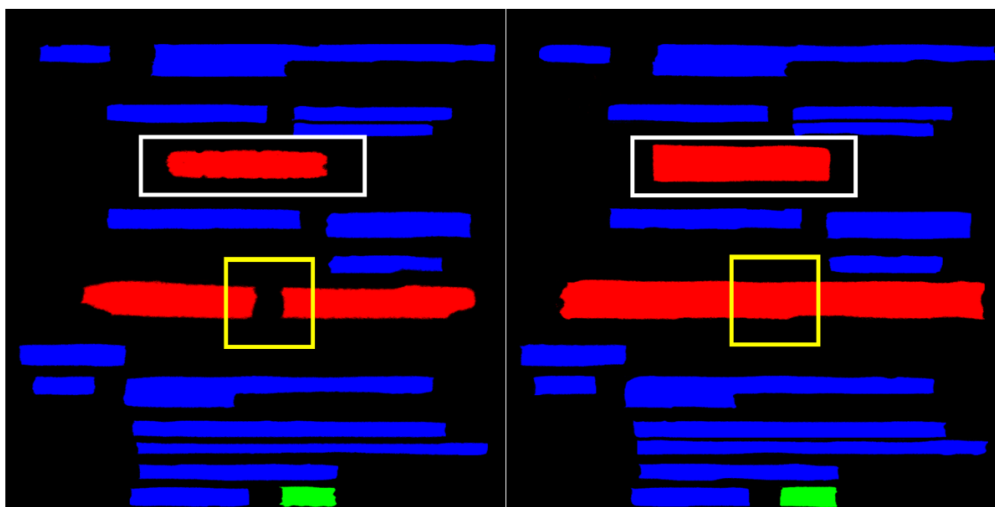


Figure 4.4: Multi-Task U-Net Model Page Segmentation Result and Attention
Multi-Task U-Net Model Page Segmentation Result

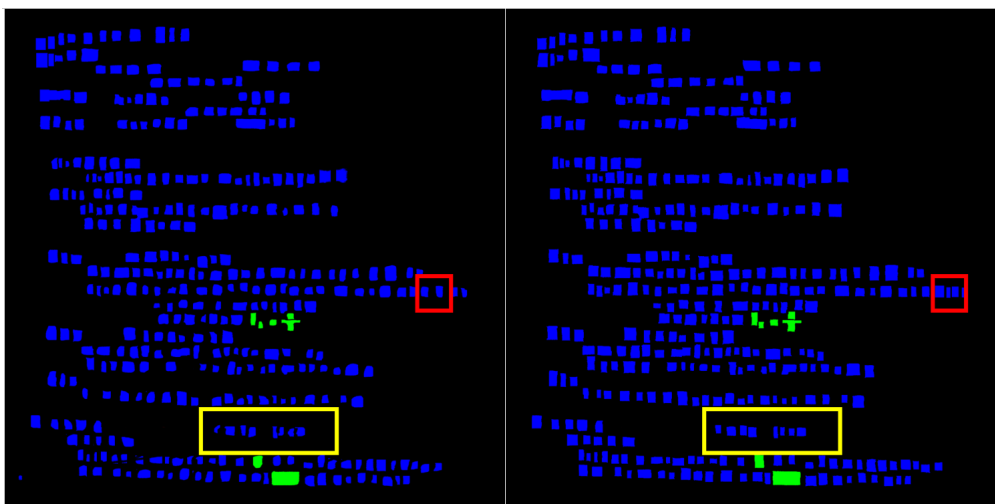


Figure 4.5: Multi-Task U-Net Model Character Segmentation Result and Attention
Multi-Task U-Net Model Character Segmentation Result

Table 4.3: Comparison with other neural networks

	Total Ac/F1	Background	Text	Formula	Graph
Attention Multi-Task U-Net	0.9425/0.9264	0.9733	0.9561	0.8951	0.8263
R-FCN	0.8237/0.7618	0.9085	0.8439	0.7628	0.5619
U-Net++	0.8659/0.8437	0.9246	0.8527	0.7839	0.6815
U-Net3+	0.8812/0.8526	0.9428	0.8773	0.7981	0.6986

the model was improved, but only by 0.33%. It was even reduced by 0.04% in the formula region. On the other hand, it is necessary to consider that the inclusion of the R2 module greatly increases the computational complexity of the model, which has a great impact on the training efficiency.

4.4 Comparison with other neural networks

The performance of four U-Net models was compared on the dataset described in Section 3.2. The AttentionR2 Multi-Task U-Net model achieved the highest accuracy; however, experiments revealed that it was highly inefficient to train. Each epoch takes about three times as long as Attention Multi-Task U-Net model. Considering the small accuracy improvement of Attention R2 model compared to Attention model while training is three times less efficient. Therefore, the Attention Multi-Task U-Net model was selected for comparison with other neural networks. As shown in Table 3. Attention Multi-Task U-Net model performs better than R-FCN [6], U-Net++ [12] and U-Net3+ [8] in the recognition of all regions.

Attention Multi-Task U-Net has obvious advantages over other models. How-

ever, the recognition accuracy of U-Net++ and U-Net3+ is comparable to that of the multi-task U-Net model and even surpasses it in the graph region. Creating a multi-task variant of U-Net3+ and incorporating an attention mechanism could yield results superior to the current model.

4.5 Conclusion

This paper focuses on an improved image-segmentation method for handwritten-notebook analysis and compares four U-Net models. The results show that the multi-task U-Net model outperforms the single-task U-Net model in terms of recognition accuracy, especially in the Text and Formula region. The recognition accuracy of the model in the Graph region is significantly improved after adding the attention mechanism, and the model accuracy is also improved to some extent after adding the R2 module.

In conclusion, this paper shows the enhancement of the character segmentation task for page segmentation. On the other hand, the model with the introduction of the attention mechanism gets a significant improvement in accuracy, especially in the graph region by 21.8%. And the F1 score has also improved. At the same time the model maintains an efficient training efficiency.

Comparison with other networks indicates that U-Net3+ performs strongly; integrating multi-task learning and an attention mechanism into this architecture could surpass the current attention multi-task U-Net model, making it a promising avenue for future research.

The improvement of model accuracy in this paper is of practical application. More accurate segmentation of students' handwritten notes will help to analyze students' learning better.

4.6 Chapter Summary

This chapter contributes significantly to the dissertation by providing a robust methodological foundation essential for subsequent analytical advancements. Specifically, the enhanced Attention Multi-task U-Net segmentation model developed here substantially improves the accuracy and reliability of content extraction from handwritten notes. These improvements are critical for enabling effective integration with semantic analysis models, which will be discussed in Chapter 5. By clearly identifying and segmenting the diverse content within educational notes, the chapter ensures that subsequent analyses can build upon structured, well-defined data. This methodological foundation directly facilitates the exploration of semantic understanding and practical educational applications detailed in later chapters, ensuring a coherent and logical progression within the dissertation.

Chapter 5

Vision–Language Fusion: A Novel Approach—Enhancing Data Extraction from Student Handwritten Notes Using Multi-Task U-Net and GPT-4

Chapter 5 builds directly upon the segmentation improvements introduced in Chapter 4. With accurate image segmentation now achievable, the dissertation moves forward by integrating these segmentation results with advanced natural language processing (NLP) techniques. This integration is essential to extracting deeper semantic insights from segmented note components, enabling a richer understanding of educational content and student learning behaviors. By combining structural segmentation with generative AI models such as GPT-4, this chapter significantly enhances the interpretability and educational relevance of automated

handwritten note analyses, forming a critical bridge between foundational image processing and practical educational applications detailed in Chapter 6.

5.1 Research Background

Handwritten notes by students are a critical data source reflecting their learning status. Traditionally, educators have had to review these notes individually to gauge students’ comprehension and mastery of course material. However, this method is time-consuming, inefficient, and often fails to capture and quantify students’ learning progress and challenges comprehensively. With advancements in technology, especially in text recognition and machine learning, new avenues have opened for automating this review process. This allows for a more efficient and systematic analysis of students’ learning situations.

For example, in this paper student handwritten notes include text, formulas, tables, and images, as shown in Figure 5.1. Students’ methods of taking notes vary. Accurately and efficiently extracting data from these notes significantly aids teachers in analyzing student learning and helps students improve notetaking and review materials. However, traditionally, teachers assess students’ learning by evaluating their notes. However, this method is time-consuming and does not provide timely feedback to students.

In earlier work, an Attention Multi-Task U-Net model was developed for page- and character-level segmentation of student notes. The segmentation results of this model provide us with the character count and the proportion of characters, formulas, and text in student notes. These data enable basic learning analytics—for example, quantifying character counts and examining whether high-performing students prefer text or diagram-based note-taking.

Based on our previous, this paper offers a method combining Attention Multi-

task U-Net and GPT-4, providing teachers a more accurate and efficient way to evaluate student notes and understand learning situations. More in-depth learning analytics requires a model capable of understanding the content of student notes. Consequently, GPT-4 was incorporated into the experimental process to leverage its natural-language-processing capabilities for understanding student notes.

The integration of GPT-4 has greatly enhanced the data extraction capabilities. It not only understands text but also connects different information fragments, providing analysis and summaries. Additionally, this technology has a wide range of applications in the field of education. These features are crucial for understanding how students integrate classroom knowledge, especially when assessing their depth of understanding and note-taking skills. The proposed method particularly aids teachers and students in achieving more effective teaching and learning. The primary motivation of this paper is twofold: first, for teachers, this system enables them to monitor student progress more effectively, through quantitative and qualitative analysis of notes, identify difficulties in learning, allowing timely adjustments in teaching strategies. Second, For students, the system not only offers timely feedback on the quality of their notes but also recommends personalized learning resources based on their notes, aiding in more targeted review and consolidation of knowledge.

5.2 Methodology

Section 5.2 will detail the method for extracting data from student notes. Section 5.2.1 introduces the image segmentation model used: Attention Multi-task U-Net. Section 5.2.2 details the development of a specialized GPT-4 model for learning analysis using GPT BUILDER. Section 5.2.3 explains the process of handling student notes, specifically how to convert them from complete images to a data format

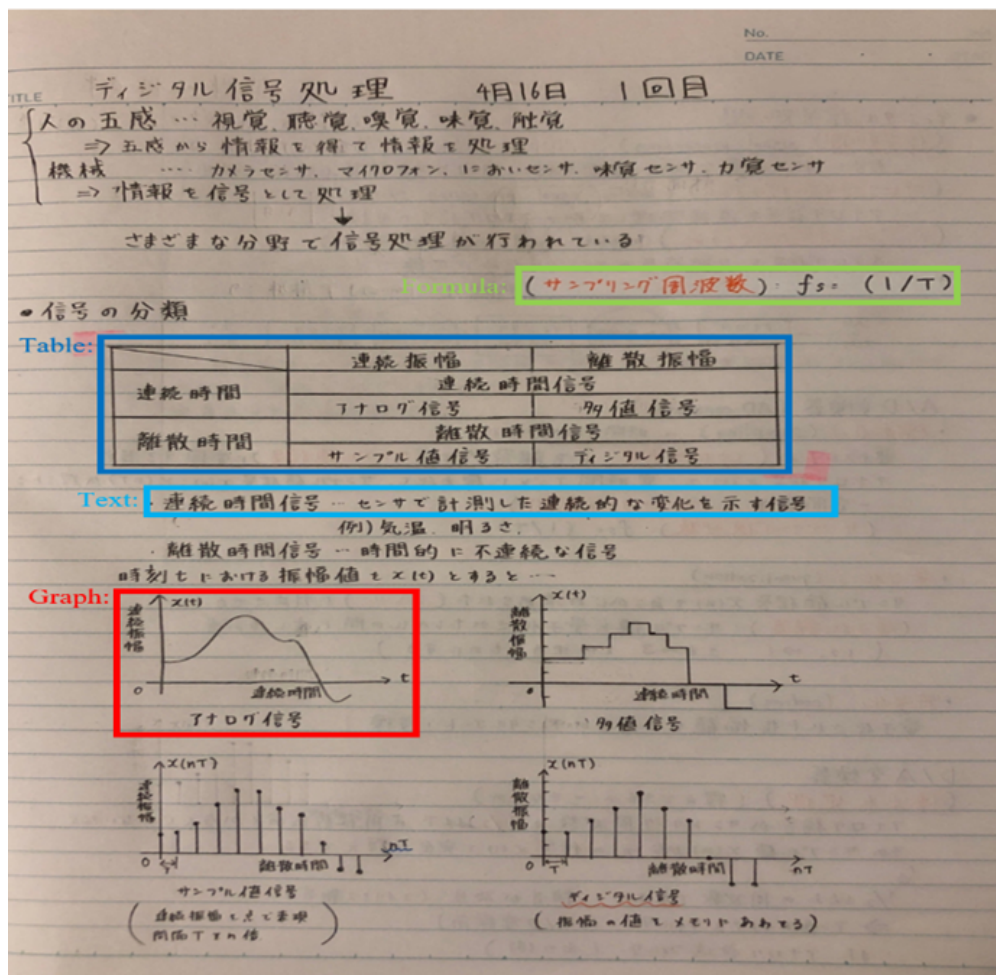


Figure 5.1: Example of handwritten note segmentation

comprehensible by GPT-4. Section 5.2.4 will present the posed questions and the final outputs.

In the previous research indicates that character segmentation significantly positively affects page segmentation tasks. When processing student handwritten notes, the model encounters difficulties in learning image area features due to the scarcity of graphic content. This discovery motivated the adoption of a multi-task learning approach. Compared to traditional single-task models, it is more efficient in image segmentation and offers complementary performance improvements across different tasks.

5.2.1 Attention Multi-Task U-Net

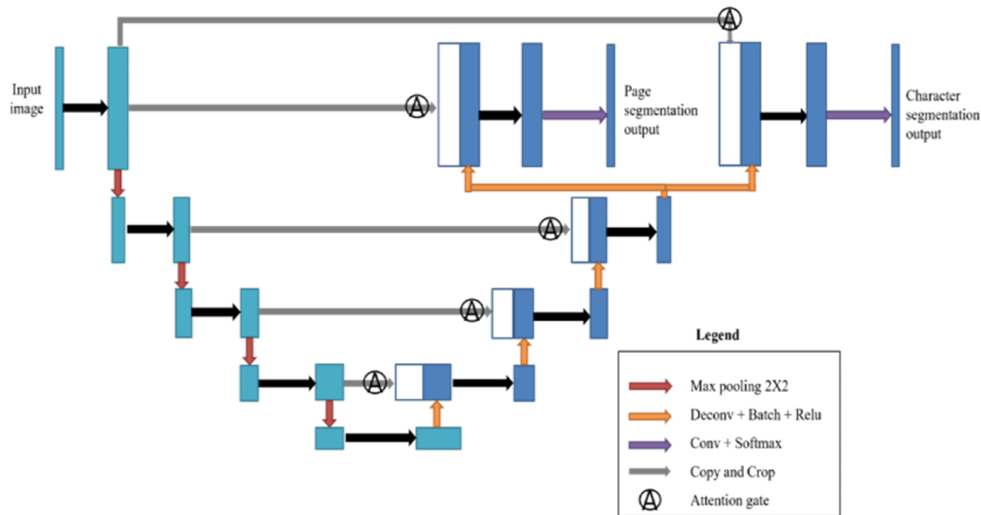


Figure 5.2: Attention Multi-Task U-Net

Figure 5.2 shows the improved Attention Multi-task U-Net model. This model extends the original U-Net architecture to support parallel processing of multiple tasks. To improve performance on image-content segmentation tasks with limited

samples, an attention mechanism was integrated into the model. This mechanism significantly improves the model’s accuracy in identifying image content in datasets, especially in situations with few samples. Additionally, the attention mechanism enhances the accuracy and efficiency of segmentation tasks by improving the model’s responsiveness to key areas.

5.2.2 GPT-4 Builder

The GPT-4 Builder used in this research is an advanced language processing framework developed based on OpenAI’s GPT-4 model. GPT-4 represents the cutting edge of current natural language processing technology, achieving deep understanding and generation of human language through extensive data pre-training. The model excels in various natural language understanding and generation tasks, capable of context analysis, logical reasoning, and complex text synthesis. Although GPT-4 has made significant progress in image understanding, experiments show limitations in its ability to parse text content within images. In particular, GPT-4’s text recognition capability is limited when dealing with non-Latin languages, such as Japanese. To compensate for this shortcoming, the research adopts a new approach to enhance GPT-4’s performance in processing handwritten Japanese notes.

As shown in Figure 5.3, the process begins with inputting student notes into Attention Multi-task U-Net for image segmentation to identify and differentiate elements such as text, tables, graphics, and formulas. Subsequently, the segmentation results are divided into text files (.txt) and image files (.jpg); the text files include text and tables, while the image files contain graphics and formulas.

This research utilizes Google Vision’s OCR API to accurately extract Japanese characters from texts and tables. Once processed by OCR, the text is saved in

.txt format for further processing and analysis. Each page of notes is saved as a separate .txt file, with chapter and page information embedded in the filename for efficient retrieval by GPT-4. Additionally, considering file upload limits and content integrity, all .jpg files of diagrams and formulas are compiled into a single .pdf file per textbook chapter. Each page of the .pdf file corresponds to one image with detailed naming information, as shown in Figure 5.4. This step ensures the number of uploaded files stays within limits and facilitates efficient processing of multi-page images by GPT-4 Builder.

The method proposed in this paper enhances GPT-4’s ability to process non-English image text content, crucial for improving the accuracy and efficiency of student note analysis. With this approach, GPT-4 can more comprehensively acquire and understand information in student notes, playing a more significant role in learning analysis. The upper half of Figure 5.4 displays an image, with its

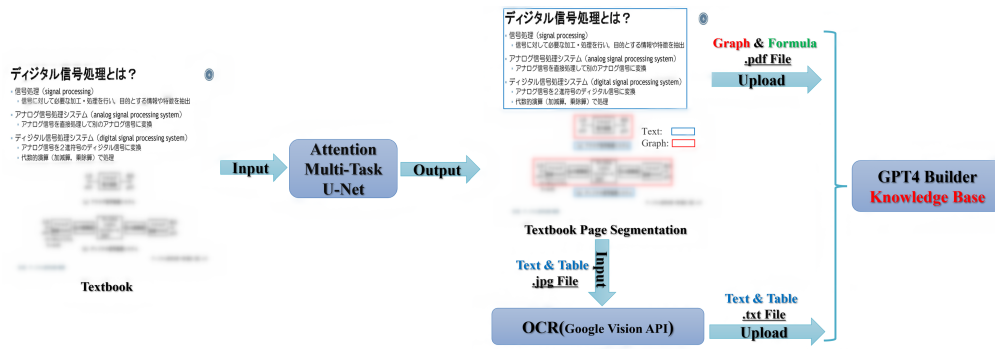


Figure 5.3: Building the GPT-4 Builder Process(Due to copyright issues, some images in the graph are made blurry.)

corresponding chapter, page number, and sequence number on the page indicated in the lower left corner. The lower half shows a photo of a formula, with information on its chapter, page number, and sequence number on the page in the lower left corner.

Following the workflow illustrated in Figure 5.3, the textbook chapter was converted into several .txt files and a single .pdf file. Finally, the created .txt and .pdf files are uploaded to the GPT4 Builder’s knowledge base, completing the GPT4 construction for a specific chapter.

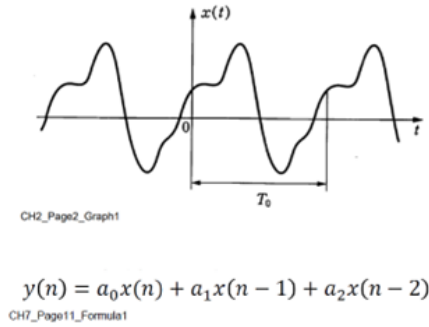


Figure 5.4: Examples of saving graphs and formulas in textbooks as .pdf files

5.2.3 Handwritten Notes Processing

This section integrates the vision and language components to analyze student notes. First, note image of each student is segmented by the Attention Multi-Task U-Net into four parts: (i) text, (ii) tables, (iii) graphics, and (iv) formulas. Next, the text and table segments are processed with Google Vision OCR, converting all handwritten characters (including Japanese text and mathematical notation) into digital text. Simultaneously, the graphic and formula segments are saved as image files. A custom pipeline—referred to as the GPT-4 Builder—is then employed to process these outputs.

The GPT-4 Builder constructs a chapter-specific knowledge base by compiling OCR-extracted text (as .txt files) and images (graphics/formulas combined into a single PDF) that correspond to the relevant textbook chapter. This organized

knowledge base is then uploaded to GPT-4, enabling the model to efficiently search and reference the content when analyzing a student’s note. In summary, the GPT-4 Builder ensures GPT-4 has access to both textual and visual materials from the notes in a structured manner, thereby enhancing its comprehension and answer accuracy.

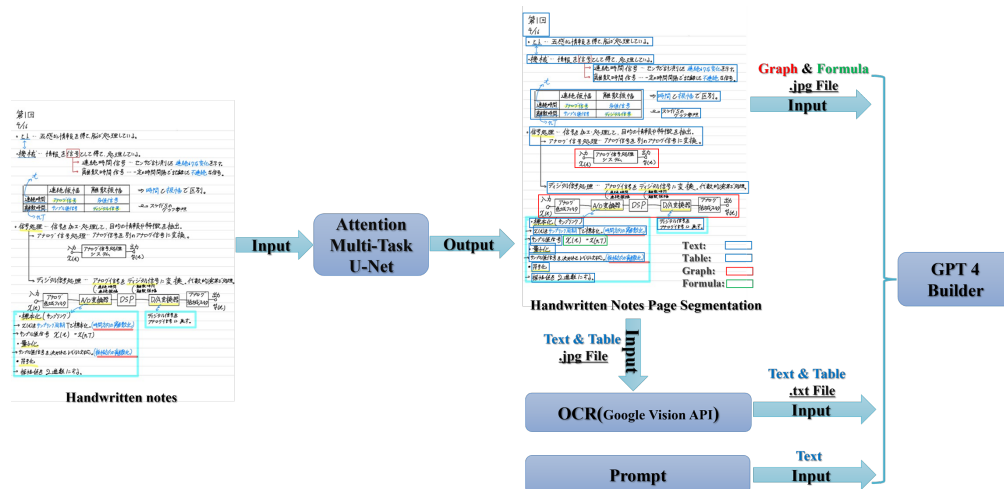
Figure 5.5 illustrates the process of processing student notes. First, student notes are input into the Attention Multi-task U-Net, and the resulting segmentation is divided into two parts. The Text and Table sections are then processed through OCR to extract text. The Graph and Formula sections are uploaded directly to GPT4 as .jpg files for questioning.

For questioning, the .txt files extracted by OCR and the .jpg files segmented by the model are uploaded to the GPT4 dialogue box as materials. To ensure GPT-4’s efficiency and accuracy, each question’s prompt is fully specified, as shown in Figure 5.6. GPT-4 is instructed to search its knowledge base and output the textbook page numbers that correspond to the student-note text. If the text from student notes is not found in the knowledge base, it is marked as [New].

For images, GPT-4 is likewise instructed to query its knowledge base and output the names of textbook images corresponding to student-note images. If the image is not found in the knowledge base, it is marked as [New]. Figure 5.5 shows the detailed process for processing student note content.

The first step of the process involves inputting the image of student notes into Attention Multi-task U-Net for page segmentation. The model’s output is divided into two types: text and tables, and graphics and formulas. Text and tables are extracted with optical character recognition (OCR) and saved in .txt format. Graphics and formulas, on the other hand, are saved as .jpg image files.

When formulating each question, the OCR-extracted text files and the model-segmented image files are uploaded to GPT-4’s interactive interface. To ensure efficient and



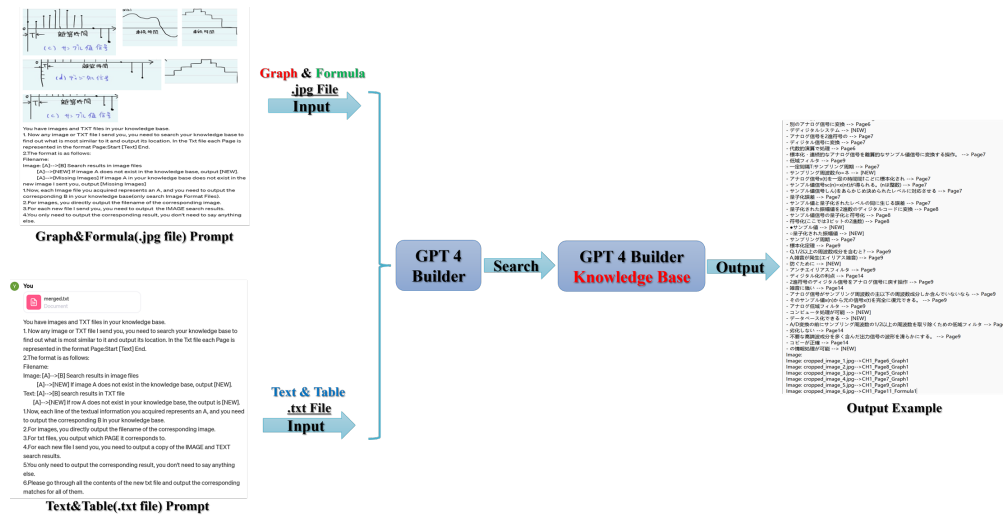


Figure 5.6: Question and Answer Process

5.2.4 Question & Answer Process

Given GPT-4’s computational limitations in handling multiple tasks within a single conversation, a step-by-step processing approach was adopted. Specifically, the questioning task was divided into two main categories—images and text—and addressed separately in two distinct conversations with the GPT-4 model. Once the model receives the prompt, it starts searching in its built-in knowledge base (as shown in Figure 5.7) and generates results based on the information found. Input data includes images and text. The model analyzes each type and searches for the most relevant information in the knowledge base, outputting it if found, or marking it as new information ([New]) if not. This process is designed to optimize the operation of the GPT-4 model, ensuring efficient and accurate information processing. This enables us to effectively use GPT-4 as a knowledge base to accurately answer various complex queries.

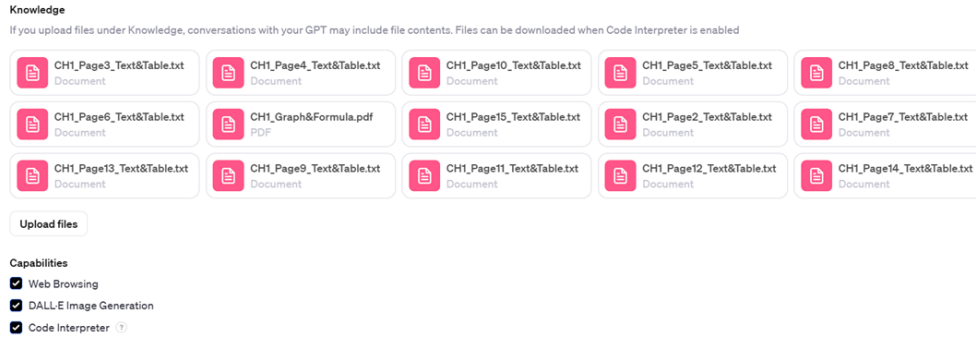


Figure 5.7: Knowledge Base

5.3 Experiments and results

Section 4.3.1 presents the dataset used in this paper. In section 4.3.2, the results of the experiments based on Figure 5.6 will be shown.

5.3.1 Dataset description

The dataset used in this study includes handwritten notes from 80 students in a Digital Signal Processing course in 2021. These notes cover all 11 chapters of the course, providing a rich dataset for the experiment. Each chapter’s notes from every student underwent the process shown in Figure 5.6 independently, ensuring the systematic and comprehensive nature of the experiment.

Preliminary analysis of the dataset revealed significant differences in the characteristics of notes across chapters. Specifically, chapters 1 and 11 focus mainly on textual descriptions, resulting in a higher volume of text and fewer diagrams and formulas in student notes. Conversely, chapters 2 to 10 contain many diagrams and formulas, leading to a higher proportion of these elements in student notes, suggesting a greater emphasis on methodology, computational processes, and graphical representation of data in these chapters.

$$Accuracy = (correctclassifications)/(allclassifications) \quad (5.1)$$

5.3.2 Evaluation Protocol

Two complementary metrics were used:

- 1)Item-level accuracy –proportion of correctly matched note fragments (text, tables, graphs, formulae) to their unique textbook counterparts.
- 2)Coverage (New-rate) –proportion of fragments labelled [New], i.e. not found in the reference textbook, which serves as a proxy for the system’s recall of novel or reformulated content.

The Attention Multi-task U-Net produced region masks, Google Vision OCR rendered text, and GPT-4 performed content matching following the pipeline in Fig. 5.6.

5.3.3 Quantitative Results

Table 5.1 summarises accuracy across chapters 1–11. Key observations:

Overall performance:Mean correspondence accuracy reached 0.861 in text-heavy chapters (1, 11) and 0.866–0.879 in math-/graph-centric chapters (2-10).

Modality gapGPT-4 was consistently 8–10 pp more accurate on Graph and Formula segments than on Text & Table. The stricter semantic equivalence required for paraphrased sentences explains the lower text score, whereas diagrams and equations admit fewer semantically “near-miss” alternatives.

Novel contentThe New column averages 0.83, indicating that roughly one-fifth of student annotations extend beyond the canonical textbook phrasing—useful evidence of student elaboration or misunderstanding.

Table 5.1: Accuracy of Chapter 1–11

Chapter	Text & Ta- ble	Graph	Formula	Corresponds	New
1	0.850	0.891	0.943	0.861	0.825
2	0.849	0.919	0.914	0.866	0.816
3	0.823	0.895	0.944	0.849	0.809
4	0.818	0.908	0.914	0.842	0.825
5	0.829	0.890	0.949	0.856	0.807
6	0.832	0.893	0.921	0.853	0.814
7	0.844	0.881	0.946	0.865	0.855
8	0.836	0.917	0.931	0.862	0.842
9	0.869	0.883	0.932	0.879	0.829
10	0.816	0.896	0.915	0.837	0.846
11	0.827	0.901	0.924	0.834	0.835

5.3.4 Chapter-wise Trends

Chapters 1 & 11 (Intro / Summary) –High text density yields the lowest overall accuracy; GPT-4 sometimes mapped concise student notes to neighbouring textbook paragraphs, failing the one-to-one criterion.

Signal-processing core (Ch. 4–9) –Rich graphical notation boosts graph/formula accuracy beyond 0.94; diagram structure provides strong visual anchors, and mathematical symbols are less ambiguous than prose.

Methodology chapters (Ch. 2–3) –Slight dip in Graph accuracy (0.90) arises from hand-drawn block diagrams whose stylised arrows differ from textbook CAD figures.

5.3.5 Discussion and Implications

The experiment confirms that the proposed vision-language pipeline can reliably align 86–88% of student note content with textbook knowledge units on unseen data, while automatically flagging novel or idiosyncratic annotations for instructor review. The modality-specific analysis highlights where different instructional media succeed or fail:

Graphs & Formula: Already near-optimal; focus should shift toward pedagogical interpretation (e.g., detecting persistent symbol misuse).

Text & Table: Requires refined semantic similarity measures or a relaxed “fuzzy match” rubric to avoid penalising legitimate paraphrases.

New Fragments: Offer a quantitative window into students’ creative elaborations and misconceptions, suggesting personalised follow-up questions.

Overall, these findings demonstrate that integrating Attention Multi-task U-Net with GPT-4 provides a scalable, modality-aware analytics tool that delivers actionable insight for both teachers and learners, thereby satisfying the professor’s

call for a deeper, data-driven discussion of experimental outcomes.

```
Results for 18723
Filename: merged.txt
Text:
- 第一回授業ノート --> [NEW]
- ヒト五感から周囲の情報を得て、それを脳などの中枢神経で処理 --> Page2
- 情報の分類 --> Page4
- 時間、距離が連続が離散によって分類される --> [NEW]
- "機械"に同様のことを行わせるには --> [NEW]
- 各五感の機能に応じたセンサーを用いて周囲の情報を「信号」として入力する必要がある --> Page2
- 離散時間サンプル値信号 --> Page4
- 各信号の形態 --> [NEW]
- アナログ信号 --> Page4, Page6
- 連続時間信号に対しては、時間幅を  $x(t)$  と表す --> Page4
- 離散時間信号も一定の時間間隔でサンプリングするため、 $n$  番目にサンプリングした振幅を  $X(nT)$  と表す --> [NEW]
- (1) アナログ信号 --> Page4
- (2) サンプル値信号 --> Page4
- デジタル信号処理ページ --> [NEW]
- デジタル信号 --> Page4, Page6
- (4) デジタル信号 --> Page4
- (3) 4の通りは振幅が連続が離散であるが、一見違いは分らない、離散振幅は振幅の値が離散的で、ある決められたレベルに対応する図の場合、整数値 --> [NEW]
- 信号処理について --> Page6
- 信号処理信号から、必要な情報、特徴を得るために、ある処理、加工を行うこと --> Page6
- (1) アナログ信号処理システム --> Page6
- アナログ信号処理システム --> Page6
- デジタル信号処理システム --> Page6
- アナログ信号 --> Page4, Page6
- アナログ信号アナログ信号 --> [NEW]
- 時間に対する離散化 --> [NEW]
- 低域フィルタ --> Page9
- アナログ信号をデジタル信号に変換する流れ --> [NEW]
- アナログ信号 --> Page4, Page6
- デジタル信号 --> Page4, Page6
- デジタル信号処理ページ --> [NEW]
- サンプル値信号=デジタル信号 --> [NEW]
- 振幅に対する離散化 --> [NEW]
- (2) デジタル信号処理システム --> Page6
- デジタル --> Page6
- 2進数デジタルコード --> Page8
- アナログ信号を一定の時間間隔でサンプリングし、サンプル値信号を得ること --> Page7
- サンプリング間隔 --> Page7
- (8-1) サンプリング周波数 --> [NEW]
- サンプル値信号の振幅もある決められたレベルに対応させ、デジタル信号を得ることの際に生じる誤差を量子化誤差という --> Page7
- デジタル信号の振幅を2進数のデジタルコードに変換すること --> Page7
- デジタル信号をアナログ信号に戻す操作 --> Page9
- 離散化定理(D/A変換で重要な定理) --> Page9
- アナログ信号がサンプリング周波数の1/2以下の周波数しか含んでいないとき元の信号(上)を完全に復元できる --> Page9
-  $X(1/2)$  以上の周波数成分を含むと、雑音が生じる --> Page9
- デジタル化の利点 --> Page14
- 雑音に強い --> Page14
- 電子コンピュータ処理可能 --> [NEW]
- デジタル信号3ページ --> [NEW]
- 応用しない --> Page14
- 正確なコピーが可能 --> Page14
Image: cropped_image_2.jpg-->CH1_Page6_Graph1
Image: cropped_image_3.jpg-->CH1_Page8_Graph1
```

Figure 5.8: Examples of Output Results

5.4 Conclusion

This paper introduces a new method that combines Attention Multi-task U-Net and GPT-4 to extract data from students' handwritten notes for more in-depth learning analysis. An Attention Multi-Task U-Net model is employed for page- and character-level segmentation, after which GPT-4's natural-language-processing capabilities are used to interpret the content of the notes. The experimental results demonstrate the effectiveness of this method, particularly in processing handwritten Japanese notes.

This research enhances the intelligence of educational analysis tools, provides teachers with in-depth student learning data, and offers personalized learning support to students. In summary, this technology has broad application potential

in education, particularly in enhancing teaching efficiency and student learning outcomes.

Given GPT-4’s computational constraints and cost, future work will focus on developing a customized model tailored to understanding and analyzing handwritten-note content. This model will focus on parsing and processing specific types of data in education, such as student notes and teaching materials. Although this customized model may not be as widely adaptable and flexible as GPT-4, it can achieve higher efficiency and accuracy in specific tasks through targeted training.

Additionally, developing customized models may reduce operational costs, making them more suitable for long-term and large-scale educational applications. With further research and improvement, this model can provide more precise and efficient learning analysis tools, significantly impacting the field of educational technology.

5.5 Chapter Summary

This chapter significantly advances the research framework by seamlessly integrating structural image segmentation and semantic analysis using GPT-4. The proposed integrated approach transforms segmented handwritten notes into educationally meaningful insights, addressing the limitations of standalone segmentation or NLP methods. This enhanced analytical capability is pivotal, as it sets a direct pathway toward practical, real-world educational analytics explored extensively in Chapter 6. The outcomes of this chapter ensure a smooth transition from foundational computational methods to applied educational practices, demonstrating a clear, progressive research trajectory.

Chapter 6

Advanced Unified Framework: Extracting Learning Data from Handwritten Notes: A New Approach to Educational Data Analysis Based on Image Segmentation and Generative AI

Chapter 6 represents the culmination of the methodological advancements introduced and developed in Chapters 4 and 5. Having established robust techniques for image segmentation and semantic analysis, this chapter applies these integrated methodologies to real-world educational contexts. It explores how these advancements can effectively inform instructional strategies, enhance personalized learning interventions, and provide actionable insights into student learning pro-

cesses and outcomes. By explicitly focusing on practical applications and empirical validations in authentic educational settings, this chapter underscores the tangible pedagogical impact and real-world utility of the developed analytical framework.

6.1 Research Background

In the field of educational data analytics, transforming handwritten notes into meaningful learning data is gaining traction. Handwritten notes offer insights into student engagement and comprehension but analyzing them manually is labor-intensive and lacks scalability. In the context of modern education, automated note analysis methods are particularly important as the amount of educational data increases and the need for personalized learning grows.

Handwritten notes contain a large amount of unstructured information, such as diagrams, formulas, and text, which can help teachers understand students' mastery of course content and possible knowledge blind spots. Therefore, it is important to study efficient and accurate handwritten note analysis methods for educational decision-making.

In recent years, developments in image segmentation and natural language processing (NLP) have gradually addressed these issues [159, 164]; however, there are still major challenges in analyzing complex, unstructured handwritten note data.

Image segmentation techniques, such as U-Net and its variants, have demonstrated good results in many image segmentation tasks. However, many existing models still suffer from a lack of accuracy when dealing with complex layouts and diverse content structures in educational notes. For example, in educational environments, notes have different formats and content layouts, including dense mathematical symbols and hand-drawn graphs, which pose a significant challenge

to traditional image segmentation models. Further improving the adaptability and accuracy of segmentation models, especially those that can simultaneously recognize elements such as text, formulas, and graphs, is crucial to improving the accuracy of handwritten note analysis.

This study employs the Attention Multi-task U-Net model, which improves page segmentation accuracy by introducing an attention mechanism while using character segmentation as an auxiliary task to improve segmentation accuracy. The multi-task feature of this model enables it to handle multiple note elements, such as text, formulas, images, and tables, simultaneously, effectively compensating for the shortcomings of existing models in segmentation.

In recent years, the integration of attention mechanisms into segmentation models has shown substantial promise in focusing on key areas within complex data. For example, attention layers can help models prioritize dense regions of text or intricate diagrams, reducing the likelihood of misclassification and ensuring more precise segmentation. This targeted focus is especially beneficial for handling the intricate structures typical of handwritten notes, where text, images, and symbols are intermingled in non-standard formats. By leveraging these advancements, the study's model is designed to efficiently dissect multifaceted content, thus providing an effective foundation for subsequent analysis.

Similarly, advances in NLP technology have made it possible to efficiently understand the content of handwritten notes. However, traditional NLP models often have limitations in accurately interpreting unstructured handwritten text. The maturity of OCR technology has laid the foundation for the automatic recognition of handwritten text, and generative AI models, especially advanced models such as GPT-4o, have made significant improvements in semantic understanding. By deeply understanding and analyzing text content through generative AI, not only can key knowledge points be identified, but potential blind spots in student

learning can also be anticipated, providing a more comprehensive view of student engagement with the material.

The research in this paper relies on the combination of OCR and generative AI, utilizing the advanced GPT-4o model, which accurately matches note content with textbook content through its powerful comprehension capabilities, identifying core knowledge points and discovering potential knowledge blind spots. GPT-4o's robust understanding of language and contextual nuances allows for a level of analysis that traditional NLP models struggle to achieve, particularly in recognizing contextual relevance between student notes and core curriculum concepts.

This combined approach of enhanced segmentation and generative NLP offers a scalable solution for extracting detailed learning insights from handwritten notes, enabling data-driven educational strategies that enhance teaching effectiveness. This method can not only be applied in actual teaching scenarios, enhancing instructors' ability to adapt lessons in real-time, but also provides new ideas for future educational data mining. As automated note analysis continues to evolve, the insights derived from student notes can form the basis for developing personalized educational feedback systems that adjust to each student's unique learning trajectory, ultimately contributing to adaptive and individualized learning environments.

Through this approach, educators can more accurately monitor students' learning progress, identify specific areas that may require additional focus, and thus develop more targeted teaching strategies that foster improved educational outcomes.

The model architecture was designed with priority given to computational efficiency and sensitivity to fine-grained features. After systematic comparisons with mainstream models (e.g., FCN, SegNet, SAM), the U-Net-based architecture was selected for its lightweight design and superior performance in capturing stroke-

level details (achieving 0.9561 text region accuracy, 6.06% higher than SAM), as demonstrated in Section IV.B.1. This choice aligns with the pipeline’s requirement for rapid preprocessing while maintaining high segmentation precision.

6.2 Related Work

6.2.1 Applications of Image Processing in Education

1) SEGMENTATION AND CLASSIFICATION OF HANDWRITTEN NOTES:

Current research has shown that image processing technologies can be effectively used to segment and classify handwritten notes, allowing for more efficient recognition of text, diagrams, and formulas. For example, Li et al. (2022) [1] applied segmentation techniques to divide notes into text, mathematical expressions, and diagrams. This segmentation enables a more refined analysis of learning behaviors, offering insights into students’ note-taking patterns and the effectiveness of their study methods. Further research has shown that models like Attention U-Net and U-Net++ are also can be applied in educational contexts, demonstrating high accuracy and multitasking capabilities in image data processing. However, the diversity and complexity of educational notes continue to challenge these techniques. This study optimizes segmentation models to address the complex demands of educational data analysis.

2) IMAGE SEGMENTATION TECHNOLOGIES IN EDUCATIONAL CONTEXTS:

Image segmentation is vital for recognizing handwritten content in educational settings. Models like Attention Multi-task U-Net, known for their high precision and multitasking capability, have recently been widely applied in educational image data processing. These models accurately segment various parts of handwritten notes, such as text and graphs, which supports subsequent data analysis and

enhances the overall understanding of students' learning progress.

3)APPLICATIONS OF DEEP LEARNING IN THE CLASSIFICATION OF HANDWRITTEN NOTES:

Rani and Suprathi (2022) proposed a deep learning-based approach for classifying handwritten notes, significantly improving the organization and analytical efficiency of students' notes. Studies like these demonstrate how deep learning can enhance the classification and management of handwritten notes, contributing to the optimization of educational resources and the systematic organization of study materials.

6.2.2 Applications of Natural Language Processing in Education

1)AUTOMATED EVALUATION AND ANALYSIS OF HANDWRITTEN CONTENT:

Natural Language Processing (NLP) has been widely applied in the automated evaluation of handwritten notes. For example, Rowtula et al. (2019) [9] developed an automatic grading system that combines Information Retrieval (IR) and NLP technologies, achieving grading accuracy comparable to human evaluators. This system demonstrates the potential of NLP in the efficient evaluation of handwritten notes, reducing the workload on educators. Moreover, the combination of OCR and generative AI is gaining attention to recognize and understand complex handwritten content. This research further leverages the advanced generative AI model GPT-4o to enhance NLP's capability in understanding and processing complex handwritten data.

2)INTEGRATION OF OCR AND NLP TECHNOLOGIES FOR HANDWRITTEN TEXT RECOGNITION:

Sanuvala and Fatima (2021) [10] introduced an automated grading system that integrates Optical Character Recognition (OCR) and NLP, significantly enhancing grading efficiency. This integration allows for the automated recognition and comprehension of handwritten content, making it a feasible solution for large-scale educational applications. **3) APPLICATION OF GENERATIVE AI IN EDUCATION:**

Generative AI has shown promising potential in content analysis and evaluation within educational contexts. Chavan et al. (2023) [11] employed GPT-4 to extract keywords and summarize answers from handwritten text, achieving accuracy comparable to manual grading. This example illustrates the capability of advanced generative AI models to understand complex, unstructured handwritten content, highlighting their role in educational assessments.

6.2.3 The Potential of Combining Image Processing and NLP in Handwritten Note Analysis

1) ADVANTAGES OF COMBINING IMAGE PROCESSING AND NLP:

The integration of image segmentation and generative AI presents new possibilities for analyzing handwritten notes. Recent approaches leverage image segmentation, such as Attention Multi-task U-Net for page division, followed by generative AI models like GPT-4o for content analysis. This combined method enables efficient content recognition and provides valuable learning feedback, improving the efficiency and accuracy of educational data analysis. However, current methods still face challenges in educational contexts, such as the difficulty of recognizing complex handwritten content and handling unstructured data. Future research could further focus on designing advanced models tailored for large-scale educational data processing to enhance accuracy and operational efficiency in analyzing

diverse handwritten content.

2) LIMITATIONS OF CURRENT METHODS AND FUTURE DIRECTIONS:

Despite their effectiveness, existing image processing and NLP methods have limitations in educational contexts. For example, accurately recognizing complex handwritten scripts and handling unstructured data can be challenging. Future research can focus on more advanced models designed for large-scale educational data processing, which could enhance accuracy and operational efficiency in analyzing diverse handwritten content.

6.3 Methodology

6.3.1 Method Overview

This study proposes a novel method that combines advanced image-segmentation techniques with generative AI to analyze handwritten notes and capture detailed insights into students' learning behaviors. The primary objective is to automatically and accurately extract key information from handwritten content—such as text, graphs, and formulas—thereby enabling educators to more efficiently assess students' comprehension, knowledge retention, and areas of focus. As educational technology progresses and the need for data mining grows, the demand for such automated methods has become increasingly urgent. Handwritten notes often contain a wealth of unstructured data, such as diagrams, formulas, and free text, making manual analysis costly. Automated methods enhance both the efficiency and accuracy of the analysis, and using AI to generate personalized feedback helps educators quickly identify knowledge gaps and learning needs.

The approach leverages two main technologies: the Attention Multi-task U-Net model for precise image segmentation and the GPT-4o model for content under-

standing. The Attention Multi-Task U-Net is utilized due to its high precision and ability to handle multiple tasks simultaneously, making it well-suited for segmenting diverse elements within student notes. This segmentation step is crucial for separating various content types, enabling a structured analysis of each note component.

Following segmentation, the extracted content is analyzed by the GPT-4o model, which excels in processing unstructured handwritten data and generating coherent analyses. By integrating these technologies, the method not only achieves high accuracy in content identification but also enables in-depth analysis of students' learning patterns through automated data extraction.

This methodology represents a significant step forward in the educational data mining field, offering a scalable, data-driven approach to understanding student engagement and knowledge acquisition.

6.3.2 Attention Multi-task U-Net [12]

The Attention Multi-Task U-Net used in this study is designed to segment various elements in handwritten notes, such as text, charts, and formulas. As shown in Fig.1, by incorporating attention and multi-task mechanisms into the U-Net architecture, the model achieves greater segmentation accuracy and efficiency, which is essential for the complex layouts often found in handwritten notes. During the image preprocessing phase, this study first uses image enhancement techniques, such as contrast enhancement and noise reduction, to improve the clarity of handwritten content. This ensures that the model can efficiently segment image content even in complex backgrounds. Subsequently, The Attention Multi-Task U-Net segments preprocessed images into distinct regions, including text, diagrams, and formulas. This segmentation provides structured data for subsequent analysis and

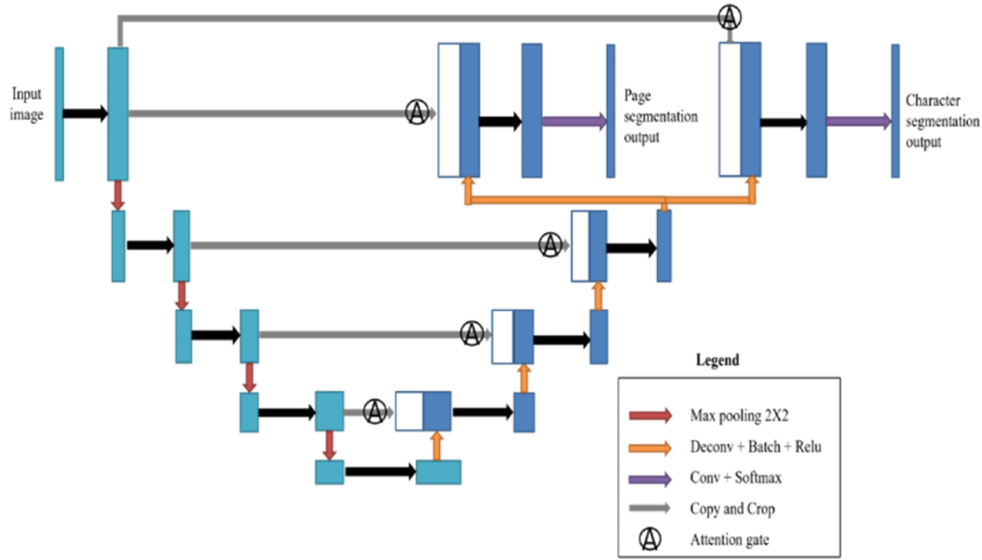


Figure 6.1: Attention Multi-task U-Net

enhances model accuracy by isolating key content types.

The process begins with pre-processing the input images to enhance contrast and reduce noise, which improves the clarity of the elements within the notes. Following this, the Attention Multi-task U-Net model performs segmentation by dividing the page into distinct regions.

Integrating an attention mechanism into the U-Net lets the network selectively concentrate on the most informative regions instead of distributing its effort uniformly across the image. By highlighting dense text, intricate diagrams, and other salient patterns while suppressing irrelevant background, the model more accurately separates textual from non-textual components. This targeted processing is especially valuable for handwritten student notes, where subtle details often carry critical meaning.

The multi-task learning feature allows the model to handle multiple segmentation tasks simultaneously, such as distinguishing between text, tables, and for-

mulas. By training the model to handle these tasks simultaneously, multitask learning enhances generalization to different content types, leading to better segmentation of various elements in educational notes. In handwritten note analysis, common content types such as text, formulas and diagrams often alternate on the same page. Multi-task learning integrates the character segmentation and page segmentation tasks into a single segmentation process, avoiding the repetition of separate processing steps in previous methods and thus greatly increasing processing speed. On the other hand, character segmentation as a side task also improves the segmentation accuracy of the page segmentation task.

Combining attention and multi-tasking in the U-Net architecture enables the segmentation model to achieve high accuracy in a single, integrated process, ensuring that the extracted elements can be effectively used in subsequent NLP and generative AI analysis.

As shown in Fig.6.2, this is an example of page segmentation for handwritten notes with text, formulas, graphs, and tables.

6.3.3 GPT-4o

Following segmentation, this study employs GPT-4o, an advanced generative AI model, to analyze the content extracted from student handwritten notes. The purpose of GPT-4o in this process is to understand and interpret text, diagrams, and formulas. The results of the segmentation of students' notes are matched with the content of the textbook to search for the keywords that are of most concern to the students, and the knowledge points that the students missed. OCR technology is first applied to the segmented text regions to convert handwritten text images into machine-readable text, improving accuracy in subsequent processing. Additionally, by constructing a knowledge base, GPT-4o can precisely match text

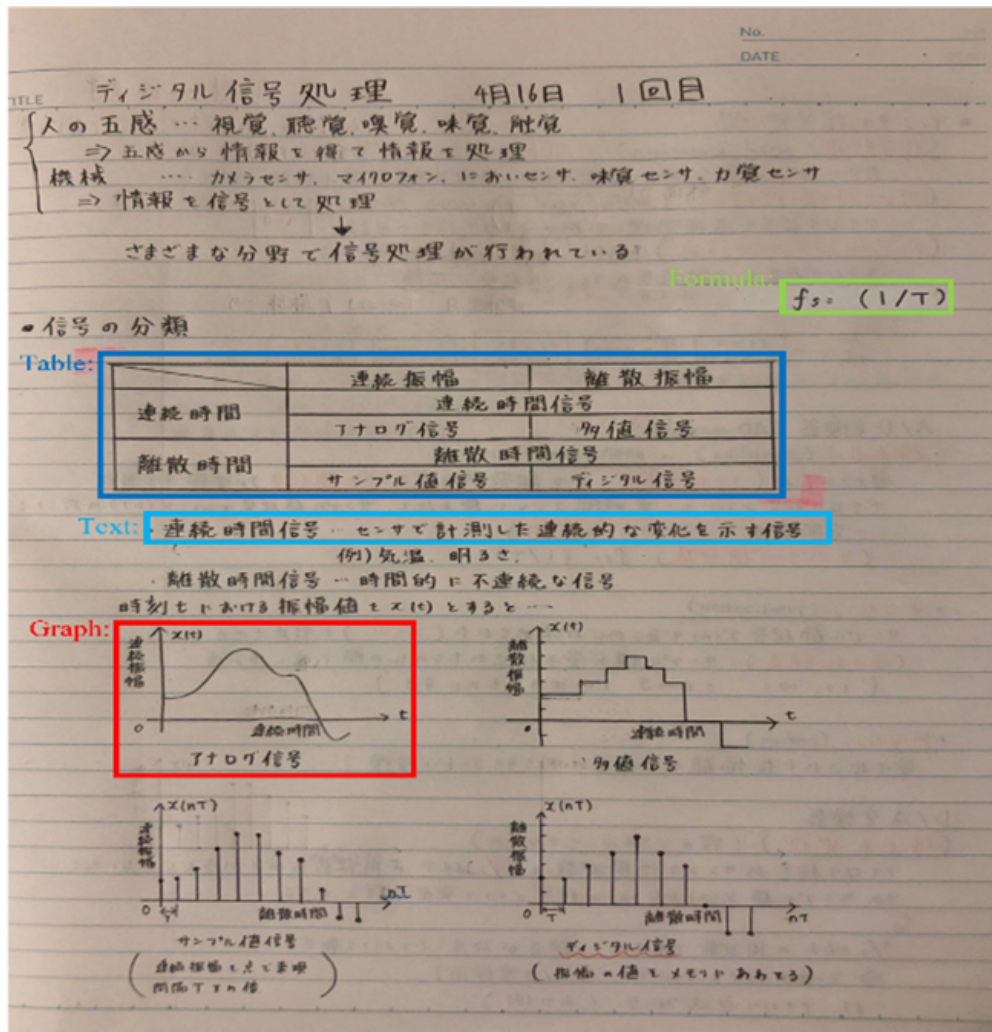


Figure 6.2: Handwritten note page segmentation example

with specific knowledge points, generating targeted learning feedback.

To prepare the content for analysis, OCR (Optical Character Recognition) is first applied to the segmented text regions to convert images of handwritten words into machine-readable text. OCR is essential in this stage because it bridges the gap between handwritten content and the NLP model, enabling accurate text extraction from various handwriting styles. Combining OCR with image segmentation ensures that only well-defined, legible text enters the NLP pipeline, thereby improving processing accuracy.

The GPT-4o model then processes the OCR-transformed text and applies its deep language understanding capabilities to analyze the content. For example, the model can extract core concepts, identify areas of repeated emphasis, and even infer knowledge gaps based on note content. Additionally, GPT-4o can generate feedback to help teachers identify students' learning challenges and weaknesses, enabling the design of personalized teaching plans.

One of the strengths of GPT-4o is its capacity to handle complex, unstructured text, which is often found in educational notes. Unlike traditional NLP models, GPT-4o can discern context, extract key points, and generate summaries, making it highly suitable for interpreting student notes. For example, the model can highlight core concepts, pinpoint areas of repeated emphasis, and even infer gaps in knowledge based on the content covered in the notes. Additionally, GPT-4o's generative capacity allows it to formulate potential feedback that educators can use to guide students effectively.

This integration of OCR with GPT-4o provides a streamlined approach to content analysis, transforming handwritten notes into structured data insights. The outputs generated from GPT-4o, such as extracted keywords, summaries, and inferred understanding levels, serve as valuable tools for educators to assess and tailor their teaching methods to better align with students' learning needs.

6.3.4 Integrated Framework

As shown in Fig.6.3 and Fig.6.4. The integrated framework combines image segmentation and generative AI to create a streamlined process for analyzing handwritten notes. This framework is designed to facilitate the automated extraction, processing, and interpretation of student notes, making it possible to gather detailed insights into learning behaviors.

Overall process: the workflow of the framework starts with the input of handwritten notes, which are subsequently processed by an image segmentation model (Attention Multi-task U-Net) to segment the different content categories. After segmentation is complete, the OCR technology will convert the text area (image format) and save it in text format (.txt). Although the current GPT model has multimodal capabilities, the experiments revealed that GPT is unable to effectively recognize Japanese text in pictures, particularly for handwritten notes, which are often written in a non-standardized manner. Accordingly, an OCR module was developed to extract text from the segmented text and table regions. This enables the Japanese text to be entered directly into the GPT, obviating the necessity for the GPT to recognize the image once more. This results in enhanced efficiency for the GPT and improved matching accuracy.

The images and formulas are saved in image formats for subsequent processing. Ultimately, the GPT-4o model, as shown in the Fig.6.5, performs a matching operation on the data to generate the matching results between student notes and textbook content. The modules within the framework engage in a structured sequence of interactions. For example, the image segmentation module outputs the segmentation results of text, formulas, graphs, and tables. The text within the segmented table is extracted by the OCR module and saved as a text file with the .txt file. Afterwards, the textual data are input into the GPT-4o. This

interaction creates a data processing framework that ensures processing efficiency and minimizes manual intervention.

Data storage and processing mechanism: all categories of data extracted from the textbook as shown in the Fig.6.4, the .txt files and image files are saved into the new constructed knowledge base of GPT-4o. Through this knowledge base, when dealing with each input of handwritten note data, GPT can directly match the search in the knowledge base, which improves the efficiency of GPT greatly. By integrating an image segmentation model, OCR, and GPT within a framework, this approach provides an efficient and accurate solution for extracting deep data from handwritten notes.

This framework design not only automates the data extraction but also prepares it for further analysis. It helps in the application of personalized learning feedback and adaptive educational recommendations.

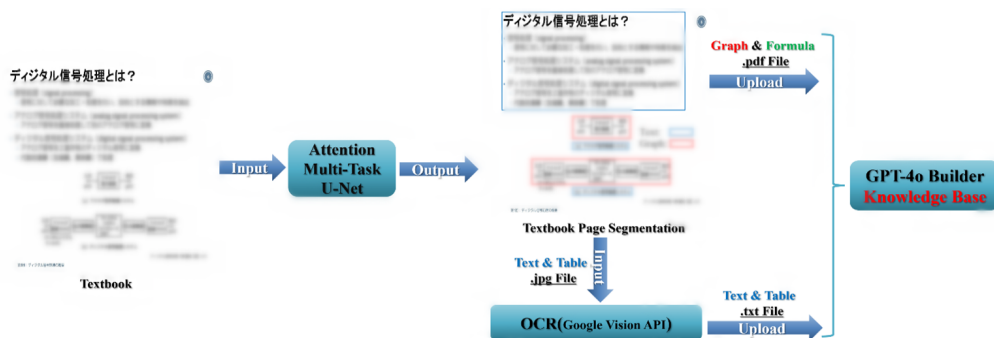


Figure 6.3: Process of building a knowledge base of textbooks

6.3.5 Innovation and Advantages

1)ATTENTION MECHANISMS AND THE INNOVATIVE COMBINATION OF MULTI-TASK LEARNING

An important innovation of this method is the introduction of the attention

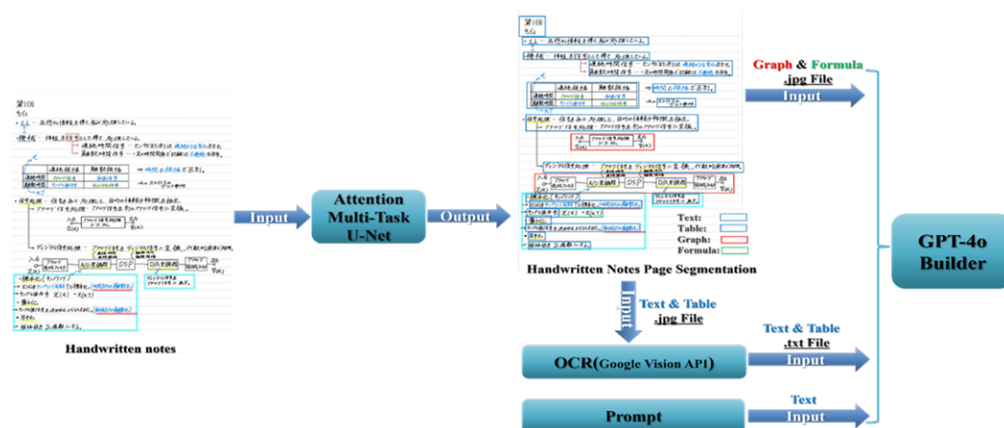


Figure 6.4: Process for student handwritten notes

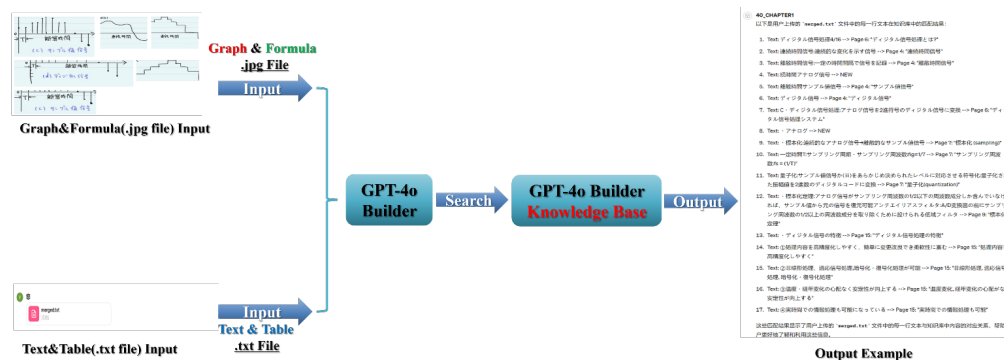


Figure 6.5: Matching result example

mechanism and multitask learning into the U-Net model. The attention mechanism enables the model to segment handwritten notes and achieve high segmentation accuracy even in complex datasets. Multi-tasking further improves the model's page segmentation accuracy, and in the previous experiments it found [12] that the character segmentation task will greatly improve the accuracy of the page segmentation task. This combination significantly enhances the efficiency and accuracy of extracting data from a complex layout dataset such as handwritten notes

2)GPT-4O FOR DEEP CONTEXT UNDERSTANDING

Another important innovation is the deep semantic analysis of segmented content using the GPT-4o model. Accurately matching handwritten notes with textbook content requires the model to understand the meanings of diverse textual expressions and align them correctly. Given GPT's strong performance in semantic understanding, it was selected for the NLP component to generate matching results and provide educators with accurate, in-depth insights into students' learning.

3)INTEGRATION OF IMAGE SEGMENTATION MODELS, OCR AND NLP MODELS

The method combines image segmentation, OCR and NLP models within a framework that provides a concise framework for handwritten notes data extraction. By integrating these processes within a single framework, the method minimizes human intervention, reduces errors that may result from human manipulation, and increases the efficiency of analyzing student handwritten note data. This integration makes it possible to process large amounts of notes quickly and has great potential to be applied to diverse educational scenarios.

4)ADVANTAGES OVER TRADITIONAL METHODS

Compared to traditional methods that typically rely on manual review, this approach is highly automated and reduces the time and effort required by instruc-

tors to assess student notes. The integrated framework provides content-matching results between students' handwritten notes and the textbook, enabling educators to better understand students' engagement and mastery of course content. In addition, the method's ability to highlight key concepts from the textbook and identify potential gaps in student knowledge supports personalized learning strategies and has the potential to become an adaptive educational tool.

5) FUTURE POTENTIAL

The methodology's scalability and flexibility make it ideally suited for future growth. As the field of education increasingly embraces digital tools, this method can be adapted to a variety of learning environments, from classrooms to online courses. Its ability to analyze a wide range of handwritten data, coupled with its adaptable framework, makes it valuable for applications in educational data mining and adaptive learning, and has great potential for future research and practice.

6.4 Experiments and Results

6.4.1 Experimental Design

This section describes the overall design of the experiment to evaluate the segmentation accuracy, content matching accuracy of the proposed method. The experimental data are handwritten notes from students 2021 and 2022 of Signal and System Processing.

Data preparation: Student handwritten notes were obtained from students' Digital Signal Processing courses to ensure data integrity. The experiment began with data preprocessing, where all note images were standardized to the same resolution to mitigate the impact of size variations. To enhance the model's robustness across diverse writing styles, samples encompassing various handwriting

styles, layout designs, and content complexities were selected.

6.4.2 Experiment Result

1)SELECTION OF MODELS

Given the dataset’s modest size (2,000 annotated images), computationally intensive architectures such as DeepLabV3+ were deliberately excluded during the model-selection phase. The primary rationale for adopting U-Net as the lightweight architecture lies in its exceptional sensitivity to fine-grained visual features—a capability that is critical for accurately segmenting handwritten student notes. The characteristic skip connections in U-Net effectively preserve spatial details through multi-scale feature fusion, which proves particularly advantageous for capturing intricate stroke patterns in free-form handwriting.

Furthermore, the computational efficiency of U-Net aligns with the pipeline design philosophy where segmentation serves as an intermediate processing step rather than the terminal objective. As shown in Table I. A quantitative comparison with FCN and SegNet demonstrates that U-Net achieves superior performance in text (0.8253) and formula (0.7346) regions. Although SegNet offers a slight advantage in graphic-element segmentation (0.5983 vs 0.5891), U-Net’s balanced accuracy across all categories and higher computational efficiency make it the most suitable choice for educational data analysis.

Furthermore, this study conducts systematic comparative analyses with mainstream segmentation models. As shown in Table I, while the baseline SAM model outperforms the traditional U-Net architecture in overall performance (e.g., Total region F1-score 0.8753 vs. 0.7869), the proposed Attention Multi-Task U-Net demonstrates significant advantages in critical metrics: achieving a text region segmentation accuracy of 0.9561 (6.06% higher than SAM) and marginally lead-

ing SAM in formula regions (0.8651 vs. 0.8571). Notably, SAM excels in graphic regions (0.8916 vs. 0.8263).

However, considering that text elements are much more prevalent than graphic elements in student handwritten notes. and considering the segmentation module’s role as a preprocessing step in the multimodal analysis pipeline, the efficiency gains from lightweight design and domain adaptability become decisive factors. The experimental results indicate that Attention Multi-Task U-Net offers superior engineering applicability in precision-efficiency trade-offs. Particularly in resource-constrained deployment scenarios, its modular design enables seamless integration with downstream LLM components, providing crucial technical support for building end-to-end intelligent note analysis systems. **2)IMAGE SEGMENTATION**

Table 6.1: Comparison of U-Net Accuracy with Other Neural Networks

Model	Total Ac/F1	Back ground	Text	Formula	Graph
FCN	0.7526/0.7135	0.8436	0.7558	0.6424	0.5017
SegNet	0.7828/0.7562	0.8706	0.7927	0.7154	0.5983
SAM	0.8926/0.8753	0.9326	0.9015	0.8571	0.8916
Original U-Net	0.8162/0.7869	0.9137	0.8253	0.7346	0.5891
Attention Multi-Task U-Net	0.9425/0.9264	0.9733	0.9561	0.8651	0.8263

MODEL (ATTENTION MULTI-TASK U-NET)

The main content of (subsection 1) is mainly quoted from my paper “Improvement of image segmentation model for handwritten notebook analytics”presented at ICIP 2023.

Model training: The dataset used in this study consists of handwritten notes

produced by students in the Digital Signal Processing course between 2021 and 2022. There are a total of 1980 images. The dataset was split it into 70% training set, 15% validation set, and 15% test set. The accuracy results shown in this paper are from the model’s performance on the test set.

Evaluation metrics: segmentation accuracy and F1 score are used to assess the overall performance of the model. Accuracy for each region is calculated using Equation 1. C represents the number of pixels that are correctly segmented, and L represents the number of pixels in the label.

$$Accuracy = C/L \quad (6.1)$$

To verify the model’s performance, comparison experiments were conducted with other U-Net variants. Comparison of the multi-task U-Net model and the original U-Net: As shown in Table , it can be clearly found that the multi-task U-Net improves the accuracy over the original single task U-Net in the text(6.74%) and formula(5.08%) region and improves 1.92% in the graph region.

This indicates that in the multi-task model, the features learned by the character segmentation task assist the model in correctly identifying text and formula regions in the page segmentation task. Table clearly demonstrates the advantages of incorporating the attention mechanism in the Attention Multi-task U-Net model, with significant improvements across regions. Notably, accuracy in the graph region increased by 21.8% compared to the multi-task U-Net model, while the text and formula regions showed respective improvements of 6.34% and 7.97%. These gains highlight the effectiveness of the attention mechanism in enabling the model to focus on specific features, particularly in regions with dense graphical elements and complex structures. This improved precision across multiple regions underscores the model’s adaptability and accuracy in handling varied and complex note layouts.

As shown in Table , the progression from the original U-Net to the Attention Multi-task U-Net illustrates a steady improvement across all evaluated regions. The table highlights that while the multi-task U-Net model provided modest gains, the attention-enhanced version of the model made substantial advancements in handling graph regions with dense content and distinguishing between various page components. This comparison affirms the enhanced segmentation capabilities of the Attention Multi-task U-Net, positioning it as a robust model for processing complex handwritten educational notes.

Table 6.2: Comparison of Accuracy Rates by Region

Model	Total Ac/F1	Back ground	Text	Formula	Graph
Original U-Net	0.8162/0.7869	0.9137	0.8253	0.7346	0.5891
Multi-Task U-Net	0.8728/0.8475	0.9406	0.8927	0.7854	0.6083
Attention Multi-Task U-Net	0.9425/0.9264	0.9733	0.9561	0.8651	0.8263
Attention R2 Multi-Task U-Net	0.9458/0.9351	0.9782	0.9608	0.8647	0.8315

3)CONTENT MATCHING ACCURACY OF GPT-4O

GPT-4o’s content-matching accuracy was evaluated by comparing its outputs with a ground-truth alignment between notes and textbook content. Specifically, each model’s matching accuracy was calculated manually using Equation 6.2.

GPT-4o’s performance was measured by the percentage of correct matches it produced. Table 6.2 reports these accuracy rates and includes a comparison with alternative models (e.g., OpenAI’s Gemini). GPT-4o achieved the highest alignment accuracy (identifying the correct textbook reference for 85% of note

segments, versus 78% for Gemini), demonstrating its superior understanding of context. These results confirm that this integrated approach yields reliable semantic matching between student notes and curricular content.

GPT-4o demonstrates strong content matching accuracy, effectively aligning handwritten notes with textbook content. This matching capability enables GPT-4o to accurately identify and associate key concepts from students' notes with corresponding curriculum materials, providing insights into student understanding and highlighting potential knowledge gaps. Such functionality is invaluable in automated educational assessments, where accurately comparing student-generated content to course material is essential for personalized feedback.

To identify the most suitable model, a preliminary comparison was conducted among GPT-4o, Gemini, and Claude, evaluating each model's accuracy in matching complex, technical content within handwritten notes. As shown in Table , GPT-4o consistently outperformed the other models, particularly in handling challenging terminology and technical expressions. This superior accuracy allowed GPT-4o to achieve higher alignment rates with textbook content, making it the preferred choice for the NLP function in this study.

The experiment analyzed handwritten notes from 145 students in a Digital Signal Processing course, covering 11 chapters of content. GPT-4o demonstrated an approximately 10 % improvement in text-processing accuracy compared with its predecessor, GPT-4. This improvement was largely achieved through enhanced prompt engineering, which optimized the model's capacity to interpret specific technical terms accurately. Furthermore, Tables IV and V provide a detailed breakdown of GPT-4o's accuracy across chapters and content types, illustrating its performance consistency. These results validate GPT-4o's applicability in educational contexts, making it a robust tool for automated analysis and feedback generation. The model's ability to align content with precision has significant im-

plications for future applications in educational data analysis and adaptive learning systems.

Table 6.3: Matching Accuracy Comparison

Chapter	Gemini	Claude	GPT-4o
1	0.908	0.904	0.938
2	0.913	0.909	0.941
3	0.890	0.883	0.916
4	0.888	0.882	0.914
5	0.901	0.896	0.925
6	0.897	0.891	0.926

$$Accuracy = (correctclassifications)/(allclassifications) \quad (6.2)$$

4) FURTHER ANALYSIS OF CONTENT MATCHING

By analyzing the results of content matching, it is possible to generate in-depth insights into student learning patterns and areas of focus. As shown in Table VI, GPT-4o was used to extract a list of keywords from the textbook content of Chapter 1, with these keywords presented in the first column. Following this, the matching results allowed us to count how many A-level students recorded each keyword. Observing the table, it is immediately apparent that the keywords most frequently recorded by students are "A/D 変換器" and "D/A 変換器," indicating a high level of attention toward these concepts.

Although teachers often have structured teaching plans that highlight the relative importance of various content areas, it is challenging for them to gauge precisely which topics capture their students' focus.

Table 6.4: GPT-4 Accuracy of Chapters 1–11

Chapter	Text & Table	Graph	Formula	Corresponds
1	0.850	0.891	0.943	0.861
2	0.849	0.919	0.914	0.866
3	0.823	0.895	0.944	0.849
4	0.818	0.908	0.914	0.842
5	0.829	0.890	0.949	0.856
6	0.832	0.893	0.921	0.853
7	0.844	0.881	0.946	0.865
8	0.836	0.917	0.931	0.862
9	0.869	0.883	0.932	0.879
10	0.816	0.896	0.915	0.837
11	0.827	0.901	0.924	0.834

Table 6.5: GPT-4o Accuracy of Chapters 1–11

Chapter	Text & Table	Graph	Formula	Corresponds
1	0.939	0.901	0.958	0.938
2	0.945	0.937	0.933	0.941
3	0.917	0.912	0.952	0.916
4	0.912	0.921	0.925	0.914
5	0.923	0.908	0.961	0.925
6	0.919	0.913	0.936	0.926
7	0.936	0.894	0.957	0.937
8	0.931	0.929	0.946	0.932
9	0.953	0.902	0.947	0.945
10	0.908	0.912	0.923	0.913
11	0.912	0.916	0.935	0.914

Table 6.6: Matching Accuracy Comparison

Chapter	Gemini	Claude	GPT-4o
1	0.908	0.904	0.938
2	0.913	0.909	0.941
3	0.890	0.883	0.916
4	0.888	0.882	0.914
5	0.901	0.896	0.925
6	0.897	0.891	0.926

Table 6.7: GPT-4 Accuracy for Chapter 1–11

Chapter				
	Text&	Graph	Formula	Corresponds
	Table			
1	0.850	0.891	0.943	0.861
2	0.849	0.919	0.914	0.866
3	0.823	0.895	0.944	0.849
4	0.818	0.908	0.914	0.842
5	0.829	0.890	0.949	0.856
6	0.832	0.893	0.921	0.853
7	0.844	0.881	0.946	0.865
8	0.836	0.917	0.931	0.862
9	0.869	0.883	0.932	0.879
10	0.816	0.896	0.915	0.837
11	0.827	0.901	0.924	0.834

Table 6.8: GPT-4o Accuracy for Chapter 1–11

Chapter				
	Text&	Graph	Formula	Corresponds
	Table			
1	0.939	0.901	0.958	0.938
2	0.945	0.937	0.933	0.941
3	0.917	0.912	0.952	0.916
4	0.912	0.921	0.925	0.914
5	0.923	0.908	0.961	0.925
6	0.919	0.913	0.936	0.926
7	0.936	0.894	0.957	0.937
8	0.931	0.929	0.946	0.932
9	0.953	0.902	0.947	0.945
10	0.908	0.912	0.923	0.913
11	0.912	0.916	0.935	0.914

By using content matching analysis, educators can quickly identify what students are concentrating on. This information allows teachers to tailor their instruction more effectively, focusing on areas where student engagement is strong or addressing topics where comprehension may be lacking. For students, this method provides a valuable tool for self-assessment, helping them to review their notes and identify key knowledge points that may have been overlooked. This proactive review can facilitate deeper comprehension and retention of essential course content.

Table VI's data allows us to explore patterns in students' understanding and note-taking behavior. Through analysis, educators can assess the breadth and depth of students' focus on different topics, thus gaining insights into where students might require additional support. For instance, a high focus on "A/D 変換器" suggests that this concept is either challenging or particularly engaging for students, signaling an area where further instructional emphasis could be beneficial. On the other hand, keywords that receive minimal attention may reflect topics students find either easier or less relevant, guiding teachers in curriculum adjustments.

Bubble charts (Figures 6.6 and 6.7) were generated from the data in Table V to illustrate student attention to these keywords in the 2021 and 2022 classes. The bubble chart visually presents keyword attention in a circular pattern, making it easy to compare the focus on each concept across different groups. Specifically, Fig.6.6 displays data for the 2021 class, while Fig.6.7 represents data from the 2022 class. This year-over-year comparison offers educators valuable insights into shifting patterns of student interest and understanding over time. For example, if certain keywords show a declining trend in attention, it may indicate the need to reinforce these topics in future sessions.

Note: The Textbook Content was too long to be fully displayed in the table,

Table 6.9: CHAPTER 1 KEYWORDS

Keyword	Textbook Content	2021	2022
ディジタル信号処理とは?	信号処理 (signal processing) …	35	21
ディジタル化の利点	雑音に強い劣化しない…	49	38
ディジタル信号処理の特徴	数式通りの処理内容を…	47	37
信号処理システムの時不変性	ある信号 $x(t)$ が時間だけ遅れた入力…	37	29
信号処理システムの因果性	システムへ信号が入力された後…	39	26
信号処理システムの線形性	線形性ある変換 $\{ \}$ を行うシステム…	31	25
A/D 変換器	標本化 (sampling) 連続的なアナログ…	53	43
D/A 変換器	2 進符号のディジタル信号をアナログ信号に戻す…	54	41
信号の形態	時刻における信号の振幅値を $x(t)$ と表現 4 種類…	48	40
情報と信号	ヒトの五感視覚, 聴覚, 嗅覚, 味覚, 触覚ヒトは…	42	35
着目点	時間成分の連続性線形性時不変性因果…	44	38

therefore the use of ellipses (···) was used in the table to replace the textbook content that could not be displayed.

Furthermore, the bubble charts enable educators to quickly observe how attention to specific knowledge points changes among different student groups. This visualization can reveal trends that help educators adapt their teaching to evolving student needs. For instance, if attention to a particular concept, like "A/D 変換器," increases significantly in one cohort compared to another, this may suggest that the curriculum or teaching approach has impacted student engagement with that topic. Such insights are useful for continuous improvement in curriculum design, enabling more focused, responsive teaching practices.

For students, these visual representations are equally beneficial. The charts provide a way to benchmark their understanding against their peers' collective focus, helping them identify topics they may have under-emphasized. By aligning their study efforts with areas that receive substantial attention from their classmates, students can reinforce learning on critical topics, fostering a well-rounded comprehension of the material. Additionally, understanding the focus trends of past student groups can help them anticipate which areas might require more attention and better prepare for exams or assessments.

In conclusion, content matching analysis and its visualization through bubble charts serve as powerful tools for enhancing educational outcomes. For teachers, they provide clear indicators of which topics may need more attention, adjustment, or support in the classroom. For students, they offer a structured method to identify and concentrate on key learning areas, promoting a more effective, data-driven approach to study and review. This dual benefit underscores the value of content matching in both guiding instructional strategies and supporting student learning practices.

6.5 Discussion

6.5.1 Key Findings

This study proposes an efficient method for analyzing handwritten note data by combining image segmentation and generative AI models. Experimental results show that the integration of the Attention Multi-task U-Net model with GPT-4o offers significant advantages in segmentation accuracy and content understanding, enabling complex handwritten notes to be processed and analyzed in an automated manner. This combination allows for the accurate extraction of student learning data, the identification of key concepts, and the detection of knowledge blind spots, thereby providing educators with a quantitative approach to analyze student learning behaviors.

One of the main advantages of this method is its ability to handle complex layouts and diverse handwriting styles, which are common in student notes. The Attention Multi-task U-Net model performs well in distinguishing between text, formulas, and graphical elements, thanks to the multi-tasking capabilities and attention mechanism that focus on specific features within each content type. This results in higher segmentation precision, even for intricate elements like diagrams and mixed-content layouts. The GPT-4o model complements this by accurately interpreting and linking segmented content to textbook material, which aids in identifying the alignment of students' notes with the curriculum. Such precision is crucial for providing teachers with insights into student comprehension and areas that might need further emphasis.

Compared to traditional manual methods, which are labor-intensive and time-consuming, this automated approach enhances both accuracy and efficiency. Educators can save significant time by using this method, allowing them to focus on instructional improvements based on data-driven insights. Furthermore, the method

provides students with targeted feedback on their understanding and knowledge gaps, helping them better prepare for assessments. In summary, this integrated approach not only offers scalability and reliability but also holds potential for broader applications in educational data analysis and personalized learning support.

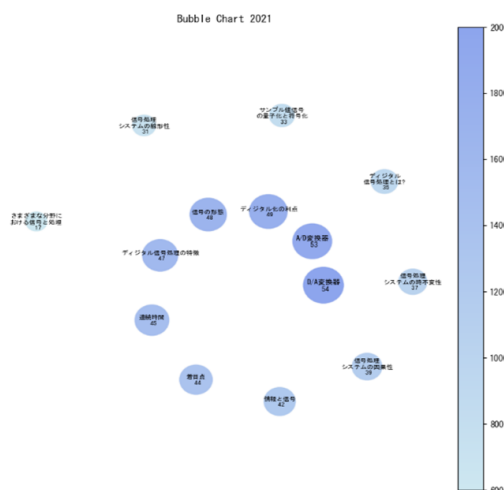


Figure 6.6: 2021 keywords Bubble Chart

6.5.2 Advantages of the Method

The main advantages of this method include:

1)Automation and Efficiency: By integrating image segmentation, OCR, and NLP, the method achieves automated analysis of handwritten notes, greatly improving processing efficiency and making it suitable for large-scale applications. Unlike manual methods, which require significant time and effort, this approach enables educators to quickly process and analyze large volumes of notes, saving valuable time and resources. This automation is especially beneficial in educational environments where student data is continuously generated, allowing for

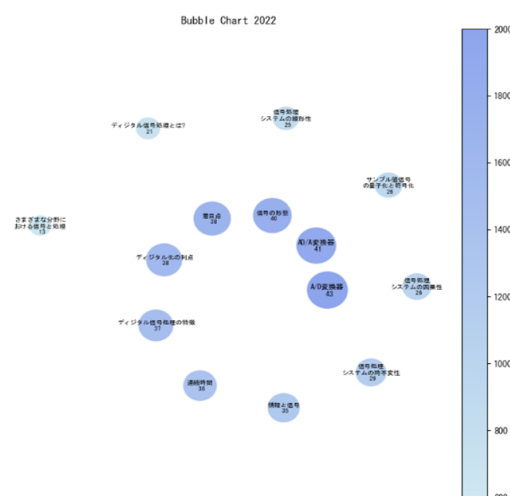


Figure 6.7: 2022 keywords Bubble Chart

more responsive and timely analysis.

2)Diverse Data Processing: The combination of Attention Multi-task U-Net and GPT-4o models allows the method to handle multiple content types, including text, charts, and formulas. This enhances the comprehensiveness of content analysis, as it accurately segments and interprets various elements that may appear in educational notes. The model’s attention mechanism specifically aids in distinguishing between different data types within a single page, ensuring that each content type is analyzed accurately. This capability allows the method to adapt to various subjects and note structures, making it broadly applicable across educational disciplines.

3)Data Insights: The integration of generative AI provides deep insights into learning behaviors, allowing educators to identify patterns in students’ study habits and knowledge gaps. By matching handwritten content with textbook material, the model enables a detailed understanding of which topics students are focusing on,

what concepts they might be struggling with, and where they may need additional support. This data-driven insight is invaluable for educators seeking to provide targeted feedback and personalized guidance to students.

4) **Lightweight model components:** The selection of the Attention Multi-task U-Net over computationally intensive alternatives (e.g., SAM) proved critical to balancing accuracy and efficiency. As shown in Table I, the model's lightweight architecture achieved a text segmentation F1-score of 0.9264 making it ideal for integration with downstream NLP components. This design choice underscores the importance of task-specific model optimization in educational data analysis pipelines.

However, the method has some limitations. For instance, OCR and NLP models may encounter difficulties with complex or poorly legible handwriting, potentially leading to errors in text recognition and interpretation. The model may also need optimization to handle different writing styles more effectively, particularly in cases where students use unconventional characters or symbols. Addressing these limitations would involve fine-tuning OCR algorithms and enhancing model adaptability, ensuring robust performance across diverse handwriting patterns.

6.5.3 Comparison with Existing Research

This study's integrated approach significantly enhances the effectiveness of traditional image processing and natural language processing (NLP) in educational data analysis. By combining multiple techniques, it not only improves analysis efficiency but also enhances content recognition accuracy. In contrast to existing methods based on manual review or single segmentation models, this study offers a more efficient and automated solution for analyzing handwritten notes.

Previous research has explored various methods for processing handwritten documents. For instance, Jo et al. (2020) [13] proposed a method that separates

handwritten and machine-printed components in documents using convolutional neural networks (CNNs). Their approach focuses on pixel-level classification to handle overlapping components, demonstrating the potential of CNNs in document analysis. However, their method primarily addresses the separation of handwritten and printed text without delving into the semantic understanding of the content.

In another study, Gholamian and Vahdat (2023) [14] introduced a novel approach to segment handwritten and printed text, particularly in scenarios where they overlap. They developed a dataset, SignalTR6K, and a model architecture that outperforms prior work in segmentation tasks. While their work advances segmentation techniques, it does not integrate NLP components for content analysis.

Furthermore, studies like the one by Chavan et al. (2023) [15] have explored the use of generative AI models, such as GPT-4, to extract keywords and generate summaries from handwritten text, achieving accuracy comparable to manual grading. This highlights the potential of generative AI in understanding handwritten content. However, their focus is primarily on content generation rather than the integration of image segmentation and NLP for comprehensive analysis.

In contrast, this study integrates advanced image segmentation models, specifically the Attention Multi-task U-Net, with generative AI models like GPT-4o. This combination allows for precise segmentation of handwritten notes into various components, such as text, formulas, and diagrams, followed by semantic analysis to extract meaningful insights. By leveraging the strengths of both image processing and NLP, our approach provides a holistic solution for educational data analysis, enabling automated extraction of student learning data, identification of key concepts, and detection of knowledge gaps.

Compared to single-model studies, this method's integration of multiple tech-

niques enhances both efficiency and accuracy. The Attention Multi-Task U-Net model improves segmentation accuracy by focusing on specific features within complex page layouts, while GPT-4o's advanced language understanding capabilities facilitate accurate content matching and analysis. This integrated approach addresses the limitations of previous methods by providing a comprehensive tool for analyzing handwritten notes, supporting educators in quantitatively assessing student learning behaviors and tailoring instructional strategies accordingly.

In summary, while existing research has made significant strides in either image segmentation or NLP for handwritten document analysis, this study bridges the gap between these domains. By combining state-of-the-art techniques from both fields, the proposed approach offers an efficient, automated solution that surpasses traditional methods, provides deeper insights into educational data, and enhances the overall learning experience.

6.5.4 Practical Applications and Educational Impact

The automation and efficiency of the method make it applicable to large-scale educational scenarios, such as learning data analysis in college classrooms, automatic collection of educational data, and personalized learning feedback generation. By automating the extraction and analysis of handwritten notes, this method enables educators to monitor student progress continuously and in real time, making it easier to identify trends in learning patterns and detect knowledge blind spots. Such insights allow teachers to make targeted pedagogical adjustments that directly address student needs, thereby improving the quality of teaching and enhancing student outcomes.

Additionally, this method holds significant potential in educational research by facilitating the collection of vast amounts of learning data across multiple institu-

tions and courses. This data can be invaluable for large-scale educational studies, helping to identify common learning difficulties and uncovering patterns that inform curriculum design. Moreover, the scalability of this approach supports its use in adaptive learning platforms, where personalized feedback can be automatically generated based on individual student notes. This integration of automated analysis into personalized learning not only supports students in self-assessment but also fosters data-driven instructional strategies, promoting advancements in educational data mining and evidence-based teaching practices.

6.5.5 Future Research Directions

To further enhance the applicability of this methodology, future research could focus on the following areas:

- 1) **Extension of Diverse Data Sources:** Expanding the application of this method to include a wider variety of learning materials, such as class notes, lab reports, and project submissions, can broaden its utility. Adapting the model to work with different types of handwriting and layout formats will make it more versatile across academic fields. Additionally, incorporating multi-language support could increase its effectiveness in diverse educational settings, allowing it to serve a broader spectrum of students and instructors.
- 2) **Model Optimization and Adaptive Enhancement:** Future research could explore adaptive learning algorithms to improve the model's ability to adjust to various handwriting styles, enhancing the accuracy of content matching across unique student writing patterns. Improved prompt engineering for generative AI models, combined with advances in AI-driven handwriting analysis, could further elevate matching accuracy. Enhanced algorithms may allow the model to distinguish subtle differences in content emphasis, enabling more precise feedback.

3)Real-Time Analysis and Feedback System: Developing a real-time analysis system based on this methodology could allow educators to monitor students' learning status more immediately, enabling timely, data-driven interventions. Such a system would facilitate live analysis during lectures or study sessions, allowing teachers to adjust their instruction dynamically and address student needs as they arise.

Through continued optimization and the integration of these advancements, this method holds promise for making substantial contributions to educational data analytics, adaptive learning, and personalized feedback generation. These improvements would provide educators with smarter, data-driven instructional support, paving the way for more interactive and responsive learning environments.

6.6 Conclusion

This study proposes an integrated approach for analyzing students' handwritten notes by combining advanced image segmentation and generative AI. By utilizing the Attention Multi-task U-Net model, the method effectively segments various content types, such as text, diagrams, and formulas, and then leverages GPT-4o to extract and analyze key information within each category. The experimental results demonstrate that this method achieves high accuracy in both segmenting and interpreting handwritten notes, providing educators with a robust tool to gain insights into students' learning patterns, knowledge retention, and focus areas. This approach supports a deeper understanding of students' comprehension, helping educators identify concepts that may need further reinforcement.

The primary contributions of this study include the automation of handwritten note analysis, which significantly reduces the time and effort required for educators to manually review and interpret student notes. This is especially valuable in

large-scale educational settings where manual review of handwritten notes is often impractical. By incorporating advanced AI techniques, the method achieves a high degree of content matching accuracy, enabling the extraction of meaningful data that can be used for personalized and adaptive learning insights. The ability to accurately link note content to curriculum topics allows educators to provide targeted feedback, helping students strengthen their grasp of core concepts and address any identified knowledge gaps.

Another notable aspect of this methodology is its scalability. The method's high accuracy and adaptability make it suitable for large-scale applications across diverse educational scenarios, from secondary education to university-level courses. By providing actionable data, this approach empowers educators to make data-informed adjustments to their teaching strategies and better support student success. Additionally, the method could be integrated into adaptive learning platforms, offering students a continuous assessment tool that helps guide their studies and aligns with their individual learning paths.

Future Impact: The successful integration of image processing and generative AI for educational data analysis represents a significant advancement in applying AI to adaptive learning environments. As educational technology continues to evolve, this methodology has the potential to expand beyond handwritten notes, analyzing other forms of student-generated data, such as lab reports, project submissions, and digital assignments. This expansion would provide a more comprehensive view of student engagement and learning outcomes, further supporting data-driven instructional methods.

Future research could focus on optimizing this method into a real-time feedback system, allowing for immediate insights into student progress and enabling responsive teaching adjustments. Additionally, enhancing the model's adaptability to handle diverse handwriting styles and expanding its scalability for broader

educational applications will be essential in maximizing its impact in the field of education.

By continually advancing and refining these AI-driven approaches, this methodology could significantly contribute to the future of personalized education, making it possible for educators to support students in a way that aligns with their unique learning needs and goals.

6.7 Chapter Summary

This chapter successfully validates the comprehensive analytical framework developed throughout the dissertation in practical educational scenarios. By applying the integrated segmentation and semantic analysis methodologies, the chapter demonstrates the framework's capability to generate actionable insights from handwritten student notes. These insights significantly enhance instructional decision-making, personalized learning, and educational analytics. Ultimately, this final research stage clearly demonstrates the practical value and applicability of the dissertation's methodological contributions, effectively bridging technical innovation with educational practice.

Chapter 7

Cross-Paper Synthesis and General Discussion

Chapter 7 synthesizes the empirical findings presented in Chapters 4, 5, and 6, integrating key results to provide a holistic view of the research. By examining the cumulative insights derived from the improved image segmentation model, the integrated vision-language framework, and their practical educational applications, this chapter consolidates theoretical and practical contributions. It sets the stage for comprehensive theoretical reflection, discusses implications for educational practice, and critically examines research limitations to guide future research directions.

7.1 Cross-Study Findings

This section integrates the findings of the three core studies to establish a cohesive understanding of how automated analysis of handwritten student notes contributes to educational data mining, content comprehension, and personalized learning analytics. The integration of vision and language technologies proved consistently

effective across contexts.

In Study I, the Attention Multi-task U-Net enhanced segmentation accuracy, especially in complex layouts. Study II revealed that integrating GPT-4 with structured note data enabled accurate semantic alignment with textbook content. Study III demonstrated the model's capacity to extract behavioral insights from notes, with meaningful correlations to academic performance.

These studies collectively support the hypothesis that a multimodal AI system can serve as a reliable proxy for assessing cognitive engagement. The pipeline proved robust across varying subjects, note formats, and writing qualities. Moreover, qualitative feedback from educators indicated the practical value of the extracted insights, particularly for formative assessment.

A deeper look into the comparative results reveals that while each model component individually offered certain advantages, it was the synergy between segmentation and language understanding that yielded the highest accuracy in identifying knowledge structures. For instance, integrating U-Net with GPT-4 significantly improved the identification of mislabeled or misunderstood concepts, which would not have been easily visible through segmentation alone. Such integrated insights facilitated more targeted pedagogical interventions.

Cross-comparison across multiple courses and learning contexts showed that notes related to STEM subjects tended to have a higher proportion of structured diagrams and equations, while those from liberal arts showed dense textual narratives. The model's ability to adapt to these different patterns confirmed its cross-disciplinary potential. Additionally, notes captured at different time points during the semester revealed shifts in students' focus, strategy, and content coverage—offering a dynamic view of learning progression.

Furthermore, the model uncovered subtle note-taking behaviors such as marginal annotations, cross-referencing with previous pages, and iterative edits over time.

These aspects, often overlooked by manual analysis, illuminated the evolving nature of student understanding and strategy refinement. It also suggested that students exhibiting recursive note updates were more likely to perform well academically, reinforcing the importance of metacognitive engagement. These insights further strengthen the argument for incorporating such AI frameworks into regular educational assessments.

7.2 Integration of Findings with Research Questions:

Collectively, the findings of Studies I–III address the research questions outlined in Chapter 1. For RQ1 and RQ2, Study I demonstrated that an attention-enhanced multi-task U-Net significantly improves segmentation accuracy, confirming the hypothesis H1 that this model outperforms baseline methods. Study II and Study III addressed RQ3 by showing that combining OCR with GPT-4o yields meaningful insights into student learning behaviors, thereby supporting hypothesis H2 regarding the added value of generative AI for semantic interpretation.

Furthermore, the outcomes across all studies provide evidence for RQ4: the integrated framework offers greater accuracy and efficiency compared to manual or fragmented approaches, validating hypothesis H4 about scalability and user-friendliness. In summary, the cross-study results affirm that the unified approach not only delivers high technical performance but also yields practical educational benefits, as envisioned at the outset of this dissertation.

7.3 Theoretical Contributions

Firstly, this dissertation reframes handwritten notes as a rich source of cognitive evidence rather than viewing them as ancillary study artifacts. This extension introduces learning analytics to a new modality by treating students' pen-and-paper notes as a data source for insights into cognition and metacognition. In addition, the study presents a novel multimodal analytical framework that bridges computer vision and NLP in education; this integrative approach provides a blueprint for future AI systems that must interpret both visual and textual information. These theoretical contributions go beyond improving accuracy metrics; they offer a new perspective on how disparate AI techniques can be unified to capture and interpret student learning processes in ways not previously possible.

The research contributes to the theoretical foundation of multimodal learning analytics by proposing a unified model architecture that operationalizes handwritten notes into machine-readable data and pedagogical indicators. Traditionally, cognitive indicators in student work have been explored through surveys or manual coding. This dissertation extends that line of inquiry by providing computational models capable of interpreting unstructured visual and linguistic artifacts.

Furthermore, the integration of attention mechanisms within the segmentation model not only contributes to computer vision literature but also shows how visual focus patterns relate to educational constructs like emphasis and salience. On the NLP side, the use of GPT-4 as a semantic bridge between extracted content and learning standards introduces a new methodological pathway for instructional alignment.

The theoretical value of this work also lies in its redefinition of student artifacts as rich, multimodal data sources rather than peripheral learning tools. It positions handwritten notes as central cognitive evidence, aligning with constructivist the-

ories of learning where knowledge is actively constructed and reflected through written expression. By framing note-taking as both a generative and diagnostic activity, this research offers a computational lens to validate long-held pedagogical assumptions.

In addition, the methodological framework paves the way for future theoretical models that link behavioral indicators (e.g., note-taking density, structure, variability) with affective and metacognitive dimensions. It creates opportunities for expanding learning analytics theory to include fine-grained, temporally ordered student data—pushing beyond static models of engagement or knowledge acquisition.

This integration of multimodal signals suggests a move toward more holistic analytics frameworks—ones that can eventually account for not just what students write, but how, when, and why they do so. This expansion into the epistemic and reflective domains further positions note analysis within a broader pedagogical and psychological context.

7.4 Practical Implications

The findings have immediate relevance for educators, instructional designers, and edtech developers. For educators, the system offers a non-intrusive method to monitor student engagement and comprehension, even in remote or hybrid settings. Automated summaries and indicators of missing or misunderstood concepts can inform real-time intervention or curriculum redesign. Generating concept maps from these analytics is identified as promising future work.

Instructional designers can use note-based analytics as a form of summative evaluation to gauge the efficacy of teaching materials. For example, if key concepts are consistently absent or distorted in student notes, this may signal the need for

pedagogical adjustment. Moreover, the framework facilitates adaptive learning by feeding insights into recommendation engines that personalize review tasks or content delivery.

In terms of system deployment, the modular architecture allows institutions to adopt components gradually, depending on their technological readiness. Cloud-based analysis also reduces the demand on local infrastructure while supporting large-scale data collection.

Additionally, the framework can be integrated into existing Learning Management Systems (LMS) such as Moodle or Blackboard through APIs, enabling a seamless flow of handwritten data into centralized analytics dashboards. Teachers could receive weekly reports highlighting students who show declining engagement, patterns of incomplete or redundant note-taking, or low alignment with textbook objectives.

Another practical use case involves peer learning facilitation. By clustering students based on their note structures or content comprehension levels, the system could support group formation for collaborative tasks, encouraging peer explanation and learning from diverse strategies.

Finally, institutions could use aggregated data to inform curriculum development. Patterns identified across multiple semesters might indicate which topics require more instructional clarity or which formats (e.g., diagrams vs. text) yield better learning retention.

Moreover, the system could be leveraged to design formative assessments that align with student-generated content, offering dynamic feedback loops. By aligning instructional material more closely with student cognition as expressed through notes, institutions could foster deeper learning, retention, and satisfaction.

7.5 Limitations

Despite its contributions, this research is not without limitations. First, the datasets used—while diverse in handwriting and subject matter—were limited to STEM courses and a Japanese university context. The generalizability of the framework to humanities disciplines or international student populations requires further validation.

Second, while the OCR component was enhanced with math recognition, certain edge cases—such as mixed-language annotations or extremely cursive writing—still pose challenges. The model’s performance also decreases when note quality is degraded due to poor scanning or low-resolution input.

Another limitation lies in interpretability. While the system generates confidence scores and visualizations, the reasoning behind certain GPT-4 inferences remains opaque. Future versions may benefit from incorporating explainability modules.

Ethical and privacy concerns also present notable limitations. Although data was anonymized, handwritten notes may contain personally identifiable information or sensitive content. Ensuring FERPA and GDPR compliance for international applications will require additional layers of consent and secure handling.

Finally, although the system allows asynchronous analysis, real-time deployment during lectures introduces latency and hardware constraints. Achieving seamless real-time feedback remains an open challenge, particularly for low-resource educational settings. Additionally, cultural biases embedded in training data or interpretation models could inadvertently affect feedback accuracy and inclusiveness.

7.6 Directions for Future Research

Future studies can explore several promising directions. First, expanding the dataset to include notes from various cultural and disciplinary backgrounds would enhance generalizability. Cross-institutional collaborations could provide a richer testbed for framework robustness.

Second, the integration of affective computing elements, such as sentiment analysis or emotional tone detection from annotations, could add a new layer of behavioral insight. This is especially relevant for understanding learning frustration or motivation.

Third, user-centered design studies involving both students and teachers could inform interface improvements and usability enhancements. This would ensure that the system's output is not only technically accurate but also pedagogically actionable.

Fourth, extending the framework to support multi-modal learning data—such as spoken reflections, video annotations, or collaborative whiteboards—would align with broader trends in holistic educational analytics. A future version of the system could simultaneously track audio annotations with note-taking patterns to detect confusion, curiosity, or conceptual shifts.

Fifth, exploring lightweight deployment strategies such as on-device inference using edge computing hardware (e.g., iPads with Apple Neural Engine or Android tablets with AI accelerators) could democratize access in resource-limited schools.

Moreover, incorporating reinforcement learning could enable the system to self-improve by learning from educator feedback over time. Real-world deployment in varied instructional settings, such as vocational training, language learning, or flipped classrooms, would further validate adaptability.

Finally, partnerships with commercial learning management systems could

pave the way for real-world deployment, turning academic prototypes into classroom-ready technologies that drive personalization at scale. Collaboration with policy-makers and edtech consortiums could help ensure ethical, scalable, and sustainable implementation pathways. An interdisciplinary advisory board comprising technologists, educators, and ethicists could further guide responsible development.

In conclusion, the synthesis of these studies confirms the viability and promise of AI-driven handwritten note analysis for advancing personalized education. By bridging vision and language, structure and semantics, cognition and computation, the proposed framework lays a solid foundation for future innovations in educational technology.

7.7 Chapter Summary

Chapter 7 provides a critical synthesis of the empirical findings and theoretical developments presented across this dissertation. By integrating insights from image segmentation, semantic analysis, and practical educational applications, this chapter highlights the holistic value of the proposed methodological framework. This synthesis underscores the dissertation's contribution to educational technology and multimodal analytics while transparently addressing limitations. The outlined future research directions clearly indicate how this foundational work can be expanded, paving the way for continued innovation in automated analysis of handwritten educational data.

Chapter 8

Conclusion and Future Work

Chapter 8 concludes this dissertation by summarizing its key contributions and discussing their broader educational implications. Building upon the comprehensive synthesis provided in Chapter 7, this chapter succinctly revisits the research objectives, highlights significant findings, and offers reflective insights on the practical applications and theoretical impacts of the proposed methods. It also clearly outlines future research opportunities, aiming to guide ongoing research efforts and inform practical implementations in educational analytics.

8.1 Summary of Contributions

This dissertation has explored the automated analysis of handwritten student notes through a unified framework that integrates image segmentation and generative language models. Grounded in three core studies, the research presents an end-to-end system that leverages Attention Multi-task U-Net for segmenting complex handwritten notes and GPT-4 for semantic understanding and knowledge alignment. The system demonstrates strong performance in identifying educationally relevant information from unstructured visual artifacts, ultimately enabling deeper

insights into learning behavior, content mastery, and pedagogical effectiveness.

Across multiple studies, the research confirms that integrating vision and language models yields higher accuracy and interpretability compared to isolated approaches. The hybrid framework not only improves segmentation precision and concept recognition but also bridges the gap between analog learning artifacts and digital analytics systems. This approach contributes new methods to the fields of educational technology, multimodal learning analytics, and AI-enhanced pedagogy.

From a theoretical perspective, the study redefines handwritten notes as rich, analyzable learning data. It supports the notion that note-taking is both a cognitive and communicative process. The system's ability to decode patterns in students' notes allows educators to observe the evolution of understanding over time, offering a dynamic and data-driven perspective on student learning.

8.2 Reflections on Research Findings

The research validates the hypothesis that handwritten notes, despite their variability and complexity, can be systematically analyzed to support teaching and learning. The experiments showed consistent success in capturing cognitive signals such as emphasis, repetition, structure, and conceptual density.

One particularly noteworthy finding is the cross-disciplinary adaptability of the framework. The models worked effectively across STEM and non-STEM domains, adapting to different notational conventions, diagram styles, and writing habits. Furthermore, educator feedback highlighted the usability of the visual outputs and concept extractions for instructional refinement.

Nevertheless, some challenges remain. Despite advances in segmentation accuracy and language comprehension, the interpretation of highly personalized note-

taking styles—such as symbolic annotations or multi-language mixing—still requires additional research. The quality of input data, particularly scanned image resolution and handwriting legibility, continues to impact model performance.

8.3 Implications for Stakeholders

This work offers practical implications for a range of educational stakeholders:

- 1) For educators, it provides tools to monitor engagement, diagnose misconceptions, and design personalized interventions.
- 2) For students, it offers potential feedback on note-taking effectiveness and content comprehension.
- 3) For administrators, the system introduces scalable metrics for curriculum evaluation and instructional improvement.
- 4) For researchers, it presents a new frontier for exploring the intersection of visual data processing and educational theory.
- 5) For edtech developers, it lays a foundation for next-generation tools that blend handwriting analysis with adaptive learning engines.

These implications suggest that the proposed framework can serve as a catalyst for transforming not only instructional support but also how institutions measure and support student progress.

8.4 Future Work

While this dissertation lays the groundwork, several avenues remain open for future exploration:

- 1) Model Generalization: Future studies could train and test the system on broader datasets that include various languages, writing tools (e.g., stylus, pencil), and age

groups.

- 2) Explainability and Trust: Building transparent model outputs, such as saliency maps or rationale traces, would improve educator trust in AI interpretations.
- 3) Real-Time Integration: Developing low-latency inference systems for classroom environments could enable real-time feedback loops for both teachers and students.
- 4) Cross-Modality Learning Analytics: Combining handwritten notes with speech, gaze, or interaction data could yield holistic learning profiles.
- 5) Ethics and Privacy: Further work is needed to establish secure and ethical protocols for collecting and analyzing student handwriting at scale.

By continuing this line of inquiry, future research can enhance the fidelity, inclusivity, and impact of AI-driven educational tools.

Another promising direction is the deployment of this framework in real classroom settings and its integration with Learning Management Systems (LMS). For instance, a scaled-up version of this system could be embedded into an LMS to automatically collect and analyze student notes as they are uploaded, providing teachers with continuous formative assessments. Such integration would facilitate practical use of this approach, closing the loop between research and classroom practice and ensuring that the system's benefits reach educators and learners directly.

8.5 Final Reflections

This dissertation demonstrates the untapped potential of handwritten notes as data-rich, cognitively meaningful resources in educational contexts. By transforming these analog traces into structured, actionable insights, it bridges a long-standing gap between student expression and educator understanding. The integrated use of vision and language AI represents a new frontier for personalized

learning analytics—one that is poised to redefine how educational progress is perceived, supported, and advanced.

Ultimately, this dissertation’s integrated approach to handwritten note analysis exemplifies how bridging multiple AI techniques can unlock new dimensions of educational data. Transforming free-form student notes into quantifiable insights takes a step toward personalized, evidence-driven education at scale. This work aims not only to close existing analytical gaps but also to inspire future innovations at the intersection of vision and language in educational technology.

Acknowledgement

First and foremost, I would like to express my heartfelt gratitude to my main advisor, Professor Chengjiu Yin, my co-advisor, Professor Masanori Yamada, and my doctoral supervisor, Professor Atsushi Shimada. Throughout my Ph.D. studies, they provided meticulous guidance, invaluable advice, and constant encouragement. Their rigorous scholarly approaches and broad academic visions have profoundly shaped both my research and personal growth.

I am sincerely grateful to my lab-mates at Kyushu University. Their unique insights during academic discussions and their assistance in the experimental process were indispensable to the successful completion of this study.

Furthermore, I wish to extend my deepest thanks to my parents. Over twenty-four years of education, their unconditional love, patience, and trust have given me the strength to overcome every obstacle and complete this academic journey.

Finally, I would like to conclude this dissertation with a heartfelt 「ありがとうございました」 to my doctoral supervisor, Professor Atsushi Shimada, and a sincere 「谢谢」 to my parents, 周春 and 刘桂梅.

Reference

- [1] B. Li, T. Minematsu, Y. Taniguchi, F. Okubo, and A. Shimada. How does analysis of handwritten notes provide better insights for learning behavior? In *Proceedings of the LAK 22: 12th International Learning Analytics & Knowledge Conference*, pp. 549–555. ACM, 2022.
- [2] R. S. Jansen, D. Lakens, and W. A. IJsselsteijn. An integrative review of the cognitive costs and benefits of note-taking. *Educational Research Review*, Vol. 22, pp. 223–233, 2017.
- [3] W. Jia, Z. Zhong, L. Sun, and Q. Huo. A cnn-based approach to detecting text from images of whiteboards and handwritten notes. In *Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR 2018)*, pp. 1–6, 2018.
- [4] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.
- [5] K. He, C. Lian, and B. Zhang. Hf-unet: Learning hierarchically inter-task relevance in multi-task u-net for accurate prostate segmentation in ct images. *IEEE Transactions on Medical Imaging*, Vol. 40, No. 8, pp. 2118–2128, 2021.

-
- [6] J. Dai, Y. Li, and K. He. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, 29, pp. 1–9, 2016.
- [7] Z. Zhou, M. M. R. Siddiquee, and N. Tajbakhsh. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, Vol. 39, No. 6, pp. 1856–1867, 2020.
- [8] H. Huang, L. Lin, and R. Tong. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1055–1059, 2020.
- [9] V. Rowtula, S. R. Oota, and J. C. Velu. Towards automated evaluation of handwritten assessments. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 426–433, 2019.
- [10] G. Sanuvala and S. S. Fatima. A study of automated evaluation of student examination papers using machine learning techniques. In *Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 1049–1054, 2021.
- [11] S. M. Chavan, M. S. Prerana, R. Bathula, S. Saikumar, and G. Dayalan. Automated script evaluation using machine learning and natural language processing. In *Proceedings of the 2023 2nd International Conference for Innovation in Technology (INOCON)*, pp. 1–7, 2023.
- [12] Y. Zhou. Improvement of image segmentation model for handwritten notebook analytics. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

-
- [13] J. Jo, H. I. Koo, and J. W. Soh. Handwritten text segmentation via end-to-end learning of convolutional neural networks. *Multimedia Tools and Applications*, Vol. 79, pp. 32137–32150, 2018.
- [14] S. Gholamian and A. Vahdat. Handwritten and printed text segmentation: A signature case study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pp. 582–592, 2019.
- [15] S. Létourneau, K. Sharma, C. Forsyth, and N. Rummel. Ai in collaborative learning: Twelve commitments for the future. *International Journal of Computer-Supported Collaborative Learning*, Vol. 16, No. 4, pp. 445–456, 2021.
- [16] X. Li, Y. Zhou, Z. Pan, and J. Feng. Partial order pruning for best speed/accuracy trade-off in neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 9145–9153, 2019.
- [17] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Trémeau, and C. Wolf. Residual conv-deconv grid network for semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC 2017)*, p. Paper 181, 2017.
- [18] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pp. 4151–4160, 2017.
- [19] K. Chen, H. Wei, J. Hennebert, R. Ingold, and M. Liwicki. Segmentation of historical handwritten document images using color and texture features.

- In *Proceedings of the 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014)*, pp. 488–493, 2014.
- [20] J. Jiang, Y.-C. Hu, C.-J. Liu, D. Halpenny, M. D. Hellmann, J. O. Deasy, G. Mageras, and H. Veeraraghavan. Multiple-resolution residually connected feature streams for automatic lung tumor segmentation from ct images. *IEEE Transactions on Medical Imaging*, Vol. 38, No. 1, pp. 134–144, 2019.
- [21] E. E. Enekwechi and M. C. Ezeanya. Study-habit as a predictor of secondary school students’achievement in chemistry in anambra state. *International Journal of Education and Evaluation*, Vol. 7, No. 2, pp. 1940–2695, 2018.
- [22] B. Flanagan, W. Chen, and H. Ogata. Joint activity on learner performance prediction using the bookroll dataset. In *Proceedings of the International Conference on Computers in Education (ICCE 2018)*, pp. 480–485, 2018.
- [23] T. Fukawa-Connelly, K. Weber, and J. P. Mejía-Ramos. Informal content and student note-taking in advanced mathematics classes. *Journal for Research in Mathematics Education*, Vol. 48, No. 5, pp. 567–579, 2017.
- [24] D. He, S. Cohen, B. Price, D. Kifer, and C. L. Giles. Multi-scale multi-task fcn for semantic page segmentation and table detection. In *Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1, pp. 254–261. IEEE, 2017.
- [25] A. Y. Q. Huang, O. H. T. Lu, J. C. H. Huang, C. J. Yin, and S. J. H. Yang. Predicting students’academic performance using educational big data and learning analytics: Evaluation of classification methods and learning logs. *Interactive Learning Environments*, Vol. 28, No. 2, pp. 206–230, 2020.

-
- [26] C. Yang and F. Gao. Eda-net: Dense aggregation of deep and shallow information achieves quantitative photoacoustic blood oxygenation imaging deep in human breast. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2018)*, pp. 246–254. Springer, 2018.
- [27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [28] O. Oktay. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [29] M. Z. Alom, M. Hasan, and C. Yakopcic. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.
- [30] A. Kundu and T. Bej. Mobile phone vs. paper–pencil mode of note-taking: A case study among the students of bankura university. *Interactive Technology and Smart Education*, Vol. 17, No. 1, pp. 82–90, 2019.
- [31] V. Zyuzin and T. Chumarnaya. Comparison of u-net architectures for segmentation of the left ventricle endocardial border on two-dimensional ultrasound images. In *Proceedings of the 2019 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT)*, pp. 110–113. IEEE, 2019.
- [32] J. Lee, H. Hayashi, W. Ohyama, and S. Uchida. Page segmentation using a convolutional neural network with trainable co-occurrence features. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1023–1028. IEEE, 2019.

-
- [33] P. A. Mueller and D. M. Oppenheimer. The pen is mightier than the keyboard: Advantages of longhand over laptop note-taking. *Psychological Science*, Vol. 25, No. 6, pp. 1159–1168, 2014.
- [34] I. Mwalumbwe and J. S. Mtebe. Using learning analytics to predict students’ performance in moodle learning management system: A case of mbeya university of science and technology. *The Electronic Journal of Information Systems in Developing Countries*, Vol. 79, No. 1, pp. 1–13, 2017.
- [35] P. Nilsson. An inferentialist perspective on how note-taking can constrain the orchestration of math-talk. *Scandinavian Journal of Educational Research*, Vol. 63, No. 7, pp. 1121–1133, 2019.
- [36] H. Cui, X. Liu, and N. Huang. Pulmonary vessel segmentation based on orthogonal fused u-net++ of chest ct images. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019)*, pp. 293–300. Springer, 2019.
- [37] H. Ogata, C. Yin, M. Oi, F. Okubo, A. Shimada, K. Kojima, and M. Yamada. E-book-based learning analytics in university education. In *Proceedings of the International Conference on Computers in Education (ICCE 2015) Workshops*, pp. 401–406, 2015.
- [38] T. Pavlidis and J. Zhou. Page segmentation and classification. *CVGIP: Graphical Models and Image Processing*, Vol. 54, pp. 484–496, 1992.
- [39] A. Piolat, T. Olive, and R. T. Kellogg. Cognitive effort during note-taking. *Applied Cognitive Psychology*, Vol. 19, No. 3, pp. 291–312, 2005.
- [40] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal, and A. Alaei. Icdar 2013 handwriting segmentation contest. In *Proceedings of the 2013 12th*

- International Conference on Document Analysis and Recognition (ICDAR 2013)*, pp. 1402–1406. IEEE, 2013.
- [41] A. L. H. Meer and F. R. Weel. Only three fingers write, but the whole brain works: A high-density eeg study showing advantages of drawing over typing for learning. *Frontiers in Psychology*, Vol. 8, p. Art. 706, 2017.
- [42] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao. Shape-robust text detection with progressive scale expansion network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 9336–9345, 2019.
- [43] F. Re, A. Shatillo, and A. Sierra. Automated joint skull-stripping and segmentation with multi-task u-net in large mouse brain mri databases. *NeuroImage*, Vol. 229, p. 117734, 2021.
- [44] X. Lu, Y. Zhong, and Z. Zheng. Multi-scale and multi-task deep learning framework for automatic road extraction. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 11, pp. 9362–9377, 2019.
- [45] X. Yang, Z. Zeng, and S. Y. Yeo. A novel multi-task deep learning model for skin lesion segmentation and classification. *arXiv preprint arXiv:1703.01025*, 2017.
- [46] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su. Dunet: A deformable network for retinal vessel segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, p. to appear, 2019.
- [47] J. Dolz, I. Ben Ayed, and C. Desrosiers. Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities. In *Proceedings of the*

- MICCAI Brain Lesion Workshop*, Vol. 11383 of *Lecture Notes in Computer Science*, pp. 271–282, 2019.
- [48] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie. Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical Physics*, Vol. 46, No. 2, pp. 576–589, 2019.
- [49] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo. Automatic brain tumor detection and segmentation using u-net–based fully convolutional networks. In *Proceedings of the Annual Conference on Medical Image Understanding and Analysis (MIUA 2017)*, pp. 506–517. Springer, 2017.
- [50] J. Zhang, Y. Jin, J. Xu, X. Xu, and Y. Zhang. Mdu-net: Multiscale densely connected u-net for biomedical image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 2018. Online, 8 pages. Available: arXiv:1811.10434.
- [51] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan. Medical image segmentation based on u-net: A review. *Journal of Imaging Science and Technology*, Vol. 64, No. 2, p. 20504, 2020.
- [52] U. G. Naidu, R. Thiruvengatanadhan, S. Narayana, and P. Dhanalakshmi. Character-level segmentation and recognition using cnn followed by a random forest classifier for npr system. *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 11, pp. 12–18, 2022.
- [53] J. Hong, B.-Y. Park, M. J. Lee, C.-S. Chung, J. Cha, and H. Park. Two-step deep neural network for segmentation of deep white matter hyperintensities

- in migraineurs. *Computer Methods and Programs in Biomedicine*, Vol. 183, p. 105065, 2020.
- [54] D. Akila, H. Garg, S. Pal, and S. Jeyalakshmi. Research on recognition of students' attention in offline classrooms based on deep learning. *Education and Information Technologies*, Vol. 28, pp. 1–29, 2023.
- [55] S. Akre, N. Palandurkar, A. Iyengar, G. Chayande, and P. Kumar. Engagedatr-vl: A multimodal engagement dataset with emotional, cognitive, and behavioral cues in virtual learning. In *Proceedings of the 8th International Conference on Pattern Recognition and Machine Intelligence (PReMI 2019)*, pp. 270–278, 2019.
- [56] R. D. Alfredo, L. Nie, P. Kennedy, T. Power, C. Hayes, and H. Chen. “that student should be a lion tamer!” designing a stress analytics dashboard for teachers. In *Proceedings of the 13th International Learning Analytics & Knowledge Conference (LAK'23)*, pp. 57–67. ACM, 2023.
- [57] D. Carpenter, W. Min, S. Lee, G. Ozogul, and J. Lester. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 403–413, 2024.
- [58] R. Y. Y. Chan, C. M. V. Wong, and Y. N. Y. Yum. Predicting behavior change in students with special education needs using multimodal learning analytics. *IEEE Access*, Vol. 11, pp. 22313–22324, 2023.
- [59] W. Chango, R. Cerezo, M. Sánchez-Santillán, R. Azevedo, and C. Romero. Improving prediction of students' performance in intelligent tutoring systems

- using attribute selection and ensembles of multimodal data sources. *Journal of Computing in Higher Education*, Vol. 33, No. 3, pp. 614–634, 2021.
- [60] W. Chango, J. A. Lara, R. Cerezo, and C. Romero. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 12, No. 2, p. e1458, 2022.
- [61] P. Chejara, L. P. Prieto, M. J. Rodríguez-Triana, A. Ruiz-Calleja, R. Kasepalu, and B. Schneider. Exploring indicators for collaboration quality in classroom settings using multimodal learning analytics. In *Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL 2018)*, pp. 60–74, 2018.
- [62] P. Chejara, L. P. Prieto, M. J. Rodríguez-Triana, A. Ruiz-Calleja, and M. Khalil. Impact of window size on generalizability of collaboration quality models developed using mmla. In *Proceedings of the 13th International Learning Analytics & Knowledge Conference (LAK'23)*, pp. 559–565, 2023.
- [63] N. Chettaoui, A. Atia, and M. S. Bouhlef. Student performance prediction with eye-gaze data in an embodied educational context. *Education and Information Technologies*, Vol. 28, No. 1, pp. 833–855, 2023.
- [64] A. H. Closser, J. A. Erickson, H. Smith, A. Varatharaj, and A. F. Botelho. Blending learning analytics and embodied design to model students' comprehension using actions, speech, and gestures. *International Journal of Child-Computer Interaction*, Vol. 32, p. 100391, 2022.

-
- [65] C. M. D’Angelo and R. J. Rajarathinam. Speech analysis of teaching assistant interventions in small-group collaborative problem solving. *British Journal of Educational Technology*, Vol. 55, No. 4, pp. 1583–1601, 2024.
- [66] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. de Weerd. A review of automated feedback systems for learners: Classification framework, challenges, and opportunities. *Computers & Education*, Vol. 162, p. 104094, 2021.
- [67] D. Mitri, J. Schneider, and H. Drachsler. Keep me in the loop: Real-time feedback with multimodal data. *International Journal of Artificial Intelligence in Education*, Vol. 32, No. 3, pp. 1093–1118, 2022.
- [68] A. Emerson, E. B. Cloude, R. Azevedo, and J. Lester. Multimodal learning analytics for game-based learning. *British Journal of Educational Technology*, Vol. 51, No. 5, pp. 1505–1526, 2020.
- [69] A. Emerson, N. Henderson, J. Rowe, W. Min, S. Lee, J. Minogue, and J. Lester. Early prediction of visitor engagement in science museums with multimodal learning analytics. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI 2020)*, pp. 107–116, 2020.
- [70] A. Emerson, W. Min, J. Rowe, R. Azevedo, and J. Lester. Multimodal predictive student modeling with multi-task transfer learning. In *Proceedings of the 13th International Learning Analytics & Knowledge Conference (LAK’23)*, pp. 333–344, 2023.
- [71] M. van Haastrecht, M. Brinkhuis, and M. Spruit. Federated learning analytics: Investigating the privacy–performance trade-off in ml for educational

- analytics. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2021)*, pp. 62–74, 2021.
- [72] C. Hou, G. Zhu, and J. Zheng. Prompt-based and fine-tuned gpt models for context-dependent and -independent deductive coding in social annotation. In *Proceedings of the 14th International Learning Analytics & Knowledge Conference (LAK'24)*, pp. 518–528, 2024.
- [73] L. Huang, T. Doleck, and B. Chen. Multimodal learning analytics for assessing teachers’self-regulated learning in planning technology-integrated lessons. *Education and Information Technologies*, Vol. 28, No. 12, pp. 15823–15843, 2023.
- [74] M. Israel, T. Liu, J. Moon, F. Ke, and I. Dahlstrom-Hakki. Understanding students’problem-solving processes and affective trajectories during game-based learning: A data fusion approach. In *Proceedings of the International Conference on Human–Computer Interaction (HCII 2020)*, pp. 201–215. Springer, 2020.
- [75] N. Ivleva, A. Pentel, and O. Dunajeva. Deep learning-based audiovisual emotion recognition in a smart learning environment. In *Proceedings of the International Conference on Interactive Collaborative Learning (ICL 2020)*, pp. 420–431, 2020.
- [76] R. Kawamura, S. Shirai, and N. Takemura. Detecting drowsy learners on e-learning platforms with multimodal learning analytics. *IEEE Access*, Vol. 9, pp. 115165–115174, 2021.

-
- [77] H. Khosravi, S. B. Shum, and G. Chen. Explainable artificial intelligence in education. *Computers & Education: Artificial Intelligence*, Vol. 3, p. 100074, 2022.
- [78] Y. Lee, H. Chen, G. Zhao, and M. Specht. Wedar: Webcam-based attention analysis via behavior recognition with a novel e-reading dataset. In *Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI 2022)*, pp. 319–328, 2022.
- [79] E. Latif and X. Zhai. Fine-tuning chatgpt for automatic scoring. *Computers & Education: Artificial Intelligence*, Vol. 6, No. 3, p. 100210, 2023.
- [80] X. Li, L. Yan, L. Zhao, R. Martinez-Maldonado, and D. Gašević. Cvpe: A computer vision approach for scalable, privacy-preserving socio-spatial learning analytics. In *Proceedings of the 13th International Learning Analytics & Knowledge Conference (LAK'23)*, pp. 175–185, 2023.
- [81] C. J. Lin, W. S. Wang, H. Y. Lee, Y. M. Huang, and T. T. Wu. Recognition of image and speech to improve learning diagnosis on stem collaborative activity for precision education. *Education and Information Technologies*, Vol. 28, No. 3, pp. 1–26, 2023.
- [82] Y. Ma, M. Celepkolu, and K. E. Boyer. Detecting impasse during collaborative problem solving with multimodal learning analytics. In *Proceedings of the 12th International Learning Analytics & Knowledge Conference (LAK'22)*, pp. 45–55, 2022.
- [83] K. Mangaroska, R. Martinez-Maldonado, B. Vešin, and D. Gašević. Challenges and opportunities of multimodal data in human learning: A student

- perspective. *Journal of Computer Assisted Learning*, Vol. 37, No. 4, pp. 1030–1047, 2021.
- [84] R. Martinez-Maldonado, V. Echeverría, and G. Fernandez-Nieto. Lessons learnt from a multimodal learning analytics deployment in the wild. *ACM Transactions on Computer-Human Interaction*, Vol. 31, No. 1, pp. 1–41, 2023.
- [85] J. Moon, F. Ke, Ž. Sokolij, and I. Dahlstrom-Hakki. Multimodal data fusion to track students’ distress during educational gameplay. *Journal of Learning Analytics*, Vol. 9, No. 3, pp. 75–87, 2022.
- [86] J. Mozafari, A. Jangra, and A. Jatowt. Triviahg: A dataset for automatic hint generation from factoid questions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*, pp. 2060–2070, 2024.
- [87] A. Nandi, F. Khafa, L. Subirats, and S. Fort. Real-time multimodal emotion classification system in an e-learning context. In *Proceedings of the International Conference on Engineering Applications of Neural Networks (EANN 2022)*, pp. 423–435, 2022.
- [88] H. Ouhaichi, D. Spikol, and B. Vogel. Research trends in multimodal learning analytics: A systematic mapping study. *Computers & Education: Artificial Intelligence*, Vol. 4, p. 100136, 2023.
- [89] H. Ouhaichi, D. Spikol, and B. Vogel. A systematic review of multimodal learning analytics design models and frameworks. In *Proceedings of the 16th International Conference on Education and New Learning Technologies (ED-ULEARN 2020)*, pp. 1007–1016, 2020.

- [90] F. Ouyang, W. Xu, and M. Çukurova. An ai-driven learning analytics method to examine the collaborative problem-solving process from a complex adaptive systems perspective. *International Journal of Computer-Supported Collaborative Learning*, Vol. 18, No. 1, pp. 39–66, 2023.
- [91] B. Pei, W. Xing, and M. Wang. Academic development of multimodal learning analytics: A bibliometric analysis. *Interactive Learning Environments*, Vol. 31, No. 7, pp. 3543–3561, 2023.
- [92] Q. Peng, N. Qie, L. Yuan, Y. Chen, and Q. Gao. Recognition of students’ multiple mental states in conversation based on multimodal cues. In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020), Revised Selected Papers*, pp. 468–479. Springer, 2020.
- [93] P. Prinsloo, S. Slade, and M. Khalil. Multimodal learning analytics—in between student privacy and encroachment: A systematic review. *British Journal of Educational Technology*, Vol. 54, No. 4, pp. 1566–1586, 2023.
- [94] A. Sabuncuoğlu and T. M. Sezgin. Developing a multimodal classroom engagement analysis dashboard for higher education. *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7, No. CSCW1, pp. 1–23, 2023.
- [95] S. K. Shankar, A. Ruiz-Calleja, L. P. Prieto, M. J. Rodríguez-Triana, P. Chajara, and S. Tripathi. Cimla: A modular data infrastructure to support context-aware multimodal learning analytics. *Journal of Universal Computer Science*, Vol. 29, No. 3, pp. 293–314, 2023.
- [96] Ö. Sümer, P. Goldberg, and S. D’Mello. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, Vol. 14, No. 4, pp. 1012–1027, 2023.

-
- [97] F. Tahiru. Ai in education: A systematic literature review. *Journal of Cases on Information Technology*, Vol. 23, No. 1, pp. 1–20, 2021.
- [98] E. Tiukhov, P. Vemuri, and N. L. Flores. Explainable learning analytics: Assessing the stability of student success prediction models via explainable ai. *Decision Support Systems*, Vol. 182, p. 114229, 2021.
- [99] C. Vatrals, M. Lee, and C. Cohn. Prediction of students’self-confidence using multimodal features in an experiential nurse training environment. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2023)*, pp. 266–271, 2023.
- [100] Y. Wang and X. Gu. Data fusion in classroom-based multimodal learning analytics: A systematic literature review. In *Proceedings of the 18th International Conference of the Learning Sciences (ICLS 2018)*, pp. 951–954, 2018.
- [101] M. Worsley, R. Martinez-Maldonado, and C. D’Angelo. A new era in multimodal learning analytics: Twelve core commitments to ground and grow mmla. *Journal of Learning Analytics*, Vol. 8, No. 3, pp. 10–27, 2021.
- [102] L. Yan, R. Martinez-Maldonado, and B. Gallo-Córdoba. Mapping proximity traces to socio-spatial behaviors and student progression in school. *British Journal of Educational Technology*, Vol. 53, No. 7, pp. 1645–1664, 2022.
- [103] L. Yan, R. Martinez-Maldonado, and D. Gašević. Generative artificial intelligence in learning analytics: Opportunities and challenges through the learning analytics cycle. *arXiv preprint arXiv:2312.00087*, 2023.
- [104] L. Yan, L. Zhao, D. Gašević, and R. Martinez-Maldonado. Scalability, sustainability, and ethicality of multimodal learning analytics. In *Proceedings*

- of the 12th International Learning Analytics & Knowledge Conference (LAK'22)*, pp. 13–23, 2022.
- [105] A. Yusuf, N. M. Noor, and S. Bello. Using multimodal learning analytics to model students' learning behavior in an animated programming classroom. *Education and Information Technologies*, Vol. 28, No. 1, pp. 775–818, 2023.
- [106] L. Zhao, V. Echeverría, Z. Swiecki, L. Yan, R. Alfredo, and X. Li. Epistemic network analysis for end-users: Closing the loop in the context of multimodal analytics for collaborative team learning. In *Proceedings of the 14th International Learning Analytics & Knowledge Conference (LAK'24)*, pp. 90–100, 2024.
- [107] Q. Zhou, A. Bhattacharya, W. Suraworachet, H. Nagahara, and M. Çukurova. Automated detection of students' gaze interactions in collaborative learning videos: A novel approach. In *Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL 2021)*, pp. 504–517, 2021.
- [108] Q. Zhou, W. Suraworachet, and M. Çukurova. Detecting non-verbal speech and gaze behaviors with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and Information Technologies*, Vol. 29, No. 2, pp. 1071–1098, 2024.
- [109] H. E. Abdelkader, A. G. Gad, A. A. Abohany, and S. E. Sorour. An efficient data mining technique for assessing satisfaction with online learning during covid-19. *IEEE Access*, Vol. 10, pp. 6286–6303, 2022.
- [110] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu. Two decades of artificial intelligence in education: Contributors, collaborations, research topics, chal-

- lenges, and future directions. *Educational Technology & Society*, Vol. 25, No. 1, pp. 28–47, 2022.
- [111] H.-C. Chu, G.-H. Hwang, Y.-F. Tu, and K.-H. Yang. Roles and research trends of artificial intelligence in higher education: A systematic review of the 50 most-cited articles. *Australasian Journal of Educational Technology*, Vol. 38, No. 3, pp. 22–42, 2022.
- [112] I. Daoudi. Learning analytics for enhancing the usability of serious games in formal education: A systematic literature review and research agenda. *Education and Information Technologies*, Vol. 27, No. 8, pp. 11237–11266, 2022.
- [113] A. Darvishi, H. Khosravi, S. Sadiq, and B. Weber. Neurophysiological measurements in higher education: A systematic literature review. *International Journal of Artificial Intelligence in Education*, Vol. 32, No. 2, pp. 413–453, 2022.
- [114] S. Doroudi. The intertwined histories of artificial intelligence and education. *International Journal of Artificial Intelligence in Education*, Vol. 32, No. 4, pp. 1–44, 2022.
- [115] A. Emerson, W. Min, J. Rowe, J. Lester, and R. Azevedo. Utilizing multimodal learning analytics to predict visitor engagement in museums. *User Modeling and User-Adapted Interaction*, Vol. 33, No. 1, pp. 1–30, 2023.
- [116] K. Fahd, S. Venkatraman, S. J. Miah, and K. Ahmed. Applications of artificial intelligence in education: A systematic review of themes in recent research. *Education and Information Technologies*, Vol. 27, No. 8, pp. 10443–10469, 2022.

-
- [117] C. S. Fichten, L. King, C. Horne, I. Mitchell, and J. Budgor. Using artificial intelligence to enhance accessibility in education for students with disabilities. *Journal of Educational Technology Systems*, Vol. 50, No. 2, pp. 220–239, 2021.
- [118] A. Gudyanga. Artificial intelligence in education in africa: Current pathways and future prospects. *SN Social Sciences*, Vol. 3, No. 4, p. 42, 2023.
- [119] W. Holmes, K. Porayska-Pomsta, and K. Holstein. Ethics of ai in education: Toward a community-wide framework. *International Journal of Artificial Intelligence in Education*, Vol. 31, No. 4, pp. 1–23, 2021.
- [120] G.-J. Hwang, K.-Y. Tang, and Y.-F. Tu. How artificial intelligence supports nursing education: Profiling the roles, applications, and trends of ai in nursing education research (1993–2020). *Interactive Learning Environments*, Vol. 30, No. 8, pp. 1504–1524, 2022.
- [121] D. Ifenthaler and J. Y.-K. Yau. Utilizing learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, Vol. 68, No. 4, pp. 1961–1990, 2020.
- [122] Z. H. İpek, A. İ. Ç. Gözümlü, S. Papadakis, and M. Kallogiannakis. Educational applications of the chatgpt ai system: A systematic review. *Educational Process: International Journal*, Vol. 12, No. 3, pp. 7–26, 2023.
- [123] M. J. Khan, I. Ahmad, M. Amanullah, and A. Pardo. Learning analytics in higher education: Current landscape and future directions. *Journal of Learning Analytics*, Vol. 9, No. 3, pp. 1–17, 2022.
- [124] M. Lan and X. Zhou. A qualitative systematic review on ai-empowered self-regulated learning in higher education. *npj Science of Learning*, Vol. 10, p. 21, 2025.

- [125] A. Létourneau, M. D. Martineau, and P. Charland. A systematic review of ai-driven intelligent tutoring systems (its) in k–12 education. *npj Science of Learning*, Vol. 10, p. 29, 2025.
- [126] D. Liaqat, Z. Irani, U. Sivarajah, and H. Yared. A decade of learning analytics: A topic modeling and bibliometric analysis (2011–2020). *Education and Information Technologies*, Vol. 27, No. 8, pp. 10517–10561, 2022.
- [127] V. I. Marin, K. Buntins, S. Bedenlier, and M. Bond. Invisible borders in educational technology research? a comparative analysis of collaboration patterns. *Education and Information Technologies*, Vol. 28, No. 1, pp. 667–690, 2023.
- [128] V. Maphosa and S. Maphosa. The potential of artificial intelligence in african higher education: Challenges and opportunities. *International Journal of African Higher Education*, Vol. 7, No. 1, pp. 69–91, 2020.
- [129] Y. Miao, K. Holstein, B. McLaren, and K. Koedinger. Integrating transformer models for automated short answer grading in classrooms. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2021)*, pp. 175–187, 2021.
- [130] J. Pan, B. Liu, J. Wang, and R. Huang. Graph-based handwriting evaluation for automated scoring of mathematical proofs. In *Proceedings of the International Conference on Educational Data Mining (EDM 2021)*, pp. 282–288, 2021.
- [131] I. A. Burney and N. Ahmad. Artificial intelligence in medical education: A citation-based systematic literature review. *Journal of Shifa Tameer-e-Millat University*, Vol. 5, No. 1, pp. 43–53, 2022.

-
- [132] S. Z. Salas-Pilco and Y. Yang. Artificial intelligence and education in latin america: A systematic review of the literature (2010–2020). *Education Sciences*, Vol. 12, No. 8, p. 790, 2022.
- [133] B. Buchanan, M. Carter, and J. Cowell. Virtual patients and ai tutors in medical and dental education: An integrative review. *Education Sciences*, Vol. 11, No. 12, p. 790, 2021.
- [134] Y. K. Dwivedi, S. Barton, and P. Kataria. Generative ai (chatgpt) for education: Opportunities, challenges, and research agenda. *International Journal of Information Management*, Vol. 71, p. 102642, 2023.
- [135] L. Zheng, W. Wang, Y. Wang, and X. Li. Large language models as ai tutors: A survey of chatgpt in education. *arXiv preprint arXiv:2308.10704*, 2023.
- [136] T.-T. Wu, H.-C. Liu, W.-C. Hou, and C.-F. Lai. Leveraging computer vision for adaptive learning in stem education: Effect of engagement and self-efficacy. *International Journal of Educational Technology in Higher Education*, Vol. 20, p. 53, 2023.
- [137] C. S. Fichten, J. V. Asuncion, J. Budd, M. N. Nguyen, and M. Barile. Higher education, digital accessibility, and covid-19: An interview-based study of students with disabilities and instructors. *Journal of Postsecondary Education and Disability*, Vol. 34, No. 3, pp. 245–260, 2021.
- [138] X. Zhai and R. Wibowo. Affect-aware virtual learning environment: Detecting and addressing student anxiety with multimodal data. *British Journal of Educational Technology*, Vol. 54, No. 1, pp. 153–171, 2023.

-
- [139] S. Bai, L. Zhang, Y. Ye, and C. Zhao. Learning engagement and academic performance: A meta-analysis. *Educational Research and Reviews*, Vol. 16, No. 6, pp. 191–200, 2021.
- [140] A. W. Qureshi, Y. Yang, and K. Hogg. Investigating the relationship between student engagement and performance in online learning environments. *Education and Information Technologies*, Vol. 26, No. 6, pp. 7307–7325, 2021.
- [141] J. G. Chaka. Prospects and challenges of artificial intelligence in education: A systematic review. *International Journal of Educational Technology*, Vol. 8, No. 1, pp. 39–52, 2021.
- [142] H. Crompton and D. Burke. Artificial intelligence in education: Emerging themes and enduring issues. *TechTrends*, Vol. 67, No. 1, pp. 16–27, 2023.
- [143] A. M. Al-Zahrani and T. M. Alasmari. Exploring the impact of ai on higher education: Ethical, social, and educational implications. *Humanities and Social Sciences Communications*, Vol. 11, p. 912, 2024.
- [144] T. Shaik, X. Tao, and Y. Li. Trends and challenges in adopting nlp for educational feedback analysis: A review. *IEEE Access*, Vol. 10, pp. 56720–56739, 2022.
- [145] S. Higgins, Z. Xiao, and M. Katsipataki. The impact of digital technology on learning: A summary for the education endowment foundation (technical report). Technical report, Education Endowment Foundation, 2022.
- [146] A. Kannan, V. Banerjee, and S. Paul. Automated evaluation of programming assignments using deep learning. In *Proceedings of the IEEE International Conference on Engineering Education (ICEED 2020)*, pp. 215–220, 2020.

- [147] K. Holstein, B. McLaren, M. Šega, and V. Aleven. Designing for human-ai complementarity in education: A case study in math tutoring. *International Journal of Artificial Intelligence in Education*, Vol. 32, No. 4, pp. 1–28, 2022.
- [148] M. Bond, H. Khosravi, and M. Laat. A meta-review of artificial intelligence in higher education: The need for ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, Vol. 21, p. 4, 2022.
- [149] A. Pinto, F. J. Rodríguez, D. Maldonado, and F. J. García-Peñalvo. Artificial intelligence in higher education: A bibliometric analysis of latin american research. *Computers & Education*, Vol. 182, p. 104463, 2022.
- [150] H. Hsu, T. Sinha, S. Bagley, and P. Dillenbourg. A video dataset for classroom group engagement recognition. *Scientific Data*, Vol. 10, p. 62, 2023.
- [151] R. Singh, A. Sharma, and D. Jha. Student engagement detection in classrooms through computer vision and deep learning. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2022)*, pp. 1818–1822, 2022.
- [152] V. Ramaswamy, A. Nagpal, and S. Srivastava. Ai-based student engagement detection: Leveraging cnns for analysis. *International Journal of Interactive Smart Education*, Vol. 4, No. 2, pp. 45–58, 2021.
- [153] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [154] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, Vol. 521, No. 7553, pp. 436–444, 2015.

- [155] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA+ML-CDS 2018)*, Vol. 11045 of *Lecture Notes in Computer Science*, pp. 3–11. Springer, Cham, 2018.
- [156] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [157] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 833–851, 2018.
- [158] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.
- [159] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies*, pp. 4171–4186, 2019.
- [160] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

-
- [161] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Technical Report, 2019. Available at <https://openai.com/blog/better-language-models>.
- [162] Raymond Smith. An overview of the Tesseract ocr engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 629–633, 2007.
- [163] Google Cloud Vision API documentation. <https://cloud.google.com/vision>, 2021. Accessed 2021-12-01.
- [164] Kelei He, Chunfeng Lian, Bing Zhang, Xin Zhang, Xiaohuan Cao, Dong Nie, Yang Gao, Junfeng Zhang, and Dinggang Shen. HF-UNet: Learning hierarchically inter-task relevance in multi-task u-net for accurate prostate segmentation in CT images. *IEEE Transactions on Medical Imaging*, Vol. 40, No. 8, pp. 2118–2128, 2021.