

LearnInsight: Class-Level Learning Assessment and Gap Detection System

Asst. Prof Nikesh Aote¹, Ayush Verma², Alok Dubey³, Chetna Khapre⁴, Durga Kashyap⁵, Om Thombre⁶
^{1,2,3,4,5,6}*Computer Science & Engineering Priyadarshini College of Engineering Nagpur, India*

Abstract—Marks and grades show how a student performed but not which concepts were misunderstood, and in large classes teachers cannot manually inspect every response to find shared weaknesses. This paper presents LearnInsight, a class-level learning gap detection system that analyses assessment responses to find weak concepts and shared misconceptions. Multiple-choice answers are checked directly, while free-text and OCR-read handwritten answers are graded by a large language model (Claude API) that names the misconception behind a wrong answer. A transparent rule engine then computes concept mastery per student and per class, flags a concept as a class-wide gap when mastery is below 50%, and flags a shared misconception when 40% or more of the class choose the same wrong answer. Flagged gaps are matched with videos, notes, tutorials and practice sets and shown on a teacher dashboard. A prototype on a sample Grade 8 class test is described, together with the planned evaluation protocol. The system is intended to support, not replace, teacher judgement.

Index Terms—Learning Analytics; Learning Gap Detection; Misconceptions; Large Language Models; Educational Data Mining; Explainable Analytics; Teacher Dashboard.

I. INTRODUCTION

A. Background

The growth of higher education strains traditional instruction. As enrolments in individual courses rise into the hundreds, particularly in Computer Science, instructors rely on one-to-many pedagogies, which reduce student–instructor interaction and the amount of tailored feedback [1]. Instructor feedback, unlike that from peers or algorithms, can correct deep-seated knowledge gaps and show students the ways of thinking foundational to a discipline. In most schools and colleges, however, students are still evaluated mainly through examination marks. Marks indicate

overall performance but do not explain which concepts were understood or misunderstood, so many students score low because of weak basic concepts that remain undetected and continue to hamper later topics.

AI-driven adaptive feedback and learning analytics promise personalised support at scale by analysing student interactions, detecting misunderstandings and returning timely feedback. Their effectiveness in closing conceptual gaps still needs empirical study, and they should complement rather than replace instructors.

B. Motivation

In a classroom with many students, manually analysing every answer sheet to locate concept-level weaknesses and common misconceptions is slow and error-prone. Gradebooks and LMS dashboards report scores and averages, and knowledge-tracing tools focus on individual learners, so a teacher still cannot easily see that, for example, most of the class chose the same wrong option for a concept. A simple, automated and explainable tool that surfaces such patterns would allow targeted re-teaching before the next topic or examination.

C. Research Gaps

- Existing systems focus mainly on marks and grades and do not analyse the reasons behind learning gaps.
- Analytics often *report* performance rather than predict or identify emerging gaps, and most work stops at student-level risk.
- Few approaches give personalised, topic-wise insight into which concepts have not been mastered; shared wrong answers are rarely treated as misconceptions.
- Few systems link a detected gap to resources a teacher can use immediately, and handwritten answers are usually excluded from automated analysis.

- Teachers need transparent, explainable and teacher-friendly methods they can trust.

D. Objectives and Contributions

The objectives are to (i) detect concept-level weaknesses and shared misconceptions at class level from assessment responses, (ii) combine LLM grading of written answers with transparent, rule-based analytics, and (iii) recommend learning resources for every flagged gap.

The main contributions are:

- A class-level gap detection pipeline with three interpretable status levels: No Gap, Class-wide Gap and Shared Misconception.
- A hybrid analysis model that pairs Claude-based grading of free text/OCR answers with an explicit rule engine (50% mastery and 40% misconception thresholds).
- Automatic matching of each flagged concept to videos, notes, tutorials and practice sets.
- A teacher-facing dashboard (Express REST API, MongoDB) and a planned evaluation protocol on unseen answers.

II. LITERATURE REVIEW

Educational Data Mining (EDM) and learning analytics have been surveyed extensively. Romero and Ventura [2] review classification, clustering and prediction techniques that reveal learning behaviour and academic performance, while Baker and Inventado [3] and Siemens and Baker [4] show that student data supports personalised learning and early identification of struggling students, though accurate prediction needs large, reliable datasets and no specific framework for classroom-level gap detection

is offered.

Early-warning work uses activity and performance data. Macfadyen and Dawson [5] mined LMS data to predict academic difficulty, but concentrated on activity rather than complete academic performance. Kotsiantis et al. [6] applied decision trees and Bayesian classification in distance learning; accuracy depended heavily on historical data quality. Romero et al. [7] used classification and association-rule mining on course-management data, with limited personalisation of interventions. Jayaprakash et al. [8] built an early-alert system for at-risk students, which identifies risk but does not deeply diagnose specific concepts. Ferguson [9] highlights privacy, data-integration and ethical challenges.

Identifying knowledge gaps at scale has also been studied. Stephens-Martinez et al. [10] proposed a semi-automated method that predicts knowledge gaps from statistical co-occurrence patterns with minimal human labelling. Recent work shows that LLMs can generate learning resources comparable to student-made ones and that prompting method strongly affects quality, yet hallucination remains a risk and instructors must guide the model [11]. Performance prediction with deep neural networks, random forests, SVMs and KNN has reached 89% accuracy [12], Moodlebased regression and cluster analysis found forum visits, quizzes, assignments and age most influential [13], and Bayes classification has been used to predict whether new students will perform well [14]. Generative AI has mostly supported teaching rather than analysing learning [15], and domain-specific learning progressions are needed to link learning processes to long-term competence [16].

Table I summarises these studies. Most identify *who* is at risk; few diagnose *which concepts* the class has not understood or why.

TABLE I. SUMMARY OF RELATED STUDIES

Study	Year	Approach	Key finding / limitation
Romero & Ventura [2]	2010	EDM review: classification, clustering, prediction	Behaviour and performance can be analysed; little real-time gap identification
Baker & Inventado [3]	2014	Data mining + learning analytics	Supports personalisation; needs large reliable data
Siemens & Baker [4]	2012	Learning analytics, behavioural data	Monitors students; no classroom-level gap framework
Macfadyen & Dawson [5]	2010	LMS data early warning	Predicts difficulty; focuses on LMS activity
Kotsiantis et al. [6]	2004	Decision trees, Bayes	Predicts poor performers; data-quality dependent

Romero et al. [7]	2008	Classification, association rules	Reveals patterns; limited personalisation
Jayaprakash et al. [8]	2014	Predictive early alert	Finds at-risk students; no concept diagnosis
Stephens-Martinez et al. [10]	2016	Semi-automated, co-occurrence	Predicts gaps with minimal labelling
Aya et al. [12]	—	DNN, RF, SVM, KNN	DNN best, 89% accuracy
Bravo-Agapito et al. [13]	2021	Regression, factor, cluster (Moodle)	Forum visits, quizzes, assignments, age most influential
Pandey & Pal [14]	2011	Bayes classification	Predicts whether new students perform well

III. METHODOLOGY

The proposed framework processes answer sheets through cleaning, preprocessing, hybrid grading, rule-based gap detection and resource recommendation (Fig. 1).

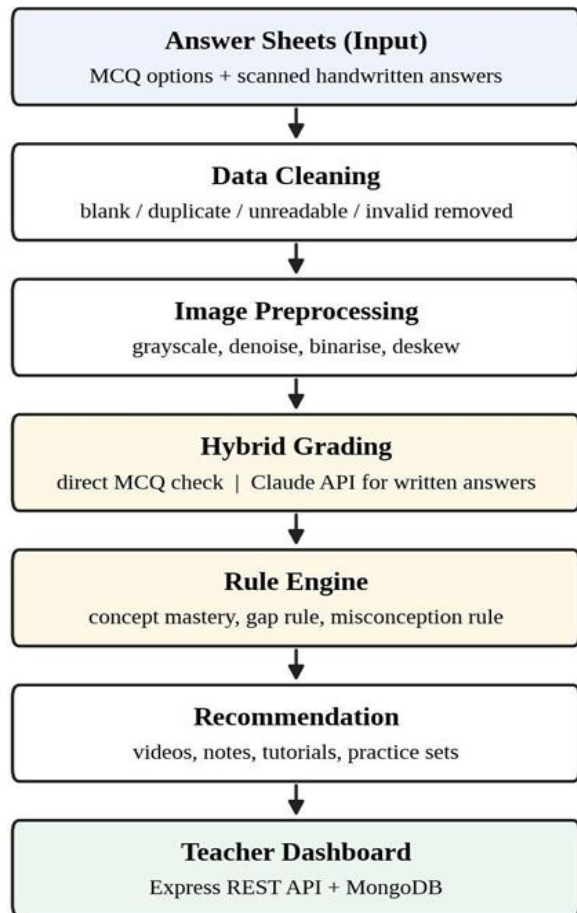


Fig. 1. Overview of the proposed LearnInsight pipeline for class-level learning gap detection

A. Dataset Description

The prototype uses a sample class-test dataset (Grade 8, Section

B) of student responses mapped to concepts, including Fractions, Photosynthesis and Newton's Second Law. Scanned handwritten answer sheets are planned as the next input type.

Records are stored in MongoDB as four collections: Student, Concept, Question and Submission.

TABLE II. DATASET DETAILS

Attribute	Value
Dataset	Class test responses, Grade 8, Section B
Data type	Multiple-choice answers; handwritten images (planned)
Sample size	6 students, 6 questions, 5 concepts
Storage	MongoDB (Student, Concept, Question, Submission)
Learning type	Rule-based analytics with AI grading of written answers

B. Data Cleaning

A Python-based pipeline reads raw answers and scans, checks quality and removes: (i) blank submissions (unanswered questions), (ii) duplicate submissions, (iii) unreadable scans that are too blurred, and (iv) invalid answers whose option is not in the question's option list. The clean dataset then proceeds to preprocessing.

C. Image Preprocessing

Selected answer-sheet images are standardised before text recognition and grading: (1) loading with OpenCV/PIL, (2) grayscale conversion to focus on ink and paper, (3) noise removal with median or Gaussian filters to reduce shadows and speckles, (4) binarisation into ink and background, and (5) deskewing of tilted pages. The output is a model-ready image for OCR and grading.

D. Data Augmentation

Only training images are augmented, so validation and test images remain untouched. Planned transformations are small rotations (about $\pm 5^\circ$), brightness changes, light Gaussian noise and blur, zooming ($0.9\times$ – $1.1\times$), and text variants such as spelling slips and paraphrased answers. Augmentation increases diversity, reduces overfitting and reduces dependence on one writing style or scan quality.

E. Data Splitting

The cleaned data is split by *student*, so no student's answers appear in two parts (split ratios are fixed at implementation time). Training data is used to design grading rubrics and misconception labels; validation data to tune the thresholds (50% gap, 40% misconception); and test data only for final evaluation.

F. Hybrid Analysis Model

Multiple-choice answers are checked directly against the correct option. Free-text and OCR-read answers are sent to the Claude API, which returns a grading result containing *isCorrect*, the named *misconception* and a *confidence* value.

A transparent rule engine then computes, for concept *c*, the mastery of student *s* and of the class:
 $M = \text{correct} / \text{answered}$, $M = \text{correct} / \text{answered}$ *s,c*
 s,c *s,c c c c* A concept is a class-wide gap when $M < 0.50$. For a wrong *c* option *o* of a question under that concept, let *p* be the share of the *o* class selecting it; if $p \geq 0.40$, it is flagged as a shared *o* misconception, since many students making the same error reveals a common misunderstanding rather than a careless mistake.

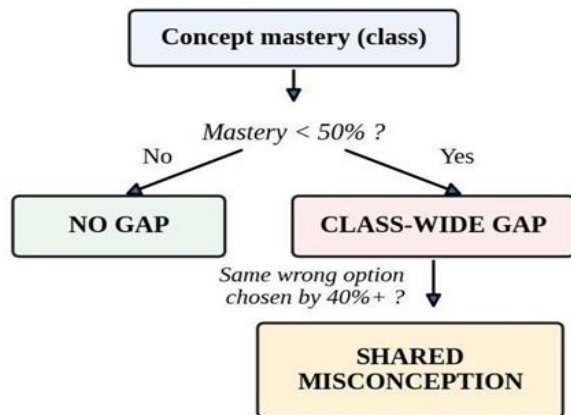


Fig. 2. Decision logic for the three prediction status levels

Concepts with neither condition are labelled no gap (Fig. 2). Flagged concepts are matched with videos, notes, tutorials and practice sets, and results are served to the teacher dashboard through an Express REST API.

G. Model Training and Refinement

Because grading is rubric- and prompt-based, “training” means preparing the system from labelled training answers: (1) answers are labelled correct or wrong, (2) a grading rubric gives the expected answer per question, (3) common wrong ideas are named as misconception labels, (4) outputs are compared with teacher marks, and (5) rubrics and prompts are refined.

H. Test-Time Augmentation

For handwritten test images, several modified versions (rotation, zoom, brightness) are created, OCR is run on each, and the readings are combined by majority vote to give a more stable final text that is sent for grading.

IV. EXPERIMENTAL SETUP

A. Software Configuration

- Backend: Express REST API with MongoDB for Student, Concept, Question and Submission data.
- Grading: Claude API for free-text/OCR answers; direct comparison for MCQs.
- Data cleaning and vision: Python pipeline; OpenCV/PIL for image preprocessing and OCR input.
- Frontend: browser-based teacher dashboard.

B. Rule Parameters

TABLE III. GAP AND MISCONCEPTION RULE PARAMETERS

Parameter	Value	Meaning
Mastery threshold	< 50%	Concept flagged as class-wide gap
Misconception threshold	$\geq 40\%$	Same wrong option chosen by this share of class
Rotation (augment.)	$\pm 5^\circ$	Small tilts for training images
Zoom (augment.)	$0.9\times$ – $1.1\times$	Scale change
Split unit	Student	No student in two parts

C. Evaluation Metrics

On unseen answers the system will be evaluated with accuracy (grading agreement with teacher marks), precision (reliability of flagged gaps), recall (share of real gaps detected), F1-score (balance of precision and recall), and a confusion matrix over the three status levels, compared against the teacher's own analysis.

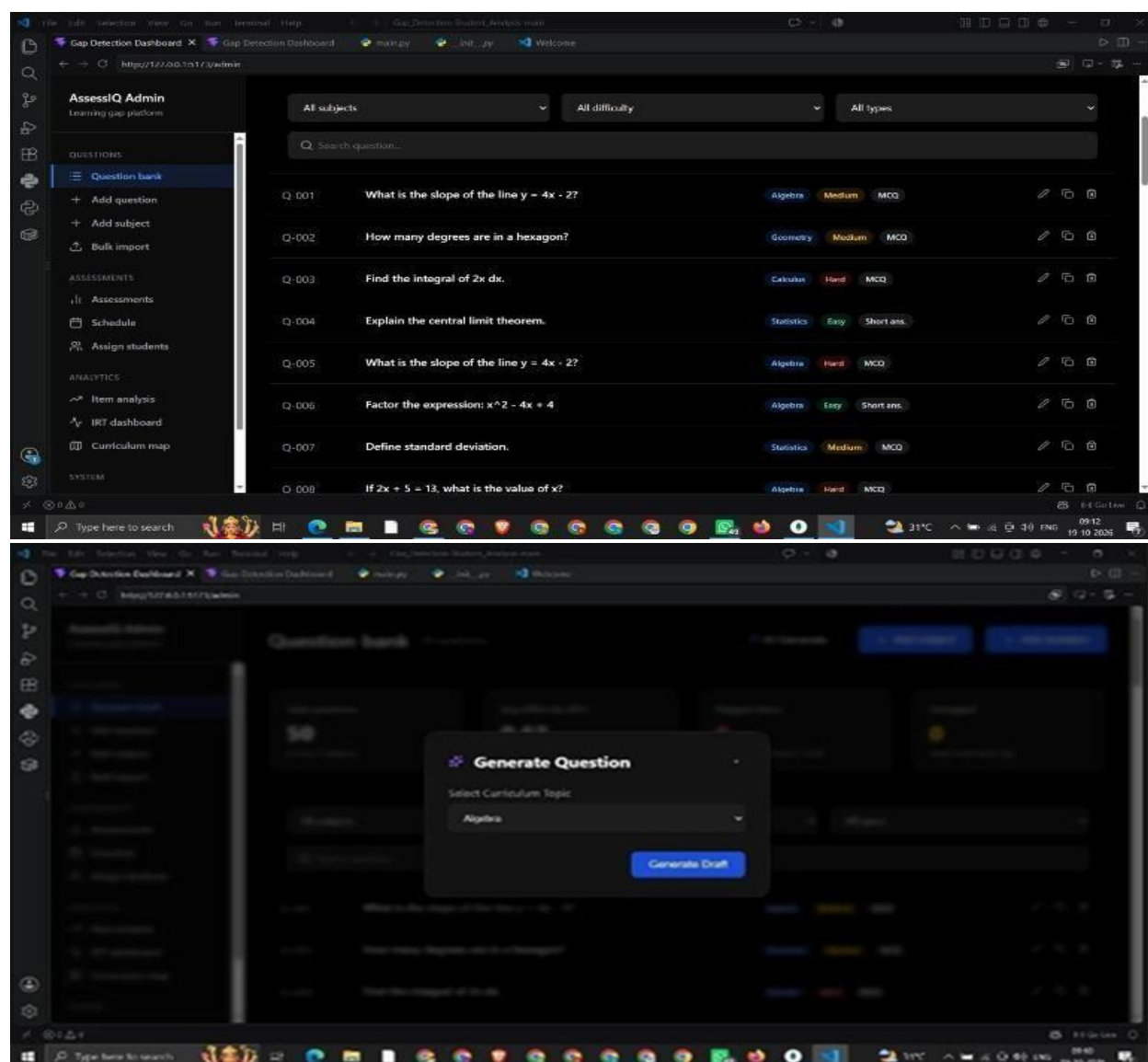
V. RESULTS

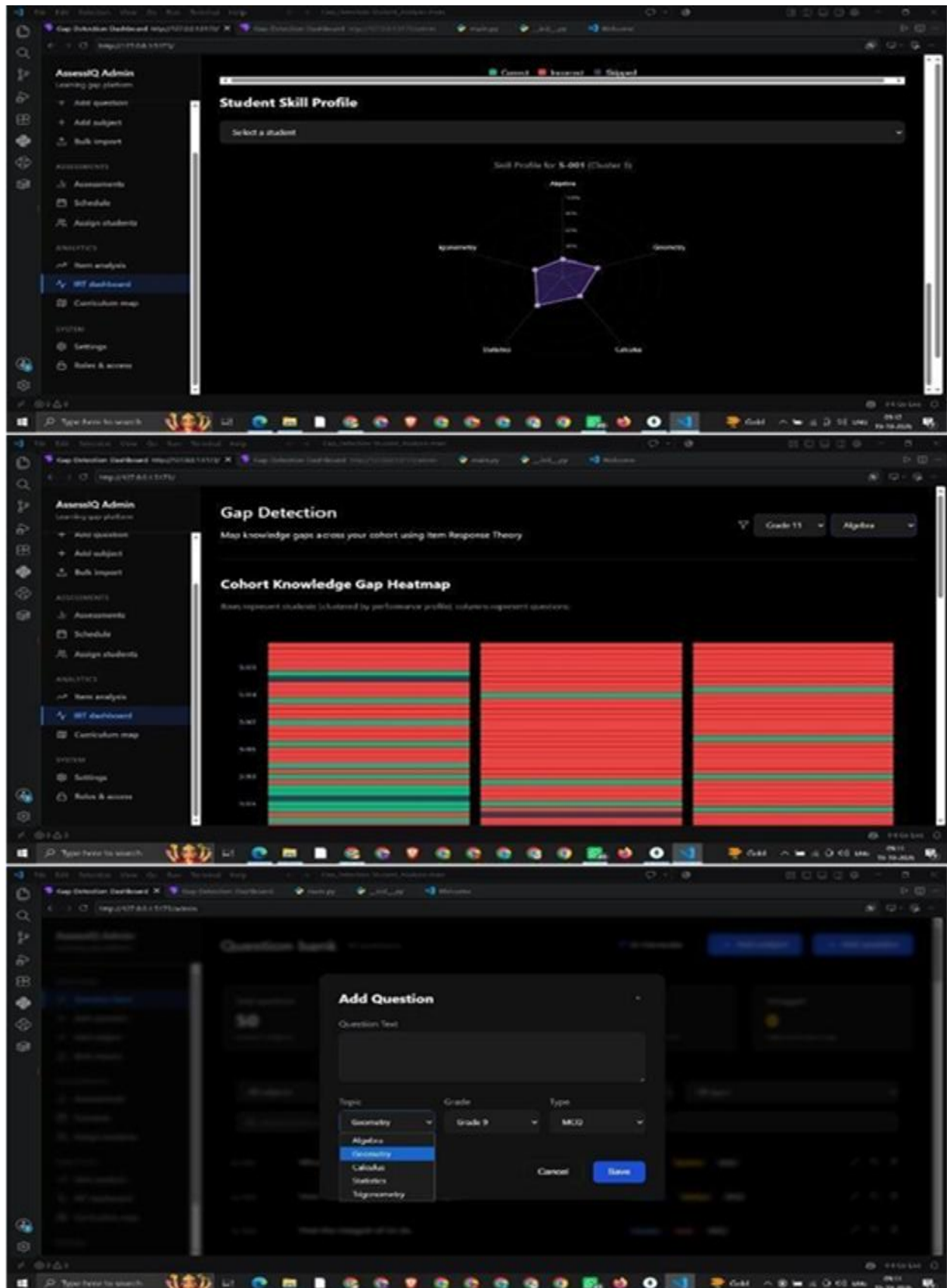
A. Prototype Output

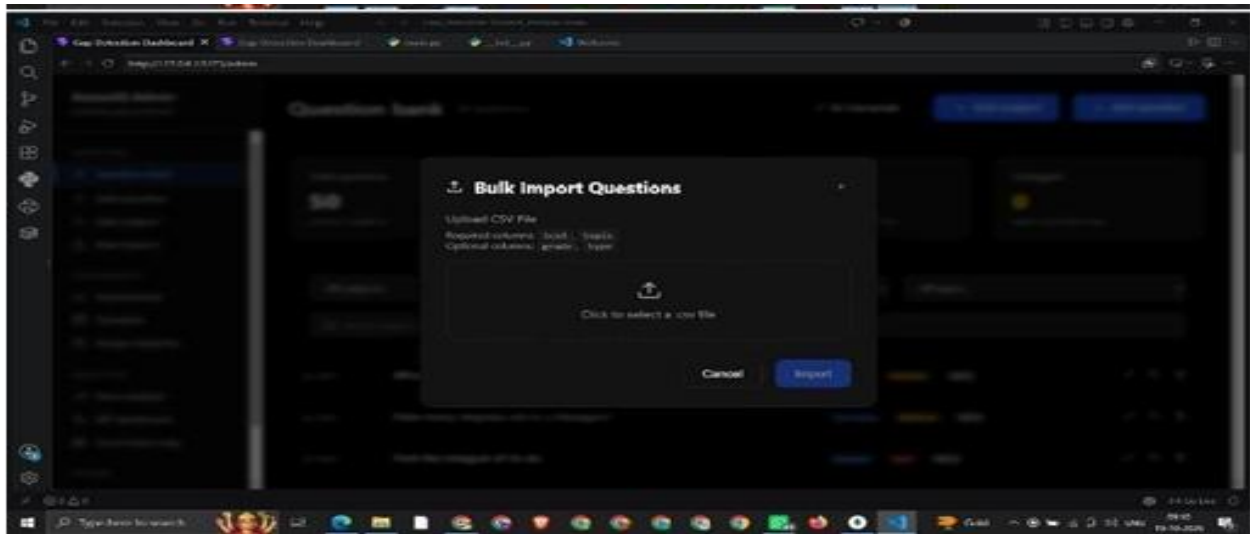
The prototype dashboard presents four views: class-wide concept mastery (concepts under 50% are flagged), detected misconceptions, recommended resources, and a student-level breakdown listing each student's weak concepts and overall mastery. For the sample class, the concept Photosynthesis: Dark Reaction was answered correctly by one-third of the class and 67% selected the same wrong answer, so it was flagged as both a classwide gap and a shared misconception and matched to a tutorial (Table IV).

TABLE IV. EXAMPLE PREDICTION FOR THE SAMPLE CLASS

Concept	Class mastery	Status	Common wrong answer	Resource
Photosynthesis: Dark Reaction	0.33	Class-wide gap	Thylakoid membrane (67%)	Calvin Cycle tutorial







B. Expected Performance

The initial target is 90%+ accuracy on unseen student answers, to be confirmed on test data. No measured accuracy, precision, recall or F1 values are reported here, because the current dataset (6 students, 6 questions) is a prototype sample and is too small for statistically meaningful evaluation. The pipeline is expected to support good performance through proper data cleaning, standardised preprocessing, suitable augmentation, Claude-based grading with transparent rules, careful rubric validation and testtime augmentation.

C. Discussion

Three design choices distinguish the approach from earlier analytics. First, it works at class level and so exposes shared misconceptions that individual-learner tools such as knowledge tracing do not show. Second, thresholds are explicit and tunable on validation data, so a teacher can see exactly why a concept was flagged. Third, each flagged gap is tied to a concrete resource, closing the loop from detection to action. Unlike purely predictive models [12]–[14], the system explains *what* is misunderstood and not only *who* is struggling.

VI. CONCLUSION AND FUTURE WORK

This paper proposed LearnInsight, an AI-based approach for finding learning gaps from student responses. Answers are cleaned, preprocessed and graded by a hybrid model combining Claude API grading with transparent gap rules, and each concept is

classified as No Gap, Class-wide Gap or Shared Misconception. The system is a teaching-support tool that helps teachers act on gaps rather than a replacement for teacher judgement; by reducing blind spots in large classes, it can help identify students needing special attention, reduce failure rates and prepare for the next examinations.

REFERENCES

- [1] Davis, G. Chen, C. Hauff, and G.-J. Houben, “Activating learning at scale: A review of innovations in online learning strategies,” *Computers & Education*, vol. 125, pp. 327–344, 2018.
- [2] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, 2010.
- [3] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” in *Learning Analytics*. Springer, 2014, pp. 61–75.
- [4] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining: Towards communication and collaboration,” in *Proc. 2nd Int. Conf. Learning Analytics and Knowledge*, 2012, pp. 252–254.
- [5] L. P. Macfadyen and S. Dawson, “Mining LMS data to develop an ‘early warning system’ for educators,” *Computers & Education*, vol. 54, no. 2, pp. 588–599, 2010.
- [6] S. Kotsiantis, C. Pierrakeas, and P. Pintelas,

- “Predicting students' performance in distance learning using machine learning techniques,” *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.
- [7] C. Romero, S. Ventura, and E. García, “Data mining in course management systems: Moodle case study and tutorial,” *Computers & Education*, vol. 51, no. 1, pp. 368–384, 2008.
- [8] S. M. Jayaprakash *et al.*, “Early alert of academically at-risk students: An open-source analytics initiative,” *Journal of Learning Analytics*, vol. 1, no. 1, pp. 6–47, 2014.
- [9] R. Ferguson, “Learning analytics: Drivers, developments and challenges,” *International Journal of Technology Enhanced Learning*, vol. 4, nos. 5–6, pp. 304–317, 2012.
- [10] K. Stephens-Martinez, A. Ju, C. Schoen, J. DeNero, and A. Fox, “Identifying student misunderstandings using constructed responses,” in *Proc. Third ACM Conf. Learning at Scale*, 2016, pp. 153–156.
- [11] P. Denny *et al.*, “LLM-generated educational resources compared with student-created resources,” 2023; S. Chan *et al.*, 2025; Y. Choi *et al.*, 2024.
- [12] A. Aya *et al.*, “Prediction of student success in programming and data structures courses using DNN, RF, SVM and KNN.”
- [13] J. Bravo-Agapito, S. J. Romero, and S. Pamplona, “Early prediction of undergraduate student's academic performance in completely online learning: A five-year study,” *Computers in Human Behavior*, vol. 115, Art. no. 106595, 2021.
- [14] U. K. Pandey and S. Pal, “A data mining view on classroom teaching language,” *International Journal of Computer Science Issues*, vol. 8, no. 2, pp. 277–282, 2011.
- [15] K. Neumann, J. Kuhn, and H. Drachsler, “Generative Künstliche Intelligenz in Unterricht und Unterrichtsforschung – Chancen und Herausforderungen,” *Unterrichtswissenschaft*, vol. 52, no. 2, pp. 227–237, 2024.