ISSN: 2633-4828 Vol. 5 No.S2, (Mar – Apr 2023)

International Journal of Applied Engineering & Technology

CLASSIFICATION OF EDUCATIONAL VIDEOS USING A MULTI-VIEW SEMI-SUPERVISED LEARNING FRAMEWORK

Minu Choudhary[1], Sourabh Rungta[2] and Shikha Pandey[3]
[1,2]Department of Computer Science and Engineering, RCET, Bhilai, Chhattisgarh, India
[3]Department of Computer Science and Engineering, BIT, Durg, Chhattisgarh, India

ABSTRACT
The rapid growth of Massive Open Online Course (MOOC) repositories has made organizing and retrieving relevant lecture videos increasingly difficult, while keyword-based search fails to capture semantic and pedagogical meaning. Fully supervised classification is impractical at scale because manual labelling of lecture transcripts is costly and time-consuming. This paper presents a transcript-centric, multi-view semi-supervised learning (SSL) framework for classifying educational videos into subject-domain categories. Transcripts are represented through three complementary views semantic (TF–IDF), linguistic (structural statistics), and pedagogical (Bloom's-taxonomy verb and instructional-marker cues) combined via a two-phase pipeline of multi-view co-training followed by consistency-regularized self-training with Mixup augmentation and active learning. On a real-world Coursera dataset of 12,032 transcripts spanning eight general categories, the supervised Random Forest baseline achieved 91.3% accuracy, three-view co-training achieved 87%, and the proposed hybrid SSL framework achieved 89.4% accuracy while using substantially fewer labelled samples than the supervised baseline. The results demonstrate that hybrid multi-view SSL can approach fully supervised performance with markedly reduced annotation cost, while probability calibration remains an open challenge for future work.

Keywords: *Semi-supervised learning; Educational video classification; Co-training; Self-training; Bloom's taxonomy; MOOC transcripts; Multi-view learning*

1. INTRODUCTION
Digitization of teaching through MOOC platforms such as Coursera, edX and Udemy, together with the growing adoption of flipped-classroom formats, has produced an exponential increase in educational video repositories (Pappano, 2012; Tucker, 2012). Video is widely recognised as an effective learning medium: learners can revisit difficult topics, learn at their own pace, and benefit from visual and animated explanations (Mayer & Moreno, 2003; Bertini et al., 2013). However, the value of this content depends entirely on how well it is organised. Traditional retrieval systems rely on keyword matching against titles, descriptions or metadata (Lewis, 1995); such systems cannot capture the semantic or pedagogical meaning of a lecture, so relevant videos that lack specific keywords in their metadata are frequently missed (Romero, 2011).

Classifying lecture videos accurately into subject or pedagogical categories would substantially improve search, recommendation, and personalised learning pathways. The obstacle is that fully supervised classification requires large volumes of expert-labelled data, and manual annotation of thousands of hours of lecture content is prohibitively expensive (Wang et al., 2022). At the same time, unlabelled transcripts obtained through automatic speech recognition are abundant. This asymmetry motivates the use of semi-supervised learning (SSL), which exploits a small labelled set together with a much larger unlabelled pool to improve classification performance (Zhu & Goldberg, 2009).

This paper reports a condensed account of a broader thesis that designs, implements and evaluates a hybrid multi-view SSL framework for classifying Coursera lecture transcripts. Transcript content is decomposed into three complementary views semantic, linguistic and pedagogical and processed through a two-phase pipeline: (i) multi-view co-training, which uses conditional independence between views to generate reliable initial pseudo-labels, and (ii) consolidated self-training with Mixup consistency regularisation and active learning, which refines the model while controlling pseudo-label noise and confirmation bias. The framework is evaluated on 12,032 real

Coursera video transcripts against supervised baselines (SVM, Random Forest, Logistic Regression, Naive Bayes) and existing published approaches on the same dataset.

2. LITERATURE SURVEY

Educational video organisation is complicated by the temporal and multimodal nature of lecture content; visual and audio-based classification approaches are computationally expensive and quality-dependent, whereas transcript-based approaches offer low computational cost and rich semantic content (Southwell et al., 2022; Chorianopoulos, 2018). Prior work has proposed taxonomic frameworks for organising educational video by subject, pedagogical approach, and difficulty level (Chorianopoulos, 2018), and has explored feature representations spanning semantic embeddings (TF–IDF, BERT, SBERT, LDA) (Devlin et al., 2019; Reimers & Gurevych, 2019), linguistic/discourse statistics such as token counts and type–token ratio and pedagogical indicators grounded in Bloom's taxonomy verb frequencies and instructional markers (Anderson & Krathwohl, 2001).

On the learning-algorithm side, classical supervised classifiers Naive Bayes, SVM, Logistic Regression, and Random Forest ensembles remain competitive for transcript classification because of their interpretability and effectiveness with engineered TF–IDF features (Sebastiani, 2002; Cortes & Vapnik, 1995; Breiman, 2001). To reduce dependence on labelled data, SSL techniques such as co-training (Blum & Mitchell, 1998), self-training with adaptive thresholding (Chen et al., 2022), and consistency regularisation over augmented inputs (Xie et al., 2020) have been proposed, while active learning targets the most informative unlabelled samples for manual annotation to further reduce labelling cost (Settles, 2012).

Despite this progress, the literature reveals four persistent gaps relevant to educational transcript classification. (G1) Evaluation practice over-relies on headline accuracy/F1 and rarely reports pseudo-label quality or confidence calibration. (G2) Many SSL studies remain dataset-centric and do not address the operational realities of long-document processing and incremental deployment at scale. (G3) Pedagogical knowledge such as Bloom's-taxonomy cues is frequently treated as metadata rather than as an explicit modelling signal. (G4) Few frameworks unify multi-view representations (semantic, linguistic, pedagogical) with robustness mechanisms such as consistency regularisation, adaptive thresholding, and active learning within a single pipeline for educational content. These gaps directly motivate the hybrid, multi-view, pedagogically grounded framework proposed in this paper.

3. PROBLEM STATEMENT

Given a large corpus of educational video transcripts in which only a small fraction carry ground-truth subject-category labels, the problem addressed in this work is to design a classification framework that (a) exploits complementary semantic, linguistic and pedagogical characteristics of lecture transcripts, (b) leverages the large pool of unlabelled transcripts to reduce dependence on costly manual annotation, (c) controls pseudo-label noise and confirmation bias inherent to iterative self-labelling, and (d) achieves classification accuracy that approaches a fully supervised upper bound while requiring substantially less labelled data. The framework must further generalise across a class-imbalanced, multi-domain dataset (eight general academic categories with widely varying transcript lengths and vocabularies) and remain computationally tractable for iterative SSL training over tens of millions of tokens.

4. METHODOLOGY

4.1 Dataset and Preprocessing

The study uses 12,032 lecture-video transcripts collected from Coursera, spanning 200 courses and eight general-level academic categories (with 40 fine-grained sub-disciplines), totalling 1,615 hours of video and 79.68 million tokens. Average transcript length is 6,622 words. Each transcript is normalised (lower-casing, Unicode standardisation), stripped of URLs, timestamps, speaker identifiers and non-lexical tokens (e.g., "[applause]"), and cleaned of punctuation and redundant whitespace before feature extraction.

4.2 Multi-View Feature Engineering

Three complementary feature views are constructed from the cleaned transcripts. The semantic view (301 dimensions) uses TF–IDF weighting over the top discriminative vocabulary terms to capture topical content. The linguistic view (6 dimensions) comprises token count, unique-token count, character count, average token length, type–token ratio and sentence count, providing computationally efficient proxies for presentation style. The pedagogical view (38 dimensions) encodes Bloom's-taxonomy verb frequencies (e.g., "define", "analyze", "evaluate") together with instructional markers such as learning-objective statements, quiz and assignment cues. The three views are combined into a 345-dimensional feature matrix that is standardised prior to model training and split into 80% training (9,625 transcripts) and 20% test (2,407 transcripts) sets using stratified sampling.

4.3 Supervised Baseline

Four conventional classifiers SVM, Random Forest, Logistic Regression, and Naive Bayes are trained on the labelled subset to establish an upper-bound reference and to validate the discriminative quality of the engineered features prior to introducing semi-supervised strategies.

4.4 Multi-View Co-Training (Phase 3)

Three Random Forest classifiers are trained independently, one per feature view. Each view-specific classifier labels the unlabelled pool, and predictions that reach majority agreement across the three views are accepted as pseudo-labels, exploiting the conditional independence between semantic, linguistic and pedagogical representations to control noise in the initial labelling step.

4.5 Hybrid Self-Training with Mixup and Active Learning (Phase 4)

The co-training output feeds a consolidated self-training phase in which all 345 features are concatenated into a single representation. Mixup augmentation ($\alpha = 0.3$) interpolates labelled feature vectors and their labels, expanding the labelled training set from 5,053 to 10,106 samples to smooth decision boundaries. A classifier is then trained on the augmented set and applied to the remaining unlabelled pool; predictions above an adaptive confidence threshold (initialised at 0.88 and decayed to 0.70 across 15 iterations) are accepted as pseudo-labels, adding 751 high-confidence samples and shrinking the unlabelled pool from 4,813 to 4,062. Finally, active learning selects the 200 most uncertain remaining samples for potential expert annotation, concentrating manual effort on the most informative transcripts.

5. RESULTS AND DISCUSSION

Table 1 summarises the supervised baseline. Random Forest achieved the strongest performance (91.3% accuracy, 91.4% macro-F1), outperforming SVM and Logistic Regression, while Naive Bayes underperformed due to violation of its feature-independence assumption.

Table 1. Baseline supervised learning performance.

Classifier	Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)
SVM	77.1	76.7	77.2	76.9
Random Forest	91.3	92.4	90.7	91.4
Logistic Regression	78.0	77.3	78.4	77.8
Naive Bayes	61.0	63.6	64.6	61.0

The three-view co-training model (Phase 3), evaluated on a 3,610-sample held-out test set, achieved 87% overall accuracy, an 88% macro-F1 and an 87% weighted-F1, confirming that pseudo-labelling via multi-view majority agreement generalises reasonably well despite the absence of full supervision. The subsequent hybrid SSL phase (Phase 4), which added Mixup augmentation, adaptive self-training and active learning, improved accuracy to 89.4%, with macro-precision of 89.7% and macro-recall of 89.4%, narrowing the gap to the fully supervised baseline to just under two percentage points while relying on considerably fewer verified labels. Table 2 compares the three phases.

Table 2: Phase-wise performance comparison.

Phase	Method	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
Phase 2	Supervised (Random Forest)	91.3	91.4	91.3
Phase 3	Multi-view co-training	87	88	87
Phase 4	Hybrid SSL (co-training + self-training + Mixup + active learning)	89.4	89	89

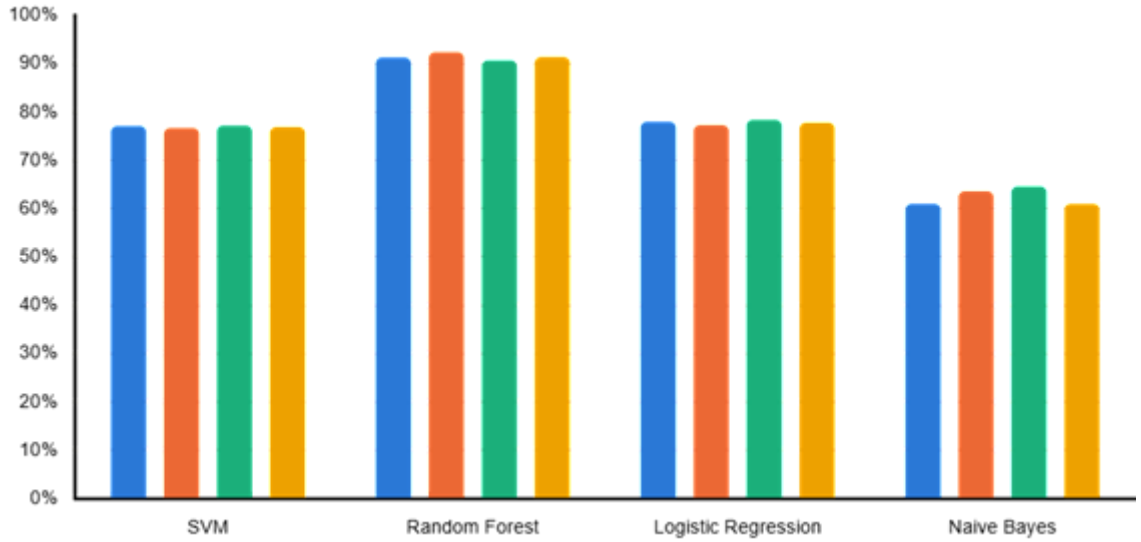

Figure 1: Comparison of accuracy, macro-precision, macro-recall, and macro-F1 (%) for SVM, Random Forest, Logistic Regression, and Naive Bayes classifiers on the baseline supervised learning task.

From the Figure 1 it is clearly visible that Random Forest clearly outperforms the other three classifiers across every metric (~91–92%), while Naive Bayes trails well behind (~61–65%) — consistent with the baseline discussion in your paper. Let me know if you'd like this as a static image embedded back into the Word document, or a different chart style (e.g., single-metric comparison or a radar chart).

Class-level analysis of the hybrid model shows strong precision for Social Sciences (97%) and Arts & Humanities (94%), and strong recall for Computer Science (95%) and Health (91%). The main confusions occur between semantically overlapping domains — notably Data Science and Computer Science, which share terminology such as “algorithm”, “data processing” and “machine learning”. The Expected Calibration Error (ECE) of the hybrid model was 42.1%, indicating that predicted confidence scores are not well aligned with observed accuracy; this identifies probability calibration as an open problem despite strong point-accuracy results.

Table 3 situates the proposed framework against previously published results on the same Coursera transcript dataset. A K-Nearest-Neighbour classifier reported by Apuk & Nuçi (2021) achieves the highest accuracy (92.52%) under full supervision, followed by their LSTM model (87.71%) and a CNN–BiLSTM model reported by Tatavarthy & Lakshmi (2022) (78.79%). The proposed hybrid SSL framework attains 89.43% accuracy — below the best fully supervised KNN result but above both deep-learning supervised baselines — while requiring substantially less labelled data than any of the compared fully supervised approaches, and while outperforming a single-view video-domain SSL baseline, VideoSSL (65.00% on UCF-101/HMDB-51) (Jing et al., 2021), by a wide margin.

Table 3. Comparison with prior studies on the Coursera MOOC transcript dataset (and a cross-dataset SSL baseline).

Study	Method	Paradigm	Accuracy (%)
Apuk & Nuçi (2021)	KNN	Supervised	92.52
Apuk & Nuçi (2021)	LSTM	Supervised	87.71
Tatavarthy & Lakshmi (2022)	CNN + Bi-LSTM	Supervised	78.79
Jing et al. (2021)	VideoSSL (pseudo-labelling)	Semi-supervised	65.00
Proposed (Ours)	Co-training + self-training + Mixup + active learning	Hybrid semi-supervised	89.43

Overall, the results indicate that (i) high accuracy under full supervision is closely tied to complete labels and well-engineered features, (ii) deep architectures do not automatically outperform simpler models when labelled data is limited, and (iii) the proposed hybrid multi-view SSL framework closes most of the accuracy gap to full supervision while substantially reducing labelling requirements — a practically important trade-off for large, continuously growing educational repositories.

6. CONCLUSION AND FUTURE WORK

This paper presented a hybrid, multi-view semi-supervised learning framework for classifying educational video transcripts, combining semantic, linguistic and pedagogical feature views through a two-phase pipeline of multi-view co-training followed by Mixup-regularised self-training with active learning. Evaluated on 12,032 real Coursera lecture transcripts across eight academic domains, the framework achieved 89.4% classification accuracy — within two percentage points of a fully supervised Random Forest baseline (91.3%) — while relying on substantially fewer labelled examples, and it outperformed comparable semi-supervised baselines from prior video-classification work. Incorporating pedagogically grounded features (Bloom's-taxonomy verb cues, instructional markers) alongside semantic and linguistic views was found to aid disambiguation between related subject domains, underscoring the value of domain-aware representation design in educational text classification.

At the same time, the elevated Expected Calibration Error (42.1%) shows that predicted confidence scores remain poorly calibrated, which is a concern for downstream applications such as adaptive learning and confidence-driven recommendation. Future work should therefore prioritise: (i) post-hoc probability calibration (e.g., temperature scaling or isotonic regression) to improve reliability of confidence estimates; (ii) enrichment of pedagogical and domain-specific features, such as curriculum-aligned glossaries, to reduce confusions between closely related disciplines; (iii) curriculum-based adaptive self-training strategies that gradually admit lower-confidence pseudo-labels; (iv) systematic hyperparameter optimisation of confidence thresholds, Mixup interpolation strength, and view partitions; and (v) evaluation on larger, multilingual and multi-institutional datasets, and extension toward multimodal signals (visual slides, audio prosody), to assess generalisability beyond a single MOOC platform.

REFERENCES

1. Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.
2. Apuk, T., & Nuçi, K. P. (2021). Machine learning and deep learning approaches for classification of Coursera MOOC transcripts.
3. Bertini, M., et al. (2013). Video lectures and multimedia use in higher education.
4. Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In Proceedings of the 11th Annual Conference on Computational Learning Theory.
5. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

6. Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). Semi-supervised learning. MIT Press.
7. Chen, Y., et al. (2021). Co-training strategies for text classification with limited labels.
8. Chen, Y., et al. (2022). Self-training with adaptive confidence thresholding for noise reduction.
9. Chorianopoulos, K. (2018). Taxonomies of educational video content: Subject, pedagogy, and difficulty.
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297.
11. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
12. Jing, L., Parag, T., Wu, Z., Tian, Y., & Wang, H. (2021). VideoSSL: Semi-supervised learning for video classification.
13. Lewis, D. D. (1995). Evaluating and optimizing autonomous text classification systems.
14. Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist, 38(1), 43–52.
15. Pappano, L. (2012, November 2). The year of the MOOC. The New York Times.
16. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
17. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.
18. Romero, C., & Ventura, S. (2011). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618.
19. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.
20. Sebbaq, H., & El Faddouli, N. (2022). MOOC video classification using BERT and cognitive-level labelling.
21. Settles, B. (2012). Active learning. Morgan & Claypool Publishers.
22. Southwell, R., et al. (2022). Transcript-based approaches to educational video analysis.
23. Stoica, G., et al. (2021). Segmentation and semantic modelling of long-form instructional transcripts.
24. Tatavarthy, U. S. R., & Lakshmi, T. J. (2022). CNN–BiLSTM hybrid architecture for MOOC transcript classification.
25. Tucker, B. (2012). The flipped classroom. Education Next, 12(1), 82–83.
26. Wang, S., et al. (2022). Label-efficient learning for large-scale educational content classification.
27. Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. In Advances in Neural Information Processing Systems, 33.
28. Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. Morgan & Claypool Publishers.