

TACL: Threshold-Adaptive Curriculum Learning Strategy for Enhancing Medical Text Understanding

Mucheng Ren, Yucheng Yan, He Chen, Danqing Hu, Jun Xu, and Xian Zeng[†]

Jiangsu Key Laboratory of Intelligent Medical Image Computing, School of Artificial Intelligence

Nanjing University of Information Science and Technology, Nanjing, China

{renm, 202412491503, chenh, danqinghu, jxu, xianzeng}@nuist.edu.cn

Abstract—Electronic medical records (EMRs) are crucial for modern healthcare, containing rich information about patient care, diagnoses, and treatments. However, their unstructured nature, domain-specific language, and complexity pose significant challenges for automated understanding. Existing methods often treat all data equally, limiting their ability to handle rare or complex cases effectively. We present TACL (Threshold-Adaptive Curriculum Learning), a novel framework that dynamically adjusts the training process based on sample complexity. Inspired by progressive learning, TACL categorizes data into difficulty levels, focusing on simpler cases early in training and gradually addressing more complex ones. A domain-specific pre-trained language model is used for difficulty assessment, considering semantic, syntactic, and contextual features. Additionally, TACL employs an adaptive training strategy to enhance task-specific performance and ensure generalization across diverse datasets. Experimental results on multilingual datasets, including MIMIC-III, MIMIC-IV, and Chinese clinical records, demonstrate TACL’s effectiveness in tasks such as ICD coding, readmission prediction, and TCM syndrome differentiation. TACL improves performance on rare and complex cases, providing a scalable and robust solution for medical text understanding.

Index Terms—Curriculum learning, Medical Text Understanding, Patient Readmission, Out-of-hospital Mortality, ICD coding.

I. INTRODUCTION

Medical texts, such as electronic medical records (EMRs), are essential for modern healthcare, capturing patient care, diagnoses, and treatments [1]. However, their unstructured, domain-specific nature and variability—from straightforward cases to complex ones involving rare conditions—make automated understanding challenging.

As shown in Figure 1(a), EMRs exhibit multi-dimensional complexity, including **narrative structure**, **clinical variability**, and **patient demographics**. Traditional NLP models follow a “one-size-fits-all” approach, often failing on rare and complex cases [2]–[7]. This motivates a key question: **can models systematically learn from easier to harder cases, akin to human learning?**

Curriculum learning [8], a training paradigm that progresses from simpler to more complex samples, has shown promise in healthcare applications like medical image segmentation [9], [10] and ICD coding [5]. However, key challenges remain: 1)

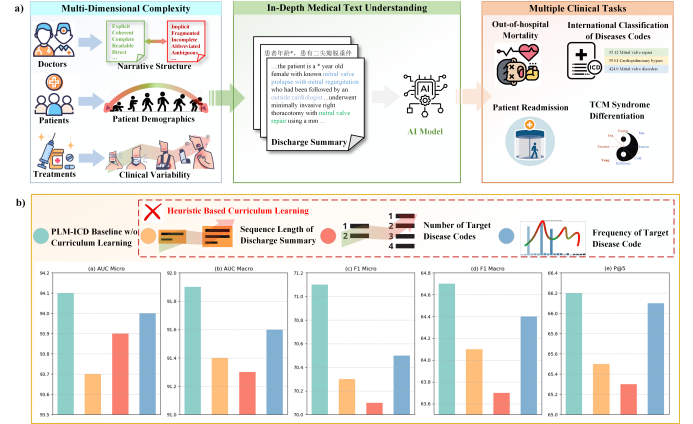


Fig. 1. a) Illustration of Multi-Dimensional Complexity of Medical Text and Various Clinical Tasks relying on deep medical text understanding. b) A comparison of the performance of PLM-ICD on MIMIC-III-50-ICD9 without curriculum learning and with heuristic-based curriculum learning.

How to effectively assess the difficulty of medical texts? and 2) **How to develop adaptive strategies to generalize across diverse tasks and datasets?** Simple metrics like sequence length or code frequency may fail to capture semantic nuances, as illustrated in Figure 1(b), necessitating more robust methods.

To address these challenges, we propose TACL (Threshold-Adaptive Curriculum Learning), a framework designed to enhance medical text understanding. TACL introduces a novel difficulty assessment method leveraging a domain-specific pre-trained language model to quantify text complexity based on semantic, syntactic, and contextual features. Furthermore, it dynamically adjusts the curriculum to optimize learning across tasks and datasets.

TACL is validated on multilingual datasets, including English (MIMIC-III [11], MIMIC-IV [12]) and Chinese clinical records, achieving state-of-the-art results in tasks like readmission prediction, out-hospital mortality, ICD coding, and TCM syndrome differentiation [13]. TACL excels in rare and complex cases, advancing medical text understanding.

Contributions

- **First application of curriculum learning in medical text understanding:** TACL introduces curriculum learn-

This work is supported by Provincial Natural Science Foundation of Jiangsu (NO. BK20250741, NO. BK20240703), and the Startup Foundation for Introducing Talent of NUIST (NO. 1523142401057, 1523142401055).

Corresponding author: xianzeng@nuist.edu.cn

ing to handle the complexity and variability of clinical narratives, offering a novel perspective for this domain.

- **Difficulty assessment with domain-specific models:** A pre-trained language model quantifies medical text complexity, enabling systematic difficulty categorization for curriculum learning.
- **Adaptive training strategy:** TACL dynamically adjusts the learning process, enhancing generalization across diverse medical datasets and improving robustness in real-world applications.

II. RELATED WORK

Advancements in NLP have addressed challenges in automated medical text understanding, particularly in **curriculum learning** and **medical text understanding**.

A. Curriculum Learning in NLP

Curriculum learning [8] structures training by starting with simpler examples and progressing to harder ones. Dynamic methods [14]–[16] adapt difficulty during training but focus on general NLP tasks, overlooking domain-specific challenges like medical texts. TACL bridges this gap with a threshold-adaptive mechanism tailored to medical data, dynamically prioritizing samples to enhance task performance and generalization.

B. Medical Text Understanding

Early rule-based approaches and traditional machine learning models required extensive feature engineering. Deep learning, including CNNs, RNNs, and transformer-based models like BERT [17], ClinicalBERT [18], and BioBERT [19], improved medical NLP by leveraging pretraining on biomedical corpora. However, these models treat all samples equally, limiting their effectiveness on rare or complex cases. TACL addresses this by integrating curriculum learning to progressively train on simpler to harder cases, improving performance and generalization.

Datasets like MIMIC-III [11] and MIMIC-IV [12] highlight the variability in medical data, demonstrating the need for adaptive frameworks like TACL to handle heterogeneous and complex distributions.

III. METHOD

This section introduces our proposed framework, illustrated in Figure 2, which consists of four key stages: (1) generating contextual representations using domain-specific pre-trained models, (2) defining data difficulty levels via clustering, (3) dynamically adjusting the curriculum with the Threshold-Adaptive Curriculum Learning (TACL) strategy, and (4) applying task-specific prediction heads for downstream tasks. These steps progressively guide the model through increasingly complex patterns within the data, ensuring robust learning across diverse medical NLP challenges.

A. Contextual Representation

To capture the semantic and terminological complexity of medical texts, we utilize domain-specific pre-trained models to generate high-dimensional contextual embeddings. Let the input dataset be denoted as:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \quad (1)$$

where x_i represents the input text, y_i is the corresponding task-specific label (e.g., ICD code, readmission risk, or TCM syndrome), and N is the total number of samples.

- **English Tasks:** For tasks such as ICD coding out-of-hospital mortality, and readmission prediction, we employ **RoBERTa** [20] pre-trained on medical corpora.
- **Chinese Tasks:** For TCM syndrome differentiation, we utilize **ZY-BERT** [13] pre-trained on traditional Chinese medicine datasets.

Thus, for each input text x_i , we extract the contextualized embedding of the **[CLS]** token, which serves as a summary representation of the entire input sequence. This embedding is obtained as follows:

$$h_i = \text{Encoder}(x_i)_{[\text{CLS}]}, \quad h_i \in \mathbb{R}^d \quad (2)$$

Here, h_i represents the output embedding corresponding to the **[CLS]** token, which is widely used as the aggregate representation of the input sequence in transformer-based models. These embeddings form the foundation for subsequent clustering and curriculum learning steps.

B. Data Clustering and Difficulty Definition

To implement curriculum learning, we define difficulty levels by clustering the contextual representations $\{h_1, h_2, \dots, h_N\}$. This process ensures that the curriculum is aligned with the intrinsic structure of the data.

a) *Clustering:* Given the embeddings $\{h_i\}_{i=1}^N$, we apply *k-means clustering* to partition the samples into k clusters, $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$. The objective of k-means clustering is to minimize the within-cluster variance (WCSS):

$$\arg \min_{\mu_1, \dots, \mu_k} \sum_{i=1}^k \sum_{h \in \mathcal{C}_i} \|h - \mu_i\|^2 \quad (3)$$

where $\mu_i \in \mathbb{R}^d$ is the centroid of cluster $\mathcal{C}_i$, defined as:

$$\mu_i = \frac{1}{|\mathcal{C}_i|} \sum_{h \in \mathcal{C}_i} h \quad (4)$$

b) *Difficulty Definition:* The difficulty of each cluster is determined by two factors:

- **Cluster Density:** A measure of how tightly packed the embeddings are within a cluster. For cluster $\mathcal{C}_i$, the density is defined as:

$$\text{Density}(\mathcal{C}_i) = \frac{1}{|\mathcal{C}_i|} \sum_{h \in \mathcal{C}_i} \|h - \mu_i\|^2 \quad (5)$$

Lower density values indicate higher cluster density, meaning the samples are more homogeneous and less ambiguous.

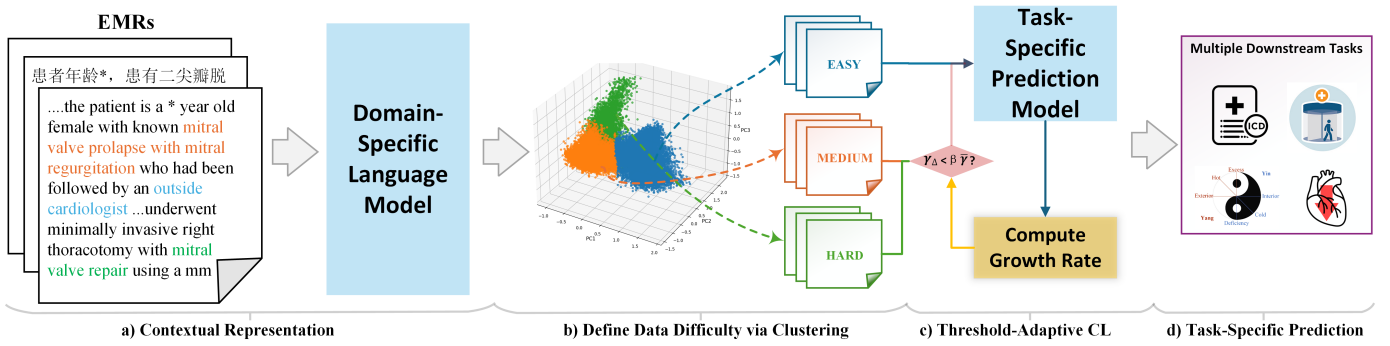


Fig. 2. Overview of our proposed TACL framework.

- **Distance from Centroid:** The average distance of samples in a cluster from the centroid. For cluster $\mathcal{C}_i$, it is defined as:

$$\text{Distance}(\mathcal{C}_i) = \frac{1}{|\mathcal{C}_i|} \sum_{h \in \mathcal{C}_i} \|h - \mu_i\| \quad (6)$$

Lower distance values indicate that the samples are closer to the centroid, representing a smaller and less diverse cluster.

Clusters are classified into three difficulty levels based on their density and distance values: *Easy* clusters have low density values (high density) and low distance values, representing homogeneous and less diverse patterns; *Medium* clusters have moderate density and distance values, reflecting patterns with some complexity and diversity but not overly challenging; and *Hard* clusters have high density values (low density) and/or high distance values, representing highly diverse and ambiguous patterns.

C. Threshold-Adaptive Curriculum Learning (TACL)

Traditional curriculum learning approaches often rely on static or heuristic criteria for transitioning between difficulty levels. In contrast, our TACL strategy dynamically adjusts the curriculum based on the model's performance growth.

a) *Growth Rate Metrics:* Let the macro-F1 score at epoch t be denoted as F_t . We define two growth rate metrics to monitor the model's progress:

- **Average Growth Rate:** Over a sliding window of the most recent N epochs, the average growth rate is computed as:

$$\bar{\gamma} = \frac{1}{N-1} \sum_{t=2}^N (F_t - F_{t-1}) \quad (7)$$

- **Instantaneous Growth Rate:** The most recent growth rate is denoted as γ_Δ and is computed as:

$$\gamma_\Delta = F_N - F_{N-1} \quad (8)$$

b) *Transition Condition:* A threshold coefficient $\beta \in (0, 1)$ is introduced to determine when the model has sufficiently learned from the current difficulty level. The model is considered *saturated* at the current stage if:

$$\gamma_\Delta < \beta \cdot \bar{\gamma} \quad (9)$$

TABLE I
DATASET STATISTICS FOR THE TASKS. THE TRAINING DATASET IS DIVIDED INTO THREE DIFFICULTY LEVELS: EASY, MEDIUM, AND HARD.

Task	Language	Datasets	Train				Val	Test
			Easy	Medium	Hard	Total		
Task 1: Readmission Prediction	English	Discharge Summary	12,256	8,472	5,517	26,245	3,027	3,063
Task 2: Out-of-hospital Mortality	English	Early Notes	22,858	13,196	11,739	47,793	5,774	5,441
		MIMIC-IV	14,437	12,823	6,227	33,487	4,186	4,186
Task 3: ICD Coding	English	MIMIC-III-full	33,024	12,963	1,732	47,719	1,631	3,372
		MIMIC-III-S0	3,437	2,389	2,240	8,066	1,573	1,729
Task 4: TCM Syndrome Differentiation	Chinese	MIMIC-IV-ICD10	49,158	38,142	1,789	89,089	13,378	19,799
		TCM-SD	33,386	5,114	4,680	43,180	5,486	5,486

When saturation is detected, the training progresses to the next difficulty stage.

c) *Curriculum Update:* Let $\mathcal{D}_{\text{easy}}$, $\mathcal{D}_{\text{middle}}$, $\mathcal{D}_{\text{hard}}$ represent the subsets of samples at each difficulty level. At each stage, training data is incrementally expanded:

$$\mathcal{D}_{t+1} = \begin{cases} \mathcal{D}_{\text{easy}}, & \text{if } \gamma_\Delta \geq \beta \cdot \bar{\gamma} \\ \mathcal{D}_{\text{easy}} \cup \mathcal{D}_{\text{middle}}, & \text{if current stage = easy and } \gamma_\Delta < \beta \cdot \bar{\gamma} \\ \mathcal{D}_{\text{easy}} \cup \mathcal{D}_{\text{middle}} \cup \mathcal{D}_{\text{hard}}, & \text{if current stage = middle and } \gamma_\Delta < \beta \cdot \bar{\gamma} \end{cases} \quad (10)$$

At the final stage (*hard*), training continues until no improvement in macro-F1 is observed for P consecutive epochs, after which training is terminated early.

D. Downstream Task Prediction

After training with TACL, the model is fine-tuned on multiple downstream tasks using task-specific prediction heads, optimized with cross-entropy loss for classification tasks and binary cross-entropy loss for multi-label tasks.

IV. EXPERIMENT

A. Datasets

We evaluate TACL on four tasks using benchmark datasets partitioned into three difficulty levels (Easy, Medium, Hard). Table I summarizes dataset statistics.

Task 1: Patient Readmission: Predicts if a patient is readmitted within a specific period using MIMIC-III [11]. Two subsets are used:

- **Discharge Summary:** Detailed discharge notes.
- **Early Notes:** Includes patient notes from 3 days earlier.

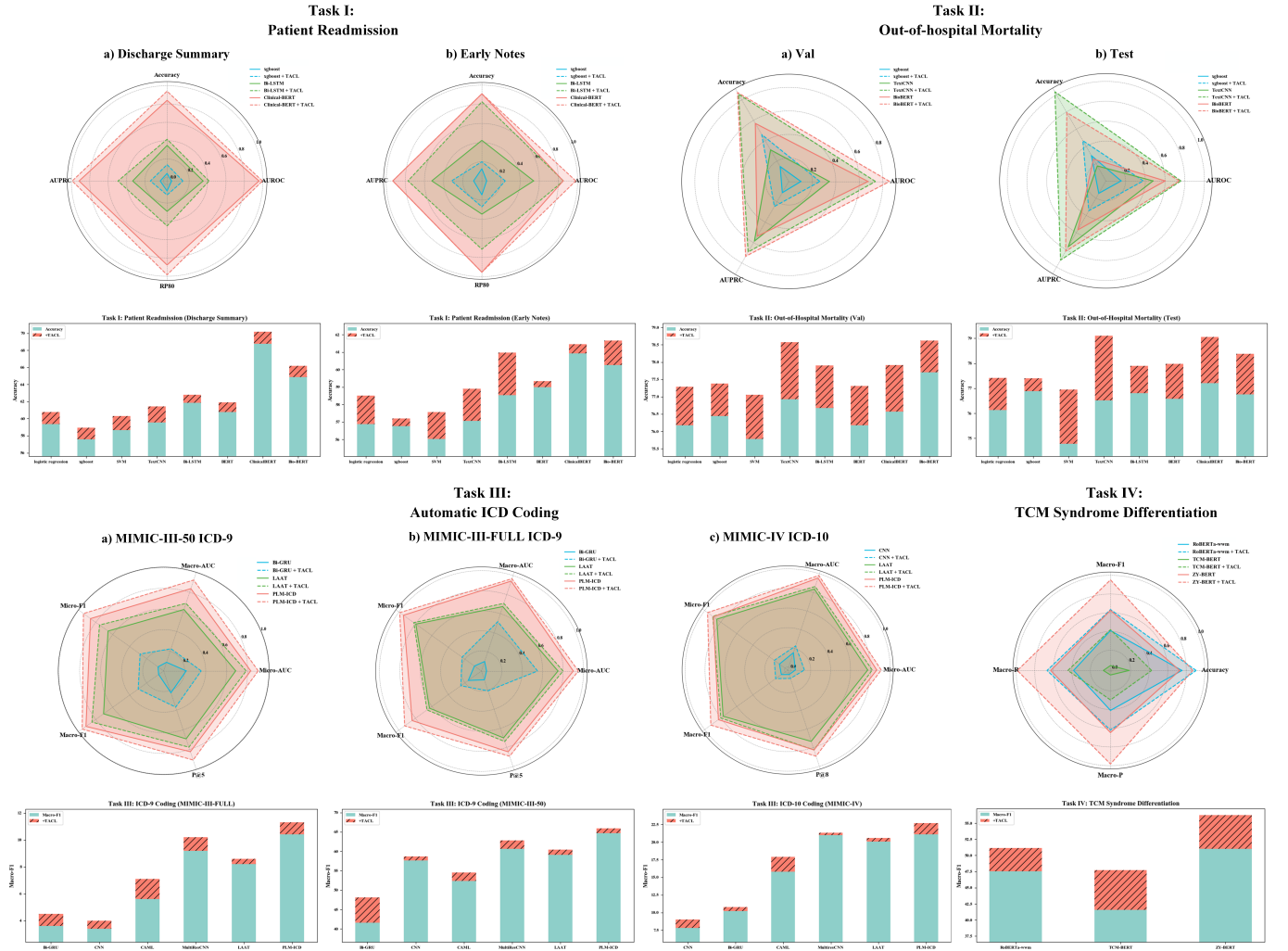


Fig. 3. Performance comparison across four tasks: Patient Readmission (Task I), Out-of-hospital Mortality (Task II), ICD Coding (Task III), and TCM Syndrome Differentiation (Task IV).

Task 2: Out-of-hospital Mortality: A binary classification task using MIMIC-IV [12] discharge summaries to predict 30-day mortality post-discharge, excluding in-hospital deaths.

Task 3: ICD Coding: Multi-label classification on:

- **MIMIC-III Full (ICD-9):** 52,723 documents, 8,929 codes (median: 14 codes/doc).
- **MIMIC-III 50 (ICD-9):** 11,368 documents, top 50 codes (median: 5 codes/doc).
- **MIMIC-IV (ICD-10):** Clinical notes annotated with granular ICD-10 codes.

Preprocessing includes lowercasing, removing non-alphanumeric characters, and truncation to 2,500–4,000 tokens.

Task 4: TCM Syndrome Differentiation: Classifies Traditional Chinese Medicine syndromes using TCM-SD [13], a Chinese-language dataset with domain-specific terminology.

B. Implementation

Experiments are conducted on an NVIDIA L40 48GB GPU with fixed random seeds. TACL thresholds range from 0.5 to 0.8, optimized on validation subsets. Metrics include F1-score, AUC, and precision at K (P@K), computed on test sets using the best model on validation. All code will be publicly available¹.

C. Baseline

We compare the proposed method against a range of strong baselines across all tasks. For tasks involving clinical and biomedical text, such as readmission prediction, out-of-hospital mortality, and ICD coding, we evaluate models like Bi-LSTM [21], which utilizes pre-trained Word2Vec embeddings and bidirectional LSTM layers, and transformer-based models like BioBERT [22] and ClinicalBERT [18], both pre-trained on large biomedical corpora. For ICD coding,

¹<https://anonymous.4open.science/r/medical-codes-354C>

TABLE II
PERFORMANCE ON PATIENT READMISSION PREDICTION.

Model	Strategy	Discharge Summary				Early Notes			
		AUROC	Accuracy	AUPRC	RP80	AUROC	Accuracy	AUPRC	RP80
Logistic Regression	Original	62.41	39.37	63.23	4.73	39.56	36.87	60.47	3.25
	+Length	62.48	59.54	63.17	2.99	59.62	56.80	60.34	3.77
	+TACL $\odot$	63.60	59.98	64.88	5.10	60.19	57.22	60.62	4.56
	+TACL $\ominus$	64.81	60.78	66.52	7.20	61.68	58.50	61.17	5.89
XGboost	Original	60.14	57.59	63.48	2.12	58.40	56.76	60.78	2.04
	+Length	60.33	57.70	63.51	3.18	59.00	56.93	61.18	1.79
	+TACL $\odot$	61.35	57.92	64.01	3.00	59.21	56.88	61.09	2.86
	+TACL $\ominus$	62.12	58.93	64.97	3.47	59.96	57.20	62.85	3.79
SVM	Original	63.26	58.67	63.21	6.59	58.49	56.03	61.17	3.16
	+Length	63.07	58.59	63.18	5.55	59.01	56.57	61.39	4.98
	+TACL $\odot$	63.77	59.09	64.02	6.88	58.97	57.02	62.71	5.50
	+TACL $\ominus$	64.29	60.29	65.94	7.14	60.41	57.57	63.24	5.49
TextCNN	Original	63.79	59.56	66.71	4.65	60.15	57.06	62.37	6.46
	+Length	63.65	59.49	66.53	3.10	60.11	56.99	62.16	5.56
	+TACL $\odot$	64.30	59.96	67.22	6.79	61.01	57.50	63.33	6.74
	+TACL $\ominus$	64.93	61.43	68.04	6.13	62.46	58.9	64.21	7.16
Bi-LSTM	Original	65.62	61.87	67.67	9.02	62.33	58.54	64.71	4.86
	+Length	66.49	61.87	68.07	12.26	63.39	59.31	65.91	7.99
	+TACL $\odot$	65.87	62.59	68.52	10.90	64.37	60.58	65.84	7.16
	+TACL $\ominus$	66.66	62.81	69.90	13.97	64.82	60.96	66.95	9.99
BERT	Original	66.37	60.76	68.89	8.46	62.67	58.99	64.55	3.52
	+Length	66.56	60.29	68.68	7.66	62.07	58.91	64.37	4.08
	+TACL $\odot$	66.72	61.03	69.11	8.94	62.95	59.02	65.05	6.46
	+TACL $\ominus$	67.65	61.90	70.01	11.61	63.40	59.33	66.58	10.43
Bio-BERT	Original	71.68	64.84	73.49	20.09	63.86	60.26	65.97	7.16
	+Length	72.01	65.18	73.67	20.58	63.85	60.19	65.79	6.33
	+TACL $\odot$	71.85	64.92	73.51	20.32	63.99	60.47	66.01	7.89
	+TACL $\ominus$	72.52	66.18	74.28	23.81	64.78	61.65	66.78	10.77
ClinicalBERT	Original	75.34	68.76	75.83	27.17	64.80	60.93	66.64	8.82
	+Length	73.73	68.46	74.81	23.34	64.04	59.77	66.24	10.67
	+TACL $\odot$	74.93	70.26	74.80	20.04	64.24	60.72	66.05	9.48
	+TACL $\ominus$	75.93	70.16	76.87	30.52	65.85	61.46	68.38	13.36

additional specialized models like CAML [23] with label attention mechanisms and MultiResCNN [24] for multi-scale feature extraction are employed, alongside PLM-ICD [25], which integrates a pre-trained BERT encoder with a label attention mechanism. For TCM syndrome differentiation, we evaluate Chinese-language models like RoBERTa-wwm [26], pre-trained with Whole Word Masking on large-scale Chinese text, and ZY-BERT [13], a TCM-specific model trained on a large unlabelled corpus, achieving state-of-the-art performance on TCM tasks.

V. MAIN RESULTS AND DISCUSSION

A. Main Results

Figure 3 summarizes TACL’s performance across all tasks, showing consistent improvements over traditional training methods in metrics such as Macro-F1, AUROC, and RP80.

TACL significantly enhances simpler models like Bi-GRU, CNN, and Bi-LSTM, particularly for readmission prediction and ICD coding, improving Macro-F1 scores and recall on rare labels. For advanced models like PLM-ICD, MultiResCNN, ClinicalBERT, and ZY-BERT, TACL provides smaller but consistent gains, especially for rare labels and imbalanced datasets, such as ICD-10 coding and TCM syndrome differentiation.

These results highlight TACL’s adaptability and scalability across models and tasks, demonstrating its potential to improve multilingual and domain-specific medical text understanding.

B. Analysis on Difficulty Definition

Defining data difficulty is critical for curriculum learning. Traditional statistical features (e.g., sequence length, code frequency) degrade performance by disrupting sample ordering (Table II). In contrast, contextual representations in TACL $\odot$ improve performance, with Bi-LSTM achieving +1.04 AUROC and +4.95 RP80 on discharge notes, highlighting the importance of task-aligned difficulty definitions.

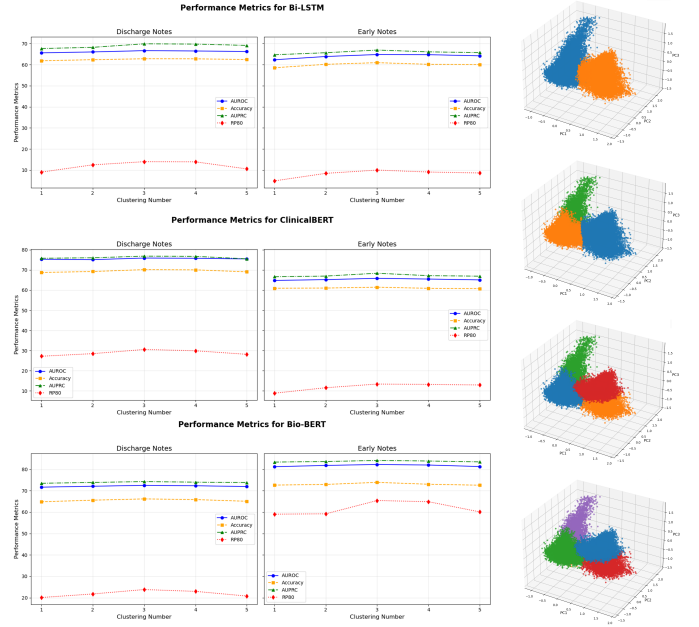


Fig. 4. Ablation study and visualization about clustering number for readmission prediction task.

C. Analysis on Data Order

Training from easy to hard ($\odot$) consistently outperforms hard-to-easy ($\ominus$) across all models and metrics (Table II). Starting with simpler samples helps models build a foundation, enabling better handling of complex cases, whereas reversed order disrupts learning progression.

D. Analysis on Clustering Number

Clustering number impacts performance and data interpretability. Metrics peak at 3 clusters (Figure 4), which balance data complexity and avoid over- or under-segmentation. Three clusters align well with inherent data structure, optimizing learning and performance.

E. Clinical Applicability of TACL

TACL enhances clinical applicability by improving prediction confidence and accuracy, particularly for rare labels and high-stakes tasks like mortality prediction. It adapts to diverse tasks and datasets, demonstrating consistent gains in AUROC, Macro-F1, and AUPRC across English and Chinese datasets. TACL’s scalability and ability to improve trust in AI predictions make it a valuable framework for advancing healthcare AI and improving patient outcomes.

VI. CONCLUSION

This paper introduces TACL, a novel framework designed to address the nuanced challenges of medical text understanding by dynamically adapting the training process to data complexity. Using contextual representation-based difficulty definitions and a curriculum learning strategy that progressively transitions from simpler to more complex samples, TACL achieves significant improvements in key medical NLP tasks, including

Discharge Summary:

admission date discharge date date of birth sex m service last name un history of present illness the patient is a year old man with a **history of hepatitis c virus** leading to **end stage liver disease post necrotic cirrhosis** with biopsy in that showed **hcc without evidence of metastatic disease** the patient has undergone courses of **radiofrequency ablation** for these lesions the patient has had no significant decompensation of fluid overload encephalopathy or gi bleeds during this time the patient has been taking lactulose off and on due to gassy pains he feels it was causing him the patient called in today by transplant staff to proceed with liver transplantation medications on admission

Decision Boundary: 0.4242

ICD-9 Code	Code Name	Is Target?	PLM-ICD	w/ TACL
070.54	Chronic hepatitis C without hepatic coma	✓	✓ (0.6741)	✓ (0.8095)
571.5	Cirrhosis of liver without mention of alcohol	✓	✗ (0.4026)	✓ (0.5353)
155.0	Malignant neoplasm of liver, primary	✓	✓ (0.5824)	✓ (0.7618)
50.59	Other partial hepatectomy	✓	✗ (0.3293)	✓ (0.4882)
571.4	Chronic hepatitis, not elsewhere classified	✗	✓ (0.4754)	✗ (0.2788)

Fig. 5. Case study of ICD-9 coding

ICD coding, readmission prediction, and differentiation of TCM syndrome. The results demonstrate TACL's ability to enhance model generalization, particularly in handling rare and complex cases, which are often the most clinically significant. Furthermore, the framework's adaptability to multilingual and domain-specific datasets, such as English clinical records and Chinese TCM texts, underscores its scalability and versatility. Beyond improving performance, TACL highlights the importance of aligning training strategies with task-specific nuances and data distributions, offering a transformative approach that bridges domain-specific challenges and scalable machine learning solutions.

REFERENCES

- [1] M. Giuffrè and D. L. Shung, "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy," *NPJ digital medicine*, vol. 6, no. 1, p. 186, 2023.
- [2] J. Edin, A. Junge, J. D. Havtorn, L. Borgholt, M. Maistro, T. Ruotsalo, and L. Maaløe, "Automated medical coding on mimic-iii and mimic-iv: a critical review and replicability study," in *Proc. 46th ACM SIGIR Conf.*, pp. 2572–2582, 2023.
- [3] G. Gomes, I. Coutinho, and B. Martins, "Accurate and well-calibrated icd code assignment through attention over diverse label embeddings," in *Proc. of the 18th EACL*, pp. 2302–2315, 2024.
- [4] C. Lu, C. Reddy, P. Wang, and Y. Ning, "Towards semi-structured automatic icd coding via tree-based contrastive learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68300–68315, 2023.
- [5] W. Ren, R. Zeng, T. Wu, T. Zhu, and R. G. Krishnan, "Hicu: Leveraging hierarchy for curriculum learning in automated icd coding," in *Machine Learning for Healthcare Conference*, pp. 198–223, PMLR, 2022.
- [6] T. Vu, D. Q. Nguyen, and A. Nguyen, "A label attention model for icd coding from clinical text," in *IJCAI*, pp. 3335–3341, 2021.
- [7] Z. Zhao, W. Lu, X. Peng, L. Xing, W. Zhang, and C. Zheng, "Automated icd coding via contrastive learning with back-reference and synonym knowledge for smart self-diagnosis applications," *IEEE Transactions on Consumer Electronics*, 2024.

- [8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.
- [9] J. Wang, Y. Tang, Y. Xiao, J. T. Zhou, Z. Fang, and F. Yang, "Grenet: gradually recurrent network with curriculum learning for 2-d medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 10018–10032, 2023.
- [10] A. Wang, M. Islam, M. Xu, and H. Ren, "Curriculum-based augmented fourier domain adaptation for robust medical image segmentation," *IEEE Transactions on Automation Science and Engineering*, 2023.
- [11] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [12] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-iv," *PhysioNet*. Available online at: <https://physionet.org/content/mimiciv/1.0/> (accessed August 23, 2021), pp. 49–55, 2020.
- [13] M. Ren, H. Huang, Y. Zhou, Q. Cao, Y. Bu, and Y. Gao, "Tcm-sd: a benchmark for probing syndrome differentiation via natural language processing," in *China National Conference on Chinese Computational Linguistics*, pp. 247–263, Springer, 2022.
- [14] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, and T. Mitchell, "Competence-based curriculum learning for neural machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 1162–1172, Association for Computational Linguistics, June 2019.
- [15] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5901–5910, 2020.
- [16] Y. Kong, L. Liu, J. Wang, and D. Tao, "Adaptive curriculum learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5067–5076, 2021.
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- [18] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.
- [19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [20] P. Lewis, M. Ott, J. Du, and V. Stoyanov, "Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop* (A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, eds.), (Online), pp. 146–157, Association for Computational Linguistics, Nov. 2020.
- [21] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [22] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.
- [23] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *NAACL* (M. Walker, H. Ji, and A. Stent, eds.), pp. 1101–1111, ACL, 2018.
- [24] F. Li and H. Yu, "Icd coding from clinical text using multi-filter residual convolutional neural network," in *AAAI*, pp. 8180–8187, 2020.
- [25] C.-W. Huang, S.-C. Tsai, and Y.-N. Chen, "PLM-ICD: Automatic ICD coding with pretrained language models," in *Proc. of the 4th Clin. NLP Workshop* (T. Naumann, S. Bethard, K. Roberts, and A. Rumshisky, eds.), (Seattle, WA), pp. 10–20, Association for Computational Linguistics, 2022.
- [26] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.