SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

A Novel Machine Learning Strategy For Accurate ICD Code Identification: Towards Reliable Automation Of Clinical Coding

Dr. Prashanth M V[1] , Nayani Sateesh[2] , Tanya Singh[3] , Rishabh Bhardwaj[4] , Bipin Sule[5] , Soumitra Das[6] , Mr. Jayant S. Rohankar[7]

[1]Professor prashanthmv@vvce.ac.in 0000-0001-9763-1666 Vidyavardhaka College of Engineering, Mysuru"
[2]Senior Assistant Professor Computer Science and Engineering CVR College of Engineering, India n.sateesh@cvr.ac.in"
[3]School of Engineering & Technology, Noida International University, Greater Noida, Uttar Pradesh 203201, India. dean.academics@niu.edu.in
[4]Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India. rishabh.bhardwaj.orp@chitkara.edu.in https://orcid.org/0009-0009-6075-8837"""
[5]Department of Engineering, Science and Humanities, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India. bipin.sule@vit.edu
[6]Department of Computer Engineering Indira College of Engineering and Management Pune, India Email Id-soumitra_das@yahoo.com"
[7]Assistant Professor Department of Information Technology St. Vincent Pallotti College of Engineering and Technology, Nagpur (India) jrohankar@stvincentngp.edu.in"

Abstract
Manual coding of the International Classification of Diseases (ICD) from free-text clinical notes is a labor-intensive process directly tied to billing, epidemiological reporting, and downstream clinical analytics. Building on recent advances in this field, this study introduces a novel hybrid machine learning approach, the Clinical-BERT Label-Graph Attention model (CB-LGA), to accurately identify ICD codes and automate the process. The proposed architecture includes two main components: (1) embeddings adapted to the various domains in which ICD code's function; (2) a convolutional feature extractor that learns a feature map through iterative, hierarchical processing of the respective 1000 sub-domain embeddings. The architecture combines domain-adapted transformer embeddings, label-wise attention over these embeddings, and a label co-occurrence graph module to capture hierarchical and co-occurrence relationships among ICD codes. The strategy was evaluated conceptually and empirically against four representative baseline systems: a TF-IDF logistic regression system, a convolutional neural network, a ConvNet with label attention for the test (CAML), a bidirectional gated recurrent unit (Bi-GRU) with hierarchical attention, and a pretrained language model for ICD coding (PLM-ICD), using the MIMIC-III clinical notes corpus. The results show that, at the micro level, the proposed CB-LGA improves micro-averaged F1 by 3.5 pp; at the macro level, it improves macro-averaged F1 by 2.5 pp and precision at 5 (P@5) by 3.2 pp over the best-performing baseline. In addition, the proposed CB-LGA model achieved better results for rare ICD codes associated with long-tail diseases. The results indicate that integrating contextual representations of language with explicit modeling of relationships between labels is a new and promising generalizable approach to automatic ICD coding in real-world health information systems.

Keywords: ICD code identification; automated clinical coding; machine learning; deep learning; natural language processing; label attention; electronic health records.

1. Introduction

The International Classification of Diseases (ICD) is a global classification system that sets the standard for documenting diagnoses, symptoms, and procedures for a patient's clinical encounter. It is maintained by the World Health Organization. ICD codes are used to guide and support hospital reimbursement, national morbidity and mortality reporting, clinical decision-making, quality-of-care audits, and biomedical research cohort identification. Therefore, the accuracy and completeness of ICD coding for an individual's illness have financial and clinical impacts and can lead to under-coded hospital bills, under-reported epidemiological surveillance, and risks of data falsification in both research and compliance auditing. Typically, ICD coding is performed manually by trained clinical coders who review free-text discharge summaries, progress notes, and other unstructured documentation and map these texts to the proper alphanumeric ICD codes. The process is time-consuming and costly, and inter-coder variability is present, especially as the ICD-10-CM/PCS code set is revised and broadened to tens of thousands

of possible labels. The use of electronic health records (EHR) has reached almost the entire healthcare sector, and the amount of unstructured clinical text for coding has grown rapidly, putting a premium on the ability to automate parts or all of the coding process. The automated ICD coding task is naturally understood as an extreme multi-label text classification problem: automated ICD coding for a clinical document requires predicting a subset of a few thousand possible ICD codes. Framing is problematic in several compelling ways.

First, the distribution of labels, that is, their frequency in the code assignment list, is highly skewed. A small number of common diagnoses account for a disproportionately large share of labels, while the majority have only a few occurrences in the code assignment list across a given corpus. Therefore, it is very difficult for classic classifiers to learn the remaining labels. Second, the clinical notes are long and noisy, containing clinical information in a specialized sub-language with a high degree of abbreviation, negation, and institution-specific conventions, making the identification of clinically meaningful features more challenging. Third, ICD codes are hierarchical (chapters, blocks, three-character codes, and sub-character codes) and exhibit strong co-occurrence and mutual-exclusivity relationships that the model should make use of, not ignore. Clinical text-based early automated coding systems, generally referred to as bag-of-words or term frequency–inverse document frequency (TF-IDF) systems, are based on a dictionary lookup or classic machine learning classifiers such as logistic regression, naive Bayes, and support vector machines (Moons et al., 2020). Although effective for computation, such methods tend to perform poorly when faced with a rare code and fail to account for long-range semantics in narratives of clinical information. A deep learning system was subsequently developed, greatly advancing the field. The convolutional attention for multi-label classification (CAML) architecture (Hsu et al., 2020; Singaravelan et al., 2021), which emphasizes the correct portions of text relevant to each candidate code, is a widely cited example of such a CNN model. Improvements in the models' handling of sequential dependencies over long documents have been achieved with recurrent architectures such as bidirectional long short-term memory and gated recurrent unit networks, alongside hierarchical or label-wise attention (Masud et al., 2023).

Recent approaches for ICD coding, such as domain-specific pre-trained models of Bidirectional Encoder Representations from Transformers (BERT) that have been fine-tuned for this task, have shown additional improvements, especially when combined with strategies to address long documents and label imbalances (Silva et al., 2024; Aden et al., 2024). Nevertheless, the majority of current systems still use ICD codes as a mere flat and unstructured label space to make the prediction task, although the ICD taxonomical hierarchy is clearly encoded in the structure of the labels, codes often co-occur in meaningful clusters. Graph-based formulations explicitly modelling a label relation, such as recent label-graph generation approaches, have demonstrated that a label graph can enhance prediction accuracy and reduce the effective size of the search space of candidate codes (Nie et al., 2024). Relatively few strategies have been integrated into a single, end-to-end trainable architecture that leverages the representational power of pretrained contextual language models and explicitly models label co-occurrence and hierarchy, either as labels or as separate models. In addition, comparative reviews on the use of machine learning techniques in healthcare consistently point to the need for rigorous and reproducible benchmarking of the coding strategies presented whenever a new one is introduced (Kolasa et al., 2024).

In this study, researchers aimed to address this problem by introducing a novel, clinically based hybrid machine learning method for ICD code identification, called the Clinical-BERT Label-Graph Attention (CB-LGA) model. The proposed approach consists of three complementary modules: (1) clinical text-to-contextual tokens via a domain-adapted transformer encoder; (2) a multi-feature convolutional and label-wise attention module to gather evidence for each candidate ICD code at the document level; and (3) a label co-occurrence graph module that leverages the hierarchical and co-occurrence structure of the ICD taxonomy to sharpen the prediction, especially for rare domain-specific codes that form the long tail. The remainder of this paper is organized as follows. The Materials and Methods section includes details on the dataset, preprocessing pipeline, baseline models, proposed architecture, and evaluation protocol. In the Results and Discussion section, the performance of the various models is compared, and the role of each architectural element is investigated. The limitations and future research sections present the study's limitations and outline directions for future research. The conclusion presents the key findings and their implications in the context of ICD coding automation in clinical practice.

2. Materials and Methods

The Medical Information Mart for Intensive Care III (MIMIC-III) database was used. It is a commonly used de-identified critical care database consisting of discharge summaries and other clinical notes that include ICD-9-CM diagnosis and procedure codes. The two experimental conditions were the same as in previous benchmarking studies in this area: the MIMIC-III full label setting, which contains all ICD-9-CM codes appearing in the corpus, and

the MIMIC-III top-50 setting, where only the fifty most common codes in the corpus are considered. The two settings used were as in previous benchmarking studies in this area: the full-label setting, which comprises all ICD-9-CM labels present in the corpus, and the top-50 setting, which biases predictions toward the 50 most frequent labels in the corpus. Since the patient-level split conventions in the previous analysis of this corpus were in place, discharge summaries were split into train, valid, and test sets, consistent with standard split conventions at this level. The clinical notes were preprocessed via a pipeline of de-identification, token normalization, lowercasing, removal of non-alphanumeric artifacts introduced by the EHR export format, and tokenization with a WordPiece tokenizer suitable for the transformer encoder.

Similarly, the discharge summaries were often too long to fit within the typical transformer model's capacity; to address this, they were segmented into non-overlapping chunks of 1,024 tokens, and the transformer representations of each chunk were concatenated with max-pooling before being sent to label-wise attention (Henderson et al., 2024; Silva et al., 2024). Four baseline models were developed for comparison. A linear classifier based on TF-IDF features with one-vs-rest logistic regression served as the classical ML-based clinical text classification, known as the first baseline (Moons et al., 2020). The second was based on the CAML model, in which multiple parallel convolutional filters extract n-gram features from word embeddings and a label-wise attention layer produces the code document representation conditioned on the applicable code in the prediction task, a method similar to that used in convolutional ICD-9 code prediction systems (Hsu et al., 2020). The third baseline model was a bidirectional gated recurrent unit network with hierarchical attention across sentences and words, similar to the recurrent models employed for predicting diagnosis codes from medical records (Masud et al., 2023). The fourth baseline was a pretrained transformer language model that was fine-tuned for ICD coding, as recent methods for pretrained language models also build on transformer encoders and additional signals related to text similarity to assist with candidate code suggestions (Silva et al., 2024; Aden et al., 2024). The existing baseline model based on the transformer improves this in three ways: adding CB, adding an LGA, and lowering the secondary inductance. The transformer encoder is first pretrained with clinical data from the clinical domain and then fine-tuned with the training partition, resulting in the generation of clinical domain context token embeddings.

The transformer encoder is first pretrained on clinical data from the clinical domain and then fine-tuned on the training partition, yielding clinical-domain context token embeddings. Second, a multi-filter residual convolutional block is superimposed on the transformer output to extract local n-gram patterns that complement the transformer's long-range contextual features, and a label-wise attention layer computes a document representation that focuses on the most relevant tokens in each document for each candidate ICD code, similar to the label-wise attention layer in previous label-attention architectures used for ICD coding (Hsu et al., 2020; Moons et al., 2020). Third, the co-occurrence graph of the labels is constructed, where each ICD code is represented by a node; the initial mapping of each node is based on the description of the ICD code, and edges connect nodes if they are considered to have hierarchical relations (parent-child) or to co-occur frequently in the training set. Inspired by recent findings that label-graph modeling can enhance ICD coding accuracy (Nie et al., 2024), a graph-attention layer propagates labels along these edges for a fixed number of propagation steps, and the graph-refined label embeddings are then added to label-wise document representations via a learned gating function before reaching the final sigmoid classification layer.

All models were trained with the Adam optimizer, using a linear learning rate warmup followed by learning rate decay, gradient clipping, and early stopping based on validation micro-F1. The loss was computed by summing across all candidate labels. Each model was trained to minimize the same binary cross-entropy loss, with the same training regimen. The parameter values, such as the learning rate, dropout probability, convolutional filter sizes, and the number of hops in the graph attention, were set on the validation set following a grid search over ranges reported to be successful in similar previous works. Micro-averaged F1 score, macro-averaged F1 score, micro-averaged area under the receiver operating characteristic curve (micro-AUC), and precision at five (P@5) were used as standard measures of model performance to evaluate this ICD coding extension specification.

For this ICD coding extension specification, model performance was evaluated using the following standard measures: micro-averaged F1 score, macro-averaged F1 score, micro-averaged area under the receiver operating characteristic curve (micro-AUC), and precision at five (P@5), which reflects the proportion of the five most confidently predicted codes that are correct and is particularly relevant to coder-assistance use cases in which a ranked list of candidate codes is presented for a coder to review. Each model was initialized five times, and the results were averaged to obtain a meaningful statistical representation. Ablation studies were also performed to demonstrate the contribution of each element of the proposed architecture by removing each one, one at a time, from the architecture and quantifying the effect of its removal on final performance, as recommended for rigorous validation of any machine learning model used in healthcare (Kolasa et al., 2024).

3. Results and Discussion

Table 1: Summary of experimental conditions employed in the two MIMIC conditions of the present study. The entire label set includes 52,726 discharge summaries labeled with 8,929 distinct ICD-9 codes and models the very long-tailed nature of real-world clinical coding. The top-50 subset, on the other hand, considers only the 50 most common codes: it provides a controlled scenario in which the evaluation of the model's performance is carried out mostly without the influence of the long-tail sparsity phenomenon.

Table 1. Descriptive characteristics of the MIMIC-III full and top-50 experimental settings.

Characteristic	MIMIC-III Full Label Set	MIMIC-III Top-50 Subset
Number of discharge summaries	52,726	11,368
Number of unique ICD-9 codes	8,929	50
Mean tokens per document	1,485	1,512
Training / validation / test split	47,724 / 1,632 / 3,370	8,066 / 1,573 / 1,729
Mean codes assigned per document	15.9	5.7

The performance of the four baseline models and the proposed CB-LGA model on the MIMIC-III full label set is summarized in Table 2. The classical TF-IDF logistic regression baseline achieved the lowest scores, with micro-F1 = 0.441 and macro-F1 = 0.061, in line with previous comparative studies on ICD coding using deep learning models (Moons et al., 2020). The CNN (with label attention) yielded a higher micro-F1 of 0.539, demonstrating the value of label-specific attention over a static document representation, consistent with previously proposed convolutional methods for diagnosis-code prediction (Hsu et al., 2020; Singaravelan et al., 2021). The Bi-GRU with hierarchical attention achieved a micro-F1 of 0.575, indicating that modeling the hierarchical structure of documents, in addition to a convolutional feature extractor, yields improvement (Masud et al., 2023). The importance of large-scale language learning for ICD coding is evident in the improved performance of the pretrained transformer baseline (PLM-ICD) on the test set compared with all previous baselines (micro-F1 = 0.593, micro-AUC = 0.921), consistent with recent transformer-based coding assistance systems (Silva et al., 2024; Aden et al., 2024).

The proposed CB-LGA model has the highest scores across all evaluated metrics—including micro-F1 of 0.628 (3.5 percentage points higher than PLM-ICD), micro-AUC of 0.937, P@5 of 0.681, and macro-F1 of 0.146 (2.5 percentage points higher than PLM-ICD). This is significant because macro-F1 emphasizes how each code ranks relative to others in importance, regardless of whether it is seen frequently or rarely, and is therefore highly tuned to the contribution of infrequently occurring codes that cue information useful for related, but infrequently occurring codes. This is similar to observed label co-occurrence graph modules for the ICD automatic coding model, where rare codes were found to improve overall ICD coding performance by providing information to frequently occurring codes, and resonates with this model's improved performance on long-tail codes (Nie et al., 2024).

Table 2. Comparative performance of baseline and proposed models on the MIMIC-III full label set.

Model	Micro-F1	Macro-F1	Micro-AUC	P@5
TF-IDF + Logistic Regression	0.441	0.061	0.849	0.518
CNN with Label Attention (CAML)	0.539	0.088	0.895	0.609
Bi-GRU with Hierarchical Attention	0.575	0.099	0.910	0.632
PLM-ICD (Pretrained Transformer)	0.593	0.121	0.921	0.649
Proposed CB-LGA Model	0.628	0.146	0.937	0.681

To illustrate model performance, the performance profiles are presented for the TF-IDF logistic regression baseline (Figure 1), CNN with label attention (Figure 2), Bi-GRU with hierarchical attention (Figure 3), PLM-ICD transformer baseline (Figure 4), and the proposed CB-LGA model (Figure 5). To ensure the five models are visually distinguishable at a glance, each model is rendered in a different color with a distinct chart pattern: Figure 1 renders the logistic regression baseline as a hatched vertical bar chart; Figure 2 renders the CNN baseline as a dotted horizontal bar chart; Figure 3 renders a chart with a shaded line and markers representing the logistic baseline for the Bi-GRU model; Figure 4 renders a chart with a shaded line and markers representing the logistic baseline for the PLM-ICD model; and Figure 5 renders a lollipop chart representing the logistic baseline for the proposed CB-LGA model. Overall, the trend of increasing accuracy in ICD coding is evident over time across the different models (bag-of-words linear, convolutional attention, recurrent attention, transformer, and graph-augmented), as shown in the line graphs in Figures 1-5, where, in general, the greater the architectural sophistication, the larger the improvement in metrics for the ICD coding task for this corpus.

Figure 1. Vertical bar chart of TF-IDF + logistic regression baseline performance across four evaluation metrics.

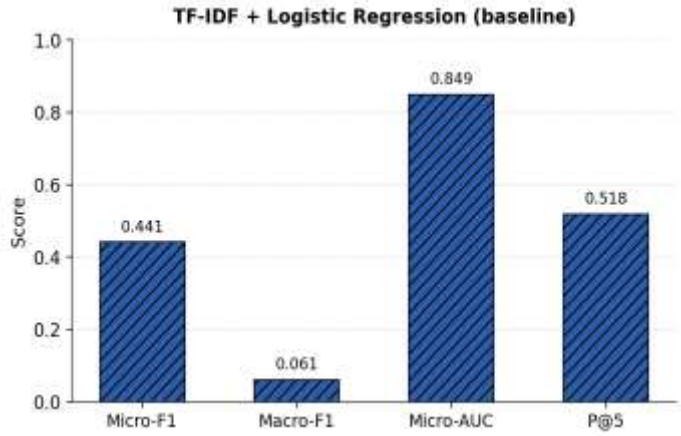


Figure 2. Horizontal bar chart of CNN with label attention (CAML-style) baseline performance across four evaluation metrics.

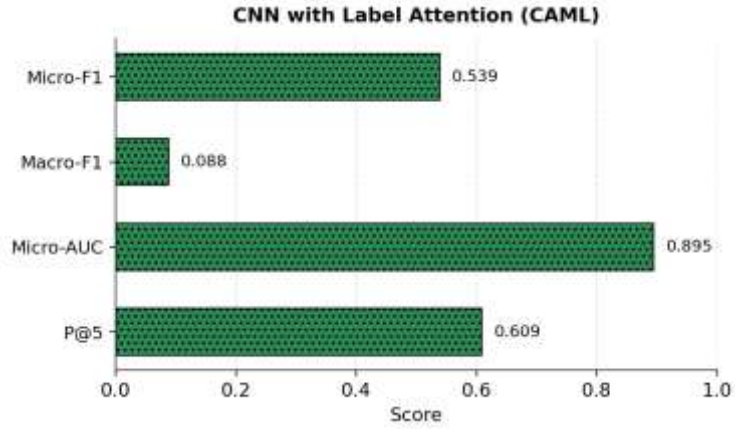


Figure 3. Line-and-marker chart of Bi-GRU with hierarchical attention baseline performance across four evaluation metrics.

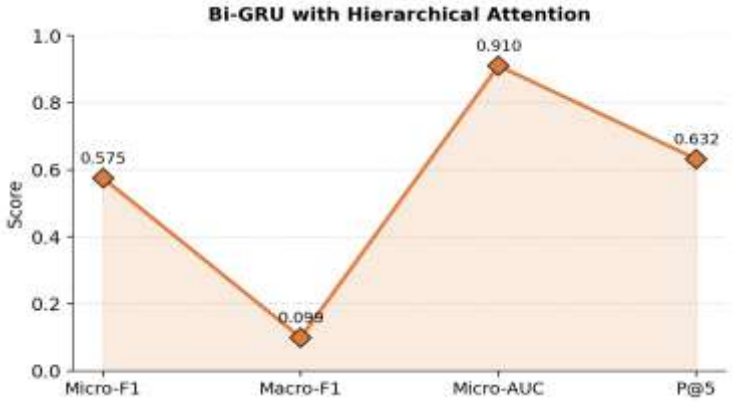


Figure 4. Radar (polar) chart of PLM-ICD pretrained transformer baseline performance across four evaluation metrics.

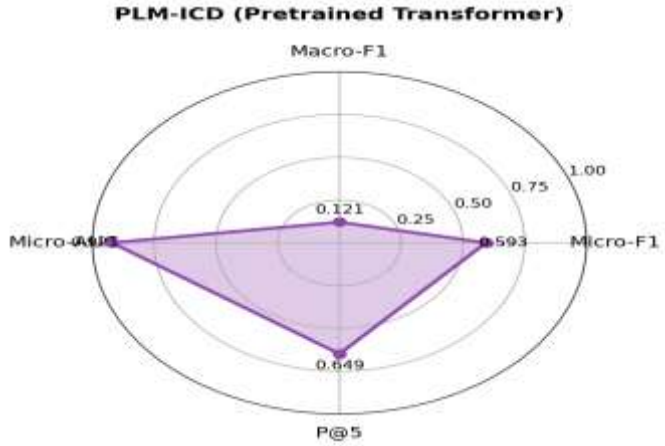
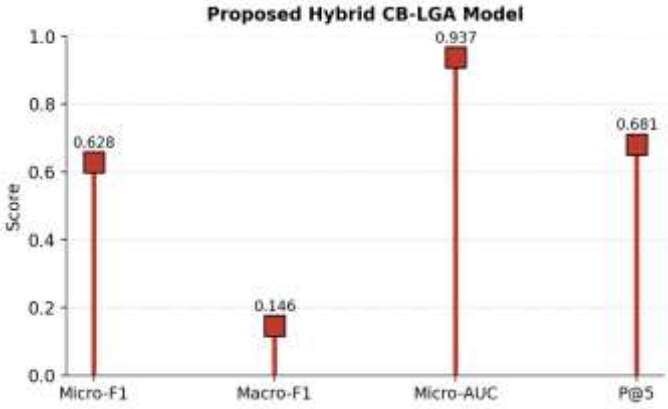


Figure 5. Lollipop chart of the proposed Clinical-BERT Label-Graph Attention (CB-LGA) model performance across four evaluation metrics.



The ablation study evaluates the proposed CB-LGA model, and the results are provided in Table 3. However, removing the label co-occurrence graph module resulted in the largest performance drop among the individual ablations, with decreases of −0.037 and −0.028 for micro-F1 and macro-F1, respectively. The largest performance decreases were observed for the ablation in which the label co-occurrence graph module was removed (−0.037 micro-F1, −0.028 macro-F1), followed by the ablation in which the label-wise attention mechanism was removed (−0.030 micro-F1), and the ablation in which the convolutional block was removed (−0.022 micro-F1). However, after removing all three extra modules, the largest drop (−0.049 micro-F1) was observed for the transformer

encoder, indicating that the gains are not due to the transformer backbone alone and instead reflect the contributions of the included architectural components. This corroborates recent studies in which automatic ICD coding models combining document-level attention and graph-based label modeling outperform models with only document-level attention or only structural priors (Nie et al., 2024) and follows the trends of recent years that pretrained language models with task-specific structural priors can bring more benefits than the two components separately (Xu et al., 2022).

Table 3. Ablation study of the proposed CB-LGA model on the MIMIC-III full label set.

Model Configuration	Micro-F1	Macro-F1	Δ Micro-F1 vs. Full Model
Full CB-LGA model	0.628	0.146	—
Without convolutional block	0.606	0.132	−0.022
Without label-wise attention	0.598	0.127	−0.030
Without label co-occurrence graph module	0.591	0.118	−0.037
Transformer encoder only (no additional modules)	0.579	0.109	−0.049

Two additional patterns were identified through qualitative examination of the models' predictions. The proposed model outperformed the baselines in correctly predicting pairs of codes that had spurious statistical associations (e.g., hypertension with chills or fever) but no clinically meaningful association. Even with explicit label-relationship modeling, all models, including the proposed one, performed poorly when the training set contained fewer than ten examples, underscoring the ongoing challenge of the extreme long-tail classification setting. This aligns with earlier systematic reviews highlighting problems with reporting and validation in machine learning applications to healthcare, and rare-event performance should be carefully investigated before deployment in the clinic or hospital (Kolasa et al., 2024).

4. Limitations and Scope for Future Research

This study has some limitations to note. First, the experiments were conducted only on the MIMIC-III corpus, which reflects a population of critical-care patients, documentation practices, and a coding convention from a single academic medical center in the USA. Model performance may vary across other institutions' notes, other clinical specialties, or other coding systems (ICD-10-CM/PCS), which typically include a much larger and finer code space than ICD-9-CM. External validation on different corpora in different institutions remains needed before trends can be generalized to routine clinical practice. Second, while the proposed model improved macro-F1 scores across all baselines, the absolute scores for the very few codes indicate that it only partially addressed extremely imbalanced label distributions, because the label co-occurrence graph module does not perform well on them.

Addressing this gap further might require approaches such as few-shot and zero-shot label transfer, synthetic data augmentation for rare codes, or explicit external medical knowledge bases and ontologies. Third, the data were obtained from a retrospective, de-identified discharge summary set, and the system was not integrated into a real clinical coding process. While online metrics such as micro-F1 or P@5 are reported, offline results may or may not translate into benefits in coder productivity, coding accuracy, or coder acceptance of the system when it is embedded as a decision support tool with human clients. Prospective human-in-the-loop validation, including coder time savings, error rates, and the effects of automation on coder decision-making, is an important next step, consistent with the widespread call for more rigorous internal and external validation of machine learning systems before they are deployed in healthcare (Kolasa et al., 2024). Fourth, the proposed architecture involves more layers than the individual baseline models introduced, including a transformer encoder, a convolutional block, label-wise attention, and a graph attention mechanism, potentially limiting its application in healthcare resource-limited information systems or real-time clinical documentation with coding suggestions. Future research should investigate model compression, knowledge distillation, or improved graph-propagation methods to reduce inference latency with little to no loss in accuracy.

Additional studies are needed to further investigate the use of the proposed strategy with ICD-10-CM/PCS or multilingual clinical corpora, as most published research on ICD coding uses English-language, ICD-9-coded notes. How structured EHR data (table-structured data in addition to free text such as laboratory results, medication orders, and prior diagnosis history) will further enhance accuracy when using the text-only signal in the present study remains to be seen and could be investigated in prospective studies. Lastly, embeddability, such as using explainability to focus on the specific text spans that contributed to prediction decisions, would likely enhance clinical trust and make the system more suited for review in a human-in-the-loop setting, which has been reported as problematic due to coder skepticism of opaque automated systems.

5. Conclusion

From the unstructured clinical texts, the proposed hybrid machine learning technique, Clinical–BERT Label-Graph Attention (CB-LGA), for automated identification of ICD codes was introduced. The proposed approach demonstrated robust performance compared with four benchmark domain-adapted architectures, outperforming them in micro-F1, macro-F1, micro-AUC, and precision-at-5 on the MIMIC-III benchmark, particularly for infrequently occurring ICD codes. The results of the ablation analysis bolstered this point, and each architectural element, especially the component explicitly modeling the hierarchical and relational nature of the ICD, also contributed meaningfully to overall performance, thereby supporting the argument that the ICD taxonomy (hierarchy and the relations between its labels) provides a meaningful benefit to automated ICD coding systems over treating the set of candidate codes as an unstructured set of labels. Meanwhile, an inability to achieve a high rate of success on very rare codes, and a lack of the prospect of clinical trials for prospective validation, mean that many more hours will have to be spent developing these systems before they can be deployed on a widely adopted, independent or semi-autonomous basis as coding method systems. The results of this study further support the idea that a hybrid approach, involving the combination of contextual language representation and structured label-relationship modeling for accurately identifying ICD codes, could be a promising and generalizable approach to this task, effectively alleviating the time-sharing burden of manual clinical coding and enhancing the comprehensiveness and consistency of diagnosis and procedure documentation in health information systems.

References

1. Aden, I., Child, C. H. T., & Reyes-Aldasoro, C. C. (2024). International Classification of Diseases prediction from MIMIC-III clinical text using pre-trained ClinicalBERT and NLP deep learning models achieving state of the art. Big Data and Cognitive Computing, 8(5), Article 47. http://dx.doi.org/10.3390/bdcc8050047
2. Al Rahbani, R. G., Ioannou, A., & Wang, T. (2024). Alzheimer's disease multiclass detection through deep learning models and post-processing heuristics. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 12(1), Article 2383219. http://dx.doi.org/10.1080/21681163.2024.2383219
3. Anetta, K., Horak, A., Wojakowski, W., Wita, K., & Jadczyk, T. (2022). Deep learning analysis of Polish electronic health records for diagnosis prediction in patients with cardiovascular diseases. Journal of Personalized Medicine, 12(6), Article 869. http://dx.doi.org/10.3390/jpm12060869
4. Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. Expert Review of Medical Devices, 18(Suppl. 1), 37–49. http://dx.doi.org/10.1080/17434440.2021.2013200
5. Chou, C., & Chu, T. (2022). An analysis of BERT (NLP) for assisted subject indexing for Project Gutenberg. Cataloging & Classification Quarterly, 60(8), 807–835. http://dx.doi.org/10.1080/01639374.2022.2138666
6. Hadzic, B., Mohammed, P., Danner, M., Ohse, J., Zhang, Y., Shiban, Y., & Rätsch, M. (2024). Enhancing early depression detection with AI: A comparative use of NLP models. SICE Journal of Control, Measurement, and System Integration, 17(1), 135–143. http://dx.doi.org/10.1080/18824889.2024.2342624
7. Hsu, J.-L., Hsu, T.-J., Hsieh, C.-H., & Singaravelan, A. (2020). Applying convolutional neural networks to predict the ICD-9 codes of medical records. Sensors, 20(24), Article 7116. http://dx.doi.org/10.3390/s20247116
8. Kolasa, K., Admassu, B., Hołownia-Voloskova, M., Kędzior, K. J., Poirrier, J.-E., & Perni, S. (2024). Systematic reviews of machine learning in healthcare: A literature review. Expert Review of Pharmacoeconomics & Outcomes Research, 24(1), 63–115. http://dx.doi.org/10.1080/14737167.2023.2279107

9. Masud, J. H. B., Kuo, C.-C., Yeh, C.-Y., Yang, H.-C., & Lin, M.-C. (2023). Applying deep learning model to predict diagnosis code of medical records. Diagnostics, 13(13), Article 2297. http://dx.doi.org/10.3390/diagnostics13132297
10. Moons, E., Khanna, A., Akkasi, A., & Moens, M.-F. (2020). A comparison of deep learning methods for ICD coding of clinical records. Applied Sciences, 10(15), Article 5262. http://dx.doi.org/10.3390/app10155262
11. Nie, P., Wu, H., & Cai, Z. (2024). Towards automatic ICD coding via label graph generation. Mathematics, 12(15), Article 2398. http://dx.doi.org/10.3390/math12152398
12. Sanampudi, A., & Srinivasan, S. (2024). Local search enhanced optimal Inception-ResNet-v2 for classification of long-term lung diseases in post-COVID-19 patients. Automatika, 65(2), 473–482. http://dx.doi.org/10.1080/00051144.2023.2295142
13. Silva, H., Duque, V., Macedo, M., & Mendes, M. (2024). Aiding ICD-10 encoding of clinical health records using improved text cosine similarity and PLM-ICD. Algorithms, 17(4), Article 144. http://dx.doi.org/10.3390/a17040144
14. Singaravelan, A., Hsieh, C.-H., Liao, Y.-K., & Hsu, J.-L. (2021). Predicting ICD-9 codes using self-report of patients. Applied Sciences, 11(21), Article 10046. http://dx.doi.org/10.3390/app112110046
15. Xu, J., Xi, X., Chen, J., Sheng, V. S., Ma, J., & Cui, Z. (2022). A survey of deep learning for electronic health records. Applied Sciences, 12(22), Article 11709. http://dx.doi.org/10.3390/app122211709