

AN IMPROVED TRANSFORMER-DRIVEN APPROACH FOR EXPLAINABLE ICD
CODE PREDICTION

A Project

Presented to the faculty of the Department of Computer Science

California State University, Sacramento

Submitted in partial satisfaction of
the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science

by

Meghana rao Kanneganti

SPRING
2024

© 2024

Meghana rao Kanneganti

ALL RIGHTS RESERVED

AN IMPROVED TRANSFORMER-DRIVEN APPROACH FOR EXPLAINABLE ICD
CODE PREDICTION

A Project

by

Meghana rao Kanneganti

Approved by:

_____, Committee Chair
Dr. Haiquan Chen

_____, Second reader
Dr. Parham Phoulady

Date

Student: Meghana rao Kanneganti

I certify that this student has met the requirements for the format contained in the University format manual, and this project is suitable for electronic submission to the library and credit is to be awarded for the project.

_____, Graduate Coordinator _____
Dr. Ying Jin Date

Department of Computer Science

Abstract

of

AN IMPROVED TRANSFORMER-DRIVEN APPROACH FOR EXPLAINABLE ICD CODE
PREDICTION

by

Meghana rao Kanneganti

Automated ICD coding is a multi-label classification task that involves assigning ICD codes to long clinical texts such as discharge summaries. It is an active research field and numerous studies demonstrate how using Natural language processing (NLP) techniques simplify this task making it cost-effective and aiding in accurate classifications. In recent years, transformers have extensively been used in NLP tasks, and offer a solution for accurately processing long texts and aid in automating ICD code classification. Transformers employ their self-attention mechanism to capture long-range dependencies. To enhance the state-of-the-art approaches for automated ICD-9 coding, we propose 2 models Long-LAT and Long-HiLAT leveraging the Longformer’s capability to process a large number of tokens. In Long-LAT, we introduce Label wise attention with the Clinical pretrained Longformer instead of solely relying on the classification head provided by the Longformer. This allows us to refine the classification process further by directing attention to the specific ICD codes. In Long-HiLAT, we make use of the Longformer’s sliding window attention to enhance the existing Hierarchical Label-wise Attention Transformer (HiLAT). This integration allows a more comprehensive analysis of the text and improves the model’s performance.

Extensive experiments using the MIMIC-III top-50 and top-5 ICD code datasets show that Long-HiLAT achieved superior performance compared to the baseline models.

_____, Committee Chair
Dr. Haiquan Chen

Date

ACKNOWLEDGMENTS

I want to express my sincere gratitude to Dr Haiquan Chen for the unconditional support and guidance he offered during every phase of my project. I would like to extend my appreciation to my second reader Dr Parham Phoulady for reviewing my work.

TABLE OF CONTENTS

Page

Acknowledgments	vii
List of Tables.....	x
List of Figures	xi
Chapter	
1. INTRODUCTION	1
2. BACKGROUND OF THE STUDY.....	4
3. DATASET.....	6
4. BASELINE MODELS	8
4.1 Longformer architecture.....	9
4.2 Labelwise attention.....	10
4.3 Clinically pretrained transformer.....	11
4.4 Hierarchical Labelwise attention transformer (HiLAT).....	12
5. PROPOSED METHODOLOGY	14
5.1 Long-LAT architecture.....	15
5.2 Long-HiLAT architecture.....	16
6. EXPERIMENTAL VALIDATION	19
6.1 Training	19
6.2 Performance Metrics	19
6.3 Results comparison.....	21
7. MODEL EXPLAINABILITY.....	25
8. WEB UI FOR ICD CODE PREDICTION.....	26

9. CASE STUDY	28
10. CONCLUSION	34
References	35

LIST OF TABLES

Tables		Page
1.	Performance comparison of pretrained clinical models in clinical document classification.....	11
2a	Performance comparison for top-5 ICD codes	21
2b.	Labelwise F1 score comparison for top-5 ICD codes	21
3a.	Performance comparison for top-5 ICD codes	22
3b.	Labelwise F1 score comparison for top-50 ICD code.....	23

LIST OF FIGURES

Figures	Page
1. Frequency distribution of top-50 ICD codes in the MIMIC-III dataset	6
2. Sample Clinical text summary	7
3. Comparing different attention models in transformers	9
4. Architecture of Hierarchical Label-wise attention transformer (HiLAT)	13
5. The architecture of Long-LAT	14
6. Architecture of Long-HiLAT	17
7. Sample attention visualization example for ICD Code 38.91	17
8. Web Interface for ICD Code prediction	27
9. HiLAT sample predictions for top-5 labels.....	30
10. Clinical Longformer sample predictions for top-5 labels.....	30
11. Long-LAT sample predictions for top-5 labels.....	31
12. Long-HiLAT sample predictions for top-5 labels.....	31
13. Attention visualization of model HiLAT for ICD code 38.93	32
14. Attention visualization of model Long-HiLAT for ICD code 38.93.....	32
15. Attention visualization of model HiLAT for ICD code 427.31	33
16. Attention visualization of model Long-HiLAT for ICD code 427.31.....	33
17. Attention visualization of model Long-HiLAT for ICD code 428.0.....	33

1. INTRODUCTION

International Classification of Diseases (ICD) is a globally used medical coding system maintained by the World Health Organization (WHO) that provides a standardized system to classify medical diagnoses. The ICD classification system provides international comparability and aids in multiple clinical tasks such as billing, medical insurance support, health statistics, and research. Therefore, it is important to maintain highly accurate and reliable medical coding systems. Currently, professionally trained medical coders use their expertise to categorize and transcribe clinical documents with appropriate ICD codes. Manual ICD coding is labor-intensive, expensive and often prone to human errors [1]. In response to these challenges, several studies prove that applying Natural Language Processing (NLP) techniques for ICD code classification produced promising results, offering a reliable solution to automate the process [2][3][4]. Automated ICD coding is a multi-label classification task where a set of ICD codes are assigned to clinical free text such as nursing notes or discharge summaries [5].

Traditionally, many researchers explored the impact of using Convolutional neural networks (CNNs) and Recurrent Neural networks (RNNs) with Labelwise attention mechanisms for automated ICD tasks and achieved state-of-the-art performance [6][7][3][4][8][9][10]. However, CNNs struggle to capture contextual information in text, and RNNs suffer from vanishing gradient problems, which makes it worse for lengthy text processing tasks, as they cannot retain long-range dependencies. These limitations hinder their effectiveness in the accurate classification of ICD codes.

The Transformer [11] is a neural network architecture with multiple layers of stacked encoder-decoder layers where each layer employs a multi-head self-attention mechanism along with fully connected feed-forward neural networks. In recent years, Pretrained transformer models achieved

great success in several NLP tasks. Several researchers studied the effectiveness of using pretrained Clinical transformers like BERT for ICD coding tasks [12][13][14][15]. The HiLAT model elaborates on the improvement seen in using the Clinical XLNET transformer over BERT [16]. Inspired by pretrained transformers, and the success of HiLAT, this project attempts to explore a hybrid architecture Long-HiLAT that takes advantage of HiLAT and Longformer. The Longformer implements sliding window self-attention mechanism to process a large number of tokens without losing long-range dependencies, thereby helping to achieve better performance while processing lengthy texts. Sliding window attention is a technique used by transformers to tackle long text in chunks [6]. The sliding window attention has been implemented in many transformers such as Sparse transformer, ETC, SWIN Transformer(images) and BigBird[17][18][19][20].

Alongside Long-HiLAT, we also explore how we can effectively improve the performance of pretrained models like Clinical Longformer using Label wise attention. Label-wise attention also aids in explaining the significance of specific words and context for the predicted codes [21][7]. We propose the model Long-LAT to explore this variation. Both models have been extensively trained and tested on the Medical Information Mart for Intensive Care (MIMIC-III) top 5 and top 50 ICD code discharge summary datasets [22].

For comparability, all the models (proposed models and baselines) discussed in this paper have been implemented and tested with possibly similar setup and configurations. Both models achieved better performance improving the baseline models, proving their superiority.

We summarize our contributions as follows:

- We proposed Long-LAT, a pretrained Clinical Longformer-based model coupled with Labelwise attention for improving the performance of pretrained transformers in multi-label classification tasks.

- We proposed the hybrid model Long-HiLAT, making use of the architecture of HiLAT incorporating layers of Longformer to utilize the advantages of sliding-window attention for processing lengthy texts.
- We conducted extensive experiments on datasets for the top 5 and top 50 ICD codes from MIMIC-III, utilizing discharge summaries. Our experiments consistently demonstrated that the Long-HiLAT model outperformed all baseline models.
- We developed a web UI for users to perform ICD code prediction using the models discussed in the project, given input free text clinical summaries. The web UI is accessible at <https://huggingface.co/spaces/meghanaraok/LongLAT>.

2. BACKGROUND OF THIS STUDY

In 1998, de Lima et al. proposed a hierarchical model for Automatic categorization of medical documents [6]. Several other studies have used classic NLP techniques for ICD code classification tasks. With the rise of Neural networks and deep learning, some studies exploiting CNNs and RNNs with Labelwise attention mechanisms demonstrated considerable performance improvement [6][3].

In 2017, with the introduction of transformer by Vaswani et al. [11], transformers have been extensively used for several natural language processing tasks and achieved incredible performance in many tasks such as Question answering classification, and language generation. Fine-tuning the transformers for domain-specific tasks and utilizing these pretrained large language models (Transfer learning) has revolutionized NLP in recent years. Clinical BERT has achieved state-of-the-art performance in clinical inference and named entity recognition tasks [12]. However, BERT cannot be applied for Long Documents due to its limited token processing and self-attention which exponentially increases memory consumption. Splitting the long text into meaningful chunks and stacking layers of BERT is one solution. Clinical XLNET with its unlimited mas token processing length also offers a similar solution and has been proven to achieve better results than BERT [16]. To overcome this problem, transformers like BigBird and Longformer implement sparse attention mechanisms like dilated attention and Sliding window attention.

Yikuan Li et al. introduced two clinically enriched long-sequence transformers - Clinical Longformer and Clinical BigBird, pretraining the models on a large clinical corpus [23]. Both models achieved optimal results in several clinical tasks. We use this version of Clinically pretrained Longformer in our Long-LAT model. M. Feucht et al. proposed a description-based Label attention classifier model (DLAC) [21]. In their study, they integrate the descriptions of the

ICD-9 codes and apply the embeddings to the text representations to obtain label-specific representations of the text for the predicted ICD codes. They evaluated this approach on BERT and Longformer and Longformer has achieved the best results. In another study [24], S. Levine and S. Abraham have pretrained the RoBERTA encoder on MIMIC-III and MIMIC-CXR datasets and converted this to the Longformer encoder by adding global attention and self-attention mechanism suitable for the Longformer encoder. This study has also produced competitive results.

The Hierarchical label-wise attention transformer model (HiLAT) model proposed by L. Liu et al produced the best results so far on the top 50 MIMIC-III Clinical dataset [16]. They pretrain the XLNET transformer on MIMIC-III Clinical notes and incorporate this transformer with a Hierarchical 2-level attention mechanism. The HiLAT model combined with ClinicalPlusXLNET achieved the best performance on the top 50 dataset.

3. DATASET

We used the MIMIC-III dataset [22] for ICD code classification. MIMIC-III is a large freely accessible critical care database comprising information of 46,520 patients admitted to ICU units, made available by Physionet for research use. From the several available unstructured free clinical texts, we used discharge summaries of patients. We created the popular MIMIC-III ICD top-50 dataset from the full dataset for comparability with other models. Along with the top-50, we also generated the top-5 ICD code dataset and evaluated the models' performance.

Figure 1 provides an overview of frequency of the top-50 ICD codes. Although the top-50 dataset exhibits lesser data imbalance compared to the full dataset, it still demonstrates a significant degree of imbalance. Certain codes, such as 401.9, exhibit very high frequency, while others are considerably rare. Figure 2 shows a sample Clinical discharge summary we used to train our models.

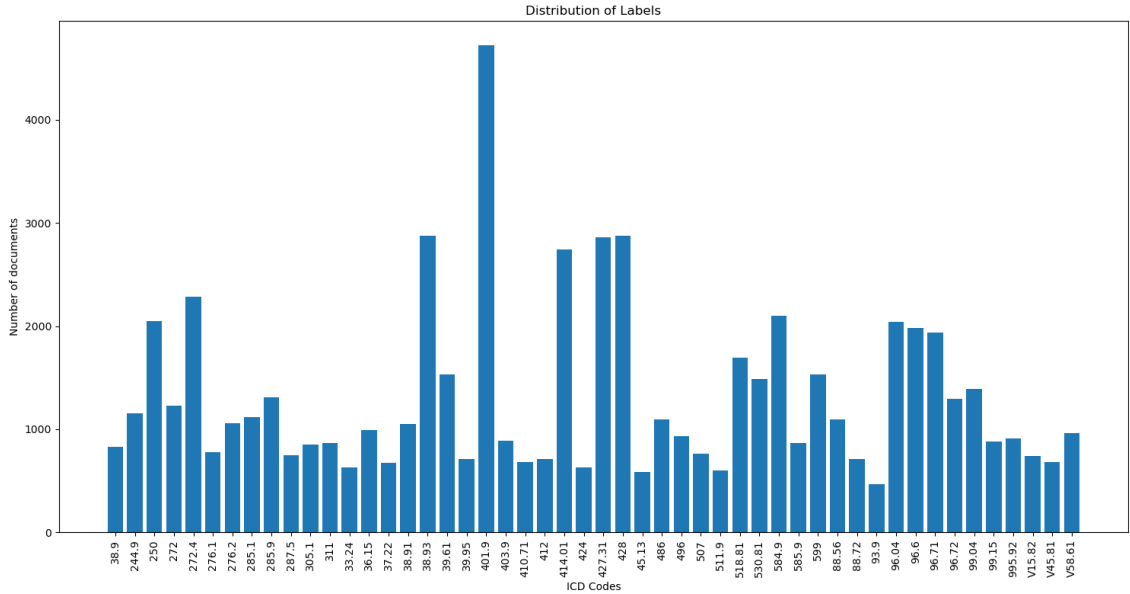


Figure 1. Frequency distribution of top-50 ICD codes in the MIMIC-III dataset

date of birth: sex: m service medicine allergies percocet erythromycin base a c e inhibitors nsaid
 attending chief complaint fevers rigors major surgical or invasive procedure dialysis bilateral hematoma and right
 lower quadrant hematoma incision and drainage left arm picc placement history of present illness year old male
 with a history of end stage renal disease of unknown etiology on hd htn prior history of hodgkin s disease gout
 peptic ulcer disease hypothyroidism depression presenting with fevers after right subclavian hd line was placed
 by transplant surgery days prior to admission mr went for his first hd treatment after new line placement day
 prior to admission with temps to he also developed rigors this morning and presented to his fever climbed to and
 was given vancomycin blood cultures from his hd line and was given fluid boluses and was then transferred to
 given that his nephrologist and transplant team are here in the ed his heart rates were in the 130s with a rectal
 temperature of he was visibly in rigors and not endorsing any pain otherwise he received tylenol and ativan for
 his rigors zosyn was added on to his vancomycin transplant team pulled his hd line and tip was sent for culture
 an ekg done in the ed showed st elevations in avf and lead iii with st depressions in v1 and v2 however these
 resolved on repeat ekg set of troponins was elevated to with no prior troponins to compare to ck mb was of
 normal fraction cards was consulted and they felt the troponin elevation was in the setting of chronic renal
 failure and demand ischemia in setting of tachycardia aspirin was given they did not recommend anticoagulation
 in setting of demand iv vancomycin was redosed and l of iv fluids were given vitals in the ed were hr bp rr satting
 ra past medical history chronic renal failure cr baseline followed by dr at no kidney biopsy hypertension hodgkin
 s disease s p chemotherapy with and radiation reportedly as a result of chemo treatment left him with numbness
 below the waist depression with one prior psych hospitalization h o suicidal attempt with narcotics overdose left
 eye blindness evaluated in ed on with functional overlay back pain right flank pain requiring chronic narcotics as

Figure 2. Sample Clinical text summary

Our dataset consists of clinical notes and the assigned ICD codes. The text was divided into different fixed-length meaningful chunks to experiment with the HiLAT and Long-HiLAT models. We experimented with chunk sizes 8 and 10. For the Clinical Longformer and Long-LAT models, we use the whole document embeddings as a single chunk of data to process with a token length of 4096. We truncate the remaining part in case of extremely long texts. The average length of a clinical discharge summary is 2188 words. The dataset was divided into train test and Validation datasets in a ratio of along with a label dictionary containing the long title for each ICD code.

4. BASELINE MODELS

4.1 Longformer architecture:

Traditional transformer models like BERT cannot process long sequences of text due to their self-attention mechanism which scales quadratically, increasing memory consumption and resources. To overcome this, multiple attention mechanisms have been proposed, such as the sparse or global sliding window and dilated window attention. Longformer is a transformer-based model proposed by Beltagy et al., which introduces the sliding-window attention mechanisms to decrease the quadratic complexity of self-attention while processing longer text sequences [25]. Longformer provides two variations of sliding window attentions - global sliding window and fixed-length sparse or dilated sliding window attentions. The local attention is still $O(n^2)$. The model has $O(n * w)$ complexity assuming w is of fixed length for all layers. Figure 3 demonstrates the global sliding window attention of Longformer which is used in our project.

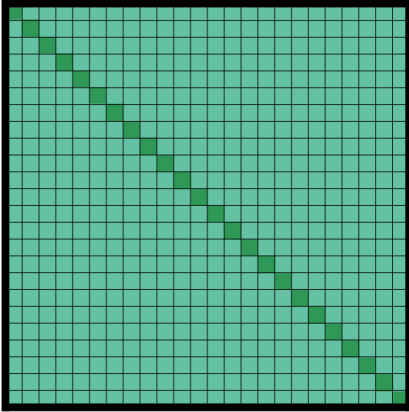


Figure 3a. Self-attention

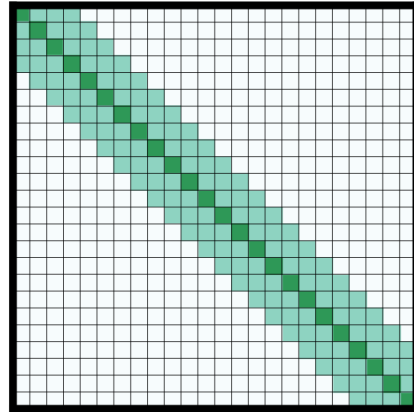


Figure 3b. Sliding window attention

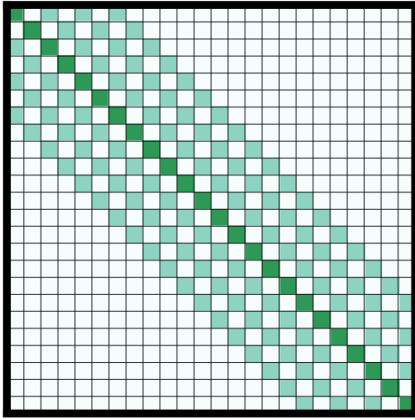


Figure 3c. Dilated Sliding window

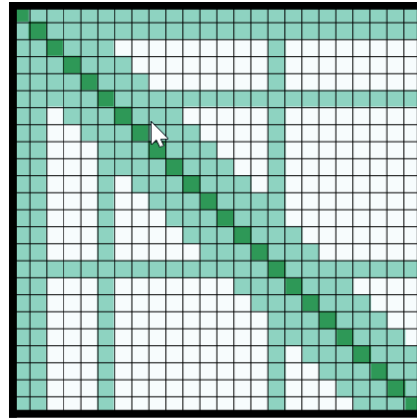


Figure 3d. Global+ sliding window

Figure 3. Comparing different attention models in transformers.[25]

For local attention, Longformer uses the Sliding window of size w where each token attends to $\frac{1}{2}w$ of tokens on each side. This reduces the complexity while processing large documents from $O(n^2)$ to $O(n * w)$. Global attention is added on a few preselected input locations. The tokens with global attention attend to all tokens in the sequence and all tokens attend to it. For the classification task, global attention is used for the [CLS] token.

Unlike traditional transformers, Longformer can process up to 4096 tokens. The Longformer has been pretrained on a large corpus of documents and finetuned for performing tasks including document classification, Question Answering, and co-reference resolution. Longformer was continually pretrained using Masked language modelling (MLM) from RoBERTa checkpoint making changes to support the attention mechanism.

The default size of Longformer sliding attention window is 512 and it is configurable. To utilize the RoBERTa's pretrained weights, Longformer's weights are initialized by copying the 512 position embeddings multiple times.

Thus, Longformer offers an efficient solution for classifying long text documents effectively.

4.2 Labelwise attention:

Labelwise attention mechanism was proposed by Vu et al. and has drastically improved the performance of ICD coding tasks [7]. The label attention model handles variable-length texts and interdependence between text fragments allowing the creation of label-specific token representations. Essentially, the label attention model takes a matrix of hidden representations $H=[h_1, h_2, \dots, h_n]$ as input and transforms it into 'L' vectors that are specific to each label, where 'L' represents the number of labels.

$$Z = \tanh(WH) \quad \dots\dots\dots(1)$$

$$A = \text{softmax}(UZ) \quad \dots\dots\dots(2)$$

$$V = HA^T \quad \dots\dots\dots(3)$$

1. Transformation of hidden representations:

Initially, the hidden representations(H) undergo a transformation to an intermediate representation (Z). This transformation involves using a hyperparameter-tuned matrix W of dimensions $R^{da \times 2u}$ applied to the input matrix H , with the hyperbolic tangent activation function.

2. Computation of Label-specific Weight Matrix:

Next, a matrix U of dimensions $R^{L \times da}$ is employed to compute Label-specific weight matrix A . This is done using softmax function, which ensures row-wise normalization. Each row in matrix A represents a weight vector corresponding to a label.

3. Calculation of Label-specific vectors:

The attention matrix A derived from the previous step is multiplied with hidden representation matrix H to obtain label-specific vectors, resulting in matrix V . Each column in V represents a label-specific representation of the document.

4.3 Baseline models

4.3.1. Clinically pretrained Longformer

Yikuan Li et al, in their work Clinical-Longformer and Clinical-BigBird: Transformers for Long Clinical Sequences pretrained the long text transformers Longformer and BigBird [23]. The difference between these transformers is in the implementation of their global attentions. As discussed earlier, Longformer’s global attention tokens are assigned to preselected tokens while BigBird uses randomized global attention tokens.

The models were pretrained on 2 million clinical notes extracted from the MIMIC-III dataset containing narratives of more than 46000 patients. During pretraining, they initialized the weights of Clinical Longformer with the pretrained weights of the base version of Longformer. It was trained for 200000 steps with batch size of 6×3 with a learning rate of $3e-5$.

The Clinical Longformer was adapted to 9 different NLP tasks such as document classification., Question answering, named-entity-recognition and natural language reference.

Pre-trained Models	OHSUMed	OpenI	MIMIC-AKI		medNLI
metrics	Accuracy	AUC	AUC	F1	Accuracy
BERT	0.717	0.952	0.514	0.293	0.776
BioBERT	0.771	0.954	0.534	0.324	0.808
ClinicalBERT	0.741	0.967	0.738	0.444	0.812
Clinical-Longformer	0.766	0.977	0.762	0.484	0.842
Clinical-BigBird	0.752	0.972	0.755	0.480	0.827

Table 1. Performance comparison of pretrained clinical models in document classification [23]

Both Clinical Bigbird and Clinical Longformer were compared to the ClinicalBERT, BERT and BioBERT models. While both the models achieved SOTA results, Clinical Longformer yielded the best performance in all the tasks demonstrating the superiority of systematic global attention over BigBird's randomized attention. Table 1 demonstrates the dominance of ClinicalLongformer in document classification [23].

4.3.2. Hierarchical Labelwise attention transformer (HiLAT):

Leibo Liu et al, in their paper “Hierarchical label-wise attention transformer model for explainable ICD coding” explored the efficiency of using pretrained XLNET transformer and achieved a significant improvement over BERT-based models [16].

They proposed a hierarchical approach for classifying clinical discharge summaries. The discharge summaries are split into multiple chunks and passed to the model. The model comprises 4 layers namely, the Pretrained transformer layer that outputs hidden representations for each chunk, the tokenwise attention layer, which produces a label-specific representation of each chunk by applying attention weights, a chunk layer, to which the representations from all chunks are stacked together and sent as input to obtain label-specific representations of the document, and a Feedforward neural network that acts as a binary classifier for each label.

They experimented with different variations in pretrained transformers and observed highest performance with ClinicalPlusXLNet which has been continually pretrained on all MIMIC-III clinical notes including discharge summaries.

HiLAT has achieved state-of-the-art performance in classifying ICD-9 clinical discharge summaries. Figure 4 demonstrates the architecture of HiLAT.

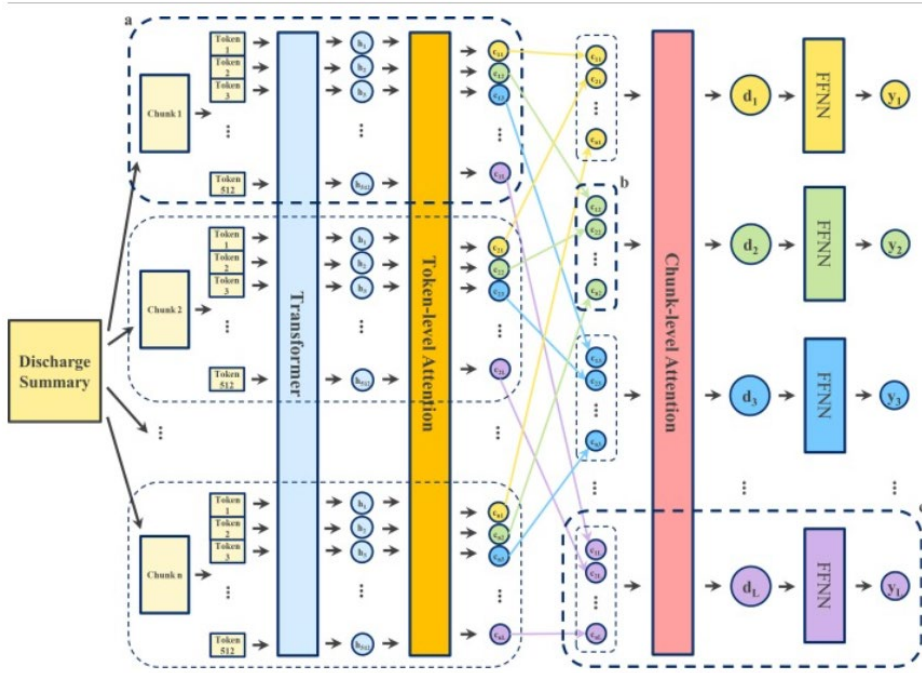


Figure 4. Architecture of Hierarchical Label-wise attention transformer (HiLAT) [16]

5. PROPOSED METHODOLOGY

Inspired by the success of HiLAT and Longformer based models in ICD coding [16][23], we proposed 2 models - Long-LAT and Long-HiLAT. The following subsections discuss the architectures of our proposed models.

5.1. Long-LAT architecture:

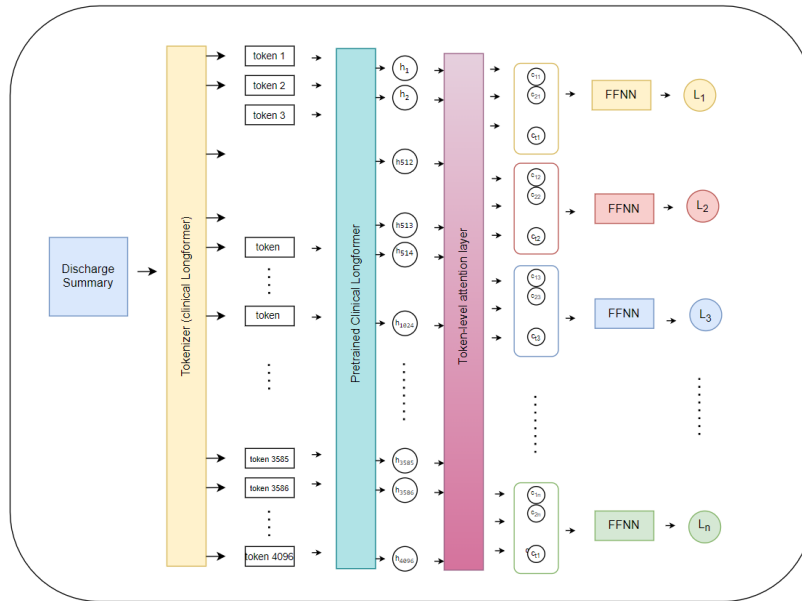


Figure 5. The architecture of Long-LAT model

Figure 5 provides an overview of the architecture of Long-LAT. The Long-LAT's architecture comprises 3 layers- the Pretrained Longformer layer, the Label-wise attention layer, and the classifier layer. We truncated the text to 4096 tokens and use a sliding window of 512 tokens. The output with hidden representations produced by the Longformer is passed as input to the token-wise attention layer, which produces label-wise attention representations of the document for each

label. We used a binary classifier (Feed forward layer) for each of these label-wise vector representations to predict the ICD codes probability using a sigmoid transformation.

5.2. Long-HiLAT architecture:

The architecture of our proposed model extends the Hierarchical Label wise attention transformer's (HiLAT) [16] architecture. Figure 6 provides an overview of the Long-HiLAT model. The clinical text is split into meaningful chunks such that each chunk contains a maximum of 512 tokens. The Long-HiLAT model has 5 components- the pretrained transformer layer, the Longformer encoder layer, the label-wise attention layer, the chunk attention layer, and a classifier layer which has multiple feed forward neural networks, which serves as a classification head for the model. Most of the skeletal code has been referenced from the paper HiLAT, with customizations, implemented in Pytorch.

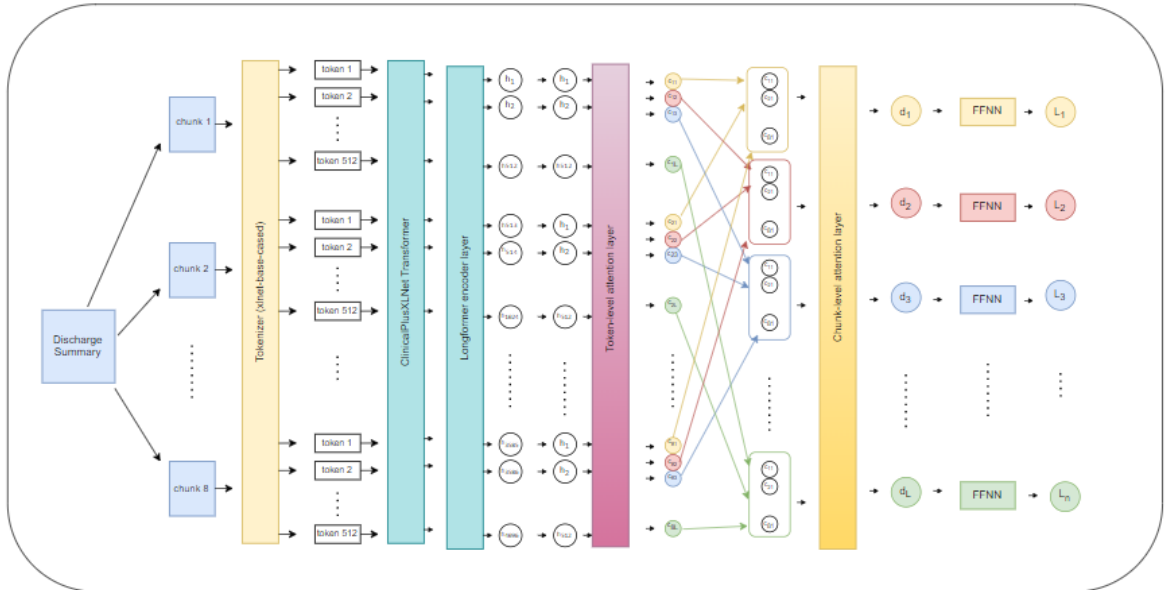


Figure 6. Architecture of Long-HiLAT

The Long-HiLAT implementation comprises of the following 5 layers -

i) The Pretrained Transformer:

For convenience, in this project, the text data is split into 8 chunks each not exceeding a length of 512, and tokenized using the “xlnet-base-cased” tokenizer. The documents are truncated to a max of 4096 tokens. The truncation is done chunk-wise meaning any chunk exceeding 512 tokens will be truncated avoiding loss of data in the last chunk. Every chunk is appended by [CLS] and [SEP] tokens. The tokens are input to the transformer layer, which has 12 pretrained xlnet encoder layers. The output will be in the form of chunkwise hidden representations.

ii) Longformer Encoder Layer:

The chunkwise representations are transformed to a single long representation adaptable to the Longformer and input to the Longformer encoder layer which has the 2 (configurable) Longformer encoder layers. In this implementation, we have limited the encoder layers to 2 due to memory constraints. With this, we incorporate the advantages of Longformer’s sliding window attention and further enhance the attention representations of the model. The outputs of the Longformer encoder layer are token representation vectors which will be separated to chunkwise representations and sent to the token-level attention layer.

iii) Token-level attention layer:

We use the label attention mechanism discussed earlier for implementing the token-level attention mechanism, but we do this for every chunk of data, forming a matrix C . To generate the input for the chunk-wise attention layer, we extract the l^{th} column (label-wise representations) of each matrix to form a new matrix M , which is then passed to the chunk-wise attention layer.

iv) Chunk-level attention layer:

We use the same attention mechanism as the previous layer, where we input M_l to produce document representations for each label l . The following equations are the operations implemented.

$$S_l = \tanh(KM_l) \dots\dots\dots(4)$$

$$o_l = \text{softmax}(v^T S_l) \dots\dots\dots(5)$$

$$d_l = M_l o_l^T \dots\dots\dots(6)$$

Here, K is used to multiply with M_l to produce S_l using the hyperbolic tangent activation function. The chunk-level attention weight vector o_l for label l is computed using a randomly initialized vector v and the matrix S_l with a softmax function. Finally, the document representation vector d_l for label l is produced using the equation (6)

v) The Feedforward Neural network layer:

When chunk attention is employed, we use the output vectors having document representations for each label and pass it through a feed forward layer which acts as a binary classifier for each label and use sigmoid to calculate the probability. We use a threshold of 0.5 to predict the binary codes for each label. In the training, we minimize loss while aiming to improve micro F1.

We also compute the token-wise and chunk-wise attention weights to demonstrate the explainability of the model.

6. EXPERIMENTAL VALIDATION

6.1. Training:

All model training and evaluations were done on the L-40 NVIDIA GPU. All of the models were trained for 10 epochs in batch sizes of 2 and 4 at a learning rate of $5e-5$. For comparability, we evaluated the performance of other baseline models with the same batch sizes and the Long-HiLAT has outperformed all other models. The evaluation metrics and the results are discussed here. All our fine-tuned models are available for use at <https://huggingface.co/meghanaraok>.

6.2. Performance metrics:

- a. **Micro-F1:** Micro-F1 is a composite metric that considers both precision and recall, calculated at the micro level across all classes. It provides an overall assessment of the model's performance in terms of correctly identifying positive and negative instances across all classes. Particularly useful when dealing with class imbalance, as it weighs each instance equally.
- b. **AUC:** AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate. It provides a single scalar value that represents the model's ability to distinguish between positive and negative classes across various thresholds. AUC is insensitive to class imbalance and provides a comprehensive evaluation of the model's performance across different operating points.
- c. **Accuracy:** Accuracy measures the proportion of correctly classified instances out of the total instances. While accuracy is intuitive and easy to interpret, it might not be suitable for imbalanced datasets, where a high accuracy can be achieved by simply predicting the majority class.

- d. **Precision:** Precision quantifies the ratio of correctly predicted positive instances to the total predicted positive instances. It reflects the model's ability to avoid false positives, i.e., instances that were predicted positive but are actually negative. Precision is crucial in scenarios where false positives are costly or undesirable, such as medical diagnosis or fraud detection.
- e. **Recall:** Recall calculates the ratio of correctly predicted positive instances to the total actual positive instances. It measures the model's ability to capture all positive instances without missing any (minimizing false negatives). Recall is important in situations where detecting all positive instances is paramount, such as disease detection or anomaly detection.
- f. **Labelwise F1 Score:** Labelwise F1 Score computes the F1 score for each class individually and then averages them, providing insights into the model's performance on each class. It is particularly useful when classes are imbalanced or when the performance of individual classes is of interest. Helps in identifying which classes are well predicted by the model and which ones need improvement.

6.3. Results comparison:

We trained all of the models on top-5 and top-50 ICD-9 code datasets. The following tables show a performance comparison of the investigated baseline and proposed models.

i) Top-5 ICD codes:

Model	F1	AUC	Accuracy	Precision	Recall
HiLAT	83.6	93.0	71.9	83.7	83.6
Long-LAT	84.5	93.5	73.2	84.9	84.1
Clinical Longformer	84.6	87.9	73.3	83.1	86.2
Long-HiLAT	85.3	93.6	74.3	84.6	85.9

Table 2a. Performance comparison for top-5 ICD codes

Label	HiLAT	Clinical Longformer	Long-LAT	Long-HiLAT
38.93	66.6	70.5	66.6	71.1
401.9	85.6	86.8	85.6	87.7
414.01	86.1	85.6	85.7	85.7
427.31	91.3	91.2	91.2	91.9
428	85.2	86.4	85.1	86.4

Table 2b. Labelwise F1 score comparison for top-5 ICD codes

Table 2a provides the evaluation metrics of all the experimented models. For top-5 labels Long-HiLAT produced significant improvement in performance scores with a 2% increase in micro F1 score. Long-HiLAT also achieved the highest accuracy and AUC scores indicating best performance. Longformer-based models- Clinical Pretrained transformer and Long-LAT also achieve higher performance on top-5 ICD-9 datasets.

Table 2b provides the Labelwise F1 scores for each label. Long-HiLAT model again achieves higher scores than other models for most of the labels.

ii) Top-50 ICD codes:

Table 3a provides the performance comparison metrics for all the models. Long-HiLAT achieves the highest F1, AUC, and Accuracy. For top-50 ICD codes.

Long-LAT also achieved competitive results and can be improved when trained with larger batch sizes eliminating noisy gradients with larger models. The Clinical Pretrained Longformer has scored the least F1 and a high precision.

Model	F1	AUC	Accuracy	Precision	Recall
HiLAT	71.6	94.2	55.7	71.2	71.0
Clinical Longformer	64.5	76.9	47.6	75.6	56.3
Long-LAT	71.1	93.7	55.2	71.2	71.0
Long-HiLAT	72.4	94.4	56.8	73.0	71.8

Table 3a. Performance comparison for top-50 ICD codes

Table 3b provides the Label-wise F1 scores for all the top 50 ICD codes. Long-HiLAT achieved better Label-wise F1 scores for most labels.

ICD Code	HiLAT	Clinical Longformer	Long-LAT	Long-HiLAT
38.9	61.3	48.6	58.9	59.9
244.9	84.9	82.8	87.2	85.9
250	79.9	68.1	77.4	80.6
272	64.3	25.8	58.9	64.9
272.4	79.9	78.2	78.7	79.9
276.1	54.2	0.0	55.6	56.3
276.2	44.5	12.9	46.2	49.1
285.1	60.1	58.2	59.1	60.8
285.9	3.1	0.0	0.8	0.00
287.5	46.3	1.4	49.0	50.8
305.1	66.3	1.1	62.3	63.3
311	59.0	0.0	63.9	63.7
33.24	70.8	48.6	71.1	74.1
36.15	97.8	95.0	97.4	97.4
37.22	69.4	61.9	66.4	69.5
38.91	37.0	13.9	25.0	30.7
38.93	61.0	57.5	59.2	58.4
39.61	97.8	96.5	97.4	98.0
39.95	88.0	88.1	92.1	87.7
401.9	83.1	81.5	83.9	84.5
403.9	72.5	62.2	75.7	74.6
410.71	71.0	56.9	66.7	70.5
412	69.6	19.7	65.6	71.2
414.01	86.8	84.0	85.6	86.8
424	70.5	15.4	64.2	72.7
427.31	91.9	90.9	91.1	91.9
428	84.8	84.1	81.8	83.9
45.13	67.0	47.8	69.7	69.7
486	66.1	52.0	64.2	71.4
496	74.7	48.5	67.9	75.1
507	62.0	60.9	64.3	66.7
511.9	44.2	0.0	44.3	48.7
518.81	69.9	59.3	67.5	68.7
530.8	75.3	71.9	78.0	78.9
584.9	68.1	69.0	70.9	70.7
585.9	61.3	50.9	64.5	60.5
599	74.3	72.0	73.3	74.6
88.56	85.4	85.3	87.9	85.0
88.72	42.5	26.1	48.5	44.1
93.9	12.8	0.0	5.4	9.8
96.04	70.6	66.4	70.9	71.4
96.6	73.2	70.2	74.1	73.7
96.71	68.7	65.6	69.4	69.3
96.72	73.9	71.6	70.1	72.4
99.04	20.9	9.9	16.7	8.3
99.15	65.9	74.1	73.8	75.9
995.92	70.3	66.9	68.2	70.3
V15.82	3.1	0.0	4.1	2.1
V45.81	85.3	83.8	81.2	86.0
V58.61	69.7	60.6	70.9	69.7

Table 3b. Labelwise F1 scores for top-50 ICD codes

7. MODEL EXPLAINABILITY

Model explainability is crucial for building trust in machine learning systems, as it provides insight into how decisions are made, helps identify biases and errors, generates valuable insights, ensures regulatory compliance, facilitates debugging and improvement efforts, enhances user experience, and promotes collaboration and education.

We compute the label-wise and chunk-wise attention weights for models Long-HiLAT and label-wise attention weights for Long-LAT. These attention weights can be used to highlight the most relevant chunks and words in the document for each predicted label, explaining the assignment of the specific code to the text.

To obtain the global contribution of each token, we compute the global token attention weights for each chunk and label by combining the token-level attention weights with the corresponding chunk-level attention weight. Since tokens in Transformer models may not directly correspond to words due to tokenization, we sum the attention weights of tokens belonging to the same word to compute word attention weights. This process involves adding up the global token attention weights associated with each word and then normalizing the sum to obtain the final attention weights for visualization. Figure 7 gives a sample visualization highlighting important tokens in the text.

Code: 38.91 - Arterial catheterization

service medicine allergies nsaid sulfa sulfonamide antibiotics attending chief complaint transfer from neurosurg to micu for acute renal failure major surgical or invasive procedure ivc filter placement central line placement arterial line placement hemodialysis intubation mechanical ventilation history of present illness 85m with prior dvt htn and ckd was admitted to nebh with decreased appetite and le swelling found to have extensive dvt and acute on chronic rf was started on heparin gtt and yesterday was noted to have a right facial droop and increased dysarthria r sided weakness and somnolence he developed what appeared to be a r sided seizure and then a grand mal seizure in the ct scanner at the osh he was intubated for airway protection and transferred to the neurosurgery service he was noted to be hypotensive after intubation without sedation prior to transfer and was started on neo an a line was placed also prior to transfer this morning the neurosurgery attending asked that the micu take over his care given the complexity of his medical problems on eval he was intubated and sedated does not follow commands not on sedation although received mg of iv ativan within the past hours for possible seizure per his son who is at his bedside he was doing well until about months ago at which point they noticed a pound weight loss and hematuria bladder cancer was discovered and he had a cystoscopic removal of tumor weeks ago his son noted that he was increasingly tired w decreased appetite and le swelling he fell and hit his head about week ago but his son noticed only a small cut and so did not have him evaluated over the week prior to admission he became unable to walk and needed a wheelchair to get around past medical history htn thoracic and abdominal aortic aneurysm h o transitional cell bladder cancer ckd h o

Figure 7. Sample attention visualization example for ICD Code 38.91

8. WEB UI FOR ICD CODE PREDICTION

Overview

Our Web UI provides an interactive interface for users to classify ICD codes based on text summaries. The key features of the UI include:

- 1. Sample Summaries:** The UI provides a set of sample summaries that the user can select from. This allows the user to quickly test the ICD code classification without having to input their own text.
- 2. Custom Summary Input:** The UI also includes a text input box where the user can enter their own custom summary for classification.
- 3. Model Selection:** The UI offers 4 different ICD code classification models that the user can choose from. This allows the user to compare the performance of different models.
- 4. Top-k Prediction Selection:** The user can select whether they want the UI to predict the top-5 or top-50 ICD codes for the selected model and summary.
- 5. Attention Visualization:** The UI displays the attention weights in the text that are responsible for the ICD code predictions. This provides the user with insight into which parts of the summary are most important for the classification.

Workflow

The user interacts with the Web UI as follows:

1. The user selects one of the sample summaries provided, or enters their own custom summary in the text input box.
2. The user selects which of the 4 ICD code classification models they would like to use.
3. The user selects whether they want the codes from top-5 or top-50 ICD codes predicted.

4. The UI then displays the predicted ICD codes along with the attention weights in the input text that contributed to the predictions.

This workflow allows the user to rapidly test and compare the performance of different ICD code classification models on a variety of text summaries, while also providing insight into the model's decision-making process through the attention visualization.

Improvised Transformer-based approach for ICD 9 code prediction

Select a model and use a sample Discharge Summary or copy your Discharge summary to predict the relevant ICD 9 codes.

The web interface is titled "Improvised Transformer-based approach for ICD 9 code prediction". It includes a sub-header "Select a model and use a sample Discharge Summary or copy your Discharge summary to predict the relevant ICD 9 codes." The interface is divided into several sections:

- Select Sample Text:** A dropdown menu with the instruction "Use the selected sample text to Predict relevant ICD9 codes".
- Use Sample Text:** A grey button.
- Enter Medical summary:** A text input field with a placeholder and a double-slash icon.
- Model:** A section with the instruction "Select the model used for prediction" and four radio button options: "HILAT", "ClinicalLongformer", "Long-LAT", and "Long-HiLAT".
- ICD Code Prediction Range:** A section with the instruction "Select the number of top ICD codes to predict: either the top 5 or the top 50" and two radio button options: "5" and "50".
- Predict:** A grey button at the bottom.

Figure 8. Web interface for ICD Code prediction

Implementation details:

Our Web UI is built using the Gradio Python library, which provides a simple and flexible way to create interactive web-based user interfaces for machine learning models.

The attention visualization is implemented by leveraging the attention mechanisms built into the underlying neural network models. The attention weights are extracted and displayed

alongside the predicted ICD codes to give the user a deeper understanding of the model's reasoning.

Our UI provides a user-friendly and informative interface for exploring and evaluating ICD code classification models explored in this project.

It is deployed on Huggingface spaces at <https://huggingface.co/spaces/meghanaraok/LongLAT>

9. CASE STUDY

Test Case: We consider the test case with the following Sample Summary and experiment with all models selecting the top-5 labels.

Sample text:

date of birth: sex: f service medicine allergies levofloxacin attending chief complaint dyspnea pneumonia major surgical or invasive procedure et tube change arterial line placement right ij line placement ir guided picc placement trach placement peg placement picc removed for fungemia single lumen picc placed history of present illness year old woman with history of asthma copd dm hypothyroidism with recent history significant for worsening dyspnea over past three months status post four courses of antibiotics and steroids for presumed copd exacerbation presenting to hospital on with acute worsening dyspnea intubated for respiratory distress and transferred to for further management as noted above she has a history of worsening dyspnea over the past few months that has been treated with antibiotics and steroids she presented to osh on with worsening dyspnea and had a fever to 103f and was started on ceftriaxone azithromycin for pneumonia and copd exacerbation she developed an increasing oxygen requirement however and initially required nasal canula on l o2 but later required face mask on l on her cxr which initially was read as negative for infiltrate progressively worsening with increasingly prominent diffuse bilateral infiltrates right greater than left she also had a negative chest cta with an oxygen requirement that was rising and worsening dyspnea she was transitioned to vanc zosyn levo and then vanc zosyn ceftaz ceftaz started levo on because of an allergy to levofloxacin it is unclear why she received double coverage for gram negatives she was also given methylprednisolone dose mg iv q8 then mg iv q8 afterward on she was admitted to the icu and intubated electively for hypoxia and sob sputum culture grew staph aureas that was mrsa per report influenza viral screen was negative phenylephrine was started because of hypotension to sbps in the 80s after being intubated and starting on propofol which had to be uptitrated for sedation she was transferred to on based on her family s wishes on arrival her ventilator settings were vt 500cc fio2 peep she was on phenylephrine at and rapidly weaned off her vs were vent past medical history copd asthma diabetes mellitus hyperlipidemia hypothyroidism gerd left breast cancer s p lumpectomy h syndrome per pcp social history she does not smoke or drink she lives with her husband family history non contributory physical exam vitals t bp p r o2 general intubated obese arousable heent ett sclera anicteric mmm oropharynx clear neck supple jvp not elevated no lad lungs soft inspiratory wheezes bilaterally no rales or rhonchi cv regularly irregular rate and rhythm normal s1 s2 no murmurs rubs gallops abdomen soft non tender non distended bowel sounds present no rebound tenderness or guarding no organomegaly ext warm well perfused pulses no clubbing cyanosis or edema pertinent results admission labs 36pm type art temp rates tidal vol peep o2 po2 pco2 ph total co2 base xs aado2 req o2 intubated intubated 22pm glucose urea n creat sodium potassium chloride total co2 anion gap 22pm calcium phosphate magnesium 22nm wbc rbc hgb hct mcv mch mchc rdw 22nm plt count 22nm pt ptt inr

True Labels:

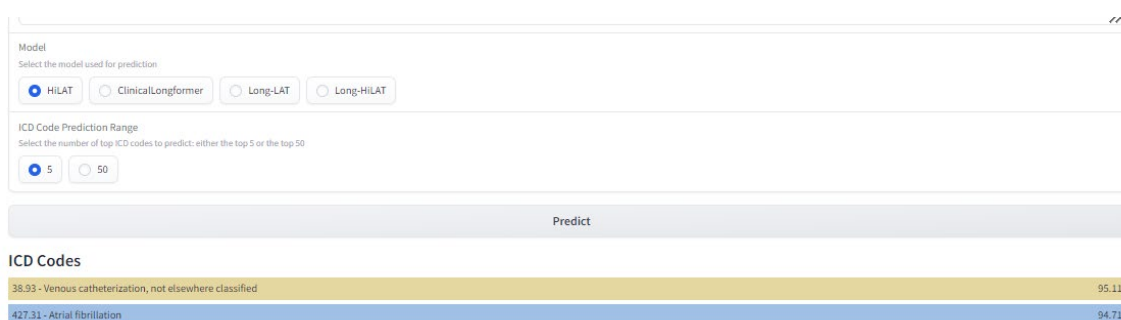
The assigned ICD codes are:

1. **38.93** - Venous catheterization, not elsewhere classified
2. **427.31** - Atrial fibrillation
3. **428.0** - Congestive heart failure, unspecified

1. Predictions:

The following images show the predictions produced by the selected models.

For the given input text, HiLAT, Clinical Longformer, Long-LAT predicted the labels 38.93 and 427.31 but failed to predict the label 428.0. Long-HiLAT successfully predicts all the 3 labels associated with the given summary.



Model
Select the model used for prediction

☒ HiLAT ☐ ClinicalLongformer ☐ Long-LAT ☐ Long-HiLAT

ICD Code Prediction Range
Select the number of top ICD codes to predict: either the top 5 or the top 50

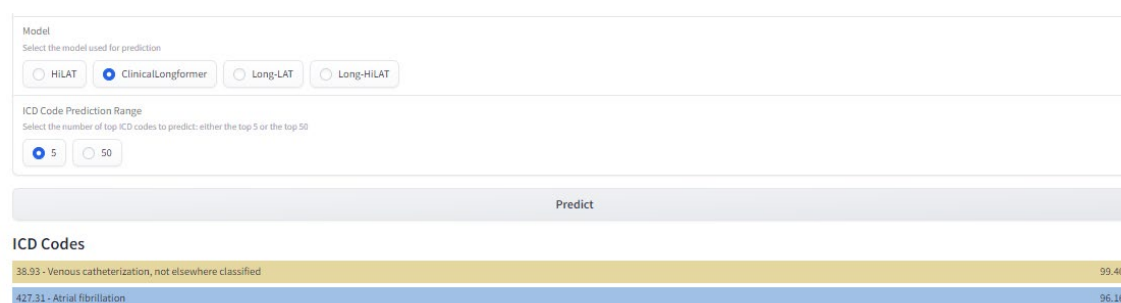
☒ 5 ☐ 50

Predict

ICD Codes

38.93 - Venous catheterization, not elsewhere classified	95.11%
427.31 - Atrial fibrillation	94.71%

Figure 9. HiLAT sample predictions for top-5 labels



Model
Select the model used for prediction

☐ HiLAT ☒ ClinicalLongformer ☐ Long-LAT ☐ Long-HiLAT

ICD Code Prediction Range
Select the number of top ICD codes to predict: either the top 5 or the top 50

☒ 5 ☐ 50

Predict

ICD Codes

38.93 - Venous catheterization, not elsewhere classified	99.40%
427.31 - Atrial fibrillation	96.16%

Figure 10. Clinical Longformer sample predictions for top-5 labels

Model
Select the model used for prediction:

☐ HiLAT ☐ ClinicalLongformer ☒ Long-LAT ☐ Long-HiLAT

ICD Code Prediction Range
Select the number of top ICD codes to predict: either the top 5 or the top 50

☒ 5 ☐ 50

Predict

ICD Codes

38.93 - Venous catheterization, not elsewhere classified	100.00%
427.31 - Atrial fibrillation	99.97%

Figure 11. Long-LAT sample predictions for top-5 labels

Model
Select the model used for prediction:

☐ HiLAT ☐ ClinicalLongformer ☐ Long-LAT ☒ Long-HiLAT

ICD Code Prediction Range
Select the number of top ICD codes to predict: either the top 5 or the top 50

☒ 5 ☐ 50

Predict

ICD Codes

38.93 - Venous catheterization, not elsewhere classified	98.90%
427.31 - Atrial fibrillation	98.71%
428.0 - Congestive heart failure, unspecified	84.02%

Figure 12. Long-HiLAT sample predictions for top-5 labels

2. Explainability:

Figure 13 and 14 gives the explainability of models HiLAT and Long-HiLAT for the ICD code 38.93.

but transitioned back to iss with glargine with good control hypothyroidism continued levothyroxine communication patient husband is h son is c medications on admission singulair mg qhs prevacid mg daily levothyroxine mg daily d flonase sprays eat nostril daily advair on inh albuterol nebs prn janumet metformin and sitagliptin discharge medications lansoprazole mg capsule delayed release e c sig one capsule delayed release e c po once a day levothyroxine mcg tablet sig one tablet po daily daily ipratropium bromide mcg actuation aerosol sig puffs inhalation q4h every hours albuterol sulfate mcg actuation hfa aerosol inhaler sig two puff inhalation q4h every hours as needed for shortness of breath or wheezing aspirin mg tablet sig one tablet po daily daily meropenem mg recon soln sig five hundred mg intravenous q8h every hours for days last dose on fluconazole in saline iso osm mg ml piggyback sig four hundred mg intravenous once a day for days last dose on fentanyl mcg hr patch hr sig one patch hr transdermal q72h every hours for days please remove am of fentanyl mcg hr patch hr sig one patch hr transdermal q72h every hours for days please place amiodarone mg tablet sig one tablet po daily daily metoprolol tartrate mg tablet sig tablet po bid times a day nystatin unit ml suspension sig five ml po qid times a day as needed for sores acetaminophen mg tablet sig tablets po q6h every hours as needed for fever chlorhexidine gluconate mouthwash sig fifteen ml mucous membrane times a day docusate sodium mg ml liquid sig one hundred ml po bid times a day hold for loose stools senna mg tablet sig tablets po bid times a day as needed for constipation lactulose gram ml syrup sig thirty ml po tid times a day as needed for constipation insulin glargine unit ml cartridge sig forty units subcutaneous once a day insulin regular human unit ml insulin pen sig sliding scale as below subcutaneous every six hours adjust as needed amp d50 units units units

Figure 13. Attention visualization of model HiLAT for ICD code 38.93.

sample was sent to the state lab and there it was positive for h1n1 the id service was consulted and recommended switching from vancomycin to linezolid and empirically completing a course of oseltamivir of note she also underwent a repeat chest cta that was negative for pe and a tte with a bubble study that was limited in quality but negative for intracardiac shunt she continued to require high peep with hypoxia if peep was lower than diuresis was tried several times with minimal change in peep linezolid was switched back to vancomycin given concern for lactic acidosis id then recommended switching to meropenem to cover resistant pseudomonal vap which she continued for a day course a **picc** line **was placed on for long term** antibiotics **oseltamivir** was discontinued on vanco was discontinued gradually able to wean down peep with improvement in bronchospastic episodes on increased sedation and a trach **was placed by ip on sedatives** eventually weaned off on and transitioned to **fentanyl** gtt and valium which too were weaned off to a mcg patch q72 hours this should be weaned further at her facility under the direction of the accepting md patient was transitioned to lasix iv boluses until cr and bun bumped then prn to keep her i os even although sputum cx with persistence of pseudomonas per id this was thought to represent colonization pt finished her day course of meropenem on **single lumen **picc** was placed after a day line holiday** afib with rvr patient developed af with rvr and hypotension chads score no evidence of pe on cta chest or right heart strain on echo normal atria she was started on hep gtt and amio loaded she remained in sinus rhythm for the majority of the rest of her stay aside from one further bout of afib heparin was discontinued at time of trach placement long term anticoagulation was not started as she is considered low risk s p conversion she was continued on amidarone and asa volume overload the patient was volume overloaded after about week in the icu lasix gtt was started to have a smoother diuresis as bolus doses of 40mg iv lasix made her transiently

Figure 14. Attention visualization of model Long-HiLAT for ICD code 38.93

We see that Long-HiLAT pays more attention to the document text across multiple chunks of the document along with locally available contributing tokens. Fig 15 and 16 visualize the importance of text tokens of models HiLAT and Long-HiLAT for the ICD code – 427.31

medications lansoprazole mg capsule delayed release e c sig one capsule delayed release e c po once a day
levothyroxine mcg tablet sig one tablet po daily daily ipratropium bromide mcg actuation aerosol sig puffs inhalation
q4h every hours albuterol sulfate mcg actuation hfa aerosol inhaler sig two puff inhalation q4h every hours as needed for
shortness of breath or wheezing aspirin mg tablet sig one tablet po daily daily meropenem mg recon soln sig five
hundred mg intravenous q8h every hours for days last dose on fluconazole in saline iso osm mg ml piggyback sig four
hundred mg intravenous once a day for days last dose on fentanyl mcg hr patch hr sig one patch hr transdermal q72h
every hours for days please remove am of fentanyl mcg hr patch hr sig one patch hr transdermal q72h every hours for
days please place amiodarone mg tablet sig one tablet po daily daily metoprolol tartrate mg tablet sig tablet po bid
times a day nystatin unit ml suspension sig five ml po qid times a day as needed for sores acetaminophen mg tablet sig
tablets po q6h every hours as needed for fever chlorhexidine gluconate mouthwash sig fifteen ml mucous membrane
times a day docusate sodium mg ml liquid sig one hundred ml po bid times a day hold for loose stools senna mg tablet
sig tablets po bid times a day as needed for constipation lactulose gram ml syrup sig thirty ml po tid times a day as
needed for constipation insulin glargine unit ml cartridge sig forty units subcutaneous once a day insulin regular human
unit ml insulin pen sig sliding scale as below subcutaneous every six hours adjust as needed amp d50 units units units
units units units units units notify m d heparin porcine unit ml solution sig units injection tid times a day discharge

Figure 15. Attention visualization of model HiLAT for ICD code 427.31

recommended switching to meropenem to cover resistant pseudomonas vap which she continued for a day course a picc line was placed on for long term antibiotics oseltamivir was discontinued on vanco was discontinued gradually able to wean down peep with improvement in bronchospastic episodes on increased sedation and a trach was placed by ip on sedatives eventually weaned off on and transitioned to fentanyl gtt and valium which too were weaned off to a mcg patch q72 hours this should be weaned further at her facility under the direction of the accepting md patient was transitioned to lasix iv boluses until cr and bun bumped then prn to keep her i os even although sputum cx with persistence of pseudomonas per id this was thought to represent colonization pt finished her day course of meropenem on single lumen picc was placed after a day line holiday afib with rvr patient developed af with rvr and hypotension chads score no evidence of pe on cta chest or right heart strain on echo normal atria she was started on hep gtt and amio loaded she remained in sinus rhythm for the majority of the rest of her stay aside from one further bout of afib heparin was discontinued at time of trach placement long term anticoagulation was not started as she is considered low risk s p conversion she was continued on amiodarone and asa volume overload the patient was volume overloaded after about week in the icu lasix gtt was started to have a smoother diuresis as bolus doses of 40mg iv lasix made her transiently hypotensive she was restarted on lasix gtt in setting of pressors which were eventually weaned off she was then transitioned to iv lasix boluses at 160mg iv tid until cr and bun bumped then transitioned to prn lasix boluses to keep i os even of note she was on acetazolamide for a few days due to metabolic alkosis but this was discontinued as bicarb normalizing please give iv lasix 160mg prn to keep fluid balance even fungemia patient began to spike fevers almost one

Figure 16. Attention visualization of model Long-HiLAT for ICD code 427.31

Both models assign high attention to the drugs administered to the patient for the predicted ICD code 427.31.

wean down peep with improvement in bronchospastic episodes on increased sedation and a trach was placed by ip on sedatives eventually weaned off on and transitioned to fentanyl gtt and valium which too were weaned off to a mcg patch q72 hours this should be weaned further at her facility under the direction of the accepting md patient was transitioned to lasix iv boluses until cr and bun bumped then prn to keep her i os even although sputum cx with persistence of pseudomonas per id this was thought to represent colonization pt finished her day course of meropenem on single lumen picc was placed after a day line holiday afib with rvr patient developed af with rvr and hypotension chads score no evidence of pe on cta chest or right heart strain on echo normal atria she was started on hep gtt and amio loaded she remained in sinus rhythm for the majority of the rest of her stay aside from one further bout of afib heparin was discontinued at time of trach placement long term anticoagulation was not started as she is considered low risk s p conversion she was continued on amiodarone and asa volume overload the patient was volume overloaded after about week in the icu lasix gtt was started to have a smoother diuresis as bolus doses of 40mg iv lasix made her transiently hypotensive she was restarted on lasix gtt in setting of pressors which were eventually weaned off she was then transitioned to iv lasix boluses at 160mg iv tid until cr and bun bumped then transitioned to prn lasix boluses to keep i os even of note she was on acetazolamide for a few days due to metabolic alkosis but this was discontinued as bicarb normalizing please give iv lasix 160mg prn to keep fluid balance even fungemia patient began to spike fevers almost one month into her hospitalization she was started on oral fluconazole on due to persistent yeast positive urine cultures despite changing out foley catheters rij placed was pulled however blood cultures from a line grew out yeast

Figure 17. Attention visualization of Long-HiLAT for ICD 428.0

Figure 17 illustrates the text visualization generated by the Long-HiLAT model for the ICD code 428.0, a prediction that couldn't be made by the HiLAT model. The visualization showcases the attention weights assigned by the Long-HiLAT model to different parts of the clinical text. Notably, the model assigns high attention weights to the drug Lasix, commonly administered to

patients with heart failure. This attention to Lasix indicates the model's understanding of the clinical context and the relevance of this medication to the prediction of ICD code 428.0. The observed attention patterns reinforce the correctness and robustness of the Long-HiLAT model's predictions, highlighting its capability to capture intricate relationships within clinical text data for accurate diagnosis and classification.

10. CONCLUSION

In this project, we explored the impact of Longformer-based models for ICD code classification using lengthy clinical free-text notes such as discharge summaries. We proposed 2 models Long-LAT and Long-HiLAT utilizing the Longformer. Extensive experiments conducted on the MIMIC-iii datasets demonstrate that our hybrid model, Long-HiLAT outperformed the all the baseline models on both top-5 and top-50 ICD-9 code datasets. We designed a web UI for experimenting the models discussed in this paper, where user can select the models and the range for predicting ICD codes giving a sample text summary.

REFERENCES

- [1] R. Kaur, J. A. Ginige, & O. Obst, (2023). "AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review," in *Expert Systems with Applications*, 213, 118997.
- [2] L R. S. de Lima, H. F. Laender, and B A. Ribeiro-Neto. 1998. "A hierarchical approach to the automatic categorization of medical documents," *Proceedings of the Seventh International Conference on Information and Knowledge Management (CIKM '98)*. Association for Computing Machinery, New York, NY, USA, 132–139.
<https://doi.org/10.1145/288627.288649>.
- [3] H. Shi, P. Xie, Z. Hu, M. Zhang, E.P. Xing, Towards automated ICD coding using deep learning, 2017, Preprint at: <https://arxiv.org/abs/171104075>
- [4] S. Wiegrefe and Y. Pinter. 2019, "Attention is not not Explanation," In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- [5] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, & J. Zou, (2019). "Gradio: Hassle-free sharing and testing of ML models in the wild [Computer software], "
<https://doi.org/10.48550/arXiv.1906.02569>
- [6] R. Qin, M. Huang, J. Liu and Q. Miao, "Hybrid Attention-based Transformer for Long-range Document Classification," *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9891918.
- [7] T. Vu, D.Q. Nguyen, A. Nguyen, A label attention model for ICD coding from clinical text, 2020, Preprint at: <https://arxiv.org/abs/2007.06351>.

- [8] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, J. Eisenstein, "Explainable prediction of medical codes from clinical text," 2018, Preprint at: <https://arxiv.org/abs/180205695>.
- [9] V. Mayya, S. Kamath, G.S. Krishnan, T. Gangavarapu, "Multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries," *Future Gener. Comput. Syst.*, 118 (2021), pp. 374-391.
- [10] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, N. Elhadad, "Multi-label classification of patient notes a case study on ICD code assignment," 2017. Preprint at: <https://arxiv.org/abs/170909587>.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [12] B. Biseda, G. Desai, H. Lin, A. Philip, "Prediction of ICD Codes with Clinical BERT Embeddings and Text Augmentation with Label Balancing using MIMIC-III," *arXiv* 2020, arXiv:2008.10492.
- [13] 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*, 36(4):1234–1240, 2020.
- [14] Z. Zhang, J. Liu, N. Razavian, BERT-XML: large scale automated ICD coding using BERT pretraining, 2020, Preprint at: <https://arxiv.org/abs/200603685>.
- [15] D. Pascual, S. Luck, R. Wattenhofer, Towards BERT-based automatic ICD coding: limitations and opportunities, 2021, Preprint at: <https://arxiv.org/abs/210406709>.
- [16] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, L. Jorm, "Hierarchical label-wise attention transformer model for explainable ICD coding," *Journal of Biomedical Informatics*, Vol 133, 2022, 104161, doi: <https://doi.org/10.1016/j.jbi.2022.104161>.

- [17] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509, 2019.
- [18] J. Ainslie, S. Ontañón, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang, and L. Yang, "ETC: encoding long and structured inputs in transformers," In EMNLP, 2020.
- [19] Z. Liu, Y. T. Lin, Y. Cao, H. Hu, B. N. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022, 2021.
- [20] M. Zaheer et al., "Big bird: Transformers for longer sequences", Proc. Conf. Neural Informat. Process. Syst., pp. 17283-17297, 2020.
- [21] M. Feucht, Z. Wu, S. Althammer, V. Tresp, "Description-based label attention classifier for explainable ICD-9 classification," 2021, doi: <https://arxiv.org/abs/2109.12026>
- [22] A.E. Johnson, T.J. Pollard, L. Shen, L. W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35. PMID: 27219127; PMCID: PMC4878278.
- [23] Y. Li, M. W. Ramsey, S. A. Faraz, H. Wang, and Y. Luo. 2022. "Clinical-longformer and clinical-BigBird: Transformers for long clinical sequences," arXiv:2201.11838. Retrieved from <https://arxiv.org/abs/2201.11838>.
- [24] S Abraham, S. Levine-Gottreich, "Clinical-Longformer: Whole Document Embedding and Classification for the Clinical Domain" <https://github.com/simonlevine/clinical-longformer>.
- [25] I. Beltagy, M.E. Peters, A. Cohan, Longformer: the long-document transformer, 2020, Preprint at: <https://arxiv.org/abs/2004.05150>.

- [26] M. T. Chiaravalloti, R. Guarasci, V. Lagani, E. Pasceri and R. Trunfio, "A Coding Support System for the ICD-9-CM Standard," 2014 IEEE International Conference on Healthcare Informatics, Verona, Italy, 2014, pp. 71-78, doi: 10.1109/ICHI.2014.1
- [27] K. Crammer, M. Dredze, K. Ganchev, P. Talukdar, S. Carroll, "Automatic code assignment to medical text," Biological, Translational, and Clinical Language Processing, ACL (2007), pp. 129-136.
- [28] J Edin, A Junge, J D. Havtorn, L Borgholt, M Maistro, T Ruotsalo, and L Maaløe. 2023. Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study, Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23). Association for Computing Machinery, New York, NY, USA, 2572–2582.
Doi: <https://doi.org/10.1145/3539618.3591918>.
- [29] K. Huang, J. Altosaar, and R. Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission," arXiv preprint arXiv:1904.05342, 2019.
- [30] F. Jaume-Santero, B. Zhang, D. Proios, A. Yazdani, R. Gouareb, M. Bjelogrić, D. Teodoro, "Cluster analysis of low-dimensional medical concept representations from electronic health records," In: International conference on health information science. 2022.
- [31] A. Singaravelan, C. H. Hsieh, Y. K. Liao, Y.-K, J. L. Hsu, "Predicting ICD-9 Codes Using Self-Report of Patients," *Appl. Sci.* **2021**, *11*, 10046.<https://doi.org/10.3390/app112110046>
- [32] C. Sen, B. Ye, J. Aslam, A. Tahmasebi. "From Extreme Multi-label to Multi-class: A Hierarchical Approach for Automated ICD-10 Coding Using Phrase-level Attention," arXiv preprint arXiv:210209136. 2021.
- [33] S. Jain, & B. C. Wallace, (2019). Attention is not Explanation. Doi: ArXiv./abs/1902.10186.