

Synonym Based Hierarchical Probability Transformer For Predicting Disease Intensity And Relationships Between Icd Codes

Neena Joseph¹ , Vinodh P Vijayan²

¹ Department of CSE, Mangalam College of Engineering, Ettumanoor
nestofneena@gmail.com [0009-0007-1657-0744]

² Department of CSE, Mangalam College of Engineering, Ettumanoor
vinodhpvijayan81@gmail.com [0000-0002-9992-8612]

Abstract

Predicting disease intensity levels and identifying relationships between the International Classification of Disease codes utilizing Large Language Models (LLMs) and Transformer Learning investigate and classify immense amounts of clinical data to acquire awareness into disease advancement and possible associations between conditions. Precisely, LLMs explore route of care for a specific patient population on a specific condition represented as series of medical codes paving the way for diagnosis, predict intensity and ascertain possible associations between diseases. Early identification as well as deterrence is pivotal for minimizing hazard of disease evolution. This study intended to introduce a novel method depend on LLM and transformer to predict disease intensity and explore the relationship between ICD codes for predicting disease progression. In this work a method called, Synonym-based Large Language Attention and Hierarchical Probability Transformer (SLLA-HPT) for predicting disease intensity and identifying relationships between ICD codes is proposed. The SLLA-HPT method is split into four sections. First, data collection and pre-processing is performed with the raw MIMIC-III dataset. Next, for predicting disease intensity levels from ICD codes Synonym-based ICD Code Attention Pre-training model is applied. Following which fine-tuning is performed using the Integrated Gradients (IG) function. Next, Hierarchical Conditional Probability and Corrective Attention Transformer are applied to the fine-tuned results for obtaining prediction results. Finally, Correlation-based Relationship Identification is applied to create ranked record of diseases through their provisional probability as well as correlative risk grade. Proposed method is estimated on MIMIC-III database as well as experimental outcomes show effectiveness of our method upon comparison to conventional techniques.

Keywords: Large Language Models, International Classification of Diseases, Synonym-based ICD Code Attention, Integrated Gradients, Hierarchical Conditional Probability, Corrective Attention Transformer.

1. Introduction

Predicting disease intensity levels from ICD codes encompasses of using NLP methods as well as transformer learning algorithms to ascertain textual information. Hence, ML, DL and transformer learning (TL) methods have employed in conventional literature for disease intensity level prediction from ICD codes. Nevertheless, prevailing systems reveal suboptimal result mainly because of their dependence on conventional embedding mechanisms as well as comprehensive language methods to bring about text embeddings. To mention this drawback, there necessitate domain-specific pre-trained language models that systematically recognize the context from ICD codes for disease intensity level prediction and identifying relationship between specific ICD codes.

A hybrid transformer architecture, combining a type of LLM called, MentalBERT and MelBERT pre-trained language models, integrated with CNN was proposed in [1] with the

intent of generating and concatenating deep features. On one hand, MentalBERT was pre-trained on widespread corpus of text information associated to mental wellbeing domain and on the other hand, MelBERT was trained on huge corpus of metaphorical information for comprehending of metaphorical appearances. Outcomes expose outstanding result in terms of both accuracy and F1-score.

In [2] a method called, eXtreme Resolution-Long Attention Transformer (XR-LATP) was proposed employing three distinct Transformer-based method for automated ICD coding. Initially, XR-LAT was used that fine-tuned pre-trained transformer model in a recursive manner. This was performed hierarchically through label-wise attention layer as well as label-wise FFNN layer employing bootstrapping and dynamic negative label sampling methods with improved true positive and minimal training time.

LD-PLAM was proposed [3] for automatic ICD coding. Here, pseudo label-wise concentration model was utilized for multi-label categorization and also one concentration mode for similar ICD codes was computed. This in turn minimized number of metrics, therefore achieving improved result of method. With this type of design resulted in improvement of memory consumption.

Leveraging EHR longitudinal information to generate actionable medical insights has dangerous demand as far as recent studies in the area of disease diagnosis is concerned. The biggest issue is that only ample data are available and hence prediction has to bring in free text diagnosis recorded immediately on admission.

Transformer based deep learning method was proposed in [4] to obtain temporal sequencing between cancer and non-cancer cells with the purpose of diagnosing therefore resulting in much more accurate results. ICD 9 diagnosis codes were used in [5] for improved prediction. Here, attention mechanism was also used to concentrate on the most influential feature towards precise predictive tasks. According to ICD, multi-label categorization aim to classify health documentation based on the set of pertinent ICD codes. Multi-label text categorization head through per-label concentration, on apex of BERT method was proposed in [6] with improved precision in English, Spanish and Swedish.

DL has considerably improved NLP as far as healthcare investigate is concerned. Nevertheless, enhancing complication of DL-basis of NLP requires transparent method for consistent decision-making. Detailed investigation of explicable as well as interpretable DL at healthcare NLP was investigated in [7]. Allocating medicinal codes for patients is mandatory for healthcare associations for billing however for precise record maintenance. However, because of affluence of disease codes, it is found to both time utilizing process and also found to be laborious for medicinal experts to allocate these codes manually. To mention this issue, a method for automatic forecast of ICD-9 codes employing two-stream DL technique was analyzed [8], therefore improving improved ICD-9 code prediction. A scoping review of LLMs in medicine was investigated in [9].

Researchers investigated utilization of NN, specifically language methods for assigning ICD-9 code in an automatic fashion. Nevertheless, imbalanced allocation of ICD-9 codes resulted in deprived result. To address on this issue a useful prior according to domain knowledge was applied in [10]. Also the usefulness of correlation was analyzed based on the correlations among ICD-9 codes as well as other medicinal codes. This in turn resulted in large performance gains. An LLM-fine tuned model based on diagnosis related group was presented in [11] with improved prediction accuracy.

According to the ICD-10, according to German outpatient claims data with diagnosis codes, an optimal vectorization embedding model was proposed in [12] for patient diagnosis. A comprehensive survey of materials and methods employing fine tuned LLMs were investigated in [13]. Existing methods utilize the current ICD codes as input aspects however not entirely explain patient progression. Additional researches applied attention mechanism with no assistance of domain information, generating distorted predictions. In [14], a hybrid LSTM-CNNself-guided attention method was proposed with the intent of predicting ICD diagnosis accurately. Yet another diagnosis code method employing supervised multi layer perceptron as a classifier was presented [15] in prediction of disease of interest.

Predicting disease intensity levels along with exploring relationship between specific ICD codes is a complicated procedure necessitating scientific research and meticulous analysis, with certain research putting forward prospective for increased accuracy, whereas others focus uncertainties and drawbacks. In earlier researchers have attempted to signify associates between statistical analysis and exploring relationship, nevertheless, for the most part by applying statistical methods. Apart from above stated issues, in our study explores the association between Large Language Model and Transformer Learning on ICD coding. The major aim is to introduce Synonym-based Large Language Attention and Hierarchical Probability Transformer (SLLA-HPT) for predicting disease intensity and hence understanding how different ICD codes relate to each other in a precise and accurate fashion.

1.1 Contributions of the work

- SLLA-HPT method for predicting disease intensity and exploring relationships between ICD codes is designed. The hybrid method of employing a LLM and Transformer learning for predicting disease intensity and exploring relationships between ICD effectiveness adopts four different processes, data collection, training model (i.e. pre-training and fine-tuning), prediction and relationship identification. The Synonym-based ICD Code Attention Pre-training model guided by the Integrated Gradients (IG) function based fine-tuning contributes to improving the prediction accuracy, sensitivity, specificity. Next, with the obtained fine-tuned results, prediction is made employing Hierarchical Conditional Probability and Corrective Attention Transformer therefore is applied to the fine-tuned results for obtaining prediction results with minimum error rate for model construction. Finally, Correlation-based Relationship Identification is applied to the predicted results that in turn based on correlation models that with a specific type of heart disease are more likely to also have other related disease that in turn aids in treatment decisions precisely.
- The MIMIC-III dataset is utilized to conduct the experimental analysis for the proposed evaluation of predicting disease intensity based on relationship between ICD codes. The best prospective combination of features that aid enhancing the prediction accuracy and precision performance has been identified. The comparative analysis of the proposed hybrid method SLLA-HPT has been evaluated based on pre-training and fine-tuning processes separately.

1.2 Structure of the work

Rest of this manuscript is organized as: Section 2 explains the recent related works related to disease intensity prediction using ICD codes, analyzing their relationship for prompt decision making using LLM, transformer learning and DL. Section 3 gives complete explanation of Synonym-based Large Language Attention and Hierarchical Probability Transformer (SLLA-

HPT). Section 4 presents the experimental setup following the presents the simulation results of the proposed to disease intensity prediction based on relationship between ICD codes along with the comparison of the traditional methods in Section 5. Lastly, Section 6 summarizes the manuscript.

2. Literature review

State of the art methods in electronic health records based predictive model achieved accurate results by means of pre-training cum fine-tuning model. Two step procedure where method was initially pre-trained on huge scale longitudinal data and then further fine-tuning the outcome prediction was proposed. Nevertheless, the pre-training objectives were found to be constrained at forecasting portion of ICD codes in every visit.

Transformer based encoder decoder method was designed [16] with limited data. Over the recent few years, prompt learning with LLMs has shown to lead generalizable results however none have investigated for rare diseases. Prompt learning for both identification and extraction of rare disease was presented in [17] to make certain phenotyping accuracy was being achieved. However in the presence of complex interactions resulted in the compromising of training time. To address on this aspect, dropout and batch normalization was applied in [18] to ensured accountable execution in clinical practice.

LLMs have manifested splendid potentialities however the constraint for clinical applications is high. Endeavors to evaluate the clinical knowledge heavily depend on evaluations automatically on the basis of constrained yardsticks. A human evaluation architecture for model answers to evaluate Pathways Language Model1 was proposed in [19] for clinical applications. Yet another multi-modal transformer was presented in [20] for molecular sub-type prediction by means of attention maps for imaging and clinical data.

ICD diagnoses information, gathered for billing as well as epidemiological uses globally designates expensive source of structured data. Nevertheless, ICD coding is demanding task. As several preceding works reported promising outcomes they frequently narrate input data specific model architectures.

In [21], a detailed study on investigating assessment and design of efficient CAC scheme by generic methods concentrating on utilization of ICD employing feed forward neural network architecture was designed. However, early diagnosis was not ensured. Self-attention based transformer network model was designed in [22] with the intent of ensuring early diagnosis along with accurate heart disease prediction.

A detailed comparative evaluation of target attention and text coder encoder was made in [23] for detailed clinical text classification. Yet another ICD code model for analyzing vital organ failure employing graph attention was presented in [24]. A review of LLM and transformer learning was investigated in [25] for medical domain.

Ten most prevalent ICD codes were first identified to ensure accurate and precise documentation in [26]. Also fine-tuning was performed using GPT model therefore ensuring maximum utility and reliability. Deep learning (DL) techniques along with ICD was modeled in [27] for healthcare and insurance domain. Also DL based CNN was applied to process the data that in turn resulted in both precision and accuracy. In [28], BERT method to allocate ICD-10 codes for reasons of bereavement, through free-text explanations at death certificates with connected autopsy information as well as medical bulletins obtained as of Portuguese Ministry of physical condition was presented. Pilot study was conducted in [29] with the intent of augmenting human coding by initially initializing lead terms following which

generation methods based on retrieval augmentation was employed for coding enhancement with improved F1-score.

From the scrutiny of related works, restricted LLM integrated with transformer learning solution have been designed to measure the intensity of disease and its relationship between ICD codes. Rather than focusing entire feature subset, applying pre-training and fine-tuning from the original features results in minimizing the error rate along with the improvement of sensitivity and specificity. Due to the reason that application of machine learning and deep learning workflow has been treated as time-consuming process, LLM integrated with transformer learning solutions are well perceived over in last few years owing to their advantages like improved prediction accuracy, minimizing training time for predicting higher-level feature representation. This in turn would ensure accurate and precise generation of predictability results. Finally, the relationship identification contributed in F1-score improvement. With this objective a method called, Synonym-based Large Language Attention and Hierarchical Probability Transformer (SLLA-HPT) is proposed.

3. Materials and Methods

LLMs and transformer-based learning are effective instruments that can be exploited for ascertaining relationships between medical conditions and predicting disease intensity by analyzing ICD codes. These LLMs and transformer-based learning models can aid healthcare professionals in making more well-considered decisions and boosting patient care. The objective of predicting disease intensity and exploring relationships between ICD codes is to extract a specific set of ICD codes ' $C = \{c_1, c_2, \dots, c_U\}$ ' from the given clinical note samples ' $S = \{S_1, S_2, \dots, S_N\}$ '. Here, 'U' and 'N' denotes number of ICD codes as well as word count of the clinical notes samples, respectively. The proposed Synonym-based Large Language Attention and Hierarchical Probability Transformer (SLLA-HPT) method for predicting disease intensity and identifying relationships between ICD codes model, as depicted in below Figure

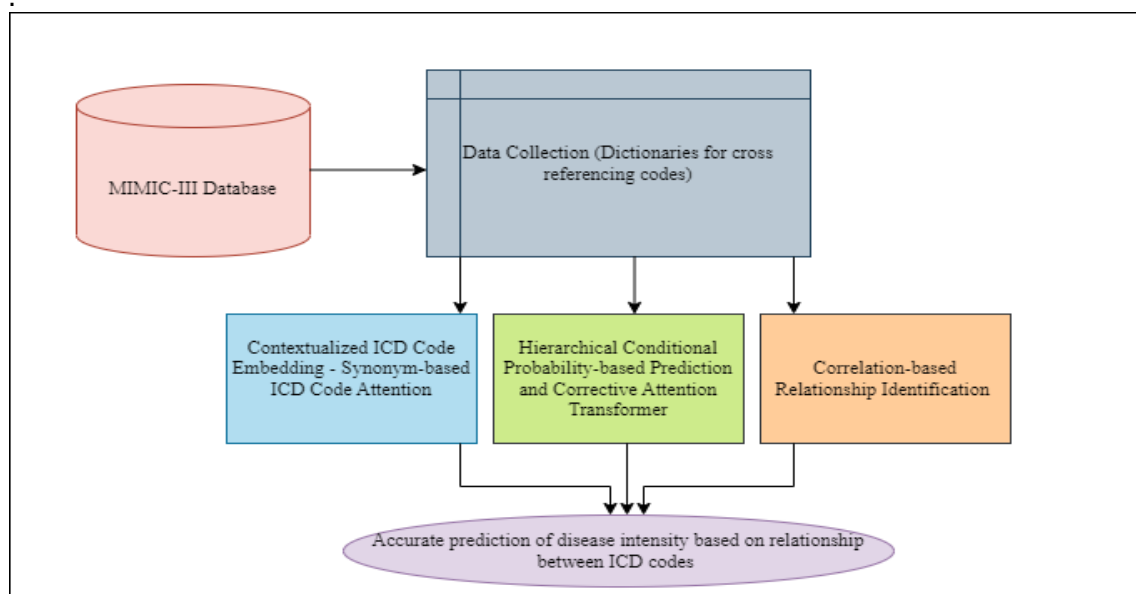


Figure 1 Structure of SLLA-HPT method

As shown in the above figure the SLLA-HPT method comprises four major steps, namely, data collection and pre-processing, contextualized ICD code embedding using

synonym-based ICD code attention model, prediction using Hierarchical Conditional Probability and Corrective Attention Transformer and Correlation-based Relationship Identification.

3.1 Data Collection and Pre-processing

The first step involved in the design of the proposed method is to gather ICD codes from MIMIC-III Clinical dataset attained as input as of <https://physionet.org/content/mimiciii-demo/1.4/> (i.e. 'DS'). Electronic Health Records (EHRs) carrying ICD codes, patient history and other appropriate data are put together and organized for predicting disease intensity levels and identifying relationships between ICD Codes. In general, five tables are utilized in defining and tracking patient stays, namely, ICU-stays, and so on. Residual tables include information involved through patient care, to name a few being, physiological measurements, and so on. So, on an overall, it's relational dataset containing 26 tables. Also numerous features 'F' are present in the dataset from different classes of data is listed in table 1.

Table 1 Classes of information available in MIMIC-III Clinical dataset

S. No	Class of data	Description
1	Billing	Coded data recorded for billing
2	Descriptive	Details of demographic, admission, discharge time
3	Dictionary	Look-up tables for cross referencing with associated labels
4	Interventions	Procedures such as dialysis, imaging studies
5	Laboratory	Blood chemistry, hematology, urine analysis, microbiology test results
6	Medications	Administration records of intravenous medications, medication orders
7	Notes	Free text notes – provider progress notes, discharge summaries
8	Physiologic	Nurse-verified vital signs, (e.g., heart/respiratory rate, blood pressure)
9	Reports	Free text reports of electrocardiogram and imaging studies.

With above different classes of information obtainable in MIMIC-III Clinical dataset, proposed method, SLLA-HPT is modeled for predicting disease intensity and identifying relationships between ICD codes in the following sub-sections.

3.2 Synonym-based ICD Code Attention Pre-training model

The second step involved in the design of the proposed method is to model training or pre-training and then fine-tune LLM built upon on the Transformer architecture on the pre-processed data and their corresponding ICD codes. The model learns to ascertain patterns and relationship between specific ICD codes. BERT type of LLM is used in our work to infer ICD-10 codes in an automatic fashion from clinical sample notes. Fine-tuning BERT and LLMs involves first obtaining a pre-trained model and further training it to boost its performance for predicting disease intensity levels from ICD codes.

In this work, a Synonym-based ICD Code Attention Pre-training model is designed with the intent of extracting contextualized ICD code features for predicting disease intensity levels from ICD codes. This process of model training or pre-training influences the knowledge gained during initial pre-training phase and reshapes the model to the variances

of the new task (i.e. exploring relationship between ICD codes). Contextualized code features in ICD denotes to the potentiality of learning models to ascertain the meaning and relationships between ICD codes within the context of a specific clinical note samples. This model intends to boost the accuracy of ICD code allocation by taking into consideration the adjoining text along with its semantic interpretation instead of depending on individual ICD code descriptions. In this work a synonym-based ICD code attention model is presented to extract contextualized ICD code features from clinical note samples by pointing out the ICD code synonym description at the outset. Figure 2 shows the architecture of Synonym-based ICD Code Attention Pre-training model to extract contextualized ICD code features.

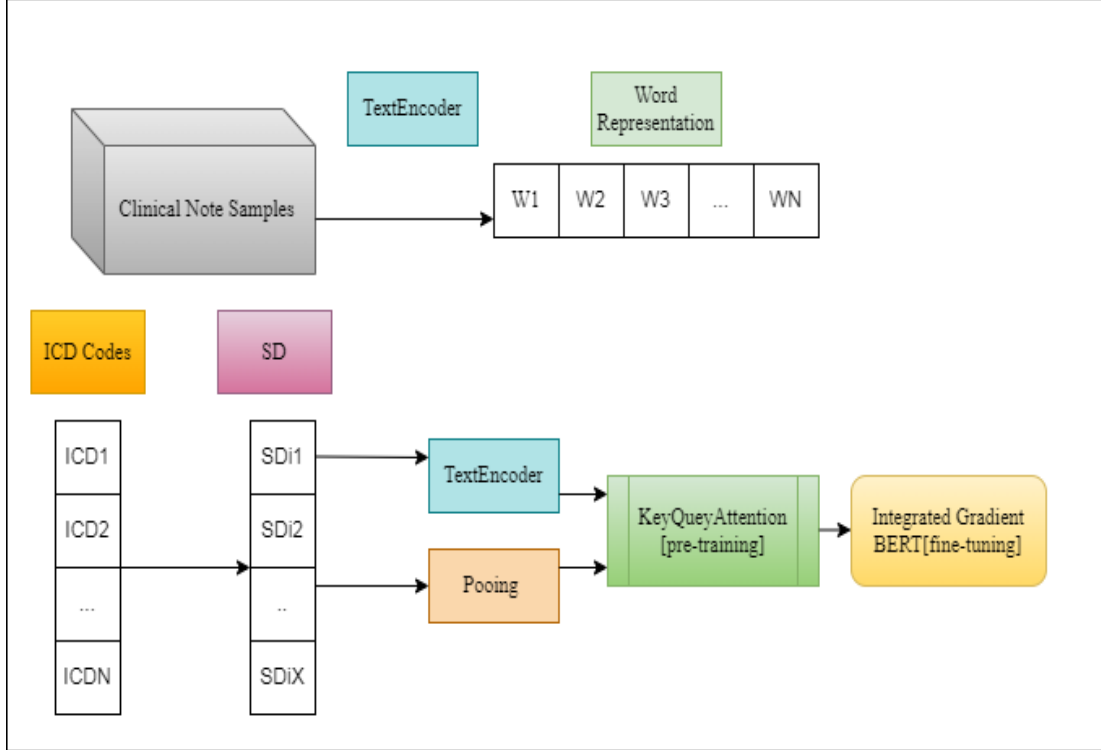


Figure 2 Architecture of Synonym-based ICD Code Attention Pre-training model to extract contextualized ICD code features

As shown in the above figure, the same text encoder is employed in encoding both the raw clinical note samples and ICD code descriptions. Hence, the clinical note sample embedding is formulated as given below.

$$\{S_1, S_2, \dots, S_N\} = \text{TextEncoder}(S) \quad (1)$$

From the above equation (1), ' S_k ' denotes the ' k -th' word in clinical note sample ' S '. Each ICD code ' $c_i \in C$ ' carries ' X ' synonyms descriptions, ' $SD_i = \{SD_{i,1}, SD_{i,2}, \dots, SD_{i,X}\}$ ' and each ' $SD_{i,j}$ ' comprises of ' W ' words. In a similar manner, for each ICD code synonym the overall clinical note sample embedding is mathematically represented as given below.

$$\{SSD_1^{i,j}, SSD_2^{i,j}, \dots, SSD_X^{i,j}\} = \text{TextEncoder}(SD_{i,j}) \quad (2)$$

Then, from the above overall clinical note sample embedding to generate overall synonym clinical note sample embedding average pooling function is utilized as given below.

$$SD_{ij} = Pool([SSD_1^{i,j}, SSD_2^{i,j}, \dots, SSD_X^{i,j}])$$

(3)

Following which a standard key-query attention layer is employed in extracting contextualized ICD code embeddings (CCE) from text embedding sequence as given below.

$$CCE_{ij} = KeyQueryAttention(SD_{ij}[S_1, S_2, \dots, S_N])$$

(4)

From the above equation (4), 'CCE_{ij}' denotes the contextualized ICD code embedding results. For each ICD code, maximum pooling is utilized in obtaining overall representation across all the obtained synonym clinical note sample embedding. This is mathematically formulated as given below.

$$CCE_i = Pool([CCE_{i,1}, CCE_{i,2}, \dots, CCE_{i,X}])$$

(5)

Following which a pre-trained Large Language Model (LLM) (i.e. synonym clinical note sample embedding using average pooling function) is fine-tuned on the prepared contextualized ICD code embeddings (CCE) clinical data using Integrated Gradients (IG) function. This involves fine-tuning the model's parameters to ascertain the specific patterns and relationships within the prepared contextualized ICD code embeddings (CCE) clinical data, specifically those associated to ICD codes using IG function.

To better comprehend the proposed method, the IG function is used in determining which drugs and diagnoses had the strongest impact on the model predictions. The IG function employed in our work make predictions by associating significance to input features along with the attributing importance to prepared contextualized ICD code embeddings (CCE) clinical data. Essentially, it evaluates a score for each input features along with the attributing importance to prepared contextualized ICD code embeddings (CCE) clinical data. This is arrived at by combining the gradients of the model's output with respect to the input from baseline input to the actual input. Next to an input sample ($S[CCE] \in \mathbb{R}^N$), the IG method necessitates a baseline input ' $S'[CCE] \in \mathbb{R}^N$ '. Following which the IGs are then approximated by summing gradients from the specified baseline to input using the following formula as given below.

$$IG_i(S[CCE])^{approx} = (S_i[CCE] - S'_i[CCE]) * \int_0^1 \frac{\partial F(S'[CCE] + \alpha(S[CCE] - S'[CCE]))}{\partial S_i[CCE]} d\alpha$$

(6)

From the above equation (6) ' ∂F ' denotes the differentiable function and according to the samples provided for simulation, the number of iterations required to approximate integrated gradients differ. Initially, IG attributions are measured for all classes of data available with regards to distinct patients in the test dataset.

Accordingly, the mean absolute attribution for each diagnosis that occurred is measured with the objective of identifying the top features. Subsequently, partial dependency score is calculated utilizing the top features. To do so, initially, the probability for each patient using fine-tuned Integrated Gradients BERT models is obtained that we refer to this probability as ' $Prob_{act}$ '. The data for each of the top features are then permuted independently by swapping the respective diagnosis with a sequence of samples. Subsequently, the modified data is then utilized as input for the proposed method to measure the probability ' $Prob_{mod}$ ' via fold change. This is represented as given below.

$$FC = \frac{Prob_{act}}{Prob_{mod}} \quad (7)$$

From the above equation (7), fold change 'FC' is measured for each feature based on the probabilities obtained for actual ' $Prob_{act}$ ' and modified ' $Prob_{mod}$ ' results to measure the influence of certain features on the model's prediction.

3.3 Hierarchical Conditional Probability and Corrective Attention Transformer based Prediction

The third step involved in the design of the proposed method is to predict the intensity or severity of a disease on the basis of a sequence of ICD codes. In this work Hierarchical Conditional Probability and Corrective Attention Transformer based Prediction model is used. This model uses a Transformer method to denote tokens in medical documents and hierarchical conditional probability and correlative risk grade-wise attention mechanism to create probability-risk specific dictionary representations for ICD code prediction. Figure 3 shows the structure of Hierarchical Conditional Probability and Corrective Attention Transformer based Prediction model.

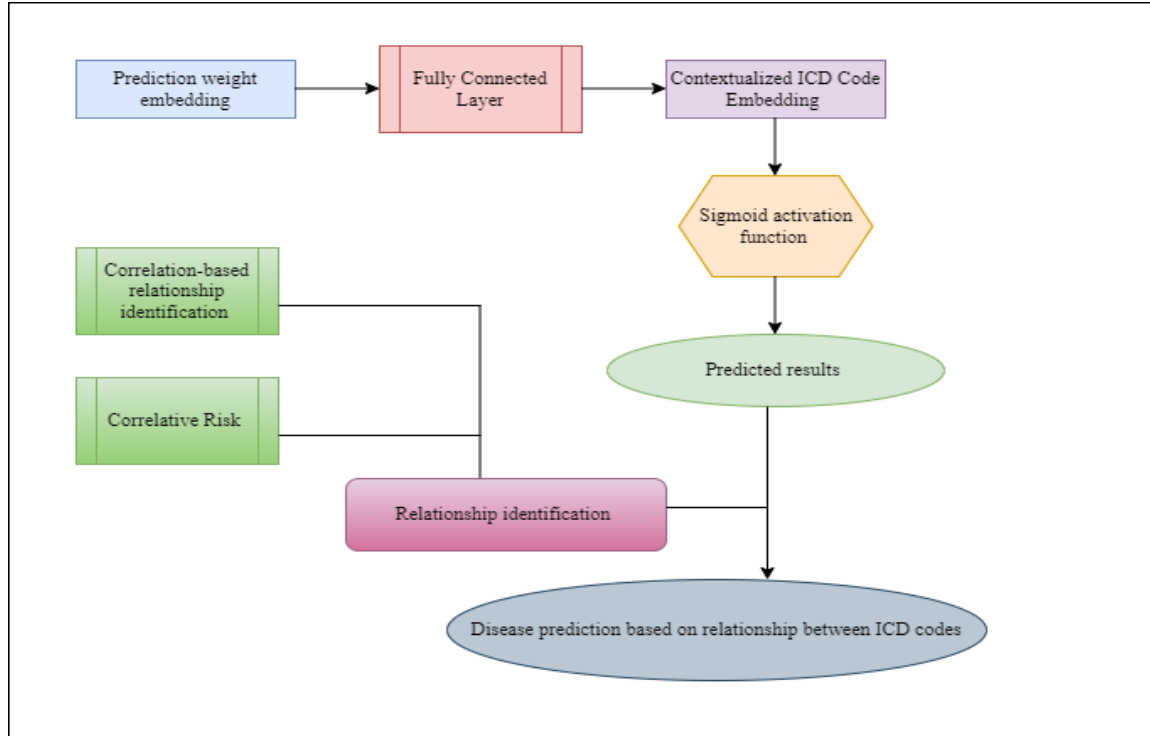


Figure 3 Structure of Hierarchical Conditional Probability and Corrective Attention Transformer based Prediction model

As shown in the above figure, initially a full connected layer is applied to obtain prediction weight embedding ' α_i ' for contextualized ICD code embedding ' CCE_i ' from the prepared contextualized ICD code embeddings (CCE) clinical data ' $CCE_{i,1}, CCE_{i,2}, \dots, CCE_{i,X}$ ' learned by equation (5).

$$\alpha_i = FC_{\alpha} \left(Pool([CCE_{i,1}, CCE_{i,2}, \dots, CCE_{i,X}]) \right) \quad (8)$$

Finally, the Hierarchical Conditional Probability is calculated by taking into consideration the inner product of prediction embedding and extracted contextualized embedding results. Along with which, sigmoid activation function ' $\sigma()$ ' is applied as given below.

$$Prob'_i = \sigma(\alpha_i \cdot CCE_i)$$

(9)

After obtaining ' i ' frequent disease sequences for contextualized ICD code embeddings (CCE) clinical data ' $CCE_{i,1}, CCE_{i,2}, \dots, CCE_{i,x}$ ', after that is to choose most widespread disease sequences depend on patient the past. To attain this, disease series are selected in such a way which had superior period of treatment as well as demonstrated progression as of diseases which patients had. To create the outcomes straightforward to appreciate, the list is ranked depend on Conditional Probabilities as well as Correlative Risk grades.

3.4 Correlation-based Relationship Identification

Finally, the proposed method is utilized in identifying possible relationships between diseases depend on co-incidence of ICD codes for medical record coding. Then, the possible relationships between diseases depend on co-incidence of ICD codes for dictionary present in the classes ' ICD_i ' given ' ICD_j ' and correlative risk grade for a sequence ' ICD_i ' and ' ICD_j ' as given below.

$$CPRI(ICD_i|ICD_j) = \frac{Prob(ICD_i, ICD_j)}{Prob(ICD_i)}$$

(10)

From the above equation (10) it is inferred that ' ICD_i ' materialized before ' ICD_j '. Then the correlative risk grade is mathematically formulated as given below.

$$CR_{ICD_i, ICD_j} = \frac{Obs}{Exp} = \frac{Prob(ICD_i, ICD_j)}{Prob(ICD_i)Prob(ICD_j)}$$

(11)

From the above equation (11), ' $Prob(ICD_i)$ ' and ' $Prob(ICD_j)$ ' denote probabilities which patient has disease with ICD code ' ICD_i ' and ' ICD_j ' respectively and ' $Prob(ICD_i, ICD_j)$ ' is probability which arbitrarily selected patient has together diseases at temporal order. Also, a ' $CR > 1.0$ ' denotes which two diseases are extremely connected with each other and vice versa. The pseudo code representation of Synonym-based Large Language Attention and Hierarchical Probability Transformer is given below.

Input: Dataset ' DS ', Tables ' $T = \{T_1, T_2, \dots, T_M\}$ ', Features ' $F = \{F_1, F_2, \dots, F_l\}$ ', Samples ' $S = \{S_1, S_2, \dots, S_N\}$ '
Output: Increased accuracy with better context understanding
1: Initialize ' $M = 26$ ', ' $N = 15000$ ' 2: Begin 3: For each Dataset ' DS ' with Tables ' T ', Features ' F ' and Samples ' S ' //Step 1 Data collection and pre-processing //Obtain tables //Obtain features from the classes of available data //Obtain samples for further processing //Step 2 Model training (i.e. pre-training)

```

4: Formulate clinical note sample embedding according to (1)
5: Evaluate overall clinical note sample embedding according to (2)
6: Evaluate overall synonym clinical note sample embedding using average pooling
function according to (3)
7: Formulate key-query attention layer for extracting contextualized ICD code
embeddings according to (4) and (5)
//Fine-tuning according to feature significance
8: Evaluate Integrated Gradients (IG) for prepared contextualized ICD code embeddings
(CCE) clinical data according to (6)
9: Measure magnitude of differences to measure the influence of certain features
according to (7)
//Step 3 Prediction
10: Obtain prediction weight embedding ' $\alpha_i$ ' for contextualized ICD code embedding
' $CCE_i$ ' via full connected layer according to (8)
11: Formulate Hierarchical Conditional Probability according to (9)
//Step 4 Relationship Identification
12: Evaluate Conditional Probability Risk Identification results according to (10) based
on the co-occurrence of ICD codes
13: Evaluate correlative risk grade according to (11) based on the co-occurrence of ICD
codes
14: If ' $CR_{ICD_i, ICD_j} > 0$ '
15: Then diseases are highly associated
16: End if
17: If ' $CR_{ICD_i, ICD_j} \leq 0$ '
18: Then diseases are not highly associated
19: End if
20: End for
21: End

```

Algorithm 1 *Synonym-based Large Language Attention and Hierarchical Probability Transformer*

As given in the above algorithm, initially, the raw dataset MIMIC-III is subjected to the data collection process and then five tables denoting dictionaries for cross-referencing codes in preparation is then used for further processing. Next, contextualized ICD code embedding using synonym-based ICD code attention model is used wherein initially, clinical note sample embedding is performed and then overall clinical note sample embedding results are obtained using average pooling function. Finally, key-query attention layer is formulated with the intent of extracting contextualized ICD code embeddings in a computationally efficient manner. Here also Integrated Gradients function is applied for fine-tuning the results therefore obtaining precise and accurate feature extracted for further processing. Third, prediction is performed using Hierarchical Conditional Probability and Corrective Attention Transformer. Here, prediction weight embedding for contextualized ICD code embedding is obtained via full connected layer therefore improving the overall prediction results. Finally, Correlation-based Relationship Identification is applied to identify relationship between ICD codes.

4. Experimental Setup

This part reports experimental results for all four techniques, Synonym-based Large Language Attention and Hierarchical Probability Transformer (SLLA-HPT) and existing Hybrid Transformer Architecture [1], eXtreme Resolution-Long Attention Transformer (XR-LATP) [2] and LD-PLAM [3]. The methods were designed and simulated by Python language with MIMIC-III database.

This work gives a technique for predicting disease intensity as well as exploring relationship between ICD codes based on MIMIC-III dataset. Similar training approach was utilized for every four techniques training sessions in this work, in spite of forecast technique. Input samples employed for every training sessions are reserved equal, as specified previous in database section for every four techniques to guarantee fair comparisons. So to recognize effectiveness of technique, 500000 samples were selected arbitrarily. Moreover, for every experiment carried out, similar database was used for every four techniques. Furthermore, numerous evaluation parameters were employed to calculate proposed technique.

5. Discussion and Inferences

5.1 Inferences

A patient with a history of asthma (ICD-10 code J45.0) followed by a lung infection (ICD-10 code J18.1) could be predicted to have a higher intensity of respiratory disease or a higher likelihood of developing complications related to their asthma due to the lung infection. Hence by combing Synonym-based ICD Code Attention Pre-training model and Integrated Gradients (IG) function for generating fine-tuned results with Hierarchical Conditional Probability and Corrective Attention Transformer in addition to Correlation-based Relationship Identification as the cornerstone aids in learning complicated patters for medical record coding resulting in more accurate prediction of disease intensity and relationships. Early Detection: By ascertaining prospective relationships between diseases, the proposed method aid in premature recognition of barriers as well as therefore boosting patient results.

5.2 Performance analysis of sensitivity and specificity

First in this section the sensitivity and specificity analysis involved in the prediction of disease intensity by exploring relationship between ICD codes is provided.

$$Sen = \frac{TP}{TP+FN}$$

(12)

$$Spe = \frac{TN}{TN+FP}$$

(13)

From equations (12) and (13), sensitivity 'Sen' and specificity 'Spe' is calculated depend on 'TP', 'TN', 'FP' and 'FN'. With disease intensity prediction by exploring relationship between ICD codes effectiveness results for sensitivity and specificity values obtained from the above equations (12) and (13) the resultant value for four methods, SLLA-HPT, Hybrid Transformer Architecture [1], eXtreme Resolution-Long Attention Transformer (XR-LATP) [2] and LD-PLAM [3] is given in table 2.

Table 2 Sensitivity and specificity results

Samples	Sensitivity	Specificity
---------	-------------	-------------

	SLLA-HPT	Hybrid Transformer Architecture [1]	XR-LATP [2]	LD-PLAM [3]	SLLA-HPT	Hybrid Transformer Architecture [1]	XR-LATP [2]	LD-PLAM [3]
50000	0.99	0.99	0.98	0.98	0.94	0.9	0.81	0.8
100000	0.98	0.93	0.92	0.9	0.92	0.88	0.82	0.72
150000	0.97	0.91	0.91	0.88	0.9	0.86	0.8	0.7
200000	0.95	0.9	0.89	0.86	0.87	0.83	0.77	0.67
250000	0.93	0.9	0.87	0.84	0.85	0.81	0.75	0.65
300000	0.92	0.91	0.85	0.82	0.83	0.79	0.73	0.63
350000	0.93	0.92	0.86	0.83	0.81	0.76	0.7	0.6
400000	0.95	0.94	0.88	0.82	0.84	0.8	0.74	0.64
450000	0.96	0.95	0.9	0.88	0.85	0.81	0.75	0.65
500000	0.97	0.94	0.92	0.9	0.88	0.84	0.78	0.68

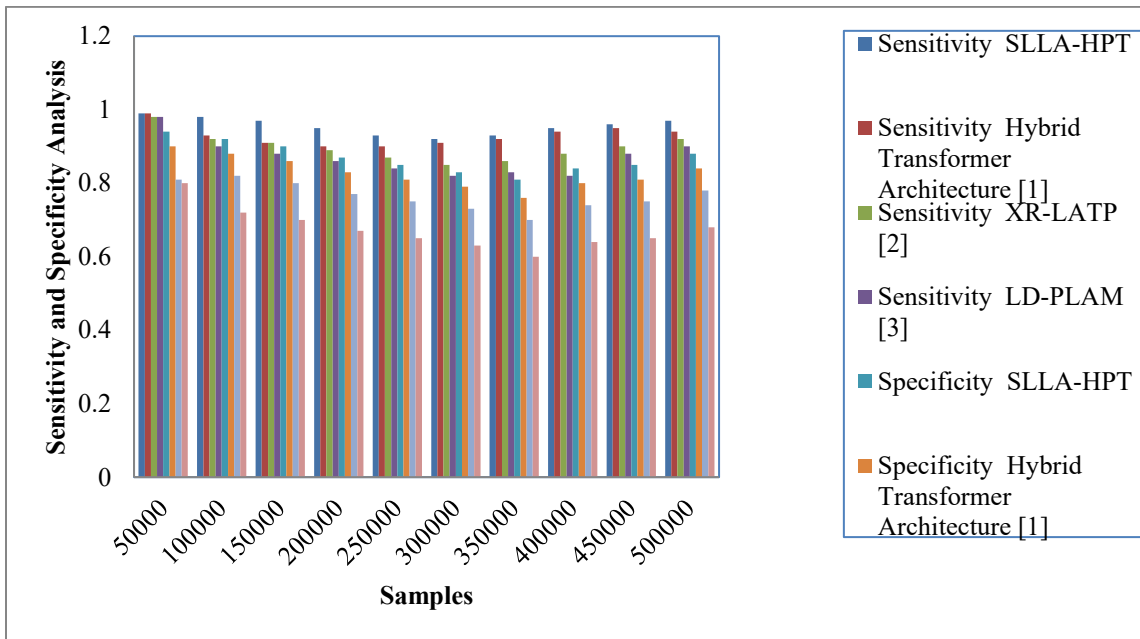


Figure 4 Graphical representations of Sen and Spe

Figure 4 depicts graphical illustration of *Sen* and *Spe* with the proposed SLLA-HPT and three existing methods, Hybrid Transformer Architecture [1], XR-LATP [2] and LD-PLAM [3]. For the first six iterations the sensitivity values were though less but were found to be increased for the rest of four iterations using all the four methods. Similarly, the specificity values for the first five iterations were though reduced then increased and decreased and so on. Hence, the increase in sample size does not have any influence on the resultant sensitivity and specificity rate. Also the sensitivity using the four methods was observed to be 0.99 for SLLA-HPT, Hybrid Transformer Architecture [1] and 0.98 using XR-LATP [2] and LD-PLAM [3]. From this simulation result it is inferred that *Sen* and *Spe* with SLLA-HPT technique was improved upon comparison to [1], [2] and [3]. The reason behind the sensitivity improvement of SLLA-HPT method over [1], [2] and [3] was due to the application of

Synonym-based ICD Code Attention Pre-training model that initially evaluated clinical note sample embedding, following which overall clinical note sample embedding results were obtained using average pooling function. This in turn enhanced the *Sen* with SLLA-HPT technique by 3% , 6% and 9% than the [1],[2], [3]. Also by fine-tuning the results using Integrated Gradients (IG) function magnitude of differences were employed to measure the influence of certain features that in turn enhanced *Spe* of SLLA-HPT technique by 5% , 12% and 23% than the [1],[2], [3].

5.3 Accuracy and F1-score

In this section accuracy and F1-score results are validated and analyzed. Accuracy or prediction accuracy is referred to as the most fundament performance metric for evaluating the effectiveness of disease intensity prediction by exploring relationship between ICD codes prediction method's overall correctness. Accuracy refers to the percentage ratio of all predictions that are correct, irrespective of whether they are positive or negative. The accuracy and F1-score is mathematically formulated as given below.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

(14)

$$F1 - score = 2 * \frac{Pre*Rec}{Pre+Rec}$$

(15)

$$Pre = \frac{TP}{TP+FP}$$

(16)

From the above equations (14) and (15) accuracy is calculated with '*TP*', '*TN*', '*FP*' and '*FN*' resultant values. F1-score is evaluated using '*Pre*' and '*Rec*' values. With the disease intensity prediction by exploring relationship between ICD codes results obtained from the equation (14), (15) and (16) the resultant accuracy and F1-score for four methods, SLLA-HPT, Hybrid Transformer Architecture [1], XR-LATP [2] and LD-PLAM [3] is given in table 3.

Table 3 Accuracy and F1-score results

Samples	Accuracy (%)				F1-score			
	SLLA-HPT	Hybrid Transformer Architecture [1]	XR-LATP [2]	LD-PLAM [3]	SLLA-HPT	Hybrid Transformer Architecture [1]	XR-LATP [2]	LD-PLAM [3]
50000	98	98	97	96	0.99	0.98	0.97	0.97
100000	96.15	94.25	91.25	90.35	0.98	0.93	0.88	0.8
150000	94.35	92.15	88.45	85.25	0.96	0.91	0.86	0.78
200000	93.15	90.35	85.35	81.15	0.95	0.9	0.85	0.77
250000	91.35	88.25	83.15	78.35	0.93	0.88	0.83	0.75
300000	90.25	85.35	80.25	75.25	0.91	0.86	0.81	0.73
350000	92.45	87.25	82.45	77.45	0.93	0.88	0.83	0.75
400000	94.55	89	84.55	79	0.94	0.89	0.84	0.76
450000	96	90	86.35	81	0.95	0.9	0.85	0.77
500000	97.35	91.35	88.15	82.35	0.97	0.92	0.87	0.7

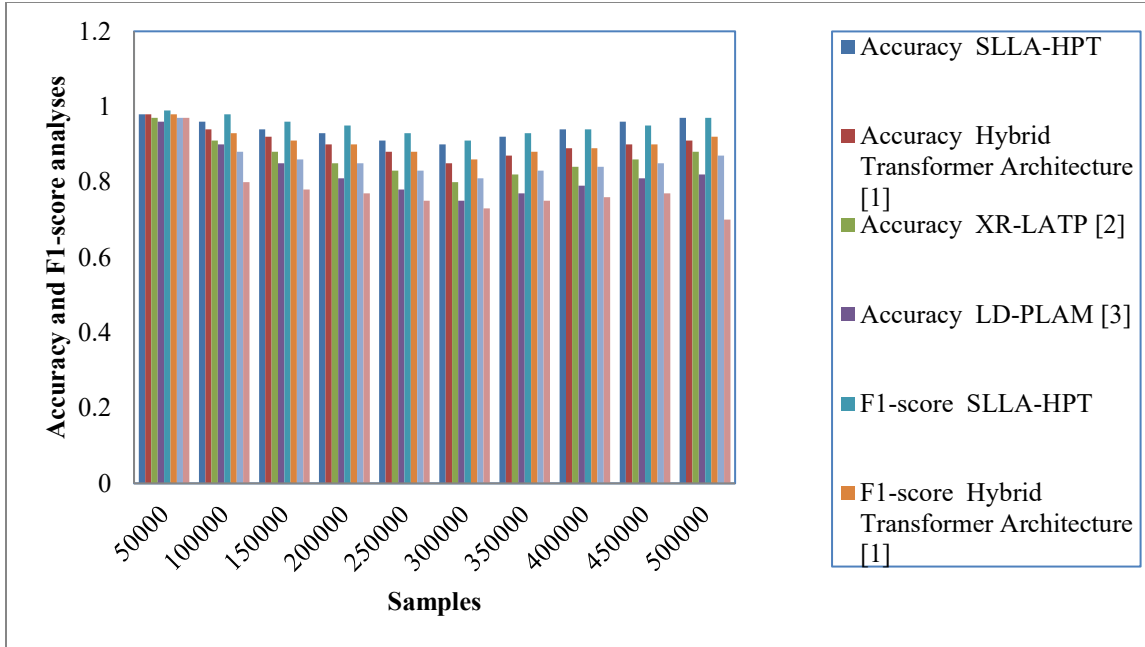


Figure 5 Graphical representations of accuracy and F1-score

Figure 5 given above illustrates the accuracy and F1-score results using the four methods, SLLA-HPT, Hybrid Transformer Architecture [1], XR-LATP [2] and LD-PLAM [3]. Though the accuracy results of all the four methods were in decreasing trend for the first seven iterations however were found in increasing trend for the rest of three iterations using all the three methods. Similar results were found for F1-score using all the four methods. However, simulation results revealed improvement in accuracy and F1-score for SLLA-HPT method upon comparison to [1], [2] and [3]. Here, the accuracy rate was found to be 0.98 for SLLA-HPT, Hybrid Transformer Architecture [1] whereas using XR-LATP [2] and LD-PLAM [3] it was observed to be 0.97 and 0.96. Similarly, the f1-score were observed to be 0.99 for SLLA-HPT, 0.98 using Hybrid Transformer Architecture [1] whereas 0.97 using XR-LATP [2] and LD-PLAM [3]. Cause behind enhancement was because of relevance of relationship identification by means of Conditional Probability Risk Identification results and correlative risk grade. By using these two function only high correlative risk grade results were retrieved for final prediction. This in turn improved the accuracy of SLLA-HPT technique by 4% , 8% and 13% than the [1],[2],[3]. Similarly, the F1-score using SLLA-HPT method was improved by 5% ,10% and 18% than the [1],[2],[3].

5.4 Error rate and training time

Error rate and training time in the overall procedure of predicting disease intensity by exploring relationship between ICD codes is evaluated as given below.

$$TT = \sum_{i=1}^N S_i * Time (CPRI + CR) \quad (17)$$

$$ER = \sum_{i=1}^N \frac{S_{IP}}{S_i} * 100 \quad (18)$$

From the above equations (17) and (18) training time 'TT' is calculated depend on samples involved in 'S_i' and time utilized in Conditional Probability Risk Identification results

'**CPRI**' and correlative risk grade '**CR**'. It is calculated in milliseconds (ms). Finally, error rate '**ER**' is calculated depend on samples considered ' S_i ' and samples inaccurately predicted with a disease though not ' S_{IP} '. It is computed in percentage (%). Table 4 lists results of **TT** and error rate by substituting the values in equations (17) and (18) for the four methods, SLLA-HPT, Hybrid Transformer Architecture [1], XR-LATP [2] and LD-PLAM [3].

Table 4 Training time and error rate results

Samples	Training time (ms)				Error rate (%)			
	SLLA-HPT	Hybrid Transformer Architecture [1]	XR-LATP [2]	LD-PLAM [3]	SLLA-HPT	Hybrid Transformer Architecture [1]	XR-LATP [2]	LD-PLAM [3]
50000	2000	2500	3500	4000	0.09	0.1	0.12	0.14
100000	2500	3000	3585	4015	0.13	0.15	0.17	0.18
150000	2850	3700	4035	4015	0.14	0.16	0.18	0.19
200000	3150	3795	4055	4095	0.16	0.18	0.2	0.21
250000	3325	4200	4085	5015	0.18	0.2	0.22	0.23
300000	3450	4250	5035	5085	0.15	0.18	0.2	0.21
350000	3815	4450	5015	5125	0.12	0.15	0.18	0.19
400000	4000	4650	5035	5235	0.1	0.12	0.15	0.17
450000	4035	4800	5055	5285	0.11	0.13	0.16	0.17
500000	4085	4850	5085	5300	0.13	0.15	0.2	0.21

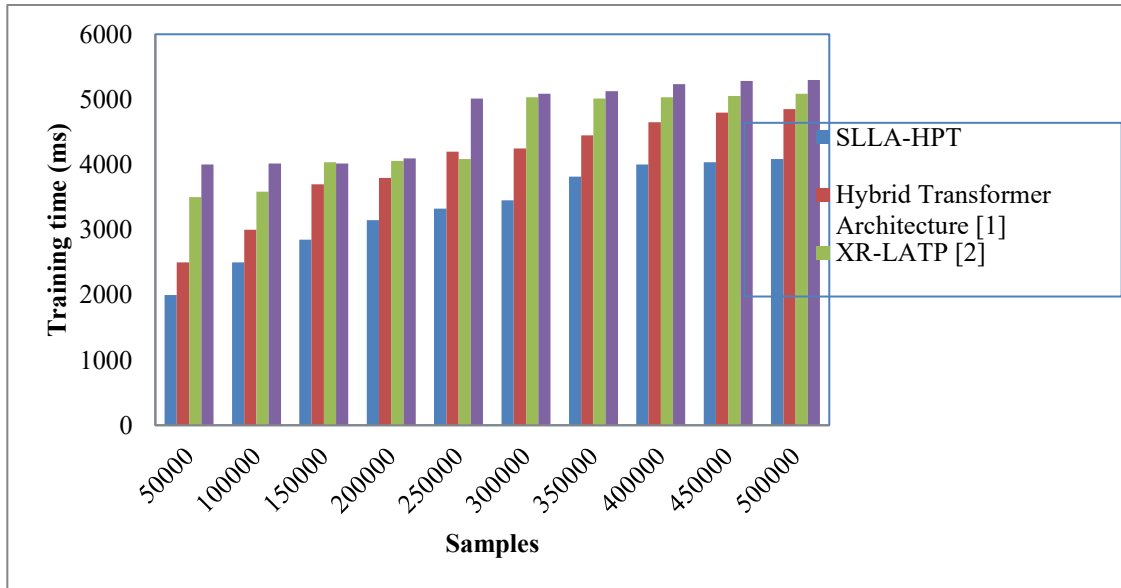


Figure 6 Graphical representation of training time

Figure 6 given above shows the results of training time in the vertical axis for 500000 samples provided in the horizontal axis. Enhancing the sample size causes enhance in clinical note samples that in turn increases the overall training time. However, sample simulations performed for 50000 samples observed training time of 2000ms using SLLA-HPT method,

2500ms using Hybrid Transformer Architecture [1], 3500ms using XR-LATP [2] and 4000ms using LD-PLAM [3] respectively.

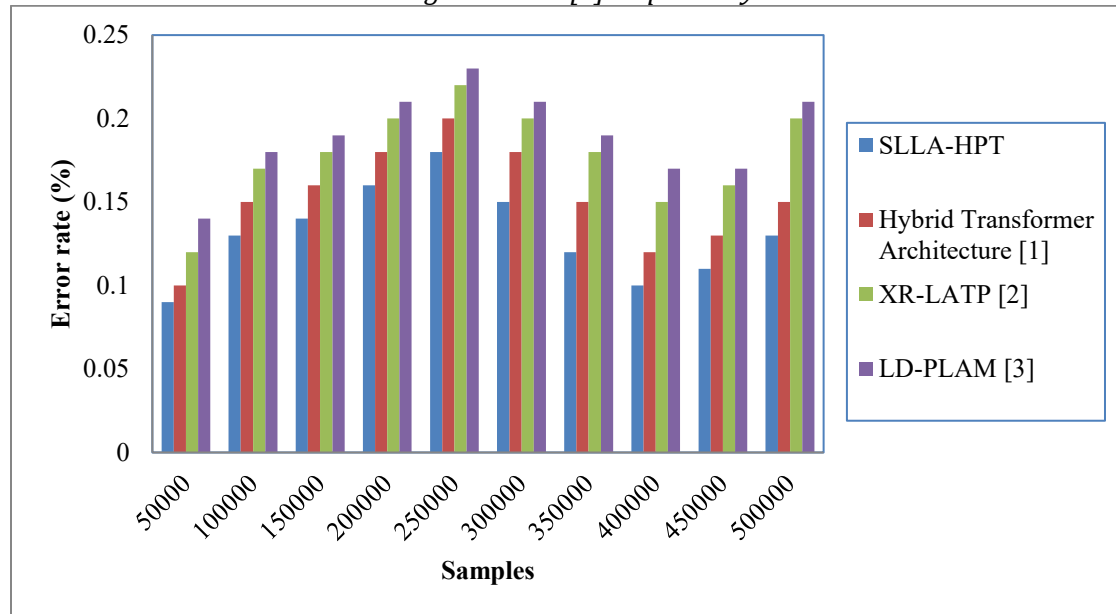


Figure 7 Graphical representation of error rate

Figure 7 given above shows the results of error rate in the vertical axis for 500000 samples provided in the horizontal axis. Increasing the sample size neither causes enhance in ER not cause decrease in ER. However, sample simulations performed for 50000 samples observed training time of 2000ms using SLLA-HPT method, 2500ms using Hybrid Transformer Architecture [1], 3500ms using XR-LATP [2] and 4000ms using LD-PLAM [3] respectively. Also the error rate for 50000 samples using the four methods was found to be 0.09, 0.10, 0.12 and 0.14 respectively. Both the training time and error rate minimization using SLLA-HPT method was observed due to the application of Hierarchical Conditional Probability and Corrective Attention Transformer based Prediction model. By applying this prediction model, inner product of prediction embedding and extracted contextualized embedding results were used via Hierarchical Conditional Probability that in turn reduced the training time of SLLA-HPT technique by 22% , 36% and 46% than the [1],[2], [3].

6. Conclusion

Employing LLMs and transformer learning, not only disease intensity can be predicted but also aids in exploring relationships between ICD codes by analyzing medical data. This process involves fine-tuning LLMs containing ICD codes with the objective of learning patterns or identifying correlations between different conditions. In this paper, the disease intensity prediction by exploring relationships between ICD codes is proposed by employing a Synonym-based Large Language Attention and Hierarchical Probability Transformer (SLLA-HPT) method. The data collection is obtained by using MIMIC-III dataset. The fine-tuning process is implemented by applying Synonym-based ICD Code Attention and Integrated Gradients (IG) function. Following which the prediction model employing Hierarchical Conditional Probability and Corrective Attention Transformer is applied to generate the results. Finally, to the generated results, Correlation-based Relationship

Identification was modeled for obtaining relationship between ICD codes for accurate prediction. With this hybrid model, the proposed SLLA-HPT prediction method provides a reasonable prediction result upon comparison with the other methods. The outcome shows that the disease intensity prediction accuracy, precision and time to be comparatively better based on a hypothesis test to that of the traditional methods with minimal error rate.

References

- [1] Adnan Karamat, Muhammad Imran, Muhammad Usman Yaseen, Rasool Bukhsh, Sheraz Aslam, Nouman Ashraf, "A Hybrid Transformer Architecture for Multiclass Mental Illness Prediction Using Social Media Text", IEEE Access, Vol. 13, Jan 2025 [Hybrid Transformer Architecture]
- [2] Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, Louisa Jorm, "Automated ICD coding using extreme multi-label long text transformer-based models", Artificial Intelligence In Medicine, Elsevier, Vol. 144, Oct 2023 [eXtreme Resolution-Long Attention Transformer (XR-LATP)]
- [3] Yifan Wu, Min Zeng, Ying Yu, Yaohang Li and Min Li, "A Pseudo Label-wise Attention Network for Automatic ICD Coding", IEEE Journal of Biomedical and Health Informatics, Volume 26, Issue 10, October 2022, Pages 5201 – 5212 [Label Description-based Pseudo Label-wise Attentional Model (LD-PLAM)]
- [4] Lan Wang, Yonghua Yin, Ben Glampson, Robert Peach, Mauricio Barahona, Brendan C. Delaney, and Erik K. Mayer, "Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data", EBioMedicine, National Library of Medicine, Vol. 110, Dec 2024
- [5] Gaspard Hareriman, Jong Wook Kim, Beakcheol Jang, "A deep attention model to forecast the Length Of Stay and the in-hospital mortality right on admission from ICD codes and demographic data", Journal of Biomedical Informatics, Elsevier, Vol. 118, Jun 2021
- [6] Alberto Blanco, Sonja Remmer, Alicia Perez, Hercules Dalianis, Arantza Casillas, "Implementation of specialized attention mechanisms: ICD-10 classification of Gastrointestinal discharge summaries in English, Spanish and Swedish", Journal of Biomedical Informatics, Elsevier, Vol. 130, Jun 2022
- [7] Guangming Huang, Yingya Li, Shoaib Jameel, Yunfei Long, Giorgos Papanastasiou, "From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?", Computational and Structural Biotechnology Journal, Elsevier, Vol. 24, Dec 2024
- [8] Mustafa Arda Ayden, Mehmet Eren Yuksel, Seniha Esen Yuksel Erdem, "A two-stream deep model for automated ICD-9 code prediction in an intensive care unit", Heliyon, Elsevier, Feb 2024
- [9] Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Yuan-Geng-Shuo Wang, Jia-ming Ji, Zifeng Qiu, "The application of large language models in medicine: A scoping review", National Library of Medicine, Elsevier, May 2024
- [10] Zihao Yang, Chenkang Zhang, Muru Wu, Xujin Chris Liu, Lavender Yao Jiang, Kyunghyun Cho, Eric Karl Oermann, "Intriguing Effect of the Correlation Prior on ICD-9 Code Assignment", ACL Anthology, Association for Computational Linguistics, Jul 2023
- [11] Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull and Jimeng Sun, "DRG-LLaMA : tuning LLaMA model to predict diagnosis related group for hospitalized patients", Digital Medicine, Jan 2024

- [12] Edgar Steiger, DPhil; Lars Eric Kroll, DPhil, "Patient Embeddings from Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework", Artificial Intelligence, JMIR, Oct 2023
- [13] Siva Sai, Aanchal Gaur, Revant Sai, Vinay Chamola Siva Sai Mohsen Guizani, Joel J. P. C. Rodrigues, "Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations", IEEE Access, Vol. 12, Mar 2024
- [14] Gaspard Harereimana Jong Wookkim, Gun Il Kim, Beakcheol Jang, "HSGA: A Hybrid LSTM-CNN Self-Guided Attention to Predict the Future Diagnosis From Discharge Narratives", IEEE Access, Vol. 11, Oct 2023
- [15] Yoon-Sik Cho, Eunsun Kim Younghoon Kwon, "Identifying Disease of Interest With Deep Learning Using Diagnosis Code", Journal of Korean Medical Science, Mar 2023
- [16] Zhichao Yang Hong Yu, Avijit Mitra, Weisong Liu, Dan Berlowitz, "TransformEHR: transformer-based encoder decoder generative model to enhance prediction of disease outcomes using electronic health records", Nature Communications, Nov 2023
- [17] Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, "Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models", Journal of Healthcare Informatics Research, Nov 2024
- [18] Fatma M. Talaat, "Revolutionizing cardiovascular health: integrating deep learning techniques for predictive analysis of personal key indicators in heart disease", Neural Computing and Applications, Springer, Jan 2025
- [19] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, "Large language models encode clinical knowledge", Nature, Jul 2023
- [20] Yunsu Byeon, Yae Won Park, Soohyun Lee, Doohyun Park, Hyung Seob Shin, Kyunghwa Han, Jong Hee Chang, Se Hoon Kim, Seung-Koo Lee, Sung Soo Ahn, Dosik Hwang, "Interpretable multimodal transformer for prediction of molecular subtypes and grades in adult-type diffuse gliomas", Digital Medicine, Mar 2025
- [21] Quentin Marcou, Laure Berti-Equille, Noël Novelli, "Creating a computer assisted ICD coding system: Performance metric choice and use of the ICD hierarchy", Journal of Biomedical Informatics, Elsevier, Vol. 151, Apr 2024
- [22] Atta Ur Rahman, Yousef Alsenani, Adeel Zafar, Kalim Ullah, Khaled Rabie, Thokozani Shongwe, "Enhancing heart disease prediction using a self-attention-based transformer model", Scientific Reports, Nov 2024
- [23] Christoph S. Metzner, Shang Gao, Drahomira Herrmannov, Heidi A. Hanson, "Attention Mechanisms in Clinical Text Classification: A Comparative Evaluation", IEEE Journal of Biomedical and Health Informatics, Vol. 28, No. 4, Apr 2024
- [24] Zhangdaihong Liu, Ying Hu, Xuan Wu, Gert Mertes, Yang Yang, David A. Clifton, "Patient Clustering for Vital Organ Failure Using ICD Code With Graph Attention", IEEE Transactions on Biomedical Engineering, Vol. 70, No. 8, Aug 2023
- [25] Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegela, Aysegul Bumin, Brandon Silva, Jessica Senae, Benjamin Shickelf, Azra Bihorac, Kia Khezelia, Parisa Rashidi, "Transformers and large language models in healthcare: A review", Artificial Intelligence in Medicine, Aug 2024
- [26] Khalid Nawab, Madalyn Fernbach, Sayuj Atreya, Samina Asfandiyar, Gulalai Khan, Riya Arora, Iqbal Hussain, Shadi Hijjawi, Richard Schreiber, "Fine-tuning for accuracy: evaluation of Generative Pretrained Transformer (GPT) for automatic assignment of International

Classification of Disease (ICD) codes to clinical documentation”, *Journal of Medical Artificial Intelligence*, Oct 2024

[27] Jakir Hossain Bhuiyan Masud, Chen-Cheng Kuo, Chih-Yang Yeh, Hsuan-Chia Yang, Ming-Chin Lin, “Applying Deep Learning Model to Predict Diagnosis Code of Medical Records”, *Diagnostics*, National Library of Medicine, Jul 2023

[28] Isabel Coutinho, Bruno Martins, “Transformer-based models for ICD-10 coding of death certificates with Portuguese text”, *Journal of Biomedical Informatics, ACM*, Vol. 136, Dec 2022

[29] Sander Puts, Catharina M L Zegers, Andre Dekker, Iñigo Bermejo, “Developing an ICD-10 Coding Assistant: Pilot Study Using RoBERTa and GPT-4 for Term Extraction and Description-Based Code Selection”, *Advancing Digital Health & Open Science*, Vol. 9, Jan 2025

[30] Johnson A., Pollard T., Mark R., “MIMIC-III Clinical Database (version 1.4) PhysioNet”, Oct 2016