Heliyon 10 (2024) e36569

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Ensemble neural models for ICD code prediction using unstructured and structured healthcare data

Alimurtaza Mustafa Merchant, Naveen Shenoy, Sidharth Lanka, Sowmya Kamath *

Healthcare Analytics and Language Engineering (HALE) Lab, Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Srinivas Nagar P.O., Mangalore, 575025, Karnataka, India

ARTICLE INFO

Keywords:
Artificial intelligence
Automated medical coding
Healthcare informatics
Label attention
Unstructured text modeling

ABSTRACT

Disease coding is the process of assigning one or more standardized diagnostic codes to clinical notes that are maintained by health practitioners (e.g. clinicians) to track patient condition. Such a coding process is often expensive and error-prone, as human medical coders primarily perform it. Automating diagnostic coding using Artificial Intelligence is seen as an essential solution in Hospital Information Management Systems and approaches built on Convolutional Neural Networks currently perform best. In this work, a neural model built on unstructured clinical text for enabling automatic diagnostic coding for given patient discharge summaries is proposed. We incorporate a structured self-attention mechanism designed to boost learning of label-specific vectors and the significant clinical text snippets associated with a certain label for this purpose. These vectors are then combined with a novel code description pipeline leveraging the descriptions provided for each standardized diagnostic code. The proposed model achieved best performance in terms of standard metrics over the MIMIC-III dataset, outperforming models based on Longformers and Knowledge graphs. Furthermore, to leverage structured clinical data to enhance the proposed model, and to enable improved diagnostic code prediction, model ensembling is considered. A neural model built on structured data by leveraging supervised machine learning algorithms such as random forest and boosting, is designed for multi-class code classification. Experimental results revealed that the proposed ensemble models show promising performance compared to traditional models that rely solely on unstructured or structured clinical data, emphasizing their suitability for real-world deployment.

1. Introduction

The International Classification of Diseases (ICD) [1] is a standardized metadata-based classification scheme used in global healthcare systems for a variety of tasks ranging from managing patient EMRs (Electronic Medical Records) [2], insurance processing, billing, etc. The assigning of standard codes to specific diagnoses and treatments carried out using information such as the medical notes or clinical notes recorded by medical practitioners upon patients' visit is a process known as ICD Coding. The World Health Organization (WHO) maintains the ICD codes, which are used worldwide with various adjustments made for particular nations. ICD codes can also be used for epidemiological studies, service billing, and other clinical research and healthcare-related activities. In addition to ensuring semantic interoperability and re-usability of recorded data, ICD coding supports resource allocation, reimbursement, guidance,

* Corresponding author.
E-mail address: alimurtaza0411@gmail.com (A.M. Merchant).

https://doi.org/10.1016/j.heliyon.2024.e36569
Received 13 February 2024; Received in revised form 17 August 2024; Accepted 19 August 2024

Available online 23 August 2024

and other use cases beyond regular healthcare activities. These codes can be used for various applications; for example, each service is identified by a common procedural technology (CPT) code if a healthcare provider bills an insurance provider for reimbursement, which corresponds to an ICD code. The insurance provider might refuse payment if the two codes do not line up correctly. Similarly, each disease has an ICD code assigned to it and based on a patient's diagnosis, one or more ICD codes are mapped to the associated medical records.

Highly trained human medical coders typically undertake the process of accurate ICD coding. As ICD codes have a pre-defined structure according to an ICD-9 labeling scheme, the categorization is performed using up to 5 digits (three digits and a decimal point to the left, followed by one or two digits to the right). The three-digit numbers to the left of the decimal are divided into chapters and sub-chapters, indicating a particular kind of disease in a fine-grained manner. For example, 001 to 139 represent infection and parasitic diseases, 290-319 represent mental disorders, 520 to 579 are used for digestive system-related diseases, and so on. The numbers to the right of the decimal sharpen the disease definition even more and are used to represent a very specific disease in the set. For example, 346 is used to denote migraine, and it has two main specific diseases, 346.0, which denotes classical migraine, and 346.1, which represents common migraine. ICD coding, when performed manually, requires a lot of effort, is time-intensive, and is prone to inevitable human errors. As it is undertaken based on key information extracted from the EMRs based on medical domain knowledge, coding errors can cause billing mistakes, underpayment, and future patient issues due to incorrect EMRs. Thus, fine-grained code assignment is a critical requirement, and an automated ICD coding system can greatly help address this problem.

The challenges associated with automating clinical coding stem from the diverse nature of data, its volume, and variety. While some of these codes are more common and are regularly used as correspond to commonly occurring diseases, many codes are infrequently employed due to the rareness of the representative disease. Thus, rule-based ML systems do not perform very well in the multi-label classification of ICD codes. Unstructured data sources, including clinical notes written by doctors and nurses, radiology reports, and discharge summaries, often lack structure and may contain notations, abbreviations, and misspellings. These sources primarily employ clinical terminology. Furthermore, pre-processing medical data is a non-trivial task. This is because medical data often lacks structure and also due to the presence of irrelevant passages, abbreviations, numbers, and misspellings, along with a clinical vocabulary. In addition to this, different sections of the clinical notes may contain information related to different diseases. Thus, effectively and efficiently using data from large-volume EMRs is a challenging task.

In this work, we propose an attention-based model to improve performance in classifying disease codes using unstructured text-based health records. The code descriptions of ICD code are utilized to incorporate the description of the disease in the model. MIMIC III (Johnson et al., 2016) discharge summaries are considered for the evaluation experiments. In addition to this, we explore various ensemble techniques to use structured data from patient lab reports to improve the ICD coding performance. The main contributions of this work are summarized below:

1. CD-LAAT, a novel label-attention based ICD coding model with a code description pipeline is proposed, which outperformed existing label-attention-based approaches [3] and achieved competitive performance with state-of-the-art baselines.
2. Design of a novel learning mechanism to incorporate code descriptions of each ICD code along with the content from clinical notes to adjust the attention given to different sections of text.
3. Development of ensemble models to effectively use structured data from patient lab reports and unstructured data from doctor notes to improve the ICD coding performance w.r.t. the CD-LAAT model.
4. Comprehensive benchmarking of the proposed models in Hospital Information Management Systems (HIMS) in comparison to existing state-of-the-art benchmarks using appropriate multi-label metrics such as AUROC, F1 and P@k scores as well as number of trainable parameters.

The remainder of this article is organized as follows. Section 2 discusses the existing works in the domain of ICD classification, highlighting the contribution and drawbacks of each. Section 3 presents a complete, in-depth discussion of the proposed approach and the associated processes. The details of the experimental setup, metrics used to evaluate the model, baseline models, and observed results obtained are presented in Section 4, followed by conclusions and possible directions for future work in section 5.

2. Related works

Related works in the ICD code prediction domain can be categorized into two categories. Works studying effective and efficient representations of clinical text and works studying feature selection and model ensembling techniques for ICD code prediction. We dive deep into both of the above categories in this study.

2.1. Convolution neural network architectures

Convolutional neural networks have been widely applied for automated ICD coding of unstructured clinical data. Li et al. [4] used a multi-filter residual CNN to overcome the flat and fixed size of the filter in convolutional architecture. Kernel sizes such as that of multi-layer filter and channel sizes were chosen empirically. Li et al. [4] achieved state-of-the-art AUC scores (macro, micro) of 0.899 and 0.928, F1 scores (macro, micro) of 0.606 and 0.670, and P@5 of 0.641 respectively.

Xie et al. [5] proposed a representation of EMR using RNN and CNN to predict the ICD codes. This research focused on leveraging the description of diagnosis in discharge summaries for code classification. They used a tree-of-sequences LSTM encoder to produce the representation of code description and diagnosis description. The representation was then passed through an adversarial learning module followed by discriminative networks to model the diagnosis description and code description and reduce the writing style differences between them.

2.2. Attention mechanism

Mayya et al. [6] proposed the LATA model (Label Attention Transformer Architecture) for the automatic assignment of ICD-10 codes and benchmarked it for the task of automatic clinical coding for multilingual medical documents. They also presented explainable interpretation through the visualization of attention weights learned by the model to reveal the associations between the clinical note and the predicted diagnostic code.

Building upon the idea of label representations using attention mechanism, label-attention models (LAAT) was developed to learn the representation for each ICD code/label [3] and achieved superior performance due to the style of representation, where each clinical note is represented by different available ICD codes generated using self-attention mechanisms.

Mullenbach et al. [7] introduced CAML (Convolutional Attention for Multi-Label classification), which incorporates attention not as a whole but to the representational vectors of a label. In CAML, descriptions were used to represent the labels and improve the prediction of rare labels through regularization.

Another effective convolution attention network with a deep convolution-based encoder as the input layer was proposed by Liu et al. [8], which was exclusively designed for Multi-Label Document Classification. They introduced ResSE, i.e., residual and squeeze-and-excitation networks, to represent the discharge summaries. In addition, to improve the performance of code prediction on rarely occurring diseases, they combined focal loss, which dynamically reduces the loss assigned to the commonly occurring labels with the regular binary cross-entropy loss. Later, an attention mechanism was applied to each label representation obtained from the ResSE networks.

2.3. Transformer based models

With the rising popularity and superior performance of transformer models in various NLP tasks, popular models like BERT (Bidirectional Encoder Representations from Transformers) have also been employed for the task of automated ICD coding. Pascual et al. [9] utilized PubMedBERT trained on biomedical data and related tasks. Different strategies were employed, such as using the first, middle to last 512 tokens for encoding. The first 512 tokens gave the best performance, further improved by combined encoded vectors of the first and last 512 tokens.

Zhang et al. [10] proposed BERT-XML where XML stands for Extreme Multi-Lable text classification to predict ICD-10 codes on de-identified notes from the NYU Langone Hospital EHR system and achieved AUC scores (micro, macro) of 97.0 and 92.7 respectively. Large computational demand for longer sequences and only marginal improvement in performance after the addition of XML to the model are the limitations of this study. However, the main difficulty faced in BERT-based transformers is fine-tuning long pieces of text, which cannot be processed at once. To improve the quality of ICD coding using BERT, there is a need to aggregate and summarize long texts effectively. Thus, to improve the performance of BERT, the encoding of the long text has to be improved. A proposed idea was to use different models to summarize and encode different text sections and combine them using the ensembling technique.

2.4. Code description embedding approaches

The Multiple Synonyms Matching Network (MSMN) proposed by Yuan et al. [11] also used a form of label attention. To match the text snippets to the specific codes, code representations were applied. It used the synonyms of the codes by aligning the codes to concepts in UMLS (Unified Medical language system) [12]. The text and the synonyms of the codes were encoded using bi-directional LSTM after word embedding. The proposed attention mechanism used a form of multi-head attention [13]. Yuan et al. [11] used the synonyms of the codes as queries to match different parts of the text for similarity, overcoming the challenges introduced while training rare codes.

2.5. Ensemble architectures

Various works have used structured and unstructured data and combine the two using ensemble technique for ICD code classification. Xu et al. [14] built an ensemble model consisting of models to process unstructured, semi-structured, and structured (or tabular) data. For structured data, multi-label classification using a simple decision tree was performed. CNN-based models were used to predict codes for unstructured and semi-structured data. The final probability of a label is the weighted sum of probabilities calculated by each model.

For ICD code prediction on structured data such as lab reports, rigorous benchmarking was attempted by [15] for various clinical tasks. These included the use of ensemble of machine learning models. Ensemble models for processing unstructured, semi-structured, and structured (or tabular) data for multi-label classification on structured data, using lab events, chart events, microbiology events, and prescriptions, were also proposed by [14]. They built different machine learning models for capturing relevant features for multiple data modalities considered. Other approaches of ensemble methods for ICD coding such as that of [16] and [15] focus on code group prediction which involves classifying medical records into a set of groups of multiple ICD codes. In these works, the ICD codes were grouped into 20 subgroups where each subgroup contained codes of a larger hierarchy of diseases. For example, ICD codes in the ranges 290-319 which denote various mental disorders were grouped as one class for prediction.

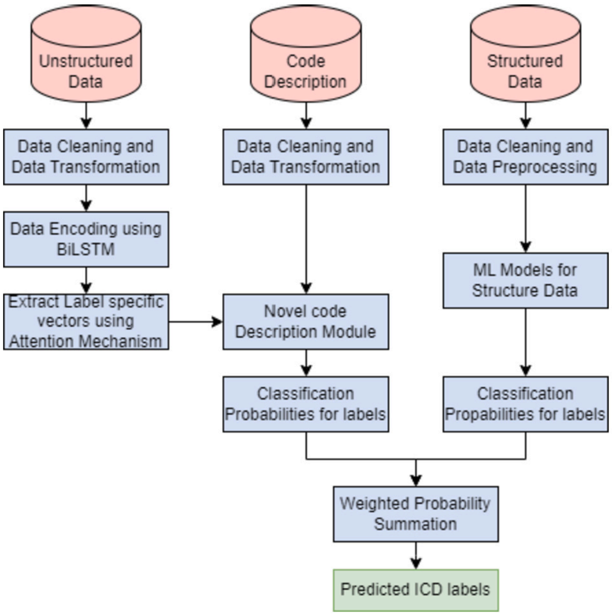
Unstructured Data
Code Description
Structured Data
Data Cleaning and Data Transformation
Data Cleaning and Data Transformation
Data Cleaning and Data Preprocessing
Data Encoding using BiLSTM
ML Models for Structure Data
Extract Label specific vectors using Attention Mechanism
Novel code Description Module
Classification Probabilities for labels
Classification Propabilities for labels
Weighted Probability Summation
Predicted ICD labels

Fig. 1. Proposed Methodology.

Some other models which has done a significant work in the field of ICD code classification are KEPTLongformer, the latest state-of-the-art [17] incorporated a contrastive learning strategy that uses the UMLS [12] medical knowledge graph to infuse knowledge of synonyms, acronyms, etc into the language model. However, a major gap is observed in the limited adoption of effective feature selection algorithms to reduce feature dimensions for tabular data and the huge complexity of the model. Apart from keyword matching, the incorporation of domain-specific knowledge to determine codes could also be investigated.

Gangavarapu et al. [18] proposed a vector space-based topic modeling approach to model unstructured clinical data based on fuzzy similarity-based techniques. They also proposed different supervised multi-label classification models to facilitate ICD-9 code group prediction.

This study aims to bridge the gaps in the above studies by proposing more efficient representations of clinical text for ICD coding of unstructured data. Additionally, we aim to design an ensemble model integrating a code description-based label attention model to classify clinical notes and a structured model to classify structured lab reports. Data sources like discharge summaries, data from Lab Events, Microbiology Events, and Prescription tables from the MIMIC-III dataset [19] are considered for the evaluation experiments. The main contributions of this work are the development of a novel ensemble model to perform ICD code prediction by effectively using structured data from patient lab reports and unstructured data from doctor notes to improve performance.

3. Methodology

Fig. 1 depicts the overall workflow of the proposed strategy of the mechanisms for handling unstructured and structured data and ensembling the models.

The proposed ensemble model for ICD code prediction is primarily composed of two parts - the structured data (which includes lab results, drugs prescribed, and microorganisms found in the body fluids of the patient, extracted from the MIMIC III dataset) and the unstructured data (comprising of the discharge summary of the patient, obtained from the MIMIC-III dataset). For modeling the structured data, machine learning models like random forest and various boosting algorithms are adopted for predicting the ICD code on structured data. An attention-based CD-LAAT (Code Description-Label Attention-based Transformer) model is designed for modeling the unstructured data, and for learning the clinical note representations as vectors representing one of the ICD codes. The code description is also incorporated in these vectors through a novel module. The two branches are combined to create various combinations of both types of data, for the task of ICD9 code prediction (shown in Fig. 7).

3.1. Data specifics

In this work, the MIMIC-III database (Medical Information Mart for Intensive Care III) [19] was used for experimental validation. It provides de-identified health data of over 40,000 patients who had been admitted between 2001 and 2012 to the Beth Israel Deaconess Medical Center's critical care units. All records in the dataset are associated with ICD-9 codes, clearly representing the diagnoses and procedures. Codes are divided into sub-codes that frequently contain particular particularized information. The collection contains 112,000 clinical report records with information like medications provided, duration of hospital stay, results from the laboratory, observations by the healthcare staff, vital signs, and other important information recorded in the data.

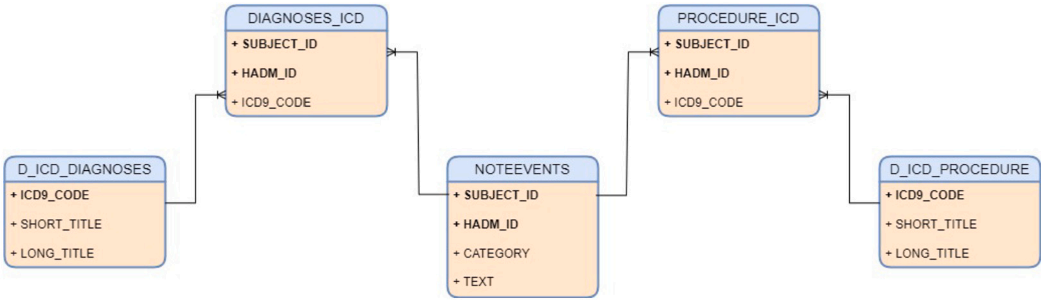

Fig. 2. MIMIC-III Unstructured Data.

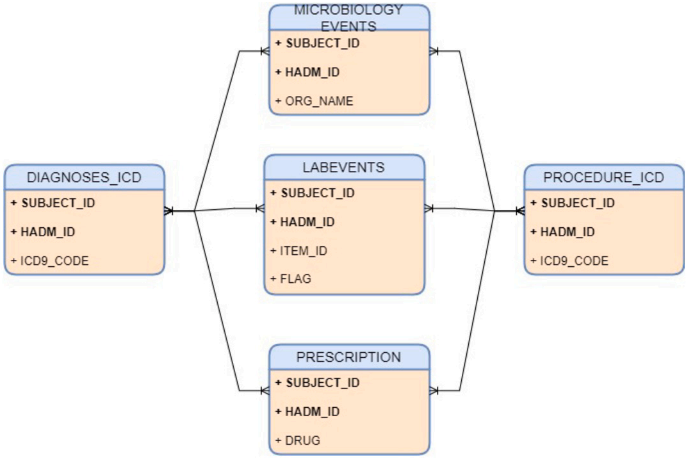

Fig. 3. MIMIC-III Structured Data.

3.2. Data preparation

We used discharge summaries from the NOTEEVENTS table, which contains most patient information as clinical text following previous works. Fig. 2 shows the entity relationship diagram of all the data tables from the MIMIC III dataset used in this work. The dataset contains a total of 52,722 unique discharge summaries, each of which is labeled with corresponding ICD-9 codes. Considering summaries containing at least one of the top-50 most common ICD-9 codes, a total of 49,414 summaries are obtained. For our work, the same summaries used in previous studies [7], [3] were considered to enable benchmarking. The final subset used consists of 11,317 discharge summaries, which were split into 8,066 for training, 1,473 for validation, and 1,729 for testing.

For generating the dataset for structured data based prediction, three data tables made available in the MIMIC III dataset are used – LABEVENTS, MICROBIOLOGYEVENTS, and PRESCRIPTIONS. The MIMIC-III dataset contains EHR data in a structured format in these tables. The relation between tables is shown as ER Diagram in Fig. 3. We omit events such as heart rate data that is present in the table CHARTEVENTS. LABEVENTS contains records of various lab test results, MICROBIOLOGYEVENTS detects the presence of various microorganisms in patient biological fluids, and PRESCRIPTIONS contain records of drugs the doctor prescribes. First, we filter the data by performing a null analysis. The LABEVENTS dataset contains multiple occurrences of NaN values of HADM_IDs and test results (FLAG). Such data points are filtered out in the null analysis. LABEVENTS gets filtered to about 27 million rows after this step comprising of 58151 unique admission ids. After performing a null analysis, we handle imbalanced nature of the tests and their results. In the LABEVENTS table, test positive results are very sparse, in the sense, for most unique tests, only a small proportion of patients test positive for the test. In the preparation of the LABEVENTS data, we only consider those lab tests where at least 5% of patient admissions give positive results on the test. Null analysis is carried out in a similar way for PRESCRIPTIONS and MICROBIOLOGYEVENTS. For PRESCRIPTIONS table, to handle imbalanced data, we only consider those drugs that have been prescribed at least 50 times. MICROBIOLOGYEBVENTS table does not require data balancing. These steps are performed in consistency with the works of [15] and [14]. and removing imbalanced data, and generating the features that were considered by an earlier study [15]. In the end, all three files are combined based on *HADM_ID* (admission ID). The final dataset has 18,536 unique *HADM_ID* and 1,530 binary features, thus forming the structured dataset for prediction.

The unstructured data comprises the discharge summaries obtained from the NOTEEVENTS file of MIMIC III. NOTEEVENTS contains a large number of clinical text-based records categorized based on their type and by the patient admission ids. We consider only discharge summaries as the category for clinical text in our ICD coding task. Thus, NOTEVENTS table is filtered using the “DISCHARGE SUMMARIES” category value. The code description is the long title taken from files D_ICD_DIAGNOSES and D_ICD_PROCEDURE. Each

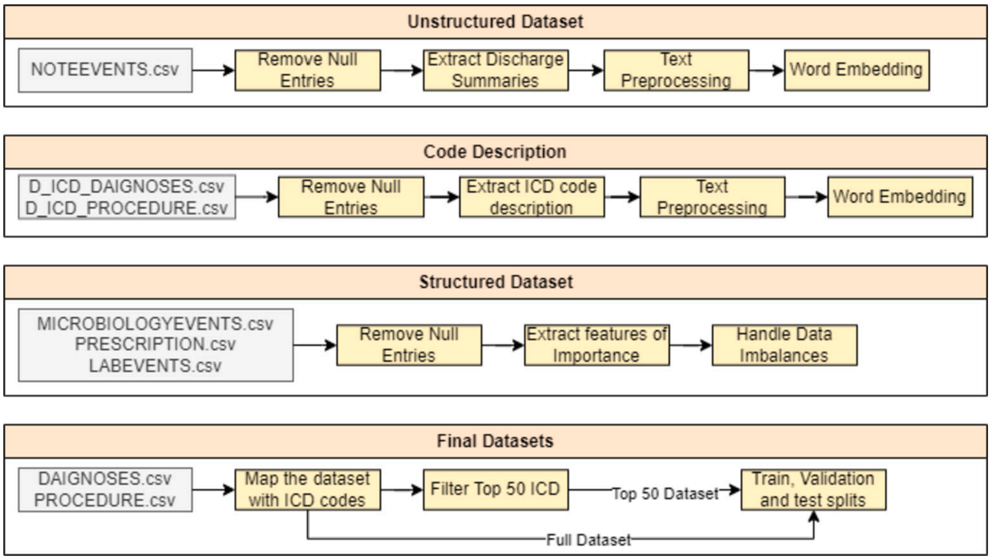

Fig. 4. Data Preparation.

ICD code is mapped to a description describing the medical diagnosis briefly. The same text preprocessing is performed on the code description, similar to the processes applied to the discharge summaries. The same subset of the patients used by previous state-of-the-art models like [7] were considered for our work also, to enable comparative evaluation. The text is processed to generate text embeddings the gensim implementation of CBOW Word2Vec model. The embedding size is set as 100, and the final data consists of 11,317 unique *HADM_ID*. The two data sources are now combined based on *HADM_ID*, to finally generate a dataset of 4245 instances. The dataset is split into *train*, *test*, and *dev* with sizes of 2979, 583, and 683 respectively. This dataset, referred to as *MIMIC50-Combined*, is used to train unstructured data based models used in the evaluation experiments. Fig. 4 illustrates the steps taken to process the various MIMIC III files used in this research.

3.3. Data preprocessing

For unstructured data, the discharge summaries from the NOTEEVENTS table are in the form of free text. The text is tokenized, and all tokens are made in lowercase. Tokens containing no alphabetic characters are removed, including punctuation. Earlier experiments by Li et al. [4] found no significant differences in performance when truncating text between lengths 2500 and 6000 tokens. Hence, we truncate all text to a maximum length of 4000 tokens in order to maintain consistency with work by Xie et al. [20]. The Word2Vec Continuous Bag of Words (CBoW) model [21] is used to obtain word embeddings of size $d_e = 100$, which were used in recent studies.

3.4. Structured data based ICD code prediction model

In order to predict ICD codes from structured data, three different classes of models were employed: Random Forest, Multi-layer Perceptron, and boosting-based classifiers. For the prediction of ICD codes, which involves multiple labels, a multi-output random forest model was utilized. This approach involved fitting unique random forest models for each ICD code, enabling more accurate predictions. A Multi-layer Perceptron Model was also employed, consisting of six fully connected layers. The hidden layers utilized the hyperbolic tangent (tanh) activation function, while the output layer employed the sigmoid function. This model aimed to explore the effectiveness of neural networks in predicting ICD codes using structured data. Boosting algorithms, namely XGBoost and Adaboost, were implemented to construct multi-output classifiers. These algorithms were tested to determine the most suitable model for predicting ICD codes using structured data from the MIMIC III dataset.

3.5. Final datasets used for testing

The following section provides an overview of the various datasets used for prediction, described in the previous sections.

3.5.1. MIMIC-III full dataset

The MIMIC-III Full dataset consists of the 52,722 unique discharge summaries from the NOTEVENTS table, where we classify diseases into each of the ICD codes present in the MIMIC-III dataset.

3.5.2. MIMIC-III Top-50 dataset

The MIMIC-III Top-50 dataset consists of 49,414 unique discharge summaries from the NOTEVENTS table where each discharge summary belongs to one out of the top-50 most commonly occurring ICD codes in the MIMIC-III dataset.

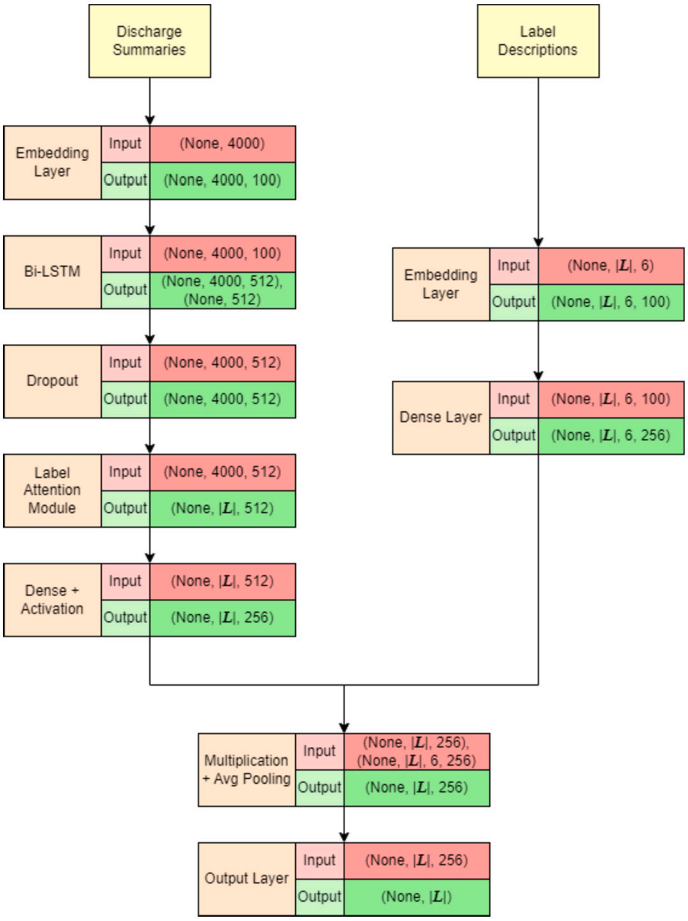

Fig. 5. Proposed Code Description based Label Attention Model Architecture (CD-LAAT).

3.5.3. MIMIC50-Combined

The MIMIC50-Combined dataset is used to test the performance of the various ensemble models for ICD code prediction. As described in 3.2, MIMIC50-Combined is formed when the MIMIC-III Top-50 dataset combined with structured dataset to get 4245 unique instances.

3.6. CD-LAAT: unstructured data based ICD code prediction model

Fig. 1 illustrates the overall processes designed as part of the proposed methodology for CD-LAAT. Fig. 5 shows the detailed architecture of the proposed model. The model is categorized into several layers, namely, the embedding layer used to represent the input clinical text. An encoding layer, where the input is passed through a language model. Further layers include an attention module to capture essential sections from the sequence output of the language model for code classification and another parallel code description module that encodes the descriptions of each class label. The output from these attention and code description layers are combined in the final classification module, which produces the final output.

In CD-LAAT's Embedding layer, each word in a clinical document D consisting of n words is represented by a pre-trained Word2Vec CBOW embedding of fixed size d_e. This document is represented in $R^{n\times d_e}$. A Bidirectional LSTM Encoder is used here for deriving contextual information from all the words which are present in D. The input words from a sequence $e_{w_1:w_n}$ of vectors $e_{w_1}, e_{w_2}, ..., e_{w_n}$ are represented using latent feature vectors, and the BiLSTMs play a major role in learning these vectors. The hidden states of all the LSTMs which correspond to the j^{th} word ($j \in 1, ..., n$) are computed as per Eq. (1), (2) and (3), where, $\overrightarrow{LSTM}$ denotes forward LSTMs while $\overleftarrow{LSTM}$ denote backward LSTMs. The final latent vector h_j is derived by concatenating the 2 vectors, $\overrightarrow{h_j}$ and $\overleftarrow{h_j}$. The size of the latent vectors h_j at $2u$, as the LSTMs hidden states had a dimensionality of u. A matrix is formed by concatenating all of the hidden state vectors of words in D to get the matrix $H = [h_1, h_2, ..., h_n] \in R^{n\times 2u}$.

$$\overrightarrow{h_j} = \overrightarrow{LSTM}(e_{w_1:w_j}) \tag{1}$$

$$\overleftarrow{h_j} = \overleftarrow{LSTM}(e_{w_j:w_n}) \tag{2}$$

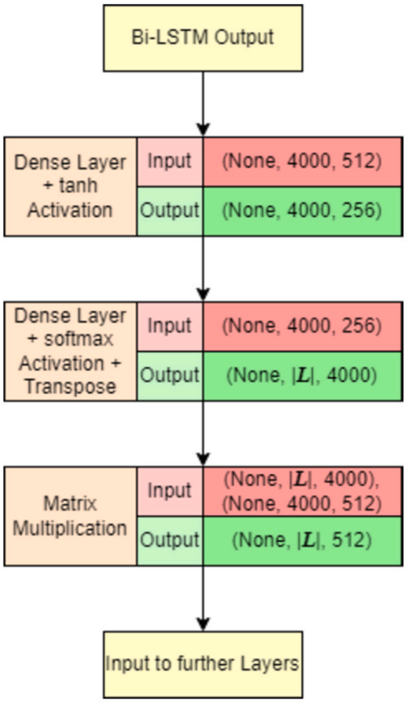

Fig. 6. Attention Module of proposed CD-LAAT model.

$$h_j = \overrightarrow{h_j} \oplus \overleftarrow{h_j} \tag{3}$$

The next component is the Attention Module integrated into the proposed model. The clinical documents are of varying lengths, each with multi-labels; the objective is to get label-specific vectors by transforming H.

This is achieved using the label attention mechanism described by Vu et al. [3]. It takes H as the input and outputs $|L|$ label-specific vectors representing the input Document D. The label-specific weights vector is computed as per Eq. (4) and (5).

$$Z = tanh(HW) \tag{4}$$

$$A = softmax(ZU) \tag{5}$$

Here, we have a matrix $M \in R^{2u \times d_a}$, where d_a is a hyper-parameter which is then tuned with the model and results in a matrix $Z \in R^{n \times d_a}$. The matrix Z is used to be multiplied with another matrix $U \in R^{d_a \times |L|}$, so the label-specific weight matrix $A \in R^{n \times |L|}$ can be computed, where each i^{th} row of A^T is a weight vector referring the i^{th} label in L.

Then, the activation function softmax is used to make the summation of weights equal to 1 at each row. Then, we multiply the transpose of matrix A by the matrix H, in which the input document D is represented by label-specific vectors as per Eq. (6). The i^{th} row of the matrix $V \in R^{|L| \times 2u}$ is a representation of D regarding the i^{th} label in L. Fig. 6 shows the architecture of the label attention module with the BiLSTM hidden size = 256, text length = 4000 tokens, and d_a = 256.

$$V = A^T H \tag{6}$$

In the proposed CD-LAAT model, a novel code description module is also incorporated, which takes the ICD code descriptions into account when predicting ICD codes. The preprocessing of the code descriptions is identical to the process adopted for the text from the discharge summaries. The label descriptions are truncated to the average length of the description of $|L|$ labels and are embedded using the same embedding layer described earlier in Section 3.6. Embeddings of size m are computed for all the label descriptions. For the i^{th} label, the embedding can be represented as per Eq. (7), where, $E_L \in R^{|L| \times m \times d_e}$ and d_e is the embedding size.

$$E_{L_i} = [e_{Li_1}, e_{Li_2}, ..., e_{Li_m}] \tag{7}$$

The embeddings are passed through the linear layer $B = E_L W_L + b_L$, where, matrix $W_L \in R^{|L| \times d_e \times d_b}$. d_b is a hyper-parameter that is hyper-tuned. The output matrix is $B \in R^{|L| \times m \times d_b}$, where, every row represents a label in $m \times d_b$ dimensions. The output from the attention layer is passed through a linear layer with a *tanh* activation function. This is done to normalize the values in the output

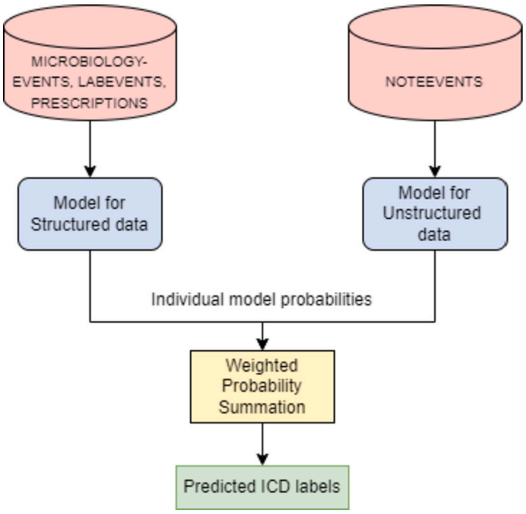
MICROBIOLOGY-EVENTS, LABEVENTS, PRESCRIPTIONS
NOTEEVENTS
Model for Structured data
Model for Unstructured data
Individual model probabilities
Weighted Probability Summation
Predicted ICD labels

Fig. 7. Proposed Ensemble Strategy.

matrix V after the matrix multiplication is performed in the attention layer. The output matrix $V \in R^{|L|\times 2u}$ and is used to compute the final label-specific representation of document D as per Eq. (8).

$$T = tanh(VW_x) \tag{8}$$

Here, the weights matrix $W_x \in R^{2u\times d_b}$ and the resulting matrix $T \in R^{|L|\times d_b}$. In T, every row represents the label-specific representation of document D. Now, for each i^{th} label-specific representation of the document in T, it is multiplied with i^{th} label description representation which $m \times d_b$. Then, we average the result from $m \times d_b$ to d_b as shown in Eq. (9), where, the output $Y \in R^{|L|\times d_b}$ and every i^{th} row in Y represents information of i^{th} label. This is passed through the classification layer to get the probabilities for each label. Fig. 5 shows the model architecture with a sample text length of 4000 tokens, average description length = 6 tokens, d_a = 256 and d_b = 256.

$$Y_i = \frac{1}{m}\sum_{j=1}^{m} B_{ij}T_i \tag{9}$$

Now, the final task is to perform label classification to obtain the predicted ICD codes. Each label-specific representation given by v_i is passed as input to the corresponding Feed-Forward Neural Network. The neural network output layer contains a single node, followed by the sigmoid activation function this produces a probability for the i^{th} label by which it can be classified for the given document. After that, thresholding is performed on the probability to predict the output, which is binary $\in \{0, 1\}$. A predefined threshold of 0.5 was used for the prediction in our experiments. The objective in training is to minimize the loss value which is calculated using a binary cross-entropy loss function given by Eq. (10), where, θ denotes all the trainable parameters.

$$Loss(D, y, \theta) = \sum_{i=1}^{|L|} y_i \times log(\overline{y_i}) + (1 - y_i) \times log(1 - \overline{y_i}) \tag{10}$$

Thus, in the proposed CD-LAAT model, we leverage label descriptions in addition to the discharge summaries for ICD-9 coding. The label description module effectively utilizes short descriptions provided for each disease from the MIMIC-III dataset by encoding and then combining them with the code representations for each ICD code. Several experiments were performed to benchmark the performance of the proposed CD-LAAT model, the details of which are presented in Section 4.

3.7. Model ensembling

The branches dealing with structured and unstructured data are combined to build an ensemble model. The ensemble model assigns weights for both models built using structured and unstructured data respectively. The sum of these weights is 1. The final probability for an ICD code is the weighted sum of probabilities assigned by the individual models. The weights for the individual models serve as hyperparameters and hence are computed using a grid search strategy. We refer to the weight obtained by the structured data model as “Alpha”. The weight obtained by the unstructured data model would then be 1 - “Alpha”. The parameter “Alpha” denotes the linear weight for the structured data model for which the ensemble model achieves the optimal results. The final classification of codes is decided based on the weighted sum of the probabilities from both the structured and unstructured models for each code.

4. Experimental results and analysis

4.1. Experimental setup

As stated earlier, the MIMIC-III dataset was used for the experimental validation of the proposed CD-LAAT and ensemble models. The implementation was undertaken using the Pytorch [22] framework. Models were trained on a 32 GB NVIDIA Tesla V100 GPU. For CD-LAAT, as stated earlier in Section 3.3, all text inputs were truncated to a maximum length of 4000 tokens in order to maintain consistency with state-of-the-art work by Xie et al. [20]. Word embeddings of size $d_e = 100$ were generated using the CBOW word2vec model. A gensim implementation of the CBoW Word2vec model [23] was used, where, min_count was chosen as 0 and n_epochs was set to 50.

4.2. Evaluation metrics

The model was evaluated using several standard metrics, including Precision at k ($P@k$ in 8, 10, 15), Micro-F1 score, Macro-F1 score, Micro AUROC, and Macro AUROC, to facilitate a fair assessment against state-of-the-art methods and prior research. The micro-F1 score is calculated by treating each *(text, code)* combination as an independent prediction, whereas, the macro-F1 score is calculated by averaging metric scores generated for each label. It is to be noted that, rare label prediction is given higher significance in the macro-F1 score.

$$Precision_{micro} = \frac{\sum_{i=1}^{|L|} TP_i}{\sum_{i=1}^{|L|}(TP_i + FP_i)} \tag{11}$$

Micro-precision (Eq. (11)) is calculated as the summation of all the true positive counts per label and normalized by the sum of the true positive and false positive total counts. Similarly, the micro-recall (Eq. (12)) is calculated by finding the summation of the true positives and dividing them by the sum of true positives and false negatives of all labels.

$$Recall_{micro} = \frac{\sum_{i=1}^{|L|} TP_i}{\sum_{i=1}^{|L|}(TP_i + FN_i)} \tag{12}$$

After the micro-precision and micro-recall values are calculated, we find the micro-F1 score using Eq. (13).

$$F1_{micro} = 2.\frac{Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{13}$$

For calculating macro-precision and macro-recall, we find the average precision and recall by summing up the values for each label and dividing them by the number of labels as shown in Eq. (14)

$$Precision_{macro} = \frac{1}{|L|}\sum_{i=1}^{|L|}\frac{TP_i}{TP_i + FP_i} \tag{14}$$

and Eq. (15) macro-recall

$$Recall_{macro} = \frac{1}{|L|}\sum_{i=1}^{|L|}\frac{TP_i}{TP_i + FN_i} \tag{15}$$

We used the macro-precision and macro-recall formulae to find the macro-F1 score as in Eq. (16) macro-F1.

$$F1_{macro} = 2.\frac{Precision_{macro} \times Recall_{macro}}{Precision_{macro} + Recall_{macro}} \tag{16}$$

True Positive Rate (TPR) (Eq. (17)) and False Positive Rate (FPR) (Eq. (18)) are plotted for each label value, and micro-AUC is calculated. For macro AUC, the ROC curve is plotted for each label separately, and the area is calculated individually, after which the average value of all the ROC curves is considered the macro-AUC value.

$$TPR = \frac{TP}{TP + FN} \tag{17}$$

$$FPR = \frac{FP}{FP + TN} \tag{18}$$

4.3. Baselines

The proposed CD-LAAT model is benchmarked against recent state-of-the-art baselines, incorporating conventional machine learning models and more complex deep learning-based models. The details of these are given below.

- **Logistic Regression [7]:** implemented a binary classifier-based Logistic Regression approach with features as a bag-of-words for each ICD code included in the MIMIC-III dataset discharge summaries.

Table 1
Experimental Results for MIMIC-III Top-50 ICD Codes Test set.

Model	AUC		F1		P@k		
	Macro	Micro	Macro	Micro	P@5	P@8	P@15
LR [7]	82.9	86.4	47.7	53.3	54.6	-	-
CNN [7]	87.6	90.7	57.6	62.5	62.0	-	-
CAML [7]	87.5	90.9	53.2	61.4	60.9	-	-
MSATT-KG [24]	91.4	93.6	63.8	68.4	64.4	-	-
BiLSTM	81.4	86.2	43.7	44.3	48.8	40.5	29.8
LAAT [3]	92.5	94.6	66.6	71.5	67.5	**54.7**	35.7
JointLAAT [3]	92.5	94.6	66.1	71.6	67.1	54.6	35.7
KEPTlongformer [17]	**92.6**	**94.7**	**68.9**	**72.8**	67.2	-	-
CD-LAAT *(proposed)*	92.5	**94.7**	68.3	71.6	**67.9**	**54.7**	**36**

Table 2
Experimental Results for MIMIC-III full Test set.

Model	AUC		F1		P@k		
	Macro	Micro	Macro	Micro	P@5	P@8	P@15
LR [7]	56.1	93.7	1.1	27.2	-	54.2	41.1
CNN [7]	80.6	96.9	4.2	41.9	-	58.1	44.3
CAML [7]	89.5	98.6	8.8	53.9	-	70.9	56.1
MSATT-KG [24]	91.4	**99.2**	9.0	55.3	-	72.8	58.1
LAAT [3]	91.9	98.8	9.9	**57.5**	**81.3**	73.8	59.1
JointLAAT [3]	92.1	98.8	10.7	**57.5**	80.6	73.5	59.0
KEPTlongformer [17]	-	-	-	-	-	-	-
CD-LAAT *(proposed)*	**92.2**	98.8	**12.0**	57.1	81.2	**73.9**	**59.3**

- **CNN [7] :** 1-D Convolutional Neural Network implemented was used on the MIMIC dataset.
- **CAML [7] :** On the MIMIC datasets, the CAML model achieved strong performance, as it uses an attention layer over a CNN, to produce label-dependent representations for each label.
- **MSATT-KG [24] :** The model that presents the best performance on MIMIC discharge summary data at present. The MSATT-KG model incorporates multi-scale feature attention to select features dynamically, as well as a fully-connected CNN that can output variable n-gram features. Graph CNN is also utilized in the model to detect the hierarchical connections between ICD codes.
- **LAAT & JointLAAT [3] :** used the hierarchical relationship between the codes for ICD code prediction in JointLAAT. The LAAT model is a non-hierarchial model used to predict ICD codes by learning label-specific vectors using the attention mechanism. LAAT and JointLAAT produced state-of-the-art performance on the ICD Top-50 coding task.
- **KEPTLongformer [17] :** To counter the limitations of BERT, the KEPTLongformer model introduced a contrastive learning approach. Further, UMLS medical knowledge graph was used to effectively inject knowledge of abbreviations, synonyms, etc. into the language model.

We also perform benchmarking experiments to evaluate the ensemble models. We evaluate our ensemble models against current works by comparing it against the work of Xu et al. Xu et al. (2019). Xu et al. (2019) considered the top-32 frequent codes considering the MIMIC-III dataset as well as the from patient diagnoses from a national hospital in the United States.

4.4. Results

4.4.1. Results for CD-LAAT model

The proposed work consists of two models, firstly, the BiLSTM model built using the same architecture as the CD-LAAT model but without the attention, code description modules, and, secondly, the CD-LAAT model. The evaluation with reference to the MIMIC-III top-50 ICD codes tests set and MIMIC-III full test set ICD code prediction are presented in terms of standard metrics like AUC (macro, micro), F1 (macro, micro), and P@k scores. Table 1 presents the results for the top-50 codes test set and Table 2 represents the results obtained for the MIMIC-III full test set.

We observe that the CD-LAAT model outperforms the Logistic Regression model, CNN-based model, CAML, MSATT, and the BiLSTM model across all metrics for over the MIMIC-III Top-50 test set. When compared to LAAT and JointLAAT, the performance of the proposed CD-LAAT model is on par in terms of AUC (micro, macro), Micro F1, and P@8 metrics. However, the proposed CD-LAAT outperforms LAAT and JointLAAT in Macro-F1, P@5, and P@15 metrics. When compared to the current state-of-the-art model, KEPTlongformer, CD-LAAT outperforms based on the P@5 metric (an improvement of 0.7) while achieving on-par performance in terms of the AUC metrics. KEPTlongformer performs slightly better in terms of the F1 metric (+0.5 for Macro-F1, +1.2 for Micro-F1 score). CD-LAAT also achieved better P@k scores when compared to LAAT, JointLAAT, and KEPTLongformer.

Table 3
Performance of proposed unstructured & structured ensemble models.

Model	AUC		F1		Alpha
	Macro	Micro	Macro	Micro	
CD-LAAT	89.75	92.49	66.16	68.92	-
CD-LAAT+RF	90.07	92.84	**66.24**	69.09	0.175
CD-LAAT+MLP	89.75	92.49	66.16	68.92	0.0
CD-LAAT+XGBoost	**90.14**	**92.88**	66.18	**69.23**	0.15
CD-LAAT+AdaBoost	89.87	92.44	66.23	69.01	0.178

RF: Random Forest, MLP: Multilayer Perceptron.

For the MIMIC-III full test set, CD-LAAT outperforms all the above baselines across all metrics except for MSATT-KG micro-AUC score and LAAT, JointLAAT for the Micro-F1 score. KEPTlongformer being a computationally heavy model, cannot be used for ICD-9 prediction on the full test set. A significant increase in performance is observed w.r.t Macro-F1 outperforming the previous best (10.7 in JointLAAT) with a 12.0 Macro-F1 score. For other metrics such as AUC(macro, micro), Micro-F1, and P@k metrics, CD-LAAT performs at par or slightly better than the baselines.

For the label attention-based models, it is observed that LAAT [3] performs well on the MIMIC-III Top-50 dataset, while, the JointLAAT [3] performs better on the MIMIC-III full dataset. CD-LAAT does not suffer from the above issue and provides the best or better performance when compared with LAAT and JointLAAT. LAAT and JointLAAT perform similarly for both AUCs, i.e Micro-AUC and Macro-AUC in both top-50 codes and full test sets. For the MIMIC-III 50 test set, CD-LAAT outperforms the previous state-of-the-art models, especially in the Macro-F1 score (+1.7) and P@5 (+0.8) metric. Performance in other metrics such as Micro-AUC, P@5, P@8, and P@15 is slightly better than the LAAT/JointLAAT scores. Performance in Micro-AUC is the same as that of LAAT and JointLAAT models. For the MIMIC-III full test set, CD-LAAT excels again in the Marco-F1 score outperforming the current best of JointLAAT (+1.3) with performance similar to the combined best of LAAT and JointLAAT across other metrics.

4.4.2. *Results for ensemble model*

The results of the experiments are tabulated in Table 3. The base CD-LAAT model trained on unstructured data only achieved macro-AUC and micro-AUC scores of 89.75/92.49, macro-F1 and micro-F1 scores of 66.16/68.92 respectively.

4.4.3. *Benchmarking experiments*

We also evaluated the proposed ensemble models against existing works. For comparison, the work of [14] was chosen, which is the current SOTA work considering both the MIMIC-III dataset as well as the patient diagnoses for training the model for top-32 frequent ICD code prediction. Among the proposed models, the best-performing ensemble model trained on structured and unstructured data was the *CD-LAAT + XGBoost* model, which outperformed other combinations in terms of almost all metrics (Refer Table I).

4.5. *Discussions*

4.5.1. *CD-LAAT*

On the MIMIC-III top-50 dataset, the improved performance of CD-LAAT in the Macro-F1 metric indicates that CD-LAAT is better at predicting infrequent codes as compared to LAAT, and JointLAAT. In addition to this, CD-LAAT's better P@k scores compared to LAAT, JointLAAT, and KEPTLongformer show that it would be better in hospital environments where a high confidence score is desirable for ICD code matching with the discharge summaries, which increases the system's trustworthiness.

On the MIMIC-III full test set, an improved performance in the Macro-F1 score over the MIMIC-III full test set by CD-LAAT is particularly useful when deployed in hospitals where predicting ICD codes for rare/infrequent diseases is necessary. Performing at or above the SOTA level in other metrics adds to its suitability for automated ICD code assignment in Medical Record Management Systems.

Both models outperforming on Macro-F1 metric shows that the models are good at assessing each disease independently, since a high macro-F1 suggests that the model achieved a high recall and precision for each predicted class. This means, the model is not particularly bad at classifying any particular disease or group of disease.

The outstanding performance on P@k metrics suggests that the model performs well in prioritizing and predicting relevant diseases when multiple diseases may be present in the instance.

4.5.2. *Scalability of CD-LAAT*

One of the aspects of evaluating the effectiveness of the proposed CD-LAAT model is to assess its size in comparison to state-of-the-art models. For top-50 ICD code prediction, the KEPTlongformer [17] is state-of-the-art, but it has limited application due to the considerable model size of 119.4M trainable parameters.

This is attributed to the use of longformers with a max limit of 8192 tokens. As stated by the authors, the model is unsuitable for ICD coding on the MIMIC-III full test set due to memory constraints [17]. This limitation may prevent the use of such a model in commercial applications where quick predictions may be expected on the go. KEPTlongformer has a total of 119.4M trainable parameters. CD-LAAT, on the other hand, has only 1.04M and 5.59M parameters when trained on MIMIC-III top-50 and full test sets

Table 4
Comparison of proposed models with Xu et al.'s [14] work.

Model	AUC		F1	
	Macro	Micro	Macro	Micro
Xu et al. (2019)	**91.77**	94.06	**61.37**	69.29
CD-LAAT + XGBoost *(proposed)*	91.18	**94.84**	56.39	**70.04**

Table 5
Comparison of trainable model hyperparameters.

Model	No. of trainable parameters
KEPTlongformer [17]	119.4 Million
Proposed CD-LAAT (Top-50 codes)	1.04 Million
Proposed CD-LAAT (All codes)	5.59 Million

respectively, as shown in Table 5. This means a 99.13% reduction in the model parameters, which significantly enhances CD-LAAT's suitability for real-world applications.

4.5.3. Ensemble model

All ensemble models except the multi-layer perceptron model improve over the results obtained by CD-LAAT. The ensemble of CD-LAAT and multi-layer perceptron resulted in an obtained weight of 0 for the multi-layer perceptron model and 1 for the CD-LAAT model. This indicates no improvement was obtained from the multi-layer perceptron model when combined with the CD-LAAT model. Of all the classifiers, CD-LAAT+XGBoost provided optimum performance with highest scores in Macro and Macro AUC and Micro F1 metrics. For all the above trained models, it is observed that optimal ensemble weights are obtained at low values of alpha. This shows the limited capacity of the structured data to predict ICD codes independently.

4.5.4. Benchmarking experiments

CD-LAAT + XGBoost is compared against the *TextCNN+LS+TD* model proposed by [14] in Table 4. Experimental observations show that the proposed *CD-LAAT + XGBoost* ensemble model outperformed [14]'s model by a respectable margin, in terms of the Micro-F1 and Micro-AUC metrics, at the "Alpha" value of 0.152. Although Macro metrics shed light on the performance of individual classes, Micro metrics better reflect the overall performance of the model. Considering the practical importance of consistent and dependable predictions for every class, the proposed model's superior Micro-F1 and Micro-AUC scores render it a more appropriate option for practical deployments in Hospital Information Management Systems (HIMS). An inspection of the "Alpha" values in Table. 3 shows that optimal ensemble performance is achieved with low values of "Alpha", which indicates that a major contribution is made by CD-LAAT in ICD code prediction. This may be due to the vastness of the structured data in terms of the number of features it contains.

5. Conclusion and future work

In this research, we proposed CD-LAAT, implementing attention mechanism aids in extracting label-specific context from unstructured clinical documents of varying lengths. At the same time, the code descriptions for each disease are used effectively to aid in extracting information from text which may not have been captured solely using the label-attention mechanism. We used discharge summaries from the MIMIC-III dataset as a benchmark. The proposed model, CD-LAAT, provides competitive performance with the current state-of-the-art models and even outperforms them across the metrics, Macro-F1 score for both MIMIC-III 50 and MIMIC-III full and the Precision at 5 (P@5) for MIMIC-III 50 with performance comparable to the state-of-the-art across other metrics.

We also developed ensemble neural models trained on clinical concepts learned from unstructured and structured clinical data to predict ICD codes. The designed ensemble model utilizes structured data in the form of patient lab reports, in addition to doctors' clinical notes to perform automated ICD coding. The proposed models achieved promising results as an improvement over the individual use of CD-LAAT for ICD coding. Utlilizing standardized medical vocabulary like the Unified Medical Language System (UMLS) may further enhance ICD coding performance. By utilizing UMLS, various expressions in discharge summaries could be mapped to medical concepts, helping overcome inconsistencies in language terminologies. Additionally, the synonyms may better match the portions of discharge summaries corresponding to the disease, increasing the probability of detecting the correct disease. As part of future work, we intend to explore this avenue and conduct experiments to evaluate the proposed models when UMLS knowledge is integrated in the CD-LAAT model. As for the structured data, we aim to explore finer feature selection and ensembling techniques that may help boost performance.

CRediT authorship contribution statement

Alimurtaza Mustafa Merchant: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Naveen Shenoy:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis,

Data curation. **Sidharth Lanka:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Sowmya Kamath:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] WHO, International Classification of Diseases: [9th] Ninth Revision, World Health Organization, 1978.

[2] E.H. Shortliffe, The evolution of electronic medical records, Acad. Med. 74 (4) (1999) 414–419.

[3] T. Vu, D.Q. Nguyen, A. Nguyen, A label attention model for ICD coding from clinical text, arXiv preprint, arXiv:2007.06351, 2020.

[4] F. Li, H. Yu, ICD coding from clinical text using multi-filter residual convolutional neural network, Proc. AAAI Conf. Artif. Intell. 34 (05) (2020) 8180–8187.

[5] P. Xie, E. Xing, A neural architecture for automated ICD coding, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1066–1076.

[6] V. Mayya, S.S. Kamath, V. Sugumaran, $\mathcal{LATA}$-label attention transformer architectures for ICD-10 coding of unstructured clinical notes, in: 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2021.

[7] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, 2018.

[8] Y. Liu, H. Cheng, R. Klopfer, M.R. Gormley, T. Schaaf, Effective convolutional attention network for multi-label clinical document classification, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 5941–5953.

[9] D. Pascual, S. Luck, R. Wattenhofer, Towards bert-based automatic icd coding: limitations and opportunities, in: BIONLP, 2021.

[10] Z. Zhang, J. Liu, N. Razavian, BERT-XML: large scale automated ICD coding using BERT pretraining, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, 2020, pp. 24–34 (Online).

[11] Z. Yuan, C. Tan, S. Huang, Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding, 2022.

[12] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (Database issue) (2004) D267–D270.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017.

[14] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, K. Maheshwari, et al., Multimodal machine learning for automated icd coding, in: Machine Learning for Healthcare Conference, PMLR, 2019.

[15] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmarking deep learning models on large healthcare datasets, J. Biomed. Inform. 83 (2018) 112–134.

[16] P. Akshara, S. Shidharth, G. Krishnan, S. Sowmya Kamath, Integrating structured and unstructured patient data for icd9 disease code group prediction, in: Proceedings of the 8th ACM IKDD CODS & 26th COMAD, Association for Computing Machinery, New York, NY, USA, 2021, p. 436.

[17] Z. Yang, S. Wang, B.P.S. Rawat, A. Mitra, H. Yu, Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding, arXiv preprint, arXiv:2210.03304, 2022.

[18] T. Gangavarapu, A. Jayasimha, G.S. Krishnan, S. Kamath, Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes, Knowl.-Based Syst. 190 (2020) 105321.

[19] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, Mimic-iii, a freely accessible critical care database, Sci. Data 3 (1) (2016) 160035.

[20] X. Xie, Y. Xiong, P.S. Yu, Y. Zhu, Ehr coding with multi-scale feature attention and structured knowledge graph propagation, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 649–658.

[21] X. Rong, word2Vec parameter learning explained, CoRR, 2014.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, et al., Pytorch: an imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÀlché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019.

[23] R. Rehurek, P. Sojka, Gensim–Python Framework for Vector Space Modelling, vol. 3(2), NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2011.

[24] X. Xie, Y. Xiong, P.S. Yu, Y. Zhu, Ehr coding with multi-scale feature attention and structured knowledge graph propagation, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 649–658.