TreeMAN: Tree-enhanced Multimodal Attention Network for ICD Coding

Zichen Liu[†], **Xuyuan Liu**[†], **Yanlong Wen**[†*], **Guoqing Zhao**[‡], **Fen Xia**[‡], **Xiaojie Yuan**[†]
[†]College of Computer Science, Nankai University
[‡]Mashang Consumer Finance Co.,Ltd.
{liuzichen,wenyl,yuanxj}@dbis.nankai.edu.cn
hsuyuanliu0204@mail.nankai.edu.cn
{guoqing.zhao02,fen.xia}@msxf.com

Abstract

ICD coding is designed to assign the disease codes to electronic health records (EHRs) upon discharge, which is crucial for billing and clinical statistics. In an attempt to improve the effectiveness and efficiency of manual coding, many methods have been proposed to automatically predict ICD codes from clinical notes. However, most previous works ignore the decisive information contained in structured medical data in EHRs, which is hard to be captured from the noisy clinical notes. In this paper, we propose a **Tree**-enhanced **M**ultimodal **A**ttention **N**etwork (TreeMAN) to fuse tabular features and textual features into multimodal representations by enhancing the text representations with *tree-based features* via the attention mechanism. *Tree-based features* are constructed according to decision trees learned from structured multimodal medical data, which capture the decisive information about ICD coding. We can apply the same multi-label classifier from previous text models to the multimodal representations to predict ICD codes. Experiments on two MIMIC datasets show that our method outperforms prior state-of-the-art ICD coding approaches. The code is available at https://github.com/liu-zichen/TreeMAN.

1 Introduction

The International Classification of Diseases (ICD), maintained by the World Health Organization, is a hierarchical classification of codes representing diseases, injuries, and so on. ICD codes have been used in diverse areas, including insurance reimbursement, epidemiology, and clinical research (Park et al., 2000).

In the hospital, when patients discharge, their electronic health records (EHRs) and all associated data are transferred to the information management department, where clinical coders manually assign the appropriate ICD codes using rigid ICD coding guidelines after reviewing records (O'malley et al., 2005). The manual code assignment is expensive, labor-intensive, and error-prone due to the large volume of medical record information and high professional requirements (Nguyen et al., 2018).

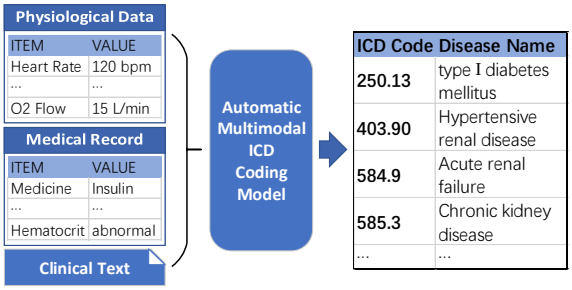


Figure 1: An example of automatic multimodal ICD coding. Model inputs include physiological data and medical records in addition to clinical text.

Since deep learning has achieved great success in lots of healthcare applications (Cai et al., 2019), many neural methods have been proposed to automate the ICD coding process by researchers (Teng et al., 2022). Recent works formulate automated ICD coding as a multi-label document classification task, using clinical notes as model input, predicting coding with a multi-label classifier, and learning text features through word embedding techniques and neural networks such as RNNs and CNNs (Mullenbach et al., 2018; Vu et al., 2020; Zhou et al., 2021). To improve the code representation learning, researchers further leverage features of ICD codes such as hierarchical structures (Cao et al., 2020) and descriptions (Mullenbach et al., 2018; Zhou et al., 2021). However, most previous methods ignore structured medical data, including physiological data collected by medical sensors and medical record information such as prescriptions and microbiology test results in EHRs. The few methods that leverage structured data are either ensemble-based approaches (Xu et al., 2019)

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 3054–3063
October 12–17, 2022.

or data-mining methods that discard semantic information (Ferrão et al., 2021).

In this work, we argue that structured medical data can improve coding accuracy by enhancing semantic representations and providing more information because clinical notes are noisy and ambiguous. For example, there are many different types of insulin, like "Insulin Aspart" and "Insulin Glargine", which are often written the same in notes but clearly distinguished by Generic Sequence Numbers (GSNs) in medical records. Considering different writing styles and polysemous abbreviations, predicting ICD codes from clinical notes is more complicated. However, automatic multimodal ICD coding (as Figure 1 shown) is challenging for the following reasons: 1) medical data is naturally heterogeneous, with data types including numerical quantities, categorical values, and derived time series such as perioperative vital sign signals (Zhou et al., 2021); 2) the feature selection method needs to be designed especially for multi-ICD codes as it's a multi-label classification task; 3) decisive information for a code in the long clinical note may be contained in short segments that are likely different for different codes (Mullenbach et al., 2018).

In this paper, we propose a novel **Tree**-enhanced **M**ultimodal **A**ttention **N**etwork named TreeMAN to address the aforementioned problems. Since it's hard to do feature engineering for structured medical data, we construct *tree-based features* from the structured medical data through decision trees that require little data preparation (Safavian and Landgrebe, 1991) instead of manually crafting features based on medical knowledge. Inspired by previous works (Wang et al., 2018; He et al., 2014), we represent the tree-based features by embedding vectors. Taking the tree-based embeddings and text representations as input, TreeMAN applies an attention mechanism to select relevant tree-based features for text representations and output fused multimodal representations that contain richer information to benefit the downstream classifier. However, our method has limitations in handling long-tailed labels as it is difficult to build a decision tree from less than 10 positive samples.

Contributions. In summary, the main contributions of our work include:

- We propose a multimodal ICD coding framework that exploits structured medical data in EHRs to construct tree-based features to enhance text representations.
- We propose TreeMAN, a tree-enhanced multimodal attention network, which fuses text representations and tree-based features into unified multimodal representations by the attention mechanism. To the best of our knowledge, it's the first model to jointly learn multimodal features for the ICD coding task.
- Experiments demonstrate the effectiveness of our proposed method. Results on two datasets show that TreeMAN outperforms previous state-of-the-art ICD coding methods.

2 Related Work

2.1 ICD Coding

Research on Automatic ICD coding can be traced back to nearly 30 years ago when Larkey and Croft (1996) proposed an ensemble algorithm to integrate different types of classifiers to assign ICD codes to inpatient discharge summaries. A series of methods based on Deep Neural Networks has been implemented on this task since this paradigm achieved colossal success in Clinical NLP. Perotte et al. (2014) built "hierarchical" Support Vector Machines (SVMs) outperforming the "flat" classifier. Mullenbach et al. (2018) built a convolutional attention model which combined the single filter CNN module and the per-label attention module. A series of network modules based on attention mechanism have been utilized after the early attempts, including multi-scale attention module (Xie et al., 2019), residual convolution module (Li and Yu, 2020). We also notice that the hierarchical structure of ICD-9 could be effectively described by a joint-classification module on different levels (Vu et al., 2020) or in the form of specific hyperbolic representation (Cao et al., 2020).

Multimodal learning methods help to integrate multiple information like test reports, nursing notes, etc., in the MIMIC-III datasets. An early attempt was made by (Xu et al., 2019), in which an ensemble-based approach was developed to integrate the structured and unstructured text of different modalities. Rajendran et al. (2021) made full use of unstructured information by effectively exploiting the geometric properties of pre-trained word embeddings.

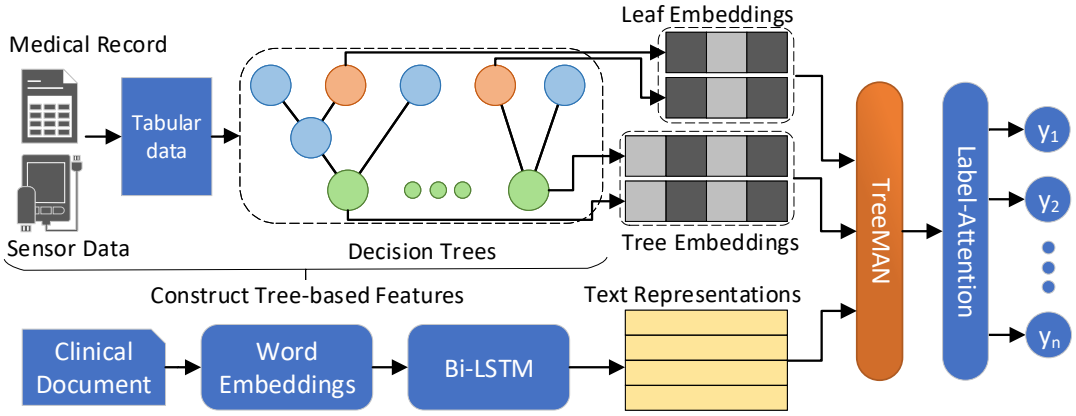


Figure 2: An overview of our proposed multimodal ICD coding framework.

2.2 Tree-based Method

Decision trees are a supervised learning algorithm broadly applied in regression and classification tasks (Quinlan, 1986). They are trained on labeled data while requiring little data preparation and domain knowledge while the preprocessed features are able to be fused with text representations easily. Multiple skills have been implemented to ensemble relatively simple decision trees to get better performance (Banfield et al., 2007; Gashler et al., 2008). Among all of these ideas, Gradient boosting decision tree (GBDT) is an important instance which introduces iterative functional gradient descent algorithms to boosting models firstly (Friedman, 2001). Significant improvement made by XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) which use different gradient information to improve accuracy and training efficiency respectively. Many attempts (Trofimov et al., 2012; Ling et al., 2017) have been made based on decision-tree boosting algorithm since people found it could generate interpretable and effective cross-feature and is easy to fix with other models. A combination of GBDT with linear model like Logistic Regression(LR) effectively helps the models to make explainable predictions by selecting top cross features (He et al., 2014). Wang et al. (2018) argued that the tree-enhanced embedding method would benefit from the explainability of tree-based models, thus improving generalization ability compared with other pure embedding ways. Incorporating decision tree learning with matrix factorization would help to extract the latent factors and get Fine-Grained embedding with rich semantic information (Kim et al., 2020), which could contribute to solve cold-start problems even further (Tao et al., 2019; Zhou et al., 2011).

3 Method

In this section, we first give an overview of our framework (Section 3.1), and then detail the key module in our framework: the tree-enhanced multimodal attention network TreeMAN (Section 3.2). Finally, we introduce the processing of structured medical data and decision tree learning (Section 3.3).

3.1 Overview

Figure 2 shows the overview of our method. Upon discharge, there are two types of data available for our model: clinical notes written by doctors and structured medical data, including physiological data collected by sensors and medical records, such as lab measurements and prescriptions. Given a clinical note and the associated structured data, two modules in the model process them separately to obtain text representations and tree-based features. Considering in the poor performance of Bert-like models on ICD coding (Zhang et al., 2020; Chalkidis et al., 2020), we train the text model from scratch instead of fine-tuning a pretrained language model. Structured medical data is first processed as tabular data and then fed into a trained decision tree to obtain tree-based features that we project into embedding vectors: the tree embeddings **T** and the leaf embeddings **L** (detailed in Section 3.3). The other module is the text encoder designed to capture the semantic information in the document and provide textual representations.

Text encoder Given an input document with N words $\{w_i\}_{i=1}^{N}$, the encoder first maps each word

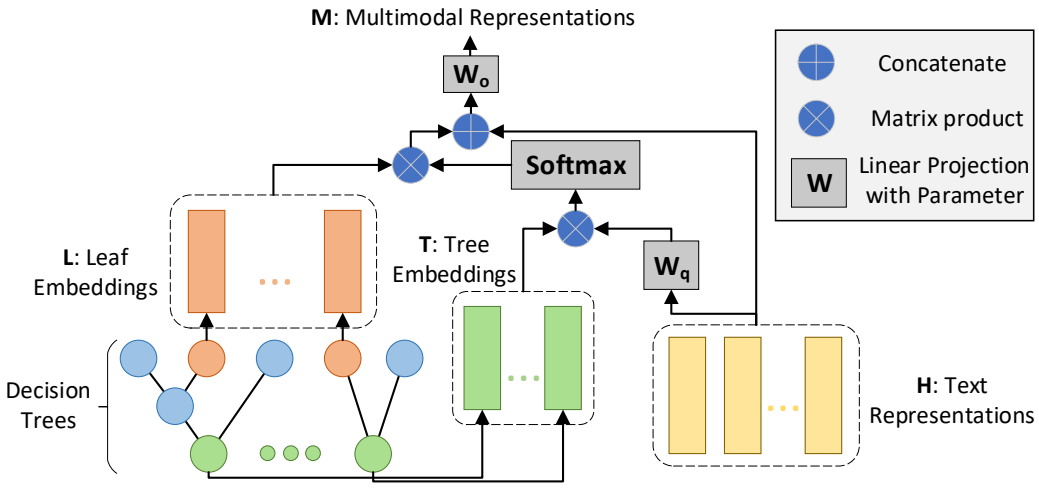


Figure 3: An illustration of our Tree-enhanced Multimodal Attention Network (TreeMAN). Green nodes and orange nodes on the decision trees respectively represent root nodes and activated leaf nodes.

w_i to a d_e-dimensional pre-trained word embedding $\mathbf{e}_i$, then concatenates embeddings into the matrix $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N]$. To capture contextual information, the word embedding matrix E is fed into a bidirectional LSTM layer to compute the text representations $\mathbf{H}$, which is a concatenation of the forward output and the backward output:

$$\begin{aligned} \overrightarrow{\mathbf{h}_i} &= \overrightarrow{LSTM}(\mathbf{e}_{1:i}), \quad \overleftarrow{\mathbf{h}_i} = \overleftarrow{LSTM}(\mathbf{e}_{i:N}), \\ \mathbf{H} &= [\overrightarrow{\mathbf{h}_1} \oplus \overleftarrow{\mathbf{h}_1}, \overrightarrow{\mathbf{h}_2} \oplus \overleftarrow{\mathbf{h}_2}, ..., \overrightarrow{\mathbf{h}_n} \oplus \overleftarrow{\mathbf{h}_n}]. \end{aligned} \tag{1}$$

Then, text representations $\mathbf{H}$ together with the tree embeddings $\mathbf{T}$ and the leaf embeddings $\mathbf{L}$ generated from tree-based features are fed to TreeMAN to obtain the multimodal representation $\mathbf{M}$ (detailed in Section 3.2):

$$\mathbf{M} = \text{TreeMAN}(\mathbf{H}, \mathbf{L}, \mathbf{T}). \tag{2}$$

In the output layer, following Mullenbach et al. (2018), we apply the per-label attention network to compute representations for each label.

Label attention The label attention network takes multimodal representations $\mathbf{M} \in \mathbb{R}^{d_m \times N}$ as input and compute the per-label representations $\mathbf{V} \in \mathbb{R}^{d_m \times |\mathcal{L}|}$ with a matrix parameter $U \in \mathbb{R}^{d_m \times |\mathcal{L}|}$, where $|\mathcal{L}|$ represents the number of labels:

$$\begin{aligned} \mathbf{A} &= \text{softmax}(\mathbf{MU}), \\ \mathbf{V} &= \mathbf{A}^T\mathbf{M}. \end{aligned} \tag{3}$$

Finally, to compute the probability $\hat{y}_i$ of the i^{th} label, the label representation $\mathbf{v}_i$ of $\mathbf{V}$ is fed into a corresponding linear layer followed by a sigmoid transformation. For training, the model is optimized to minimize the binary cross-entropy loss between the prediction $\hat{y}$ and the target y:

$$Loss = \sum_{i=1}^{|\mathcal{L}|} -y_i log(\hat{y}_i) - (1-y_i) log(1-\hat{y}_i). \tag{4}$$

3.2 TreeMAN

TreeMAN, a tree-enhanced multimodal attention network, is designed to fuse tree-based features and text representations and provides enhanced multimodal representations for multi-label classification. We argue that the critical information in text representations is respective and fragmented because decisive information for different labels in the document is likely contained in different short segments (Mullenbach et al., 2018). Therefore, we use the attention mechanism to learn the relevant features for each text vector and then fuse tree-based features and text information into a unified multimodal representation.

An illustration of TreeMAN is shown in Figure 3. Specifically, for each text vector $\mathbf{h}_i \in \mathbb{R}^{d_h}$ in $\mathbf{H}$, we first project it to a query vector $\mathbf{q}_i$ by a learnable parameter $\mathbf{W}_q \in \mathbb{R}^{d_t \times d_h}$:

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{h}_i. \tag{5}$$

The vector $\mathbf{q}_i \in \mathbb{R}^{d_t}$ is used to generate the attention weight $\boldsymbol{\alpha}_i$ by computing with the tree embeddings $\mathbf{T} \in \mathbb{R}^{d_t \times |\mathcal{T}|}$, where $|\mathcal{T}|$ represents the number of decision trees:

$$\boldsymbol{\alpha}_i = \text{softmax}(\mathbf{T}^T \mathbf{q}_i). \tag{6}$$

The attention vector $\boldsymbol{\alpha}_i$ is then multiplied with the leaf embeddings $\mathbf{L} \in \mathbb{R}^{d_l \times |\mathcal{T}|}$ to produce the special representation $\mathbf{s}_i$ for the text vector:

$$\mathbf{s}_i = \mathbf{L}\boldsymbol{\alpha}_i. \tag{7}$$

To fuse the text information and tree-based features, we concatenate the special representation with text vector and then apply a linear projection:

$$\mathbf{m}_i = \mathbf{W}_o[\mathbf{h}_i||\mathbf{s}_i], \tag{8}$$

where $\mathbf{W}_o \in \mathcal{R}^{d_m \times (d_l + d_s)}$ is a learnable parameter. All the multimodal vectors are concatenated to formulate the output matrix $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_N] \in \mathbb{R}^{d_m \times N}$.

3.3 Construction of Tree-based Features

In this section, we introduce how we construct the tree-based features from structured medical data. Based on the characteristics of the data, we divide the structured medical data into three types: 1) *derived time series data* such as perioperative vital sign signals; 2) *multivalued vertical data* denotes data with multiple records for one admission, such as lab measurements and prescriptions; 3) *single horizontal data* indicates data with only a single record for an admission, such as admission type and patient age. We process different types of data into tabular data in different ways: 1) for *derived time series data*, we compute mean, maximum, and minimum values for each class of data; 2) for *multivalued vertical data*, we convert it into binary vector to indicate whether a test is abnormal or a medication is prescribed; 3) for *single horizontal data*, we directly put it into the table as it is.

Then, we use the processed tabular data to construct decision trees by applying decision trees, which are trained with ICD codes as the target, using one-versus-all strategy for multi-label classification. Formally, we get a set of decision trees, $\mathbb{Q} = \{Q_1, ..., Q_{|\mathcal{T}|}\}$, where each tree maps the tabular data $\mathbf{x}$ to a leaf node, which can be represented by a one-hot vector. The representation of tree-based features is a multi-hot vector $\mathbf{q}$ which is a concatenation of one-hot vectors:

$$\mathbf{q} = [Q_1(x), ..., Q_{|\mathcal{T}|}(x)]. \tag{9}$$

Therefore, there are $|\mathcal{T}|$ elements of value 1 in $\mathbf{q}$ indicates activated leaf nodes.

Inspired by the success of TEM (Wang et al., 2018), we project $\mathbf{q}$ into an embedding matrix $\mathbf{L}$ as leaf embeddings. For the attention computation in Section 3.2, we also generate a tree embedding matrix $\mathbf{T}$ based on the number of decision trees.

	MIMIC-III 50	**MIMIC-II 50**
Vocubulary Size	51,917	30,688
# Samples	11,371	3,726
# *Drugs	2350	52
# *Lab Items	245	217
# *Organism	183	135
# *Specimen	74	63
# *Antibiotic	30	30
# *Chart Items	200	-
Mean # labels per document	5.7	3.4
Mean # tokens per document	1530	1014

Table 1: The statistics of the two MIMIC datasets and the structured medical data used therein, where "#" indicates "the number of" and "*" denotes the number of classes is counted.

4 Experiments

4.1 Datasets

To make a fair and all-round comparison with former SOTA models, we evaluate our model on two widely used **M**edical **I**nformation **M**art for **I**ntensive **C**are (MIMIC) datasets: MIMIC-III (Johnson et al., 2016) and MIMIC-II (Saeed et al., 2002). Because it's hard for our method to be implemented on ICD codes with less than 10 positive samples, we filter out records not relative to the top 50 most frequent ICD codes (denoted as MIMIC-III 50, MIMIC-II 50) to train and evaluate our method.

MIMIC-III 50. Except for structured medical data, we use the same experimental setup including the same splits as previous works (Mullenbach et al., 2018; Cao et al., 2020; Vu et al., 2020). For structured medical data, we use the following tables in MIMIC-III dataset [1] : 1) *Admissions* contains patients' admission information such as admission time; 2) *Patients* contains patients' basic information such as date of birth; 3) *Chartevents* contains charted data including patients' routine vital signs; 4) *Labevents* contains laboratory measurements such as pH of blood; 5) *Microbiologyevents* contains microbiology information such as organism test information; 6) *Prescriptions* contains medications related to order entries including the Generic Sequence Number (GSN) of drugs.

[1] A detailed introduction to MIMIC-III tables can be found at https://mimic.mit.edu/docs/iii/tables.

MIMIC-II 50. We subset the MIMIC-II full used in previous works (Mullenbach et al., 2018) based on the 50 most frequent labels and use a set of 3,726 admission samples in which there are 2980 samples for training, 373 for validation and 373 for test. For structured medical data, we use the following tables in MIMIC-II dataset [2]: 1) *Admissions*; 2) *Patients*; 3) *Medevents* is similar to *Prescriptions* in MIMIC-III; 4) *Labevents*; 5) *Microbiologyevents*.

Basic statistic information of all the datasets shows on Table 1.

4.2 Implementation Details

Following the preprocessing schema of previous works (Mullenbach et al., 2018; Li and Yu, 2020; Xie et al., 2019), we lowercase all tokens and remove tokens that contain unrelated alphabetic characters like numbers and punctuations. We implement the word2vec CBOW (Mikolov et al., 2013) method to pre-train word embeddings and truncate all discharge summary documents to the maximum length of 4,000 tokens. We employ XGBoost [3] to implement the decision trees in our approach. There is only one decision tree built for each label where the learning rate and the maximum depth of the tree are set as 0.99 and 5, respectively, while the rest of settings follow the default. The sizes of the tree embedding **T** and the leaf embedding **L** are 128 and 30, respectively. We set the size of multimodal representations to be the same as that of text representations.

For the baseline methods we reproduced on the MIMIC-II 50 dataset, we used the same implementations used by the authors on MIMIC-III 50. To reduce randomness, we repeated all experiments 5 times with different random seeds and report the average performance.

4.3 Metrics

To compare with previous and potential future work thoroughly, we measured our model mainly on indicators of macro-averaged and micro-averaged F1, macro-averaged, and micro-averaged AUC (area under the ROC curve) and Precision@k (P@k). Among these metrics, the "micro-averaged" method takes every single decision into consideration by pooling all text-code pairing and then calculating an effectiveness indicator on the pooled data. And "macro-averaged based" metrics would provide statistics from the perspective of label instead of pair-relationship. Furthermore, we rank predictive probabilities to compute the precision of the top-k predicted labels, denoted as P@k. We set k to be five on MIMIC-III 50 dataset and three on MIMIC-II 50 dataset for the average discharge summary has 5.7 labels in MIMIC-III 50 while 3.4 in MIMIC-II 50. We believe a full comparison of all the above metrics will provide insight into our work.

4.4 Baselines

We compare our model TreeMAN with the following baseline; all of them were SOTA when they were proposed initially.

CAML **C**onvolutional **A**ttention network for **M**ulti- **L**abel classification (CAML) and **d**escription **R**egularized CAML was proposed by Mullenbach et al. (2018), which combined a single-layer CNN with attention layer to generate ICD coding for given text.

LAAT&Joint-LAAT **LA**bel **AT**tention and Joint **LA**bel **AT**tention model was proposed by Vu et al. (2020). It encodes the input text with BiLSTM layer and implements self-attention mechanism to learn label-specific vectors representation. A hierarchical joint learning architecture is utilized to improve performance in the second model.

HyperCore **Hyper**bolic and **Co**-graph Representation was proposed by Cao et al. (2020). It leveraged hierarchical structure of ICD code in hyperbolic space and used graph convolutional network(GCN) to capture co-occurrence correlation of labels.

ISD **I**nteractive **S**hared Representation Network with Self-**D**istillation Mechanism was proposed by Zhou et al. (2021), they implemented a self-distillation learning mechanism to alleviate the noisy text and only focus on noteworthy part of text.

4.5 Results

Table 2 reports $mean \pm standard\ deviation$ of TreeMAN's results on two datasets, the performance of baselines on MIMIC-III 50 and the results of our implementation of baselines on MIMIC-II 50. Compared with previous text methods, our multimodal approach achieves the best results on all metrics on both datasets. It indicates that our

[2] A detailed introduction to MIMIC-II tables can be found at https://archive.physionet.org/mimic2/UserGuide/UserGuide.pdf .

[3] https://xgboost.readthedocs.io.

Model	MIMIC-III 50					MIMIC-II 50				
	AUC		F1		P@5	AUC		F1		P@3
	macro	micro	macro	micro		macro	micro	macro	micro	
CAML	0.875	0.909	0.532	0.614	0.609	0.871	0.902	0.426	0.553	0.552
HyperCore	0.895	0.929	0.609	0.663	0.632	-	-	-	-	-
LAAT	0.925	0.946	0.666	0.715	0.675	0.874	0.908	0.436	0.557	0.556
Joint LAAT	0.925	0.946	0.661	0.716	0.671	0.875	0.908	0.434	0.547	0.560
ISD	0.935	0.949	0.679	0.717	**0.682**	-	-	-	-	-
TreeMAN	**0.937** ±0.002	**0.953** ±0.000	**0.690** ±0.002	**0.729** ±0.002	**0.682** ±0.001	**0.883** ±0.002	**0.916** ±0.002	**0.479** ±0.001	**0.574** ±0.001	**0.605** ±0.004

Table 2: Results on MIMIC-III 50 dataset, MIMIC-II 50 dataset and $mean \pm standard\ deviation$ of each indicator gained from replicated experiments with random initial states. Baseline scores are from the corresponding papers in Section 4.4.

Model	AUC		F1		P@5
	macro	micro	macro	micro	
text	92.6	94.5	67.4	71.4	66.6
maxpooling	93.1	94.9	68.4	72.3	67.5
average	93.4	95.1	68.9	72.7	67.6
TreeMAN	93.7	95.3	69.0	72.9	68.2

Table 3: Results of ablation experiments on the MIMIC-III 50 dataset (in %).

model benefits from the rich information contained in structured medical data. Furthermore, the small standard deviations demonstrate that the good results our model achieved are stable. We also observe more significant improvements in the f1-marco and f1-micro metrics compared to other ranking-based metrics. Since the binary output is produced by a fixed threshold 0.5, a possible reason for the disparity is that the sigmoid function of our model in the final layer outputs more dispersed probabilities due to the decisive information provided by structured medical data.

4.6 Ablation Experiment

To testify the effectiveness of the different modules in TreeMAN, we perform a series of ablation experiments on the MIMIC-III 50 dataset, design following experiments, and report the results in Table 3.

The Effect of Structured Medical Data To study the effectiveness of the information captured from structured medical data, we remove the tree-based features in TreeMAN and directly feed the unfused text representations to the multi-label classifier (*text* in Table 3). The experimental results of all metrics decreased significantly compared to TreeMAN, demonstrating the importance of the tree-based features constructed based on structured medical data. It's also a comparison between the text representations and the multimodal representations, which proves that TreeMAN is capable of learning multimodal features.

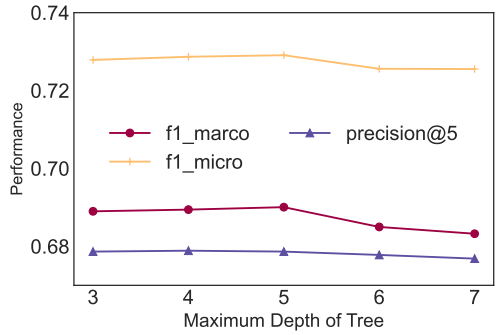
The Effect of Attention Mechanism To examine the effectiveness of the attention mechanism in TreeMAN, we design two experiments by replacing the attention network with the max-pooling layer (*maxpooling* in Table 3) and the average layer (*average* in Table 3) on leaf embeddings. Formally, we change the Equation 7 as:

$$\begin{cases} \text{maxpooling: } \mathbf{s}_i = max\_pool_{\mathbf{l}\in\mathbf{L}}(\mathbf{l}), \\ \text{average: } \mathbf{s}_i = \frac{1}{|\mathcal{T}|}\sum_{\mathbf{l}\in\mathbf{L}}(\mathbf{l}), \end{cases} \tag{10}$$

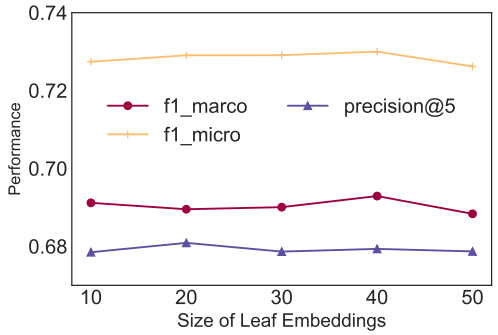
where $\mathbf{l}$ and $\mathcal{T}$ represent a vector in the leaf embeddings $\mathbf{L}$ and the number of decision trees, respectively. As shown, the experimental results of *maxpooling* and *average* are both better than *text* and worse than TreeMAN. Thus, the attention mechanism improves TreeMAN's ability to learn multimodal information and the information captured by tree-based features is robust to learn.

4.7 Parameter Studies

We have already analyzed the efficacy of our proposed model, and now we want to conduct a series of experiments to test the effect of two critical hyper-parameters in the TreeMAN module: the maximum depth of the decision tree and the size of leaf embeddings $\mathbf{L}$. The former decides how tree-based features are constructed from structured medical data, and the latter is the representation format of the tree-based features. Various metrics of different settings would help us to demonstrate how



(a) Performance of TreeMAN with different maximum Depth of Decision Tree



(b) Performance of TreeMAN with different Leaf Embedding Dimensions

Figure 4: Results of different maximum tree depth and leaf embedding size on MIMIC-III 50.

TreeMAN extracts information from multimodal data:

The Effect of Maximum Depth of the Decision Trees The maximum depth of the decision tree would decide the number of feature extracted and the properties of the leaf embedding layer, for each of the leaf nodes represents a tree-based feature. For example, if we set the maximum depth to 3, we would get 401 leaves and 4752 leaves for a 7-layer tree. A shallow decision tree cannot extract enough features to represent the latent information of initial input. However, a too deep tree would risk over-fitting as well as colossal costs in the training process. Based on this assumption, we make a complete comparison of different pre-set depths of the decision tree. As the Figure 4 (a) shows, a tree of depth 5 outperforms other decision trees, especially on the indicator of f1_marco and f1_micro because of the improvement in the aspect of recall ratio. Furthermore, we also notice that changes in this hyper-parameter don't seriously affect the performance of our module, proving the robustness of our method.

The Effect of Leaf Embeddings As we project multimodal information gained in the decision tree to leaf embeddings $\mathbf{L}$, we need the proper capacity of this layer to collect and store them. Thus we experiment with the leaf embedding size ranging from 10 to 50 to study the effect of the setup. Figure 4(b) shows that a vector with 30 dimensions is a proper choice because short vectors would abandon helpful information, while long ones would carry redundant information. Taking note of the limited size of datasets, relatively simple architecture could be a practical solution. These results also indicate that TreeMAN has learned an operative and steady pattern to learn from various types of multimodal information.

5 Conclusion

In this paper, we proposed a tree-based multimodal method for the ICD coding task, which constructs tree-based features by decision trees learned from structured medical data and fuses the tree-based features and text representation by a novel tree-enhanced multimodal attention network (TreeMAN). Experimental results on two MIMIC datasets show that our method outperforms state-of-the-art methods. Further ablation studies demonstrate that structured medical data and the attention mechanism in TreeMAN have improved the performance.

For future work, we plan to investigate the interpretability of our method since tree-based methods are naturally interpretable. We are also interested in exploring a generalized and robust way to construct the tree-based features to capture more generalized medical information from structured medical data.

6 Acknowledgement

This research is supported by Chinese Scientific and Technical Innovation Project 2030 (No. 2018AAA0102100), National Natural Science Foundation of China (No. U1936206, 62077031, 62272250). We appreciate the reviewers for their constructive comments. We thank Shuyun Deng for her help in writing the paper.

References

Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer, and W. Philip Kegelmeyer. 2007. A comparison of decision tree ensemble creation techniques. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):173–180.

Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. 2019. A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access*, 7:133583–133599.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Hypercore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3105–3114. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7503–7515. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

José Carlos Ferrão, Mónica Duarte Oliveira, Filipe Janela, Henrique MG Martins, and Daniel Gartner. 2021. Can structured ehr data support clinical coding? a data mining approach. *Health Systems*, 10(2):138–161.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232.

Mike Gashler, Christophe Giraud-Carrier, and Tony Martinez. 2008. Decision tree ensemble: Small heterogeneous is better than large homogeneous. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 900–905.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. 2014. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, ADKDD 2014, August 24, 2014, New York City, New York, USA*, pages 5:1–5:9. ACM.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sundong Kim, Yu-Che Tsai, Karandeep Singh, Yeonsoo Choi, Etim Ibok, Cheng-Te Li, and Meeyoung Cha. 2020. Date: Dual attentive tree-aware embedding for customs fraud detection. KDD '20, page 2880–2890, New York, NY, USA. Association for Computing Machinery.

Leah S. Larkey and W. Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 289–297. ACM.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8180–8187. AAAI Press.

Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun. 2017. Model ensemble for click prediction in bing search ads. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 689–698, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1101–1111. Association for Computational Linguistics.

Anthony N. Nguyen, Donna L. Truran, Madonna Kemp, Bevan Koopman, David Conlan, John O'Dwyer, Ming Zhang, Sarvnaz Karimi, Hamed Hassanzadeh, Michael Lawley, and Damian J. Green. 2018. Computer-assisted diagnostic coding: Effectiveness of an nlp-based approach using SNOMED CT to

ICD-10 mappings. In *AMIA 2018, American Medical Informatics Association Annual Symposium, San Francisco, CA, November 3-7, 2018*. AMIA.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Jong-Ku Park, Ki-Soon Kim, Tae-Yong Lee, Kang-Sook Lee, Duk-Hee Lee, Sun-Hee Lee, Sun-Ha Jee, Il Suh, Kwang-Wook Koh, So-Yeon Ryu, et al. 2000. The accuracy of icd codes for cerebrovascular diseases in medical insurance claims. *Journal of Preventive Medicine and Public Health*, 33(1):76–82.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

J. R. Quinlan. 1986. Induction of decision trees. *Mach. Learn.*, 1(1):81–106.

Pavithra Rajendran, Alexandros Zenonos, Joshua Spear, and Rebecca Pope. 2021. Embed wisely: An ensemble approach to predict ICD coding. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part II*, volume 1525 of *Communications in Computer and Information Science*, pages 371–389. Springer.

M. Saeed, C. Lieu, G. Raber, and R.G. Mark. 2002. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in Cardiology*, pages 641–644.

S. Rasoul Safavian and David A. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.*, 21(3):660–674.

Yiyi Tao, Yiling Jia, Nan Wang, and Hongning Wang. 2019. The fact: Taming latent factor models for explainability with factorization trees. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 295–304, New York, NY, USA. Association for Computing Machinery.

Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. 2022. A review on deep neural networks for icd coding. *IEEE Transactions on Knowledge and Data Engineering*.

Ilya Trofimov, Anna Kornetova, and Valery Topinskiy. 2012. Using boosted trees for click-through rate prediction for sponsored search. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, ADKDD '12, New York, NY, USA. Association for Computing Machinery.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3335–3341. ijcai.org.

Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. TEM: tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1543–1552. ACM.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. 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 2019, Beijing, China, November 3-7, 2019*, pages 649–658. ACM.

Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K. Khanna, Jacek B. Cywinski, Kamal Maheshwari, Pengtao Xie, and Eric P. Xing. 2019. Multimodal machine learning for automated ICD coding. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2019, 9-10 August 2019, Ann Arbor, Michigan, USA*, volume 106 of *Proceedings of Machine Learning Research*, pages 197–215. PMLR.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: large scale automated ICD coding using BERT pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 24–34. Association for Computational Linguistics.

Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. 2011. Functional matrix factorizations for cold-start recommendation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 315–324, New York, NY, USA. Association for Computing Machinery.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic ICD coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5948–5957. Association for Computational Linguistics.