Interpretable Disease Prediction from Clinical Text by Leveraging Pattern Disentanglement

Malikeh Ehghaghi
Computer Science Department
University of Toronto
Toronto, ON, Canada
malikeh.ehghaghi@mail.utoronto.ca

Pei-Yuan Zhou
System Design Engineering
University of Waterloo
Waterloo, ON, Canada
p44zhou@uwaterloo.ca

Wendy Yusi Cheng
Computer Science Department
University of Toronto
Toronto, ON, Canada
wendy.cheng@mail.utoronto.ca

Sahar Rajabi
Electrical and Computer Engineering Department
University of Tehran
Tehran, Iran
sahar.rajabi@ut.ac.ir

Chih-Hao Kuo
Computer Science Department
University of Toronto
Toronto, ON, Canada
h.kuo@mail.utoronto.ca

En-Shiun Annie Lee
Computer Science Department
Ontario Tech University
University of Toronto
Toronto, ON, Canada
enshiun.lee@utoronto.ca

***Abstract*—For artificial intelligence (AI) systems to be adopted in high stake human-oriented applications, they must be able to make complex decisions in an understandable and interpretable manner. While AI systems today have grown leaps and bounds in predictive power using larger datasets with more complex architectures, existing models remain ineffective at generating interpretable insights in the clinical setting. In this paper, we address the challenge of discovering interpretable insights from the clinical text for disease prediction. For this purpose, we apply the clinical notes from the electronic health records (EHR) available in the Medical Information Mart of Intensive Care III (MIMIC-III) dataset, which are labeled with the international classification of diseases (ICD9) codes. Our proposed algorithm combines interpretable text-based features with a novel pattern discovery and disentanglement algorithm. Specifically, our approach encompasses the following: (1) uncovering strong association patterns between clinical notes and diseases, (2) surpassing baseline clustering algorithms in effectively distinguishing between disease clusters, and (3) demonstrating comparable performance to baseline supervised methods in predicting diseases. Our results validate the model's capability to strike a balance between interpretability and outcome prediction accuracy. By unveiling insightful patterns between clinical notes and diseases, our approach upholds a reasonable level of diagnostic accuracy.**

***Clinical relevance*—This paper proposes a novel all-in-one clinical Natural Language Processing (NLP) knowledge base, which can be applied in healthcare systems to discover interpretable insights from the clinical text for predicting medical conditions.**

***Index Terms*—Interpretability, Electronic Health Records, Pattern Discovery, Pattern Disentanglement, Clinical Notes**

I. Introduction

The complex and opaque nature of Artificial Intelligent (AI) systems is often a hurdle to their widespread adoption and acceptance in high stake human-oriented applications such as health care. Therefore greater transparency for explaining predictions and decisions is in demand to meet critical scientific, medical, legal, and social needs [1]. Interpretability is frequently defined as the degree to which a human can understand the cause and reason of decisions from domain knowledge [2]. However, even though some AI models can also provide various degrees of interpretability, they generally sacrifice interpretability for predictive power [3].

Therefore, in this paper, we focus on the task of predicting diseases from clinical text found in electronic health records (EHR) in an interpretable manner. Although deep learning black-box models achieve state-of-the-art results [4], their decision-making process remains challenging to interpret without posthoc analysis as they lack the capability to directly observe and understand the internal mechanisms of the prediction model.

Hence, to address the issue of interpretability of EHR, we created a novel two-stage model (Figure 1), leveraging interpretable features of text such as topic models [5] and pattern discovery and disentanglement (PDD) algorithm [6], to discover strong association patterns, thus revealing their relationships with the diagnosed disease, and clustering patients into specific groups. The output is clustering groups and an interpretable knowledge base. Our method outperforms baseline supervised and unsupervised algorithms, which were also trained on topic features. We quantitatively analyzed keywords in the top 20 topics as well as topic-class associations to discover some keyword-disease associations.

The main contributions of the paper are threefold: 1) Interpretability: a novel algorithm focusing on word-based features for interpretation of free-text clinical notes; 2) Unsupervised Learning: the grouping (i.e., clustering) of records based on the discovered associations revealing characteristics of records via unsupervised learning; 3) Knowledge Base: generating a centralized representation to link the knowledge (hierarchical clusters), patterns (characteristics of records), and data (patients' records) together to show 'what' (disease), 'who/where' (tracking patient records back) and 'why' (discovered patterns) to interpret the clinical text to aid better decision making [7].

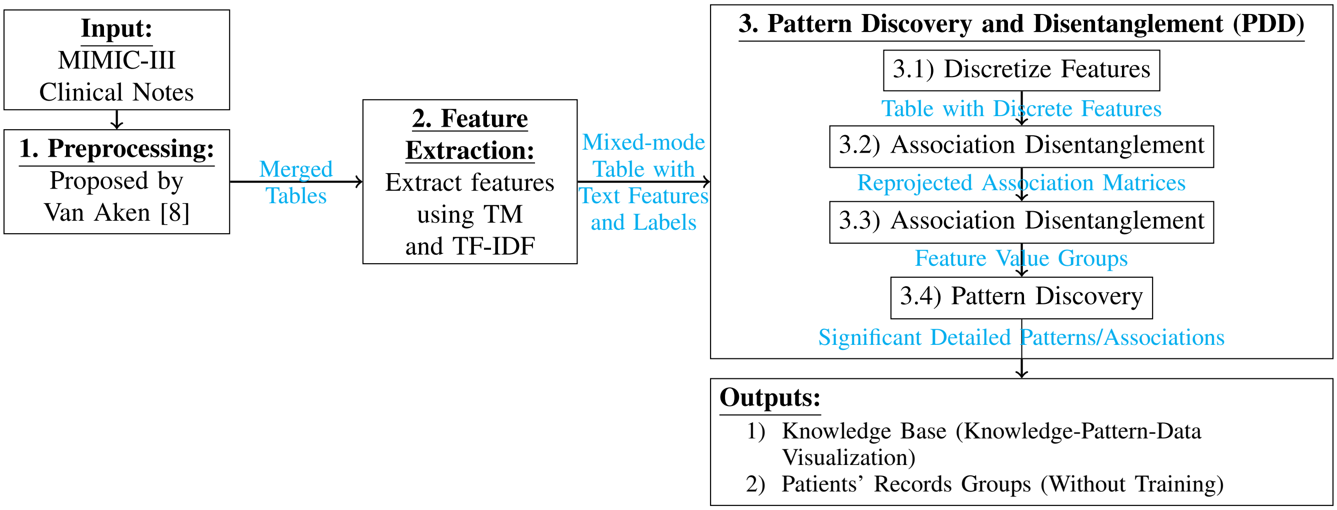
Input:
MIMIC-III
Clinical Notes
1. Preprocessing:
Proposed by
Van Aken [8]
Merged
Tables
2. Feature
Extraction:
Extract features
using TM
and TF-IDF
Mixed-mode
Table with
Text Features
and Labels
3. Pattern Discovery and Disentanglement (PDD)
3.1) Discretize Features
Table with Discrete Features
3.2) Association Disentanglement
Reprojected Association Matrices
3.3) Association Disentanglement
Feature Value Groups
3.4) Pattern Discovery
Significant Detailed Patterns/Associations
Outputs:
1) Knowledge Base (Knowledge-Pattern-Data Visualization)
2) Patients' Records Groups (Without Training)

Fig. 1. The overview of the proposed process for Pattern Discovery and Disentanglement for Clinical Insights

II. Related Work

Existing research in the field of clinical data analysis has predominantly concentrated on structured data, disregarding complementary text data [9]. Therefore, there is a need to explore methods and techniques that can effectively leverage the wealth of information present in unstructured clinical texts to enhance clinical decision-making and improve patient care.

Latent Dirichlet Allocation (LDA) [10] is a method of topic modeling (TM), which has been applied in prior studies to predict clinical outcomes from the clinical notes in EHRs [11]–[13]. The topic features cluster terms into a small set of semantically related groups, which is proved useful in text classification and categorizing clinical reports [5], [14], [15]. For example, Horng et al. [16] combined structured and unstructured data for sepsis prediction using text modeling involving topic models. Furthermore, Gangavarapu et al. [17] proposed a vector space and TM-based approach applied to structure the raw clinical data by exploiting the data in the nursing notes. Hence, we use TM in this study to transform unstructured free clinical notes into structured tabular data with topic features.

With the developments in neural networks, variants of pre-trained large language models (LLMs) such as bidirectional encoder representations from transformers (BERT) [18] have widely been applied to clinical domains (e.g., BioBERT [19], ClinicalBERT [20]). However, LLMs have been shown to have issues in the clinical domain, such as failure to interpret numerical values like body temperature [21], as well as issues with the misinterpretation of medical-specific negation (e.g., ”abstinence from alcohol” becomes ”alcohol dependence syndrome”) [8]. Unlike topic models or term-frequency-inverse document frequency (TF-IDF), a BERT vector does not contain any explicit semantic information that can be easily interpretable by a person.

While there have been applications of black-box neural networks in conjunction with posthoc interpretability methods on free clinical texts [22], [23], posthoc methods do not have the ability to directly observe the inner workings of the prediction model like an all-in-one solution such as PDD can. For example, Decision Trees can generate a rule set between features and class labels for interpretable prediction, but the rules need to be trained by relying on labeled classes. On the other hand, Frequent Pattern Mining [24], [25] can discover knowledge in the form of association rules from relational data [25] [8] but a manual threshold needs to be set for calculated likelihood, support, or confidence [8]. Furthermore, the discovered patterns may be overwhelmed [26] by overlapping/redundant patterns, which requires some posthoc analysis approaches such as pattern pruning or pattern summarizing [26].

III. Materials and Methods

A. Clinical Dataset

In this study, we applied the Medical Information Mart of Intensive Care III (MIMIC-III) v1.4[1] dataset, which is a freely available extensive database comprising de-identified EHRs about patients admitted to Intensive Care Unit (ICU) of Beth Israel Deaconess Medical Center between 2001 and 2012 [27]. MIMIC-III includes English clinical notes written in free text format. Our research focuses specifically on discharge summaries and the associated International Classification of Disease (ICD9) codes for each admission.

To examine the pattern-discerning capability of PDD algorithm on clinical text, we only used the top 4 most frequent codes. Increasing the number of ICD9 codes results in higher data imbalance and decreased prediction performance. Hence, we may not clearly examine the pattern-discerning capability of PDD when prediction performance is low [28]. The dataset consists of 11,537 patient records labeled based on the prevalence of the top four diseases indicated by their respective ICD9 codes: 414 (chronic ischemic heart disease),

[1] https://physionet.org/content/mimiciii/1.4/

038 (septicemia), 410 (acute myocardial infarction), and 424 (diseases of the endocardium). The distribution of instances across the four classes is as follows: 3502 (30.35%) for the first class, 3184 (27.6%) for the second class, 3175 (27.52%) for the third class, and 1676 (14.53%) for the fourth class.

B. Preprocessing and Feature Extraction

We first applied the preprocessing pipeline proposed by Van Aken et al. [8]. We then extracted features from the clean clinical notes by TF-IDF and TM methods (Details in Appendix V-A and V-B respectively).

C. Baseline Supervised and Unsupervised Approaches

In order to assess the predictive capability of our method, we conducted a comparative analysis with baseline unsupervised (i.e., K-Means) and supervised (i.e., Random Forest (RF) and Convolutional Neural Network (CNN)) learning models for the ICD9 prediction task (Details in Appendix V-C). For evaluating the performance of the models, we divided the dataset into a 70% train set and a 30% test set. Considering the imbalanced nature of our dataset, we employed evaluation methods outlined in [8]. Accordingly, we utilized the 'Balanced Accuracy' and 'Weighted F1-Score' metrics to evaluate the effectiveness of the selected models in predicting ICD9 codes.

D. Pattern Discovery and Disentanglement

1) Pattern Disentanglement: Firstly, we convert the values of numerical features into categorical features by using the Equal Frequency Discretization[2], which distributes the values into equal size bins. We denote categorical values of features as Attribute Value (AV) [6].

Secondly, in order to measure the strength of the association between each pair of AVs (i.e., the specific values of one attribute co-occurring with the value of another attribute), we construct an association matrix using the value of adjusted Standardized Residual (SR) [6] to represent the statistical weights of the AV pair, which is denoted as SR($AV_1 \leftrightarrow AV_2$) (shorten as SR($AV_{12}$)) and calculated by Eqn. (1) below:

$$SR(AV_{12}) = \frac{Occ(AV_{12}) - Exp(AV_{12})}{\sqrt{Exp(AV_{12})}} \times (1 - \frac{Occ(AV_1)}{T}\frac{Occ(AV_2)}{T}) \tag{1}$$

where $Occ(AV_1)$ and $Occ(AV_2)$ are the numbers of occurrences of AV; $Occ(AV_{12})$ is the total number of co-occurrence for two AVs in a AV pair; and $Exp(AV_{12})$ is the expected frequency, and T is the total number of records.

Hence, an association matrix, treated as a vector space, is generated to represent the strength of associations between each pair of AVs. Each row of the matrix, corresponding to a distinct AV, represents an AV-vector with SRs between that AV associated with all other AVs corresponding to the column vectors as its coordinates.

Then, we use Principal Component Analysis (PCA) to decompose the association matrix into principal components that are ranked according to the weights of the associations (eigenvalues). We then reproject the principal components onto the association matrix again. We refer to the reprojected association matrix as disentangled space. The above process is called *Pattern Disentanglement*, which allows us to take the reprojected components/vectors from PCA and use the reprojected values as new measurements/criteria to represent the strength of associations between AVs in different orthogonal disentangled spaces. Lastly, in order to obtain only the significant pairs of AV associations, we filter out statistical residual values greater than 1.96 in our newly reprojected association matrix (i.e., association matrix with disentangled associations)

2) Pattern Clustering: In an unsupervised manner, we cluster the associations. Typically, the number line of one projected principal component has two opposite sets of AV. However, when such opposing sets do not exist, we only use AV sets from one side of the PC. To reveal further characteristics of the records of the disentangled patterns, we separate the above sets into several subsets by clustering them. The similarity measure we used for clustering is the percentage of the overlapping records covered by each AV subcluster, and we denote each AV subgroup by a three-digit code [#Principal Component (PC), #Attribute Value Group (Group), #Attribute Value Sub-Group (SubGroup)]. The AV sets or subsets can reveal the characteristics of the records corresponding to disentangled patterns in order to provide statistical evidence for downstream clustering or prediction. The patient records are obtained according to their particular characteristics (disentangled patterns) from the AV groups or subgroups.

3) Output: The output of PDD is organized into a representational framework (PDD Knowledge Base) with three parts: a Knowledge Section showing the hierarchical clusters such that each cluster unveils distinct characteristics of a related group of records; a Pattern Section listing patterns showing detailed associations between AVs; and the Data Section listing the record ID, which links the patient to the knowledge and pattern sections. This is shown in Table II.

IV. Results

A. Comparison of PDD Classification Performance with the Baseline Approaches

PDD performance in predicting ICD9 codes is compared with baseline supervised (i.e., K-Means) and unsupervised learning (i.e., RF and CNN) models in Table I based on balanced accuracy and weighted F1-score metrics. The results are reported for four different feature sets containing $TF{-}IDF_{40}$, and $TM_{5,20,30}$. As for the topic features, we only included the number of topics with the highest topic coherence.

[2] Equal Frequency Discretization is a technique that divides a continuous variable into equal-sized intervals, ensuring an equal number of observations in each interval.

(a) Unsupervised Methods

	$TF-IDF_{40}$		TM_5		TM_{20}		TM_{30}	
	K-means	PDD	K-means	PDD	K-means	PDD	K-means	PDD
Balanced Acc.	0.48	0.45	0.62	**0.78***	0.50	0.74*	0.51	0.73*
Weighted F1	0.42	0.41	0.57	**0.78***	0.54	0.72	0.56	0.71

(b) Supervised Methods

	$TF-IDF_{40}$		TM_5		TM_{20}		TM_{30}	
	RF	CNN	RF	CNN	RF	CNN	RF	CNN
Balanced Acc.	0.81	**0.85***	0.62	0.62	0.72	0.70	0.71	0.70
Weighted F1	0.82	**0.84***	0.65	0.66	0.74*	0.72	0.73*	0.72

TABLE I

COMPARISON OF ICD9 PREDICTION PERFORMANCE BETWEEN DIFFERENT UNSUPERVISED (A) AND SUPERVISED (B) LEARNING APPROACHES USING TF-IDF AND TM FEATURE SETS.

(a) Compares two unsupervised methods, K-Means and PDD, on every feature set. (b) Compares two supervised methods, RF and CNN, on every feature set. In (a) and (b), **bold** texts represent the best value achieved for a specific metric among all feature sets and methods, while underlined texts specify the better value achieved for each feature among different methods. '*' specifies the best overall value of each metric among methods, supervised or unsupervised, for each feature set.

PDD Knowledge Base													
Knowledge Space			Pattern Space (Attributes i.e., Topics in this study)										Data Space
PC	Group	SubGroup	Residual	ICD9	Topic 0	Topic 1	Topic 2	...	Topic 6	Topic 7	...	Topic 19	Records ID
1	1	1	19.76	424	[0.01 0.42]	[0.03 0.54]	[0.03 0.44]	...	[0.00 0.01)	[0.00 0.04)	...		#1,#9,#13,...
2	1	1	24.46	424	[0.01 0.42]		[0.00 0.03)	...		[0.00 0.08)	...	[0.02 0.04)	#1,#9,#13,...
1	1	2	9.39	410	[0.01 0.42]		[0.03 0.44]	...	[0.00 0.01)	[0.00 0.04)	...		#2,#4,#5,#7,...
2	2	1	15.28	410		[0.00 0.01)	[0.03 0.44]	...		[0.08 0.47]	...		#2,#4,#5,#7,...
1	1	3	26.59	414	[0.01 0.42]	[0.00 0.01)	[0.03 0.44]	...	[0.00 0.01)	[0.00 0.04)	...		#3,#6,#16,...
2	1	2	33.81	414	[0.01 0.42]	[0.03 0.54]	[0.00 0.03)	...			...	[0.02 0.04)	#3,#6,#16,...
1	2	1	50.27	38	[0.00 0.01)	[0.00 0.01)	[0.00 0.03)	...	[0.01 0.31]	[0.08 0.47]	...		#9,#12,#16,...
Note: PC=Principal Component; Group=Attribute Value Group; SubGroup=Attribute Value Sub-Group													

TABLE II

THE PDD KNOWLEDGE BASE WHEN TOP 20 TOPICS ARE USED AS INPUT.

As it is shown in Table I, both supervised learning approaches outperform the unsupervised learning techniques when trained on TF-IDF features. One potential reason is that the top 40 features are selected based on ICD9 classification performance using feature importance ranking with RF.

It can also be observed that the performance of PDD is comparable to that of supervised learning approaches when trained on TM features. Notably, PDD with 5 topic features exhibits the best performance among all the selected models trained on various numbers of topic features.

B. Discussion on Interpretability of PDD Model

From a clinical perspective, the generated topic models reasonably aligned with each ICD9 code. In the model with 20 topics, septicemia, a widespread infection of the body, was predicted by topics containing relevant words including 'infection', 'bacteria', and 'culture'. Topics with words like 'ventricular' or 'aorta' contributed to the prediction of heart-related diseases. Additionally, the model was able to discern the heart-related diagnoses from one another: dividing acute myocardial infarction (410) from the more chronic and congenital diseases (414, 424). The algorithm could have discerned the words representing severe prognoses or procedures, such as 'angioplasty', 'emergency', and 'death' were more correlated with acute myocardial infarction.

Table II shows the partial PDD knowledge base on 20 topics. In each principal component, two opposite groups are discovered; one ICD9=4XX, which represents heart-related diseases, and one ICD9=038, which represents septicemia disease. Also, in the ICD9=4XX group of the first principal component, we observe three subgroups that could distinguish between 424 (diseases of the endocardium), 414 (chronic ischemic heart disease), and 410 (acute myocardial infarction), three different types of heart diseases; while these subgroups were not discovered using the 5 topic features, which is shown in Table III in the Appendix. In Table II, you can observe similar patterns between 424 (diseases of the endocardium) and 414 (chronic ischemic heart disease), like high probabilities in topics 1 and 2 (cardiovascular/surgery) or low probabilities in topics 6 and 7 (status/consciousness). In contrast, 038 (septicemia) shows opposite patterns, such as high probability in topic 7, along with low probabilities in topics 1 and 2.

In this work, we demonstrated that the integration of topic modeling with PDD presents an interpretable method for effectively predicting ICD9 diagnoses using unstructured clinical text. This approach ensures a balance between interpretability and outcome prediction accuracy, offering insightful patterns while maintaining reasonable diagnostic accuracy.

V. Conclusions

In this work, we propose a novel two-step algorithm using interpretable word-based features with unsupervised PDD to predict diseases. Our method outperforms K-means, especially when applied to the dataset extracted by TM features. In addition, clustering results of PDD based on the discovered patterns reflects the functional sources of the original dataset instead of class labels. Our method is a global interpretable white-box model (from the input, throughput to the output) that can provide clinicians with an explainable knowledge base that synchronizes self-correcting classification and clustering results in summarized and comprehensive forms to provide interpretability and traceability [7]. For future work, we plan to compare PDD against popular text interpretability methods.

References

[1] B. Kim, "Interpretability," 2021. [Online]. Available: https://beenkim.github.io/

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[3] R. B. Ghannam and S. M. Techtmann, "Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring," *Computational and Structural Biotechnology Journal*, 2021.

[4] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes," *Computer methods and programs in biomedicine*, vol. 177, pp. 141–153, 2019.

[5] J. Chen, J. Lalor, W. Liu, E. Druhl, E. Granillo, V. G. Vimalananda, and H. Yu, "Detecting hypoglycemia incidents reported in patients' secure messages: using cost-sensitive learning and oversampling to reduce data imbalance," *Journal of medical Internet research*, vol. 21, no. 3, p. e11990, 2019.

[6] A. K. Wong, P.-Y. Zhou, and Z. A. Butt, "Pattern discovery and disentanglement on relational datasets," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[7] A. K. Wong, P.-Y. Zhou, and A. E.-S. Lee, "Theory and rationale of interpretable all-in-one pattern discovery and disentanglement system," *npj Digital Medicine*, vol. 6, no. 1, p. 92, 2023.

[8] B. Van Aken, J.-M. Papaioannou, M. Mayrdorfer, K. Budde, F. A. Gers, and A. Löser, "Clinical outcome prediction from admission notes using self-supervised knowledge integration," *arXiv preprint arXiv:2102.04110*, 2021.

[9] P. Culliton, M. Levinson, A. Ehresman, J. Wherry, J. S. Steingrub, and S. I. Gallant, "Predicting severe sepsis using text from the electronic health record," *arXiv preprint arXiv:1711.11536*, 2017.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[11] R. A. Bright, S. K. Rankin, K. Dowdy, S. V. Blok, S. J. Bright, and L. A. M. Palmer, "Finding potential adverse events in the unstructured text of electronic health care records: Development of the shakespeare method," *JMIRx Med*, vol. 2, no. 3, p. e27017, 2021.

[12] Z. Huang, W. Dong, and H. Duan, "topic model for clinical risk stratification from electronic health records," *Journal of Biomedical Informatics*, vol. 58, pp. 28–36, 2015.

[13] Y. Wang, Y. Zhao, T. M. Therneau, E. J. Atkinson, A. P. Tafti, N. Zhang, S. Amin, A. H. Limper, S. Khosla, and H. Liu, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *Journal of biomedical informatics*, vol. 102, p. 103364, 2020.

[14] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and lda topic models," *Expert Systems with Applications*, vol. 80, pp. 83–93, 2017.

[15] E. S. Kayi, K. Yadav, and H.-A. Choi, "Topic modeling based classification of clinical reports," in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, 2013, pp. 67–73.

[16] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro, and L. A. Nathanson, "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning," *PloS one*, vol. 12, no. 4, p. e0174708, 2017.

[17] T. Gangavarapu, A. Jayasimha, G. S. Krishnan, and S. Kamath, "Predicting icd-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes," *Knowledge-Based Systems*, vol. 190, p. 105321, 2020.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[20] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[21] E. Wallace, Y. Wang, S. Li, S. Singh, and M. Gardner, "Do nlp models know numbers? probing numeracy in embeddings," *arXiv preprint arXiv:1909.07940*, 2019.

[22] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset," vol. 12, no. 1, p. 7166, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-022-11012-2

[23] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical intervention prediction and understanding with deep neural networks," in *Proceedings of the 2nd Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 68. PMLR, 18–19 Aug 2017, pp. 322–337. [Online]. Available: https://proceedings.mlr.press/v68/suresh17a.html

[24] S. Naulaerts, P. Meysman, W. Bittremieux, T. N. Vu, W. Vanden Berghe, B. Goethals, and K. Laukens, "A primer to frequent itemset mining for bioinformatics," *Briefings in bioinformatics*, vol. 16, no. 2, pp. 216–231, 2015.

[25] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data mining and knowledge discovery*, vol. 15, no. 1, pp. 55–86, 2007.

[26] A. K. Wong and G. C. Li, "Simultaneous pattern and data clustering for pattern cluster analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 7, pp. 911–923, 2008.

[27] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[28] W. J. Murdoch, C. Singh, K. Kumbier, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019. [Online]. Available: https://doi.org/10.1073/pnas.1900654116

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[31] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[32] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[33] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

PDD Knowledge Base										
Knowledge Space				Pattern Space (Attributes i.e., Topics in this study)						Data Space
PC	Group	SubGroup	Residual	ICD9	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Records ID
1	1	1	24.99	410	[0.00 0.01)		[0.03 0.17)	[0.13 0.95]	[0.07 0.36)	#2,#11,#44,#53,#63,...
1	1	1	11.71	414	[0.00 0.01)		[0.17 0.94]	[0.13 0.95]	[0.00 0.07)	#62,#88,#93,...
1	1	1	13.64	424	[0.00 0.01)	[0.42 0.97]	[0.17 0.94]		[0.00 0.07)	#1,#63,#184,...
1	2	1	51.07	38		[0.18 0.42)	[0.00 0.03)	[0.03 0.13)	[0.36 0.97]	#35,#53,#77,#80,...
1	2	1	86.06	38	[0.01 0.84]	[0.00 0.18)	[0.00 0.03)		[0.36 0.97]	#84,#96,#99,...
1	2	1	56.5	38	[0.01 0.84]	[0.00 0.18)		[0.03 0.13)	[0.36 0.97]	#84,#126,#130,...
2	1	1	10.55	424		[0.42 0.97]	[0.17 0.94]		[0.00 0.07)	#1,#63,#176,...
2	2	1	85.89	38		[0.00 0.18)	[0.00 0.03)	[0.03 0.13)	[0.36 0.97]	#12,#83,#84,...
3	1	1	18.99	424		[0.42 0.97]	[0.00 0.03)	[0.03 0.13)	[0.00 0.07)	#206,#225,...
3	2	1	19.1	410	[0.00 0.01)	[0.18 0.42)	[0.17 0.94]		[0.07 0.36)	#8,#64,#75,...
3	2	1	31.56	410	[0.00 0.01)	[0.00 0.18)		[0.13 0.95]	[0.07 0.36)	#2,#42,#53,...
Note: PC=Principal Component; Group=Attribute Value Group; SubGroup=Attribute Value Sub-Group										

TABLE III
THE PDD KNOWLEDGE BASE WHEN TOP 5 TOPICS ARE USED AS INPUT.

APPENDIX

A. *TF-IDF Definition and Implementation*

TF-IDF can be computed as:

$$\text{tf-idf(t,d)} = \text{tf(t,d)} \times \text{idf(t)}$$

where tf refers to the term frequency (proportion of a particular term t over all terms); and

$$\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1$$

where n is the total number of documents in the set and df is the number of documents containing the term t.

We set TF-IDF configurations to the default parameter settings of the `TfidfVectorizer` feature extraction package of Scikit-learn library [29] and converted each clinical note into a vector of TF-IDF features. We determined the optimal number of TF-IDF features for ICD9 code prediction by employing feature importance ranking with random forest (RF). We evaluated the performance based on balanced accuracy and weighted F1-score by training the RF model on various numbers of top TF-IDF features, ranging from 1 to 50. The results demonstrated an improvement in performance as the number of features increased, peaking at 40. However, performance started to degrade beyond that point. Thus, we selected the top 40 words as the input feature set when utilizing TF-IDF.

B. *Topic Modeling Implementation and Parameter Tuning*

1) Topic Modeling: We used `models.ldamodel` package of Gensim open-source topic modeling library [30] to derive the topic features from the clinical notes. To determine the optimal number of topics, we calculated the coherence score [31] of the topic clusters across a range of topic numbers, spanning from 1 to 40. The coherence score peaks when the number of topics is set to 5, 20, and 30, and therefore, we only extracted topic features with those respective parameters.

Figure 2 illustrates the coherence score per number of topics for topic clusters across a range of topic numbers (1-40). Optimal topic selections are indicated at 5, 20, and 30 topics.

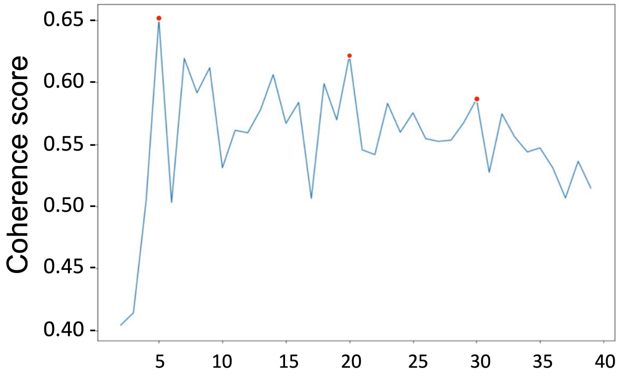

Fig. 2. Visualization of coherence score per k number of topics for topic clusters derived from MIMIC-III clinical notes with the red dots being the local maximums for selecting the best k.

C. *Baseline Implementations*

To implement the PDD algorithm, we utilized the PDD web interface[3]. To implement K-Means clustering, we employed the `sklearn.cluster.KMeans` package from the Scikit-learn library [29], using its default parameter setting.

We used the Keras [32] deep learning package written in Python to implement CNN architecture [33]. The architecture

[3] http://pdd.uwaterloo.ca/

of the model consists of multiple layers, including an input layer, a 1D convolution layer, a batch normalization layer, a dropout layer, and a 1D max-pooling layer. These layers are then followed by a fully connected classification layer specifically designed for the ICD9 prediction task. During the training process, we employ the Adam optimizer with a learning rate of 0.001. The model is trained over 25 epochs, with a batch size of 32. To implement the RF model, we applied the `RandomForestClassifier` package of the Scikit-learn library [29] with its default parameter setting.