

Predicting ICD-9 Code Groups with Fuzzy Similarity based Supervised Multi-Label Classification of Unstructured Clinical Nursing Notes[☆]

Tushaar Gangavarapu^{a,1,*}, Aditya Jayasimha^b, Gokul S Krishnan^b,
Sowmya Kamath S^b

^a*Amazon.com, Inc., Bangalore, Karnataka, India*

^b*Healthcare Analytics and Language Engineering (HALE) Lab,
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, Mangaluru, India*

Abstract

In hospitals, caregivers are trained to chronicle the subtle changes in the clinical conditions of a patient at regular intervals, for enabling decision-making. Caregivers' text-based clinical notes are a significant source of rich patient-specific data, that can facilitate effective clinical decision support, despite which, this treasure-trove of data remains largely unexplored for supporting the prediction of clinical outcomes. The application of sophisticated data modeling and prediction algorithms with greater computational capacity have made disease prediction from raw clinical notes a relevant problem. In this paper, we propose an approach based on vector space and topic modeling, to structure the raw clinical data by capturing the semantic information in the nursing notes. Fuzzy similarity based data cleansing approach was used to merge anomalous and redundant patient data. Furthermore, we utilize eight supervised multi-label classification models to facilitate disease (ICD-9 code group) prediction. We present an exhaustive comparative study to evaluate the performance of the proposed approaches using standard evaluation metrics. Experimental valida-

[☆]© 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>. The final authenticated version is available online at: <https://doi.org/10.1016/j.knosys.2019.105321>.

*Corresponding author.

Email addresses: tusgan@amazon.com (Tushaar Gangavarapu),
15it103.aditya@nitk.edu.in (Aditya Jayasimha), gsk1692@gmail.com (Gokul S Krishnan),
sowmyakamath@nitk.edu.in (Sowmya Kamath S)

URL: <http://infotech.nitk.ac.in/faculty/sowmya-kamath-s> (Sowmya Kamath S)

¹T. Gangavarapu completed most of this work at Healthcare Analytics and Language Engineering (HALE) Lab, Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangaluru, India.

tion on MIMIC-III, an open database, underscored the superior performance of the proposed Term weighting of unstructured notes *AG*gregated using fuzzy Similarity (*TAGS*) model, which consistently outperformed the state-of-the-art structured data based approach by 7.79% in AUPRC and 1.24% in AUROC.

Keywords: Clinical Decision Support Systems, Disease Prediction, Healthcare Analytics, ICD-9 Code Group Prediction, Machine Learning, Natural Language Processing.

1. Introduction

Disease prediction and quantification of patients’ health data have been shown to have significant contributions in improving clinical care and management [70]. Every year, over 30 million patients visit hospitals in the United States alone [22], and 83% of these hospitals utilize the Electronic Health Record (EHR) system [36]. EHRs have seen widespread adoption due to the stipulations of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [1]. Over recent years, with the rise in EHR implementation in the hospitals of developed countries, application of machine and deep learning models to patient data for the prediction of clinical outcomes such as causal effect inference and survival analysis has sparked widespread interest [100, 81, 28, 92]. Owing to the availability of large, de-identified, public healthcare databases such as MIMIC (Medical Information Mart for Intensive Care II [55] and III [43]), mining patient data to accurately assess the severity of illness and determining diagnostic measures for augmenting healthcare policies has become a prominent area of research [44, 57, 16]. Healthcare data accessible via structured EHRs is widely used in the existing Clinical Decision Support Systems (CDSSs) [10, 46, 67]. However, there is limited adoption of these structured EHRs in developing countries, thus leaving clinicians in such countries with no choice but to resort to manual consumption of available clinical notes for causal effect inference and decision-making [48].

Clinical notes maintained by caregivers like nurses, record subjective assessments and crucial information concerning a patient’s state, which is mostly lost when transcribed into structured EHRs [29]. Mining and modeling such nursing notes for extracting rich patient data and utilizing this to predict clinical events and outcomes with machine learning models is a challenging process, owing to their rawness, high-dimensionality, sparsity, complex temporal and linguistic structure, and presence of rich medical jargon and abbreviations [29, 42]. The efficacy of using such raw clinical notes largely depends on the ability to extract and consolidate the information embedded in them effectively [91]. Furthermore, there is often a need for multiple-label assignment (from a large set of

potential labels) to a patient record [3] due to the manifold nature of disease symptoms. Disease prediction (ICD-9² code group prediction [30]) and risk assessment via nursing notes can help in taking effective measures at the earliest signs of patient distress. Recognition of the onset of disease and the determination of its risk using clinical nursing notes, followed by effective communication and response by interdisciplinary care team members could be both time- and cost-efficient [25], which can also lead to reduced hospital mortality rate [20].

Early works [87, 31, 60, 34, 19] applied machine learning techniques to structure patient data in forecasting the length of stay in Intensive Care Units (ICUs) and mortality prediction. In recent years, practical progress in clinical machine and deep learning is benchmarked using MIMIC databases, for clinical prediction tasks such as in-hospital, short-term, and long-term mortality prediction, length of stay prediction, phenotyping, and ICD-9 code group prediction [35]. Johnson et al. [44] extracted a set of features from the MIMIC-III database for the prediction of ICU mortality and compared several existing works against Logistic Regression (LR) and gradient boosting models. More recently, Purushotham et al. [70] reported their performance on five clinical prediction tasks (on MIMIC-III database) using deep learning models and compared the performance with existing state-of-the-art methods and scoring systems.

Although some state-of-the-art methods benchmark machine and deep learning models for several clinical prediction tasks on MIMIC, they have neglected the rich patient information available in the unstructured clinical nursing notes. In this paper, the applicability of vector space models (with term weighting [80] and Doc2Vec [53]), topic modeling (Hierarchical Dirichlet Process (HDP) [84] and Latent Dirichlet Allocation (LDA) [6] with Topic Coherence (TC) [77]) is studied to model this data. Our objective is to measure their effectiveness in vectorizing and accurately modeling the semantic relationships between the textual features of unstructured nursing notes, for accurately predicting the ICD-9 code groups. A fuzzy similarity based data cleansing approach was designed to derive optimal data representations and eliminate redundant information in the nursing notes, thus improving the causal effect inference. We experimented with eight supervised multi-label classification approaches including K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), One-vs-Rest (OvR) with KNN, OvR with LR, OvR with Support Vector Machines (SVM), Random Forest (RF), Hard Voting Ensemble (HVE), and Stacking Ensemble (SE), to accurately predict the ICD-9 code groups. Furthermore, we present an exhaustive study to evaluate a variety of data cleansing (using similarity) and modeling (using machine learning) approaches across several standard evaluation metrics. The key

²International Classification of Diseases, ninth revision.

70 contributions of our work are summarized below:

- 71 • Design of a fuzzy token-based similarity matching approach for unstruc-
72 tured clinical data. This is used for deriving optimal data representations
73 and eliminating anomalous or redundant data, due to which the cognitive
74 burden is reduced, and an improvement in the clinical decision-making
75 process is observed.
- 76 • Leveraging vector space and topic modeling to extract the rich patient-
77 specific information available in unstructured clinical nursing notes to pre-
78 dict ICD-9 code groups accurately. Experimental results show that our
79 proposed supervised learning models consistently outperformed the state-
80 of-the-art models built on structured data.
- 81 • Design of an approach that utilizes unstructured clinical text for the de-
82 velopment of CDSSs, thus eliminating the dependency on the availability
83 of structured EHRs. This can be crucial in countries where structured
84 EHR adoption is not widespread.

85 The rest of this paper is organized as follows: Section 2 provides an overview
86 of the related work and reviews their advantages and limitations. Section 3
87 describes the MIMIC-III database and the preprocessing steps designed to gen-
88 erate optimal representations from the clinical nursing notes. The experiments,
89 evaluation, and results are discussed in great detail in Section 4. Finally, Sec-
90 tion 5 concludes this paper with highlights on future research possibilities.

91 2. Related Work

92 An extensive body of research on using machine and deep learning models for
93 clinical predictions is available in the existing literature. In this section, we
94 discuss a few of these works to provide an overview of the existing models and
95 state-of-the-art methods built on large healthcare datasets. In this discussion,
96 we also highlight the importance of accurate ICD-9 code group prediction in
97 modern healthcare systems.

98 Buchman [9] compared statistical and connectionist models for the prediction of
99 clinical trajectory, including resource and outcome utilization in surgical ICUs.
100 However, much of this work formulated the task of identifying patients at risk as
101 binary classification rather than regression. Other early works [11, 21] showed
102 that machine learning models provide promising results in predicting medical
103 risk, mortality, and in forecasting the length of stay in ICU. Early works [11, 12]
104 also established that feed-forward neural networks almost always outperformed
105 severity scores and logistic regression in mortality risk prediction among hos-
106 pitalized patients. With recent advances in machine and deep learning, there

107 is widespread interest in applying these models to predict healthcare outcomes
 108 accurately [52, 63, 15]. Dabek and Caban [24] reported that several psychological
 109 conditions, including depression, post-traumatic stress disorder, and anxiety,
 110 could be improved using a neural network model. Che et al. [14] designed a
 111 scalable feed-forward deep learning framework for disease diagnosis that learns
 112 relevant clinical features based on the prior knowledge from medical ontologies.

113 Some works that aimed at multi-label prediction of the diagnostic codes from
 114 clinical time series used feed-forward neural networks [52], temporal Convolutional
 115 Neural Networks (CNNs) [74], and Long Short Term Memory (LSTM)
 116 networks [17] to capture the co-morbidities in the hidden layers implicitly. Other
 117 recent works [56, 33, 69] modeled clinical time series and disease data by leveraging
 118 the power of deep learning approaches. In 2016, novel deep learning
 119 architectures were proposed to model survival analysis as a time-to-event regression
 120 task [96, 72]. Luo [58] used sentence and segment LSTM models with
 121 word embeddings to classify the relations in the nursing notes. More recently,
 122 Rajkomar et al. [71] showed that novel neural network based architectures including
 123 LSTM perform well in the prediction of an extended length of stay,
 124 30-day unplanned re-admission, inpatient mortality, and diagnoses on general
 125 EHR data. Krishnan and Kamath [48] used extreme learning machine architecture
 126 with Word2Vec embedding for mortality prediction using unstructured
 127 ECG text reports. Khin [45] developed a bi-directional LSTM with deep contextualized
 128 word embeddings and variational dropouts, and empirically validated the model's
 129 superiority in terms of performance and convergence. These previous works demonstrate
 130 the power and efficacy of machine and deep learning models in large healthcare applications.

132 The availability of large public healthcare databases such as MIMIC-II and
 133 MIMIC-III has enabled healthcare researchers to benchmark the developed machine
 134 and deep learning models in the effective prediction of clinical events and
 135 outcomes. In 2016, Pirracchio [66] presented that the super learner algorithm
 136 which is an ensemble of various machine learning models outperforms severity
 137 scores such as SOFA (Sepsis-related Organ Failure Assessment) [89], SAPS-II
 138 (Simplified Acute Physiology Score) [54], and APACHE-II (Acute Physiologic
 139 Assessment and Chronic Health Evaluation) [47] in ICU mortality prediction.
 140 The author's work underscored the superiority of machine learning models over
 141 traditional prognostic scores but the author did not benchmark the obtained
 142 results against most recent machine and deep learning models.

143 Recently, Johnson et al. [44] presented a case study on clinical mortality prediction
 144 task, highlighting the challenges in replicating results reported by related
 145 and recent publications on MIMIC-III. They reviewed 28 key existing works and
 146 compared the reported performance against LR and gradient boosting models

147 using an extracted set of features from MIMIC-III. Furthermore, the authors
 148 stressed the need for an improvement in the way of reporting the performance of
 149 clinical prediction tasks, to account for the substantial heterogeneity in the stud-
 150 ies and to ensure fairer comparison among approaches. Harutyunyan et al. [35]
 151 proposed a comprehensive deep learning approach using multitask Recurrent
 152 Neural Networks (RNNs) and empirically benchmarked their outcomes using
 153 four different clinical prediction tasks on the MIMIC-III database. Their work
 154 showed promising results for using deep learning models in clinical prediction.
 155 However, the authors only compared their obtained results against standard LR
 156 model and LSTM deep learning model [38], and excluded the comparison with
 157 machine learning models (specifically, super learner) or severity scoring systems.
 158 Purushotham et al. [70] presented an exhaustive set of benchmarking results on
 159 several clinical tasks including the length of stay, phenotyping, multiple versions
 160 of in-hospital mortality predictions, and ICD-9 code group predictions using the
 161 MIMIC-III database. They used LSTM-based deep architectures and compared
 162 their performance with traditional machine learning approaches and severity
 163 scoring systems on these tasks.

164 In 2019, Krishnan and Kamath [50] proposed a novel hybrid metaheuristic ap-
 165 proach with genetic algorithm and extreme learning machine for patient-specific
 166 mortality prediction that outperformed various severity scoring systems and ma-
 167 chine learning models. However, their study uses large-scale structured lab event
 168 data for the clinical prediction task. In a parallel work [49], ICU mortality pre-
 169 diction task was performed using Word2Vec, Glove, and FastText embeddings
 170 of MIMIC-III nursing notes. They used the RF classifier, and their data pro-
 171 cessing and feature extraction are quite different from the approaches followed
 172 in this paper. Stone [83] discussed the opportunities of improving the triage
 173 accuracy in CDSSs, to effectively assist the medical personnel in drawing in-
 174 ferences in high-pressure situations with many distractions, where the patient
 175 history concerning the sustained trauma is limited. This work extends the ef-
 176 forts of the author by utilizing the patient-centric information to identify high-
 177 risk patients, thus aiding the underlying CDSS with increased triage accuracy,
 178 optimized patient outcomes, and minimized risk of clinical deterioration. To
 179 automate the process of ICD-9 coding, Zeng et al. [97] proposed a multi-scale
 180 deep neural transfer framework which employs the transfer of (Medical Subject
 181 Headings (MeSH) domain knowledge to improve the coding process. Huang et
 182 al. [40] employed state-of-the-art deep neural models, including CNN, LSTM,
 183 and Gated Recurrent Unit (GRU) to predict (top-10) ICD-9 code categories.
 184 However, these works utilize discharge summaries of the MIMIC-III database
 185 rather than the nursing notes—clinician’s notes are more rich, informative, and
 186 patient-centric. Moreover, modeling nursing notes can facilitate reliable billing,
 187 effective clinical decision support, and revising healthcare policies, while mod-

¹⁸⁸ eling discharge summaries is only useful only in billing.

Table 1: Comparison of this work with the state-of-the-art works in the prediction of clinical outcome(s) using the MIMIC-III database.

Work	Data			Approach(es)		Modeling and classification		Performance evaluation	
	Data source(s)	Structure	Volume			Classification type(s)	Feature modeling	Classifier(s)	Evaluation metric(s)
Harutyunyan et al. [35]	Chart and lab events data	Structured	42,276 ICU stays	In-hospital mortality prediction, decomposition prediction, length of stay prediction, and phenotyping	Mortality: binary; decomposition: binary; length of stay: multi-class; phenotyping: multi-label	17 selected clinical variables (1)	Deep supervision, multitask standard LSTM, and multitask channel-wise LSTM (3)	LR, standard LSTM, and channel-wise LSTM (3)	AUROC, AUPRC, Kappa, and mean absolute difference (4)
Punushodham et al. [70]	Lab, input, output, and chart events data, and prescriptions	Structured	35,627 admissions	In-hospital mortality prediction, short- and long-term mortality prediction, length of stay prediction, phenotyping, and ICD-9 code group prediction	Mortality: binary; length of stay: multi-class; phenotyping: multi-label; ICD-9 code group: multi-label	Three feature sets of 17, 20, and 135 features respectively (3)	MLP, multimodal deep learner, and RNNs (2)	Scoring methods and super learner (2)	AUPRC and AUROC (2)
Huang et al. [40]	Discharge summaries	Unstructured	59,652 summaries	Prediction of (top-10) ICD-9 code categories using state-of-the-art deep learning models	Multi-label classification via deep learning approaches	TF-IDF, Word2Vec, and word sequencing with an embedding matrix (3)	CNN, LSTM, and GRU (3)	Prakash et al. [68], LR, RF, and MLP (4)	ACC, micro F1, AUPRC, precision@5, and hamming loss (5)
Zeng et al. [97]	Discharge summaries	Unstructured	58,929 summaries	ICD-9 code assessment via deep transfer learning framework	Multi-label classification via deep neural networks	Word embeddings (1)	Transferring MeSH domain knowledge with sequential CNN (1)	Hierarchy-based SVM, flat SVM, and segmented CNN (3)	Micro-average precision, micro-average recall, and micro-average F-measure (3)
This work	Nursing notes	Unstructured	223,556 notes	Term weighting of voluminous nursing notes aggregated using the fuzzy similarity of the raw clinical text for effective ICD-9 code group assessment	Multi-label classification via machine learning approaches	Term weighting, Doc2Vec (500 and 1,000), HDP with BoW, HDP with term weighting, and LDA with TC (6)	KNN, MLP, KNN as Ovr, LR as Ovr, SVM as Ovr, RF, HVE, and SE (8)	Punushodham et al. [70], Doc2Vec (500 and 1,000), HDP with BoW, HDP with term weighting, and LDA with TC (and their respective variants of naive aggregation) (12)	Accuracy, MCC, AUROC, AUPRC, F1, CE, and LRL (7)

Many hospitals in developed countries, including the United States, employ ICD-10 diagnostic coding systems, and hence there is a need for the translation of legacy ICD-9 codes into more specific ICD-10 concepts. Hernandez-Ibarburu et al. [37] studied the incompatibilities between ICD-9 and ICD-10 coding schemes. They presented a way of improving the translation of legacy data (that employs ICD-9 codes) with an extended version of ICD-10 codes generated using selected ICD-9 codes, in turn improving the mapping reliability. To achieve the mapping, they employed general equivalence mappings and integration of certain ICD-9 concepts within the hierarchical relations of ICD-10 codes. Angiolillo et al. [2] also studied the effect of coding terminology transitions on healthcare quality analysis. They reported that the legacy metrics across ICD generations could be bridged through equivalence mapping of ICD-9 concepts. Furthermore, they hypothesized that developing novel metric definitions could mitigate the complexity arising from equivalence mapping.

Our work explores a much-neglected, but an abundant source of patient information, i.e., unstructured clinical notes, and advances the state-of-the-art methods in the literature by using the rich information present in them, which is so often lost in the structured EHR generation process. By utilizing the patient-centric information to identify high-risk patients, this work enhances the underlying CDSS with optimized patient outcomes, increased triage accuracy, and minimized risk of clinical deterioration. Furthermore, our work presents an exhaustive comparative study to evaluate the performance of various data cleansing and modeling approaches across a variety of machine learning models in the multi-label prediction of ICD-9 code groups. Table 1 shows a detailed comparison of our proposed work with the state-of-the-art works in the area of prediction of clinical outcome(s) using the MIMIC-III database.

2.1. Motivation

In hospitals, especially in ICUs, a high patient-to-staff ratio and advanced medical equipment are utilized for continuous support and monitoring of critically ill patients. However, critical care patients are often susceptible to varied complications arising from advanced medical interventions, that can adversely affect their mortality and morbidity [85]. Common infections include central line-related bloodstream infection, ventilator-related pneumonia, and catheter-related urinary tract infection, that arise from the usage of invasive devices in ICUs. Surgical site infections resulting from prior procedures performed on patients and acute renal failure due to unrecognized drug interactions are also potential risks [85]. Ventilator support provided to critical care patients is often related to several complications including barotrauma, short and long-term intubation, weaning errors, and gastrointestinal tract bleeding [94]. Additionally,

ICU patients pose a risk of acid-base problems, nutritional complications, and psychological disturbances [94]. Furthermore, ICU survivors are known to suffer from neuro-psychiatric, quality of life, and long-term physical impairments [27]. The minute variations in the condition of ICU patients is recorded and monitored regularly by the trained nursing staff. Hence, nursing notes are very data-rich voluminous resources containing continuously documented subjective and objective assessments concerning a patient’s state. Effective modeling of such clinical text to aid in the early identification of high-risk patients is of utmost importance, to provide prioritized care and prevent further complications.

Due to practical constraints, the availability of resources including medical equipment and staff in ICUs is, more often than not, limited [32]. There is often a lack of accurate knowledge of the etiology of ICU complications, leading to the inability of accurate risk assessment and prevention of resulting complications; as a result of which, in most cases, adequate clinical care can only be provided after a complication develops. ICD-9 codes are designed to code diseases into categories, essential in epidemiological studies [73], cost-effectiveness analysis, and determining healthcare policies [18]. ICD-9 code group prediction is a preliminary step to ICD-9 code prediction, requiring high prediction performance. Since the patient encounters are grouped by diagnoses, ICD-9 code groups facilitate research, along with tracking and billing, by reporting on severity, symptoms, and use of resources across agencies. Furthermore, disease-specific staging systems could be beneficial towards capturing the severity, symptoms, and use of resources within a single code group. However, the existing state-of-the-art model [70] built on structured EHR data reported modest performance in ICD-9 code group prediction with an AUROC score of 0.7772 and AUPRC score of 0.6008. Thus, there is a need for the development of an effective modeling strategy to facilitate accurate ICD-9 code group prediction, in turn aiding in the accurate determination of ICD-9 codes.

3. Materials and Methods

In this section, we first discuss in brief, the statistics of the MIMIC-III database. The detailed overview of the Natural Language Processing (NLP) pipeline architecture used in the task of ICD-9 code group prediction is shown in Figure 3. Then, we elucidate on the preprocessing steps employed to extract features for ICD-9 code group prediction as a multi-label classification task.

3.1. Dataset Description and Cohort Selection

MIMIC-III is a freely accessible large database developed by the Massachusetts Institute of Technology Lab for Computational Physiology. It encompasses di-

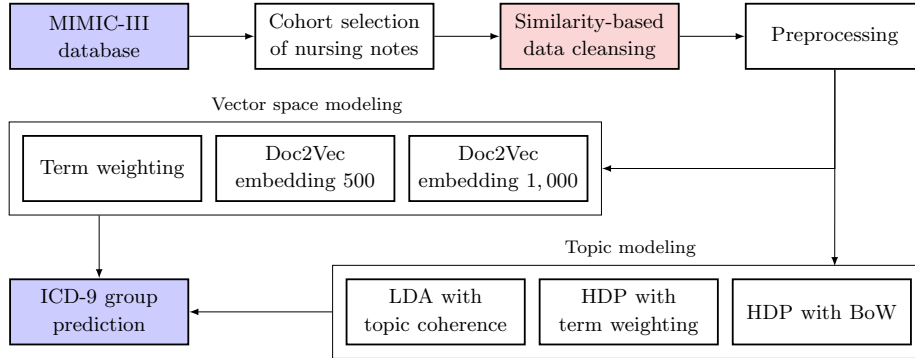


Figure 1: NLP pipeline used to predict the ICD-9 code group using unstructured clinical nursing notes.

verse and comprehensive de-identified health-related data of over 40,000 critical care patients at the Beth Israel Deaconess Medical Center, Boston, Massachusetts between June 2001 to October 2012. The database contains crucial patient information including vital sign measurements, demographics, laboratory test results, medications, procedures, imaging reports, caregiver (nursing) notes, and in and out of hospital mortality.

MIMIC-III database contains 2,083,180 note events, of which 223,556 are nursing notes of 7,704 distinct ICU patients (subjects). Details of the nursing note text corpus are summarized in Table 2. At present, we considered two criteria to select the MIMIC-III subjects in the preparation of our datasets. Firstly, the subjects with age less than 15 (neonates) were identified using the age at the time of admission to the ICU. Based on the existing literature [44, 70], only adult subjects (age 15 or above) are considered for the study. Secondly, for each MIMIC-III subject, only their first admission to the hospital was considered, and all later admissions were discarded. This was done to ensure the prediction with the earliest detected conditions (faster risk prediction), to avoid any information loss, and to ensure similar experimental settings as in existing literature [44, 70, 48]. Figure 2c outlines the distribution of the number of code group mismatches across patients’ first admission to their later admissions. From Figure 2c it can be observed that the code groups in the later admissions of over 94% of the patient nursing notes are the same as those occurring in their first hospital admission. Owing to this, we decided to consider only the first admission of a MIMIC-III subject to a hospital, with no loss of information.

Table 2: Statistics of the clinical nursing note text corpus.

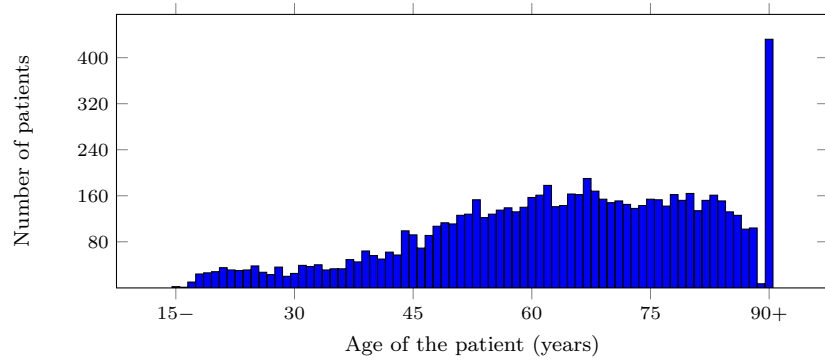
Parameter	Total	Average
Clinical nursing notes	223,556	—
Sentences in the nursing notes	5,244,541	23.46
Words in the nursing notes	79,988,065	357.80
Unique words in the nursing notes	715,821	3.20

3.2. Data Extraction

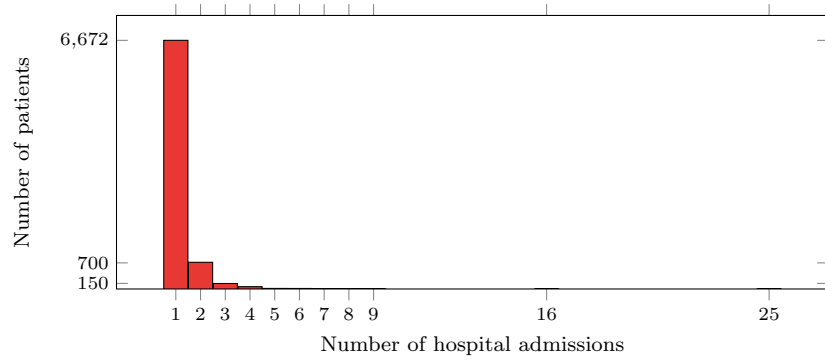
The MIMIC-III (v1.4) database consists of 26 relational tables in total. For the purpose of this study, the following four tables were used to extract the selected cohort data: **noteevents** consisting of several kinds of reports and notes including ECG reports, radiology reports, nursing notes, and discharge summaries in an unstructured text form; **admissions** reports information concerning the patient’s admission to the hospital and is used for the time of the subject’s admission to the ICU; **patients**, containing the charted data for all critical patients, from which the patients’ date-of-birth is obtained for the computation of the age of patients; **diagnoses_icd**, comprises the ICD-9 diagnoses of the patients. Most relevant healthcare features and data is present in these tables, and therefore these tables are selected to prepare datasets for the task of ICD-9 code group prediction. The statistics of the data extracted from the MIMIC database is shown in Figure 2. With the patient cohorts presented in Section 3.1, the dataset extracted from the selected tables contained nursing notes corresponding to 7,638 patients with the median age of 66 years (Quartile $Q_1 = 52$ years, Quartile $Q_3 = 78$ years).

3.3. Data Cleansing, Aggregation, and Preprocessing

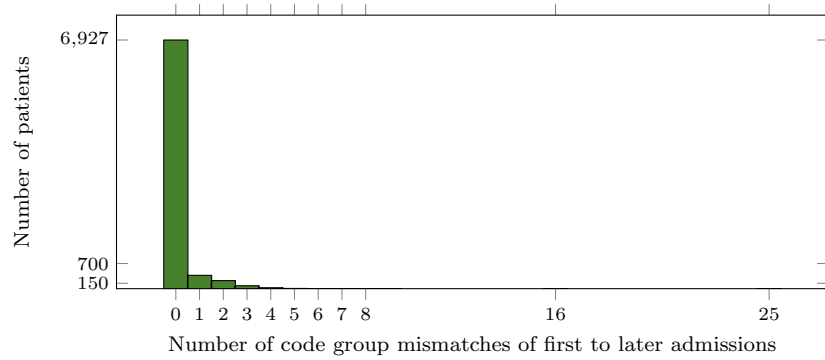
Due to various factors including outliers, noise, missing values, incorrect or duplicate records, and others, the data extracted from the MIMIC-III database has erroneous entries. The following three issues with the extracted data were identified and handled accordingly. Firstly, the erroneous entries in nursing notes with the *iserror* attribute of the **noteevents** table set to one were identified and removed. Secondly, some subjects that had duplicate records were identified, and the duplicate entries were deduplicated. The resulting data obtained by handling erroneous entries corresponded to 6,532 MIMIC-III subjects. Finally, a MIMIC-III subject had multiple nursing notes with different ICD-9 code groups, which were merged or purged using a fuzzy token-based similarity approach.



(a) The distribution of the age of MIMIC-III patients.



(b) The distribution of the hospital admissions of MIMIC-III patients.



(c) The distribution of the code group mismatches across MIMIC-III patients' first and later admissions.

Figure 2: Statistics of the data extracted from the MIMIC-III database.

3.3.1. Fuzzy Token-based Similarity Merging

Multiple nursing notes of a MIMIC-III subject have to be merged to enable multi-label ICD-9 code group classification. Figure 3 shows the heavy-tailed distribution of nursing notes across various patients. It can also be observed that the extracted MIMIC-III patient cohort has an average of 176.49 nursing notes per patient, with 4,183 patients having more than fifty nursing notes composed of over 17,890 words on an average. Such voluminous nursing notes often include many similar terms which could significantly affect the vector representations. To handle the voluminosity and near-duplicate nursing notes of a patient, Monge-Elkan (ME) [61], a token-based fuzzy similarity scoring scheme is integrated with Jaro [41] internal scoring scheme and used as a decision-making mechanism. ME similarity is used to handle clinical abbreviations, alternate names, and medical jargon. Jaro similarity is used as an internal scoring scheme to handle typographical errors and to obtain a normalized similarity score between 0 and 1. Given two nursing notes η_i and η_j with $|\eta_i|$ and $|\eta_j|$ tokens ($\mathcal{C}_k^{(i)}$ s and $\mathcal{C}_l^{(j)}$ s) respectively, their ME similarity score with Jaro is,

$$\text{ME}_{\text{Jaro}}(\eta_i, \eta_j) = \frac{1}{|\eta_i|} \sum_{k=1}^{|\eta_i|} \max_{l=1}^{|\eta_j|} \left\{ \text{Jaro}(\mathcal{C}_k^{(i)}, \mathcal{C}_l^{(j)}) \right\} \quad (1)$$

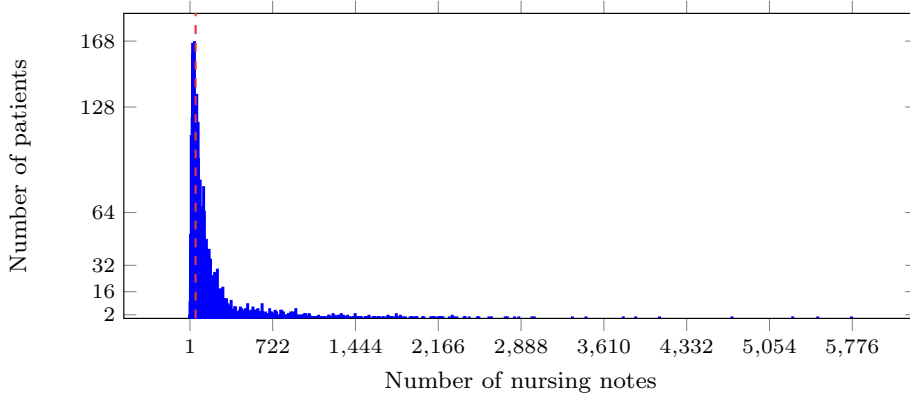


Figure 3: The distribution of nursing notes across various MIMIC-III subjects (red dashed line exhibits the distribution at 50 nursing notes).

where the Jaro similarity score of two given clinical terms (tokens) \mathcal{C}_i of length $|\mathcal{C}_i|$ and \mathcal{C}_j of length $|\mathcal{C}_j|$ with m matching characters and t transpositions is,

$$\text{Jaro}(\mathcal{C}_i, \mathcal{C}_j) = \begin{cases} 0, & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|\mathcal{C}_i|} + \frac{m}{|\mathcal{C}_j|} + \frac{2m-t}{2m} \right), & \text{otherwise} \end{cases} \quad (2)$$

The nursing notes of a patient are processed in the order of oldest to the most recent. Based on the predetermined similarity threshold (θ) ranging between 0 and 1, a pair of nursing notes ($\eta_i^{(k)}$, $\eta_j^{(k)}$) corresponding to a patient ($\mathcal{P}^{(k)}$) are merged only if $\text{ME}_{\text{Jaro}}(\eta_i^{(k)}, \eta_j^{(k)})$ is less than θ , else $\eta_j^{(k)}$ is retained and $\eta_i^{(k)}$ is purged, thus maintaining only the latest of the two nursing notes. Note that, similarity merging and purging applies only to nursing notes and not to the ICD-9 code groups. Corresponding ICD-9 codes across various nursing notes of a patient are merged to enable multi-label classification. The resultant nursing note for a patient $\mathcal{P}^{(k)}$ after merging is hereafter referred to as the *aggregate nursing note* of that patient. For the purpose of this research, we have empirically determined the fuzzy-similarity θ to be 0.825 using grid search.

Consider two sample nursing notes ($\eta_i^{(p)}$ and $\eta_j^{(p)}$) of a patient (p) extracted from the MIMIC-III database, recorded at times T (shown in Figure 4a) and $T' > T$ (shown in Figure 4b) respectively. It can be observed that both the recorded nursing notes are quite similar—the nursing note recorded at time T' records all the details in nursing note $\eta_i^{(p)}$, along with additional ‘response’ concerning the patient’s state. To handle the voluminosity of the nursing notes and delete the near-duplicate nursing notes, we compute the ME similarity (with internal Jaro similarity scoring) score using Equation 1. The nursing notes shown in Figure 4 have an ME similarity score of 0.85, which is higher than the preset threshold

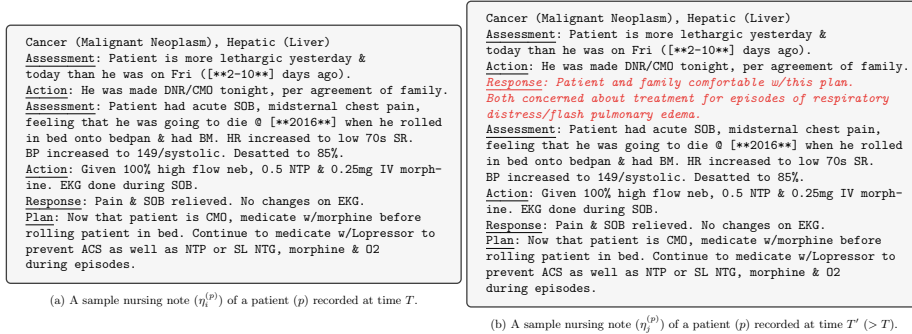


Figure 4: Two sample de-identified nursing notes from the MIMIC-III database. The two nursing notes are quite similar, while the only new content is the updated response (indicated as red italicized text).

355 of 0.825. Thus, note $\eta_j^{(p)}$ is retained, and note $\eta_i^{(p)}$ is purged.

356 3.3.2. Preprocessing

357 The next phase in the NLP pipeline is to preprocess the nursing notes to achieve
 358 data (text) normalization. Transformation of text into a canonical form allows
 359 for the separation of concerns and helps maintain consistency. Preprocessing
 360 essentially includes tokenization, stopword removal, and stemming/lemmatiza-
 361 tion. First, multiple spaces, special characters, and punctuation marks are re-
 362 moved. During tokenization, the clinical notes' text is split into several smaller
 363 tokens (words). Stopwords from the generated tokens are removed using the
 364 NLTK English stopwords corpus [5]. Furthermore, character case folding is per-
 365 formed, and references to images (file names such as '*scanImage.png*') are re-
 366 moved. It is to be noted that, token-length based token removal was not per-
 367 formed to avoid the loss of important medical information (such as '*CT*' in
 368 '*CT Scan*'). Finally, stemming was performed for suffix stripping, followed by
 369 lemmatization to convert the stripped tokens to their base forms. To eliminate
 370 overfitting and lower the computational complexity, the tokens appearing in less
 371 than ten nursing notes were removed before any further processing.

372 3.4. Feature Extraction

373 Let \mathcal{P} be the set of all patients. A patient ($\mathcal{P}^{(k)} \in \mathcal{P}$) has a sequence of nursing
 374 notes, $\mathbb{S}^{(k)} = \{\eta_i^{(k)}\}_{i=1}^{N^{(k)}}$, with $N^{(k)}$ total nursing notes ($\eta_i^{(k)}$ s).

Each nursing note constitutes a variable length of tokens from a sizeable vocab-
 ulary \mathbb{V} , and each patient has a variable number of such notes, thus making $\mathbb{S}^{(k)}$
 very complex. Thus, the transformation (T) of unstructured clinical text ($\mathbb{S}^{(k)}$)
 into an easier-to-use form (such as fixed length vector of tokens) is critically
 important. Thus, an effective mapping from the \mathbb{S} space to \mathbb{R} is attempted.

$$T : \mathbb{S}^{(k)} \longrightarrow \mathbb{R}^d \quad (3)$$

375 The patient information is transformed into a machine processable form, $\mathcal{P}^{(k)} =$
 376 $T(\mathbb{S}^{(k)})$, $\mathcal{P}^{(k)} \in \mathbb{R}^d$. To tackle the curse of dimensionality [4], usually $d \ll |\mathbb{V}|$.
 377 Although traditional dictionary and rule-based NLP transformations show good
 378 performance in certain applications, they are not automated and need manual
 379 effort to adapt them in various domains [48]. To improve the performance and
 380 effectiveness of the classification models, optimized vector representations of the
 381 underlying corpus is mandatory. To enable an exhaustive comparative study,
 382 we use six data modeling approaches as described below.

383 3.4.1. Vector Space Modeling of Aggregated Clinical Notes

A prominent transformation of the Bag of Words (BoW) that weighs each token in an unsupervised way, is the term weighting scheme. It is a numerical statistic that captures both the importance and specificity of a term in the given vocabulary. The weight ($W_m^{(i)}$) of a term $w_m^{(i)}$ (of total $|w^{(i)}|$ terms) in a nursing note η_i (of total N nursing notes) occurring $f_m^{(i)}$ times is given by,

$$W_m^{(i)} = \begin{cases} \left(1 + \log_2 f_m^{(i)}\right) \left(\log_2 \frac{N}{|w^{(i)}|}\right), & \text{if } f_m^{(i)} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

384 The weight of every term in a patient’s aggregate nursing note ($\mathcal{P}^{(k)}$) is com-
 385 puted to obtain a vector $\mathcal{V}^{(k)} \in \mathbb{R}^{|\mathbb{V}|}$. Now, the patient information in machine
 386 processable form, $\mathcal{P}_{\text{term_weighting}}^{(k)} = \mathcal{V}^{(k)}$.

387 Due to the one-hot encoding of every word in BoW models, the resulting mod-
 388 els suffer from high dimensionality and sparsity. Moreover, BoW models do
 389 not capture the intuition of semantically similar nursing notes having similar
 390 representations. For example, two terms with a close semantic relationship (as
 391 in ‘Cancer’ and ‘Melanoma’) could be mapped to two entries with large dis-
 392 tance. Vector space embeddings cope with these shortcomings by efficiently
 393 learning the term representations in a data-driven manner. An influential work
 394 in this domain is the Doc2Vec or Paragraph Vector (PV) network. Doc2Vec
 395 aims at numerically representing variable length documents as fixed length low
 396 dimensional document embeddings (vectors). Doc2Vec is essentially a neural
 397 network with one shallow hidden layer that learns the distributed representa-
 398 tions, to provide a content-related measurement. It incorporates semantic tex-
 399 tual features obtained from the nursing notes text corpus. The PV Distributed
 400 Memory (PV-DM) variant of Doc2Vec was chosen over PV Distributed BoW
 401 (PV-DBoW) due to its ability to preserve the word order in the nursing notes
 402 and its comparatively superior performance [53]. The implementations in the
 403 Python Scikit-learn [65] and Gensim packages [75] were used to extract term
 404 weighting and Doc2Vec style textual features on the transcribed clinical words
 405 (extracted from aggregate nursing notes). For an exhaustive analysis, Doc2Vec
 406 dimension sizes of 500 (trained for 25 epochs) and 1,000 (trained for 50 epochs)
 407 were used.

408 3.4.2. Topic Modeling of Aggregated Clinical Notes

409 Topic modeling can be used for finding a set of terms (topics) from a collection
 410 of documents that best represents the documents in the corpus. Traditional
 411 models of information retrieval such as Latent Semantic Analysis (LSA) [93]

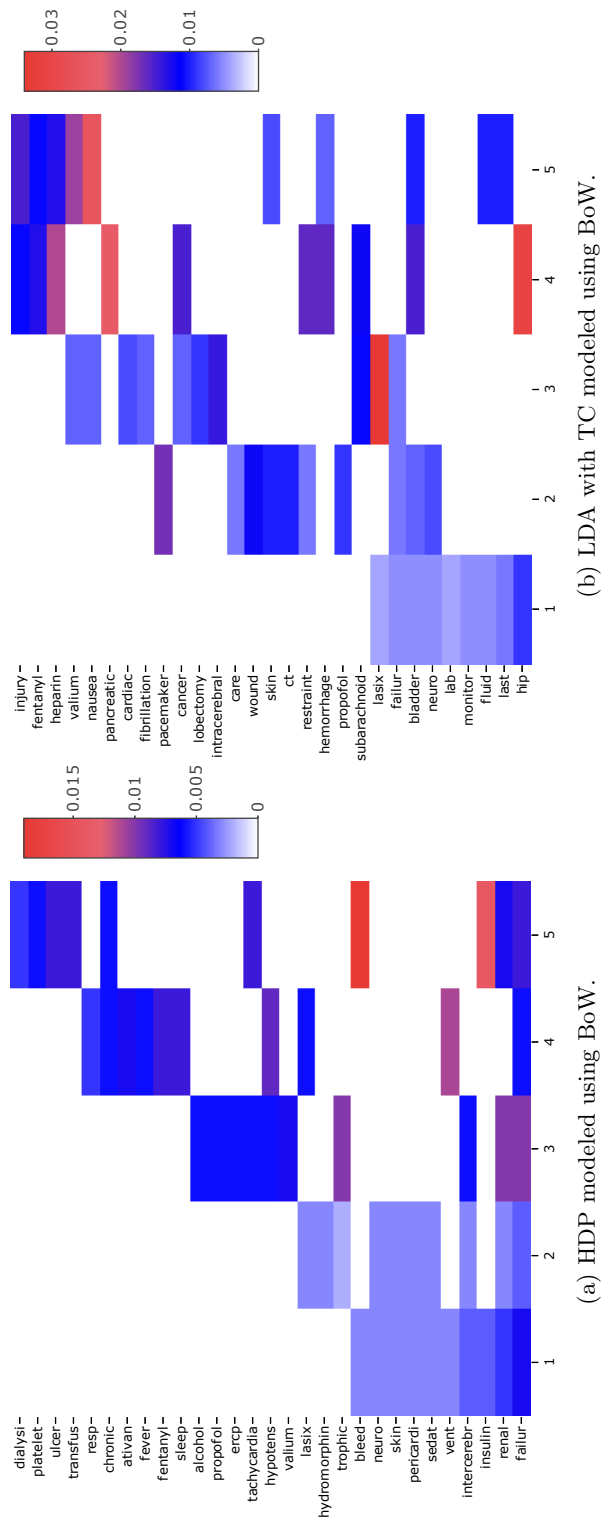
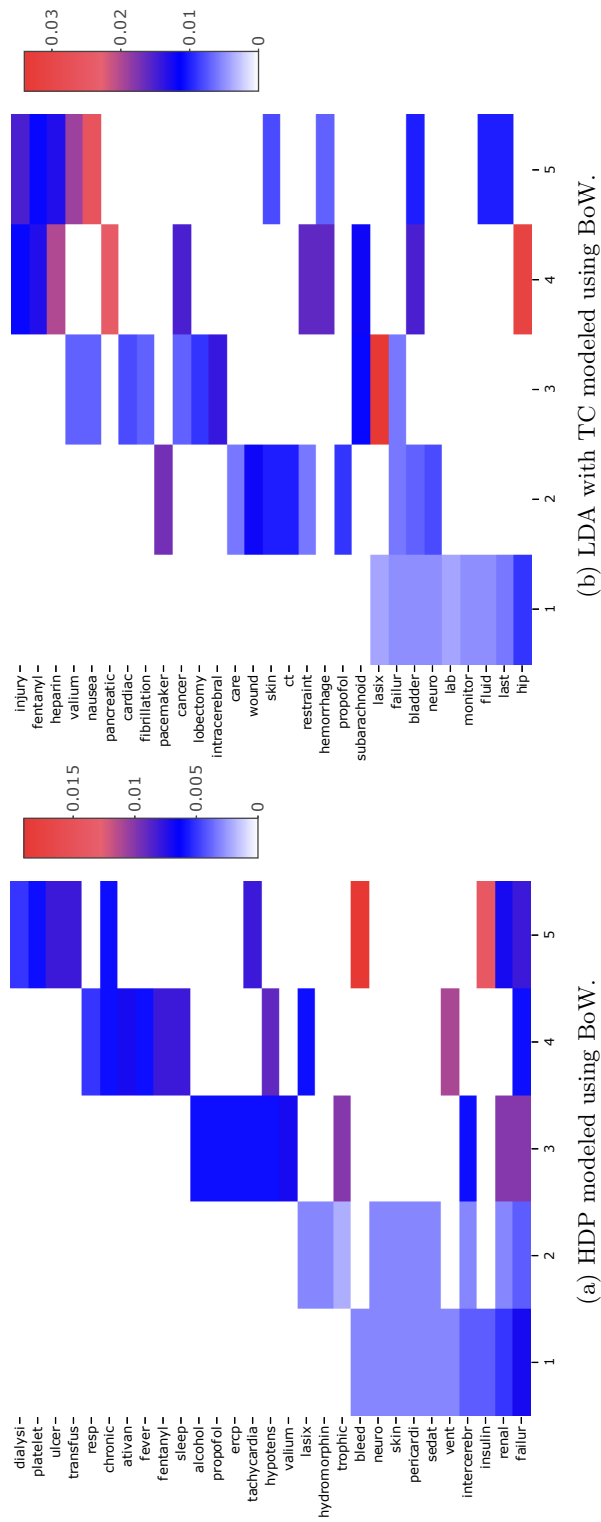


Figure 5: Correlations between top ten terms’ membership in $d = 5$ topic modeling clusters obtained using aggregated nursing notes (using fuzzy similarity $\theta = 0.825$).

412 use a low approximation of BoW/term weight matrix by calculating the singular
 413 value decomposition of the matrix. Such models usually deal with complex
 414 matrix computations. A variant of the LSA is the probabilistic LSA [39] that
 415 combines co-existing and implicit topic data into probabilistic statistics to find
 416 potential relationships among terms.

A popular cluster analysis approach, LDA is a generative topic model based on the Bayesian framework of a three-layer structure (documents, topics, and terms). LDA generates a soft probabilistic and flat clustering of terms into topics and documents into topics. LDA posits that each (aggregate) nursing note $\eta_i^{(k)}$ of a patient $\mathcal{P}^{(k)}$ and each term belongs to a set of d ($\ll |\mathbb{V}|$) clusters (topics) \mathcal{T} , with some probability ρ . Thus, each nursing note is transformed as,

$$\eta_i^{(k)} \longrightarrow \mathcal{T}_i^{(k)} \in \left[\rho_{ij}^{(k)} \right]_{j=1}^d \text{ where } \sum_{j=1}^d \rho_{ij}^{(k)} = 1 \text{ and } \rho_{ij}^{(k)} \geq 0 \ \forall j \quad (5)$$

417 Similar to other clustering approaches, there is no simple way to determine
 418 the correct number of d LDA clusters. To cope with this issue, more complex
 419 models such as Hierarchical Bayesian Non-parametric (HDP) which automati-
 420 cally determine the number of clusters through posterior inference can be used.
 421 HDP is a hierarchical Bayesian non-parametric model that can model mixed-
 422 membership data with potentially infinite terms, in an unsupervised way. In
 423 LDA, only the mixture of topics is drawn from the Dirichlet distribution, while
 424 in HDP, a Dirichlet process is used to capture the uncertainty in the number of
 425 terms. For the ease of interpretation, the top ten terms' membership with five
 426 HDP clusters is shown in Figure 5a.

427 Probabilistic models are commonly evaluated by measuring the log-likelihood of
 428 unseen documents. As an alternative to HDP, the methods of average similarity,
 429 perplexity [90], and TC between topics can also be used to derive the optimal
 430 number of topics. Perplexity measures the quality and generalization ability of
 431 the model. However, perplexity may not always correlate with human judgment
 432 and some times the two are anti-correlated [13]. TC is a way to evaluate topic
 433 models with a much greater guarantee of human interpretability. In this paper,
 434 we adopt LDA with TC as it accounts for the semantic similarity between high
 435 scoring terms. C_v , a variant of coherence measurement is used in this study, as
 436 it accounts for high correlation with all the available human ranking data [77].
 437 First, C_v segments each of the topic's top K tokens into token pairs. Then,
 438 it incorporates a Boolean sliding window approach in which for every window
 439 of size s sliding at one token per step, a virtual document is created. Token
 440 or token pair probabilities are computed from the total count of virtual docu-
 441 ments. To some degree, the sliding window approach captures the proximity

between tokens. Then, a confirmation (similarity) measure is used to quantify how strongly a token set supports another token set. Normalized point-wise mutual information [7] is used in this paper as a confirmation measure due to its high correlation with human interpretability. All the confirmation measures are averaged to obtain the final coherence score. The higher the coherence value, the stronger is the model’s human interpretability and generalization ability. For the ease of interpretation, the top ten terms’ membership with five LDA (with TC) clusters is shown in Figure 5b.

The implementations available in the Python Gensim package were used to implement LDA with TC and HDP models. To provide exhaustive analysis, HDP with truncation level set to 150 was modeled with both BoW and term weighting. Alternatively, LDA (set to 100 topics) with TC was modeled with BoW representations. Furthermore, the number of LDA topics was determined by comparing the TC scores of several LDA models obtained by varying the number of LDA topics from 2 to 500 in the increments of 100.

4. ICD-9 Code Group Prediction

ICD-9 codes are a taxonomy of diagnostic codes that are used by doctors, public health agencies, and health insurance companies across the world to classify diseases and a wide variety of infections, disorders, symptoms, causes of injury, and others. Owing to the high granularity of ICD-9 codes, researchers suggested differentiating between category-level (group) predictions and full-code predictions [51]. Each ICD-9 code group includes a set of similar diseases, and almost every health condition can be represented with a unique ICD-9 code group. In this study, we focus on ICD-9 code group predictions as a multi-label classification problem, with each patient’s nursing note mapped to more than one group. All the ICD-9 codes assigned to a patient’s admission are grouped into 19 diagnosis classes³. In this study, the Ref and V codes are classified into the same code group to lower the computational cost of training. Table 3 presents the statistics of ICD-9 code group labels extracted from MIMIC-III nursing notes.

4.1. ICD-9 Disease Code Group Prediction

In this section, we discuss the prediction algorithms employed to achieve the task of ICD-9 code group multi-label classification. We experimented with eight different prediction models conforming to various algorithmic classes including

³http://tdrdata.com/ipd/ipd_SearchForICD9CodesAndDescriptions.aspx.

Table 3: Statistics of the ICD-9 code group labels extracted from MIMIC-III nursing notes.

ICD-9 group	ICD-9 code range	Diagnosis	#Patients (out of 6,532)
1	001 – 139	Parasitic and infectious diseases	1,856
2	140 – 239	Neoplasms	1,319
3	240 – 279	Endocrine, immunity, metabolic, and nutritional	4,785
4	280 – 289	Blood-forming organs and blood	2,705
5	290 – 319	Mental disorders	2,614
6	320 – 389	Sense organs and nervous system	2,611
7	390 – 459	Circulatory system	5,393
8	460 – 519	Respiratory system	3,301
9	520 – 579	Digestive system	2,903
10	580 – 629	Genitourinary system	2,912
11	630 – 677	Childbirth, pregnancy, and puerperium	31
12	680 – 709	Subcutaneous tissue and skin	781
13	710 – 739	Connective tissue and musculoskeletal system	1,637
14	740 – 759	Congenital anomalies	269
15	780 – 789	Symptoms	2,432
16	790 – 796	Nonspecific abnormal findings	647
17	797 – 799	Unknown or ill-defined causes of mortality and morbidity	299
18	800 – 999	Poisoning and injury	2,978
19	Ref and V codes	Reference codes and supplemental V codes	4,853

algorithm adaptation based, problem transformation based, and ensemble models. The implementations available in the Python Scikit-learn package were used to make predictions.

4.1.1. Algorithm Adaptation Classification Models

The models in this class adapt existing machine learning algorithms for the task of multi-label classification. We used two models including K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP), for the prediction of ICD-9 code groups. KNN [99] is a non-parametric instance-based (non-generalizing) lazy learner used in regression and classification tasks. In KNN classification, the output class membership is determined by the majority vote of its K closest neighbors. In the sense of multi-label classification, KNN first identifies the K

closest neighbors and then, based on the statistical inferences gained from the neighboring class label sets, maximum a posteriori principle is used to determine the class label set of an unseen instance. Let $\mathbb{S} = \{\eta^{(i)}\}_{i=1}^{|\mathcal{P}|}$ be the set of all aggregate notes of $|\mathcal{P}|$ patients, and \mathbb{Y} denote the set of all possible class labels. Each nursing note $\eta^{(i)}$ is mapped to a class label set $\mathcal{Y}^{(i)} \subseteq \mathbb{Y}$. For an unseen instance $\eta^{(m)}$, let $K(m)$ denote the K closest neighbors. Membership counting function for c^{th} class label ($c \in \mathbb{Y}$), based on K -closest neighbors can be computed as,

$$\text{Count}_m(c) = \sum_{n=1}^{K(m)} \mathcal{Y}^{(n)}(c), \text{ where } \mathcal{Y}^{(n)}(c) = \begin{cases} 1, & \text{if } c \in \mathcal{Y}^{(n)} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Let $E(\text{Count}_m(c))$ denote the event ($E(\cdot)$) that $\text{Count}_m(c)$ neighbors of $\eta^{(m)}$ belong to the c^{th} class. Then, using the maximum a posteriori principle, we obtain the membership of a class label (c) as,

$$\mathcal{Y}^{(m)}(c) = \arg \max_{s \in \{0,1\}} \mathbf{P}(H_s^{(c)} | E(\text{Count}_m(c))), \quad H_s^{(c)} = \begin{cases} E(c \in \mathcal{Y}^{(m)}), & \text{if } s = 1 \\ E(c \notin \mathcal{Y}^{(m)}), & \text{otherwise} \end{cases} \quad (7)$$

Thus, finding all class membership values will help in obtaining the multi-label classification of an unseen nursing note. In our work, 15 closest neighbors were considered (empirically determined using grid search), where closeness is weighted as the inverse of the distance between instances.

MLP (vanilla neural network) [98] is a feed-forward neural artificial network with an input layer, one or more hidden layers, and one prediction layer at the top, for classification. The first layer takes $\eta^{(m)}$ with p' clinical terms as the input and uses the output of each layer as the input to the following layer. The transformation from a layer l with the output $\mathcal{O}^{(l)}$ to the following layer with weights $W^{(l+1)}$ and biases $b^{(l+1)}$ can be represented as,

$$\mathcal{O}^{(l)} \longrightarrow W^{(l+1)}\mathcal{O}^{(l)} + b^{(l+1)} \longrightarrow \mathbf{g}(W^{(l+1)}\mathcal{O}^{(l)} + b^{(l+1)}) \longrightarrow \mathcal{O}^{(l+1)} \quad (8)$$

where \mathbf{g} is a non-linear activation function such as a tanh, logistic sigmoid, or ReLU [62]. In training, to update the weights and biases, MLP uses a supervised approach called Backpropagation (BP) [78]. BP is used to calculate the gradient of the loss function to update weights, which aids the MLP to learn the internal representations, allowing it to learn any arbitrary mappings within the network. In the case of multi-label classification, while the forward pass remains the same, the classical BP algorithm uses a global error function that addresses the dependencies between the class labels. Figure 6 shows a one hidden layer

491 feed-forward MLP network for multi-label classification. In this study, we use
 492 vanilla neural networks with one hidden layer of 75 nodes and a ReLU activation
 493 function, empirically determined using grid search.

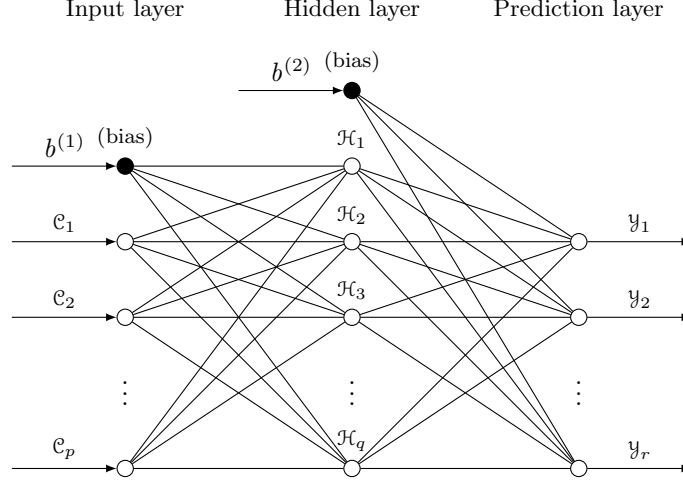


Figure 6: Multi-label classification neural network model with p input clinical terms (\mathcal{C}_i s), a hidden layer with q nodes (\mathcal{H}_i s), and r possible ICD-9 code groups (\mathcal{Y}_i s).

4.1.2. Problem Transformation Classification Models

These classification models aim at transforming an existing multi-label task into one or more single-label regression or classification tasks. Three classifiers including KNN, LR, and SVM were utilized as OvR classifiers in the prediction of ICD-9 diagnosis code groups. LR or maximum-entropy classification [23] is a discriminative model that models the probabilities of possible outcomes using a logistic function. The model posits that,

$$\mathbf{P}(\mathcal{Y}^{(i)}|\rho^{(i)}) = \rho^{(i)\mathcal{Y}^{(i)}} (1 - \rho^{(i)})^{1-\mathcal{Y}^{(i)}}, \text{ where } \rho^{(i)} = \frac{1}{1 + \exp(-x_i\beta)} \quad (9)$$

495 where $\mathcal{Y}^{(i)}$ is a single outcome variable corresponding to x_i and following a
 496 Bernoulli probability distribution, that draws a value of 1 with ρ_i probability.
 497 The unknown parameter $\beta = (\beta_0, \beta_1)'$ is an $(m \times 1)$ vector, where β_0 is the
 498 scalar intercept (constant term), and β_1 is an $(m - 1 \times 1)$ vector with elements
 499 corresponding to $m - 1$ explanatory variables of x_i . To achieve fast convergence
 500 to the optimal solution, we used the stochastic average gradient solver.

SVM [88] is also a discriminative approach that classifies by constructing hyperplane(s) in a high-dimensional space. For a given set of linear separable

training instances, SVM finds a linear rule that maximizes (optimizes) the geometric margin (street width). In practice, most of the training sets are not usually linearly separable. Now, a trade-off between minimizing prediction error and maximizing the geometric margin must be incorporated. Kernels such as tanh, sigmoid, Radial Basis Function (RBF) [64], and others are generally used to transform from the linearly inseparable space to a higher dimensional space where the points could be separated. The RBF kernel on two samples $\eta^{(i)}$ and $\eta^{(j)}$ can be defined as,

$$\mathbb{K}_{\text{RBF}}(\eta^{(i)}, \eta^{(j)}) = \exp(-\gamma \|\eta^{(i)} - \eta^{(j)}\|^2) \quad (10)$$

501 where γ measures the spread of the kernel. The RBF kernel defines a space
 502 that is larger than linear or polynomial kernels and has properties such as being
 503 stationary, isotropic, and infinitely smooth. Thus, in this analysis, we used SVM
 504 with an RBF kernel with γ set to $1/\text{\#features}$.

505 OvR [76] prediction strategy essentially transforms the multi-label classification
 506 problem into multiple binary relevance tasks. OvR trains a classifier for each
 507 class ($c \in \mathbb{Y}$), with the samples (aggregate nursing notes, $(\eta^{(i)}, \mathcal{Y}^{(i)})$) of that
 508 class as positive ($c \in \mathcal{Y}^{(i)}$) and the remaining samples as negative ($c \notin \mathcal{Y}^{(i)}$). The
 509 base classifiers produce a real-valued confidence score for the prediction decision.
 510 Then, for an unseen instance, the combined model predicts all the class labels
 511 for which the corresponding base classifiers predicted a positive result.

512 4.1.3. Ensemble Classification Models

Ensemble learning approaches help in the improvement of the prediction performance by combining several learning models. Three ensemble prediction approaches including Random Forest (RF), Hard-voting Ensemble (HVE), and Stacking Ensemble (SE) were employed in the classification of ICD-9 diagnostic code groups. RF or decision tree ensembles [8] predict by constructing multiple Classification And Regression Trees (CARTs) during training and predict the output class as a function of the outputs of individual trees for the test data. At each node of the CART, a random subset of input parameters (usually of size $\sqrt{\text{\#features}}$) are chosen, and the best feature is selected based on the splitting condition. The splitting conditions are based on the threshold which is determined by optimizing a cost function (such as information gain or Gini index). In multi-label classification, multiple labels are present in the tree leaves, and the entropy is calculated as the sum of entropies of each label,

$$\text{Entropy} = - \sum_{c \in \mathbb{Y}} \rho_c \log_2(\rho_c) + (1 - \rho_c) \log_2(1 - \rho_c) \quad (11)$$

513 where ρ_c is the probability of class c (\in the set of possible labels (\mathbb{Y})). The
 514 predictions of multiple base CARTs are combined using a simple voting scheme
 515 (such as probability distribution or majority vote). In this research, we use RF
 516 with 100 CARTs of maximum depth 2, and bagging was used to obtain diversity
 517 among the base CARTs.

HVE aggregates the predictions of multiple diverse classifiers using a majority rule. Given a set of diverse classifiers (N_i s) with prediction sets \mathcal{Y}_i s, where each \mathcal{Y}_i a subset of \mathbb{Y} (set of all class labels), then the presence of a class (c) in an unseen instance ($\eta^{(m)}$) can be estimated as,

$$\mathcal{Y}^{(m)}(c) = \begin{cases} 1, & \text{if } \sum_{i=1}^N \mathcal{Y}_i^{(m)}(c) > \lceil \frac{N}{2} \rceil \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

518 Thus, using the majority voting principle, the possible class label set for the
 519 unseen instance can be predicted. Many variations on the classifiers used in
 520 HVE were tried, starting with KNN, MLP, LR, LR as OvR, SVM as OvR,
 521 and KNN as OvR. After much experimentation, only MLP, LR as OvR, and
 522 SVM as OvR were used, due to their superior performance. Additionally, the
 523 plurality voting scheme was also tested; however, the majority voting scheme
 524 outperformed the plurality voting scheme. In this paper, we only present the
 525 performance recorded using the majority voting scheme.

526 SE [95] also combines discrete learning algorithms using a meta-classifier. In the
 527 first phase, all the base classifiers (N_i s) are applied to the training data which
 528 generate the predictions (\mathcal{Y}_i s). Then, in the second phase, a meta-level dataset
 529 is created by replacing every trained record ($\eta^{(k)}$) with the predictions for that
 530 record ($\mathcal{Y}_i^{(k)}\big)_{i=1}^N$. Then, another learning algorithm (L) is used to classify the
 531 meta-level dataset. On an unseen testing instance η_m , the predicted class set
 532 is $L(\mathcal{Y}_i^{(m)}\big)_{i=1}^N$. In this study, MLP, LR as OvR, and SVM as OvR are used as
 533 first-level classifiers, and MLP is used as the second-level classifier. In contrast
 534 to voting, SE learns at the meta-level, when combining multiple classifiers.

535 4.2. Experimental Validation and Discussion

536 To validate the proposed approach, we performed extensive experiments over
 537 the nursing notes data obtained from the MIMIC-III database. The primary
 538 challenge is the multi-label classification, where a set of ICD-9 code groups are
 539 predicted for a given nursing note. Let \mathbb{Y} denote the set of all possible labels,
 540 $\mathcal{Y}_{\text{true}}$ denote the ground truth class labels, $\mathcal{Y}_{\text{pred}}$ denote the predicted class
 541 labels, and $\mathcal{Y}_{\text{score}}$ denote the target scores which are either confidence values or

542 probability estimates of the true class or binary decisions ($\mathcal{Y}_{\text{pred}}$). In this work,
 543 binary predictions were used as the target scores, where, pairwise comparison
 544 of predicted values and true values is performed. Seven standard evaluation
 545 metrics were used to assess the performance of each prediction algorithm with
 546 reference to each data modeling approach.

Accuracy (ACC): This metric computes the average number of correct predictions over given samples. In the case of multi-label classification, the function uses a pairwise label matching to estimate the accuracy, as per Equation 13.

$$\text{ACC}(\mathcal{Y}_{\text{true}}, \mathcal{Y}_{\text{pred}}) = \frac{1}{s} \sum_{i=1}^s I(\mathcal{Y}_{\text{true}_i}, \mathcal{Y}_{\text{pred}_i}) \quad (13)$$

547 where s is the total number of samples, and $I(x, y)$ is the indicator function and
 548 returns one only when $x = y$.

549 *Area Under the ROC Curve (AUROC)*: The ROC curve is a graphical plot cre-
 550 ated by plotting sensitivity against the fall-out ($1 - \text{specificity}$). The AUROC
 551 metric [26] indicates the probability that a prediction model will rank a ran-
 552 domly chosen true instance higher than a randomly chosen false instance. A
 553 greater AUROC score indicates greater performance.

554 *Area Under the Precision-Recall Curve (AUPRC)*: The PR curve is a graphical
 555 plot created by plotting precision against the recall. When dealing with highly
 556 skewed datasets, the AUPRC [26] metric provides a more informative insight
 557 into the performance of the prediction algorithm. Higher the AUPRC, the better
 558 is the model's performance.

559 *MCC Score*: The Matthews correlation coefficient (ϕ -coefficient) [59] presents
 560 the essence of the correlation between the observed and the predicted binary
 561 classifications. It is a balanced score that takes into account the true/false
 562 positives and negatives. The higher the MCC score, the better the prediction
 563 is (Range = $[-1, 1]$).

F1 Score: Balanced F-measure or F1-score [82] is an indicator of the prediction accuracy, interpreted as a weighted average of precision and recall. F1 score reaches a perfect recall and precision at 1 (Range = $[0, 1]$) and is computed as,

$$F_{\beta} = (1 + \beta^2) \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \beta^2 \cdot \text{Precision}}, \text{ where } \beta = 1 \quad (14)$$

Coverage Error (CE): This metric [86] evaluates the average number of labels to be included in order to cover all the true labels of the instance. It can be related to precision at the level of perfect recall, and the lesser the value of CE, the better the performance. CE is calculated as,

$$CE(\mathcal{Y}_{\text{true}}, \mathcal{Y}_{\text{score}}) = \frac{1}{s} \sum_{i=1}^s \max_{j: \mathcal{Y}_{\text{true}ij}=1} rank_{ij} \quad (15)$$

564 where s is the total number of samples, and $rank_{ij} = |\{k : \mathcal{Y}_{\text{score}ik} \geq \mathcal{Y}_{\text{true}ij}\}|$
 565 ($|\cdot|$ is the cardinality of the set).

Label Ranking Loss (LRL): LRL [86] computes the average number of label pairs that are incorrectly ordered. The lower the LRL, the better the performance (Min = 0). LRL can be computed as,

$$LRL(\mathcal{Y}_{\text{true}}, \mathcal{Y}_{\text{score}}) = \frac{1}{s} \sum_{i=1}^s \frac{|\{(j, k) : \mathcal{Y}_{\text{true}ij} = 1, \mathcal{Y}_{\text{true}ik} = 0, \mathcal{Y}_{\text{score}ik} \geq \mathcal{Y}_{\text{score}ij}\}|}{\|\mathcal{Y}_{\text{true}i}\|_0 (\|\mathcal{Y} - \mathcal{Y}_{\text{true}i}\|_0)} \quad (16)$$

566 where s is the total number of samples, $|\cdot|$ denotes the cardinality of the set,
 567 and $\|\cdot\|_0$ denotes the l_0 norm.

568 4.3. Experimental Results

569 In this section, we report an exhaustive comparative study of the performance
 570 of various data and modeling approaches on the nursing notes of the MIMIC-III
 571 database. For the prediction task of ICD-9 code group classification, 10-fold
 572 cross-validation was performed. Furthermore, the mean and standard errors
 573 (of the mean) of the performance scores are presented. Table 4 shows the
 574 performance of all data modeling approaches and all prediction models using

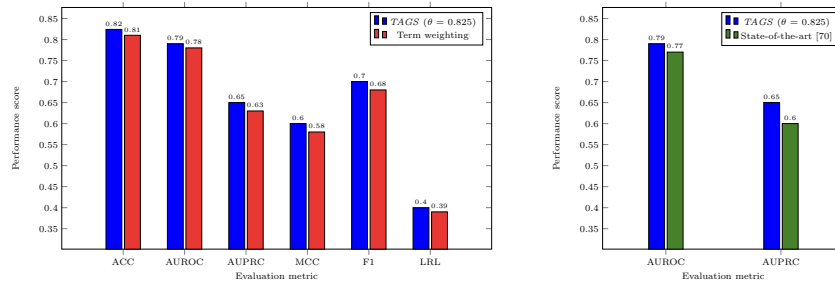


Figure 7: Comparative evaluation of the best performing models (with and without fuzzy similarity modeling) and the state-of-the-art model.

Table 4: ICD-9 code group prediction using nursing notes of MIMIC-III (using fuzzy similarity with $\theta = 0.825$).

Data model	Classifier	Performance scores						
		ACC	AUROC	AUPRC	MCC	F1	CE	LRL
TAGS (6,532 × 14,650)	KNN	0.7857 ± 0.0011	0.7681 ± 0.0010	0.5904 ± 0.0016	0.5286 ± 0.0019	0.6688 ± 0.0017	18.0936 ± 0.0501	0.4181 ± 0.0018
	MLP	0.7947 ± 0.0009	0.7677 ± 0.0013	0.5987 ± 0.0018	0.5366 ± 0.0020	0.6664 ± 0.0018	18.2327 ± 0.0574	0.4226 ± 0.0024
	KNN as OvR	0.7725 ± 0.0018	0.7645 ± 0.0011	0.5738 ± 0.0021	0.5108 ± 0.0024	0.6619 ± 0.0017	17.9385 ± 0.0791	0.4204 ± 0.0020
	LR as OvR	0.8239 ± 0.0011	0.7868 ± 0.0011	0.6476 ± 0.0011	0.5953 ± 0.0018	0.6981 ± 0.0016	18.2849 ± 0.0643	0.3978 ± 0.0021
	SVM as OvR	0.7413 ± 0.0014	0.6801 ± 0.0011	0.5249 ± 0.0014	0.4007 ± 0.0024	0.5207 ± 0.0019	19.5542 ± 0.0206	0.5880 ± 0.0018
	RF	0.7630 ± 0.0012	0.6926 ± 0.0009	0.5486 ± 0.0014	0.4388 ± 0.0022	0.5450 ± 0.0016	19.5678 ± 0.0238	0.5728 ± 0.0014
	HVE	0.8171 ± 0.0010	0.7781 ± 0.0007	0.6367 ± 0.0007	0.5786 ± 0.0007	0.6837 ± 0.0009	18.5659 ± 0.0614	0.4132 ± 0.0014
	SE	0.7972 ± 0.0009	0.7698 ± 0.0015	0.6027 ± 0.0021	0.5421 ± 0.0016	0.6701 ± 0.0017	18.2673 ± 0.0630	0.4195 ± 0.0029
Doc2Vec 500 (6,532 × 500)	KNN	0.7399 ± 0.0020	0.6628 ± 0.0027	0.5247 ± 0.0021	0.3949 ± 0.0041	0.4802 ± 0.0055	19.5644 ± 0.0278	0.6363 ± 0.0058
	MLP	0.7368 ± 0.0009	0.7102 ± 0.0012	0.5240 ± 0.0020	0.4150 ± 0.0023	0.5911 ± 0.0021	18.8039 ± 0.0450	0.5078 ± 0.0021
	KNN as OvR	0.7377 ± 0.0016	0.6674 ± 0.0024	0.5206 ± 0.0015	0.3888 ± 0.0030	0.4902 ± 0.0052	19.5144 ± 0.0269	0.6197 ± 0.0055
	SVM as OvR	0.8059 ± 0.0013	0.7666 ± 0.0010	0.6184 ± 0.0012	0.5514 ± 0.0022	0.6743 ± 0.0015	18.7379 ± 0.0462	0.4273 ± 0.0017
	RF	0.7484 ± 0.0013	0.6787 ± 0.0010	0.5356 ± 0.0010	0.4142 ± 0.0021	0.5190 ± 0.0018	19.6208 ± 0.0225	0.5991 ± 0.0019
	HVE	0.8013 ± 0.0014	0.7636 ± 0.0011	0.6084 ± 0.0016	0.5407 ± 0.0024	0.6691 ± 0.0012	18.6652 ± 0.0149	0.4312 ± 0.0015
	SE	0.8047 ± 0.0014	0.7652 ± 0.0011	0.6164 ± 0.0008	0.5482 ± 0.0023	0.6715 ± 0.0014	18.7367 ± 0.0483	0.4296 ± 0.0017
		0.7322 ± 0.0018	0.6543 ± 0.0030	0.5104 ± 0.0016	0.3741 ± 0.0036	0.4650 ± 0.0062	19.6614 ± 0.0478	0.6494 ± 0.0072
Doc2Vec 1,000 (6,532 × 1,000)	MLP	0.7458 ± 0.0011	0.7170 ± 0.0013	0.5307 ± 0.0011	0.4291 ± 0.0025	0.5089 ± 0.0015	18.8467 ± 0.0374	0.4988 ± 0.0021
	KNN as OvR	0.7376 ± 0.0017	0.6712 ± 0.0029	0.5189 ± 0.0013	0.3883 ± 0.0035	0.5020 ± 0.0057	19.5014 ± 0.0415	0.6074 ± 0.0068
	LR as OvR	0.7735 ± 0.0015	0.7414 ± 0.0017	0.5667 ± 0.0015	0.4845 ± 0.0030	0.6374 ± 0.0019	18.7376 ± 0.0526	0.4623 ± 0.0029
	SVM as OvR	0.8067 ± 0.0012	0.7693 ± 0.0013	0.6187 ± 0.0012	0.5542 ± 0.0021	0.6762 ± 0.0016	18.6286 ± 0.0472	0.4227 ± 0.0023
	RF	0.7464 ± 0.0012	0.6760 ± 0.0010	0.5334 ± 0.0014	0.4102 ± 0.0020	0.5136 ± 0.0018	19.6269 ± 0.0248	0.6045 ± 0.0020
	HVE	0.7904 ± 0.0015	0.7562 ± 0.0018	0.5922 ± 0.0017	0.5201 ± 0.0033	0.6566 ± 0.0022	18.6607 ± 0.0545	0.4413 ± 0.0033
	SE	0.8052 ± 0.0015	0.7680 ± 0.0013	0.6164 ± 0.0009	0.5510 ± 0.0025	0.6738 ± 0.0016	18.6683 ± 0.0402	0.4249 ± 0.0023
		0.7718 ± 0.0009	0.7422 ± 0.0009	0.5723 ± 0.0017	0.4892 ± 0.0018	0.6318 ± 0.0014	18.7632 ± 0.0514	0.4629 ± 0.0014
HDP with BoW (6,532 × 150)	MLP	0.7912 ± 0.0011	0.7557 ± 0.0012	0.5974 ± 0.0014	0.5255 ± 0.0019	0.6502 ± 0.0019	18.6689 ± 0.0330	0.4464 ± 0.0022
	KNN as OvR	0.7682 ± 0.0008	0.7397 ± 0.0010	0.5661 ± 0.0019	0.4822 ± 0.0018	0.6275 ± 0.0014	18.7482 ± 0.0380	0.4666 ± 0.0016
	LR as OvR	0.7815 ± 0.0010	0.7417 ± 0.0011	0.5850 ± 0.0014	0.5017 ± 0.0020	0.6251 ± 0.0016	19.9294 ± 0.0476	0.4729 ± 0.0020
	SVM as OvR	0.7511 ± 0.0011	0.6875 ± 0.0008	0.5410 ± 0.0015	0.4245 ± 0.0019	0.5284 ± 0.0017	19.4253 ± 0.0279	0.5827 ± 0.0015
	RF	0.7574 ± 0.0015	0.6915 ± 0.0014	0.5486 ± 0.0017	0.4359 ± 0.0028	0.5412 ± 0.0023	19.5291 ± 0.0314	0.5751 ± 0.0026
	HVE	0.7826 ± 0.0013	0.7404 ± 0.0015	0.5869 ± 0.0008	0.5029 ± 0.0022	0.6229 ± 0.0020	18.9688 ± 0.0626	0.4767 ± 0.0029
	SE	0.7851 ± 0.0008	0.7453 ± 0.0008	0.5874 ± 0.0014	0.5083 ± 0.0013	0.6317 ± 0.0006	18.7915 ± 0.0498	0.4660 ± 0.0014
		0.7116 ± 0.0015	0.6723 ± 0.0018	0.4887 ± 0.0023	0.3479 ± 0.0034	0.5254 ± 0.0031	19.3027 ± 0.0297	0.5724 ± 0.0028
HDP with term weighting (6,532 × 150)	MLP	0.7409 ± 0.0016	0.6779 ± 0.0027	0.5245 ± 0.0014	0.3997 ± 0.0028	0.5158 ± 0.0056	19.5698 ± 0.0277	0.5940 ± 0.0069
	KNN as OvR	0.7076 ± 0.0014	0.6689 ± 0.0018	0.4842 ± 0.0024	0.3399 ± 0.0034	0.5213 ± 0.0031	19.2999 ± 0.0269	0.5764 ± 0.0028
	LR as OvR	0.7458 ± 0.0014	0.6780 ± 0.0010	0.5310 ± 0.0019	0.4082 ± 0.0027	0.5161 ± 0.0017	19.5929 ± 0.0258	0.5987 ± 0.0019
	SVM as OvR	0.7413 ± 0.0014	0.6801 ± 0.0011	0.5249 ± 0.0014	0.4007 ± 0.0024	0.5207 ± 0.0019	19.5542 ± 0.0206	0.5880 ± 0.0018
	RF	0.7557 ± 0.0010	0.6880 ± 0.0008	0.5376 ± 0.0012	0.4257 ± 0.0017	0.5359 ± 0.0015	19.4695 ± 0.0268	0.5800 ± 0.0015
	HVE	0.7414 ± 0.0017	0.6801 ± 0.0013	0.5249 ± 0.0015	0.4007 ± 0.0029	0.5207 ± 0.0023	19.5542 ± 0.0083	0.5880 ± 0.0022
	SE	0.7414 ± 0.0017	0.6801 ± 0.0013	0.5249 ± 0.0015	0.4007 ± 0.0029	0.5207 ± 0.0023	19.5542 ± 0.0083	0.5880 ± 0.0022
		0.7883 ± 0.0016	0.7512 ± 0.0015	0.5939 ± 0.0014	0.5201 ± 0.0029	0.6440 ± 0.0021	18.7220 ± 0.0465	0.4554 ± 0.0025
LDA with TC (6,532 × 100)	MLP	0.8037 ± 0.0010	0.7657 ± 0.0014	0.6181 ± 0.0016	0.5544 ± 0.0021	0.6663 ± 0.0020	18.5933 ± 0.0463	0.4341 ± 0.0026
	KNN as OvR	0.7838 ± 0.0012	0.7479 ± 0.0010	0.5859 ± 0.0008	0.5098 ± 0.0017	0.6389 ± 0.0015	18.7532 ± 0.0544	0.4593 ± 0.0019
	LR as OvR	0.8018 ± 0.0010	0.7644 ± 0.0012	0.6157 ± 0.0014	0.5505 ± 0.0019	0.6624 ± 0.0017	18.6514 ± 0.0467	0.4361 ± 0.0023
	SVM as OvR	0.7773 ± 0.0014	0.7272 ± 0.0013	0.5852 ± 0.0016	0.4949 ± 0.0026	0.5999 ± 0.0021	19.1559 ± 0.0464	0.5087 ± 0.0025
	RF	0.7569 ± 0.0014	0.6945 ± 0.0011	0.5531 ± 0.0013	0.4415 ± 0.0023	0.5462 ± 0.0019	19.4421 ± 0.0404	0.5694 ± 0.0022
	HVE	0.8018 ± 0.0011	0.7633 ± 0.0011	0.6160 ± 0.0011	0.5498 ± 0.0017	0.6607 ± 0.0012	18.6970 ± 0.0587	0.4384 ± 0.0020
	SE	0.7983 ± 0.0012	0.7570 ± 0.0010	0.6096 ± 0.0012	0.5408 ± 0.0017	0.6504 ± 0.0013	18.7473 ± 0.0621	0.4513 ± 0.0017
		0.7883 ± 0.0016	0.7512 ± 0.0015	0.5939 ± 0.0014	0.5201 ± 0.0029	0.6440 ± 0.0021	18.7220 ± 0.0465	0.4554 ± 0.0025

575 nursing notes processed using fuzzy token-based similarity with $\theta = 0.825$. Ta-
576 ble 5 tabulates the performance of all data modeling approaches and all predic-
577 tion models using nursing notes processed without similarity. We observe that
578 the Term weighting of unstructured (nursing) notes AGgregated using fuzzy
579 Similarity (TAGS) model, modeled with LR as OvR, consistently outperforms
580 more complex vector space and topic models. Furthermore, it can be observed
581 from Figure 7 that, the model’s performance is higher when nursing notes are
582 processed with similarity, than when processed without similarity.

Table 5: ICD-9 code group prediction using nursing notes of MIMIC-III (without similarity modeling).

Data model	Classifier	Performance scores						
		ACC	AUROC	AUPRC	MCC	F1	CE	LRL
Term weighting (6,532 × 14,665)	KNN	0.7866 ± 0.0012	0.7689 ± 0.0016	0.5920 ± 0.0025	0.5306 ± 0.0032	0.6697 ± 0.0021	18.0463 ± 0.0691	0.4168 ± 0.0027
	MLP	0.7962 ± 0.0011	0.7694 ± 0.0015	0.6009 ± 0.0026	0.5400 ± 0.0029	0.6685 ± 0.0024	18.2134 ± 0.0530	0.4199 ± 0.0026
	KNN as OvR	0.7741 ± 0.0017	0.7662 ± 0.0014	0.5764 ± 0.0027	0.5144 ± 0.0032	0.6639 ± 0.0020	18.1744 ± 0.0644	0.4179 ± 0.0023
	LR as OvR	0.8143 ± 0.0014	0.7804 ± 0.0017	0.6378 ± 0.0032	0.5845 ± 0.0035	0.6874 ± 0.0030	18.2934 ± 0.0389	0.3985 ± 0.0030
	SVM as OvR	0.7414 ± 0.0015	0.6801 ± 0.0015	0.5249 ± 0.0026	0.4007 ± 0.0036	0.5207 ± 0.0028	19.5542 ± 0.0368	0.5880 ± 0.0024
	RF	0.7653 ± 0.0011	0.6951 ± 0.0013	0.5517 ± 0.0024	0.4449 ± 0.0031	0.5484 ± 0.0023	19.5449 ± 0.0387	0.5695 ± 0.0022
	HVE	0.8064 ± 0.0014	0.7782 ± 0.0014	0.6369 ± 0.0031	0.5788 ± 0.0032	0.6832 ± 0.0026	18.5193 ± 0.0489	0.4132 ± 0.0023
	SE	0.7971 ± 0.0013	0.7693 ± 0.0018	0.6017 ± 0.0032	0.5412 ± 0.0034	0.6682 ± 0.0029	18.2290 ± 0.0363	0.4207 ± 0.0030
Doc2Vec 500 (6,532 × 500)	KNN	0.7134 ± 0.0013	0.5986 ± 0.0021	0.4719 ± 0.0024	0.3111 ± 0.0040	0.3323 ± 0.0059	19.9011 ± 0.0208	0.7824 ± 0.0048
	MLP	0.7370 ± 0.0011	0.7081 ± 0.0017	0.5217 ± 0.0022	0.4113 ± 0.0029	0.5885 ± 0.0026	18.8870 ± 0.0421	0.5113 ± 0.0028
	KNN as OvR	0.7177 ± 0.0013	0.6091 ± 0.0020	0.4783 ± 0.0020	0.3167 ± 0.0035	0.3627 ± 0.0054	19.8782 ± 0.0171	0.7533 ± 0.0048
	LR as OvR	0.7970 ± 0.0007	0.7586 ± 0.0009	0.5999 ± 0.0020	0.5291 ± 0.0016	0.6659 ± 0.0016	18.6661 ± 0.0346	0.4382 ± 0.0017
	SVM as OvR	0.8068 ± 0.0010	0.7678 ± 0.0012	0.6206 ± 0.0024	0.5527 ± 0.0025	0.6774 ± 0.0018	18.7267 ± 0.0269	0.4245 ± 0.0021
	RF	0.7490 ± 0.0014	0.6801 ± 0.0016	0.5351 ± 0.0027	0.4142 ± 0.0037	0.5232 ± 0.0029	19.6314 ± 0.0357	0.5942 ± 0.0027
	HVE	0.8011 ± 0.0006	0.7627 ± 0.0008	0.6083 ± 0.0024	0.5387 ± 0.0013	0.6701 ± 0.0011	18.6705 ± 0.0216	0.4318 ± 0.0014
	SE	0.8054 ± 0.0009	0.7659 ± 0.0010	0.6179 ± 0.0028	0.5489 ± 0.0022	0.6740 ± 0.0018	18.7635 ± 0.0400	0.4279 ± 0.0018
Doc2Vec 1,000 (6,532 × 1,000)	KNN	0.7141 ± 0.0016	0.6058 ± 0.0026	0.4754 ± 0.0028	0.3192 ± 0.0045	0.3520 ± 0.0069	19.8945 ± 0.0179	0.7643 ± 0.0058
	MLP	0.7442 ± 0.0011	0.7159 ± 0.0017	0.5312 ± 0.0024	0.4270 ± 0.0030	0.5995 ± 0.0027	18.8172 ± 0.0321	0.4992 ± 0.0028
	KNN as OvR	0.7162 ± 0.0018	0.6112 ± 0.0034	0.4781 ± 0.0037	0.3219 ± 0.0058	0.3671 ± 0.0091	19.8661 ± 0.0200	0.7493 ± 0.0076
	LR as OvR	0.7749 ± 0.0005	0.7425 ± 0.0007	0.5698 ± 0.0018	0.4864 ± 0.0017	0.6418 ± 0.0015	18.7278 ± 0.0397	0.4592 ± 0.0010
	SVM as OvR	0.8071 ± 0.0009	0.7684 ± 0.0012	0.6194 ± 0.0027	0.5528 ± 0.0026	0.6768 ± 0.0022	18.6731 ± 0.0429	0.4239 ± 0.0020
	RF	0.7455 ± 0.0014	0.6760 ± 0.0014	0.5313 ± 0.0023	0.4077 ± 0.0032	0.5138 ± 0.0025	19.6283 ± 0.0375	0.6034 ± 0.0025
	HVE	0.7915 ± 0.0009	0.7559 ± 0.0014	0.5943 ± 0.0037	0.5200 ± 0.0035	0.6588 ± 0.0029	18.6419 ± 0.0225	0.4410 ± 0.0022
	SE	0.8061 ± 0.0011	0.7674 ± 0.0013	0.6179 ± 0.0035	0.5508 ± 0.0032	0.6750 ± 0.0025	18.6649 ± 0.0241	0.4256 ± 0.0022
HDP with BoW (6,532 × 150)	KNN	0.7778 ± 0.0011	0.7505 ± 0.0014	0.5792 ± 0.0024	0.5033 ± 0.0027	0.6407 ± 0.0019	18.5832 ± 0.0558	0.4502 ± 0.0024
	MLP	0.7946 ± 0.0013	0.7574 ± 0.0016	0.6026 ± 0.0031	0.5336 ± 0.0036	0.6518 ± 0.0028	18.6202 ± 0.0417	0.4467 ± 0.0028
	KNN as OvR	0.7733 ± 0.0013	0.7476 ± 0.0017	0.5726 ± 0.0030	0.4949 ± 0.0037	0.6367 ± 0.0026	18.5783 ± 0.0456	0.4536 ± 0.0027
	LR as OvR	0.7878 ± 0.0016	0.7453 ± 0.0020	0.5932 ± 0.0030	0.5183 ± 0.0042	0.6307 ± 0.0033	18.7679 ± 0.0444	0.4723 ± 0.0033
	SVM as OvR	0.7623 ± 0.0014	0.6926 ± 0.0017	0.5510 ± 0.0029	0.4450 ± 0.0038	0.5411 ± 0.0032	19.5415 ± 0.0398	0.5776 ± 0.0029
	RF	0.7619 ± 0.0015	0.6982 ± 0.0017	0.5535 ± 0.0029	0.4468 ± 0.0039	0.5563 ± 0.0030	19.5531 ± 0.0314	0.5606 ± 0.0030
	HVE	0.7886 ± 0.0011	0.7438 ± 0.0016	0.5941 ± 0.0027	0.5183 ± 0.0029	0.6286 ± 0.0024	18.8647 ± 0.0482	0.4759 ± 0.0031
	SE	0.7886 ± 0.0006	0.7431 ± 0.0011	0.5935 ± 0.0023	0.5172 ± 0.0017	0.6288 ± 0.0018	18.8853 ± 0.0417	0.4766 ± 0.0022
HDP with term weighting (6,532 × 150)	KNN	0.7108 ± 0.0010	0.6718 ± 0.0018	0.4885 ± 0.0025	0.3476 ± 0.0030	0.5262 ± 0.0026	19.3230 ± 0.0378	0.5728 ± 0.0027
	MLP	0.7413 ± 0.0014	0.6783 ± 0.0016	0.5253 ± 0.0029	0.4009 ± 0.0037	0.5167 ± 0.0033	19.5623 ± 0.0396	0.5934 ± 0.0046
	KNN as OvR	0.7067 ± 0.0012	0.6685 ± 0.0020	0.4837 ± 0.0028	0.3393 ± 0.0036	0.5221 ± 0.0029	19.3410 ± 0.0392	0.5767 ± 0.0030
	LR as OvR	0.7455 ± 0.0016	0.6779 ± 0.0016	0.5301 ± 0.0030	0.4072 ± 0.0041	0.5161 ± 0.0030	19.5868 ± 0.0369	0.5984 ± 0.0026
	SVM as OvR	0.7414 ± 0.0015	0.6801 ± 0.0015	0.5249 ± 0.0026	0.4007 ± 0.0036	0.5207 ± 0.0028	19.5542 ± 0.0368	0.5880 ± 0.0024
	RF	0.7559 ± 0.0012	0.6862 ± 0.0018	0.5386 ± 0.0030	0.4259 ± 0.0039	0.5313 ± 0.0033	19.4848 ± 0.0370	0.5854 ± 0.0030
	HVE	0.7444 ± 0.0023	0.6789 ± 0.0012	0.5286 ± 0.0038	0.4058 ± 0.0049	0.5179 ± 0.0023	19.5742 ± 0.0588	0.5948 ± 0.0031
	SE	0.7413 ± 0.0016	0.6800 ± 0.0010	0.5248 ± 0.0025	0.4007 ± 0.0031	0.5206 ± 0.0024	19.5566 ± 0.0507	0.5882 ± 0.0015
LDA with TC (6,532 × 100)	KNN	0.7872 ± 0.0011	0.7517 ± 0.0012	0.5937 ± 0.0023	0.5197 ± 0.0027	0.6449 ± 0.0024	18.7065 ± 0.0454	0.4539 ± 0.0020
	MLP	0.8039 ± 0.0011	0.7669 ± 0.0014	0.6182 ± 0.0025	0.5547 ± 0.0028	0.6681 ± 0.0023	18.5665 ± 0.0489	0.4311 ± 0.0025
	KNN as OvR	0.7824 ± 0.0008	0.7482 ± 0.0013	0.5851 ± 0.0022	0.5087 ± 0.0026	0.6392 ± 0.0021	18.7217 ± 0.0364	0.4581 ± 0.0021
	LR as OvR	0.8018 ± 0.0013	0.7639 ± 0.0014	0.6152 ± 0.0027	0.5497 ± 0.0033	0.6626 ± 0.0025	18.6916 ± 0.0466	0.4367 ± 0.0024
	SVM as OvR	0.7778 ± 0.0016	0.7297 ± 0.0015	0.5858 ± 0.0028	0.4961 ± 0.0036	0.6050 ± 0.0027	19.1415 ± 0.0275	0.5024 ± 0.0025
	RF	0.7587 ± 0.0015	0.6962 ± 0.0013	0.5527 ± 0.0027	0.4424 ± 0.0032	0.5487 ± 0.0024	19.4452 ± 0.0393	0.5655 ± 0.0022
	HVE	0.8009 ± 0.0009	0.7613 ± 0.0009	0.6141 ± 0.0022	0.5469 ± 0.0020	0.6584 ± 0.0018	18.7753 ± 0.0523	0.4423 ± 0.0019
	SE	0.7975 ± 0.0011	0.7566 ± 0.0013	0.6078 ± 0.0027	0.5388 ± 0.0023	0.6509 ± 0.0025	18.7774 ± 0.0599	0.4510 ± 0.0029

4.4. Discussion

In clinical tasks such as disease prediction, capturing true/false positives and true/false negatives is of utmost importance, due to the critical nature of the task itself. As can be seen from the results in Tables 4 and 5, the AUROC metric captures the hit and miss rates, while AUPRC captures the number of true positives from positive predictions. AUPRC, unlike AUROC, varies with the change in the ratio of target classes in the data, and hence is more revealing while evaluating imbalanced data [79]. From Table 3, it can be observed that the dataset is highly class imbalanced, and hence AUPRC is more informative than AUROC. It can be seen that our approach outperforms the existing state-

of-the-art method [70] in these metrics, indicating the significant decrease in the false positives and false negatives. F1-measure captures both precision and recall of the prediction, while MCC score serves as a balanced measure even with class imbalance, as it takes into account true positives, false positives, and false negatives. More specifically, in healthcare applications like disease or diagnosis prediction, false negatives (prediction miss, i.e., a disease which is present, but not diagnosed) are likely to cause more harm than false positives (false alarm) and CE captures these false negatives. LRL performs a pairwise label comparison to determine the loss of prediction. Existing works have benchmarked their performance using only AUROC and AUPRC metrics. Since all the metrics used in this research are very relevant and essential in understanding the proposed model’s predictive power, we benchmark these promising results for MIMIC-III database.

Furthermore, the state-of-the-art work by Purushotham et al. [70] is built on structured EHRs that are modeled in the form of feature sets to make clinical predictions. It is a fact that the richness and abundance of information captured by unstructured nursing notes are often lost in the structured EHRs coding process [29]. Our proposed *TAGS* model combines the fuzzy similarity based data cleansing and aggregating approach with a term weighting scheme that captures the importance and rarity of clinical concepts, to model the informally written clinical nursing text into a clinically relevant and usable format effectively. From the results, it can be seen that more complex data modeling approaches such as Doc2Vec and HDP, in contrast to the *TAGS* model, fail to capture all the discriminative features of the clinical nursing notes needed for the machine learning classifier to learn and generalize. We observe that using the *TAGS* model, risk stratification can be achieved well in advance, with an overall accuracy of 82.4%. Also, it can be noted that token-based similarity processing of nursing notes yields higher performance in comparison to that processed without similarity. These promising results emphasize the need for reduction in redundancy and anomalous data for relieving the cognitive burden and improving the clinical decision-making process. CDSSs built on the predictive capabilities of *TAGS* could be suitable for patient-centric and evidence-based treatments, resulting in reduced mortality rates and better risk assessment.

5. Concluding Remarks

In this paper, vector space and topic modeling approaches for multi-label classification of unstructured nursing notes were presented, which capture the semantic information in the nursing notes effectively and leverage such information for disease prediction. The nursing notes were aggregated using a fuzzy

token-based similarity matching approach, on which several classification models were built. Exhaustive benchmarking experimentation results on the nursing notes of the MIMIC-III database were presented. We demonstrated that fuzzy token-based similarity processing of nursing notes provides optimal data representation and eliminates anomalous and redundant data, in turn, improving the clinical decision-making process. Furthermore, we observed that the *TAGS* model consistently outperformed other complex vector space and topic modeling approaches by effectively capturing the discriminative features of the nursing notes. The *TAGS* model also achieved superior predictive performance when benchmarked against the state-of-the-art method with 7.79% improvement in terms of AUPRC and 1.24% improvement in terms of AUROC.

The improvement in prediction accuracy though small, is still significant, as our model utilizes unstructured clinical text, in contrast to the state-of-the-art model. Thus, the dependency on availability of structured EHRs for building CDSSs can be eliminated, which is advantageous in countries with low EHR adoption rates. The experimental results highlight the richness of information that our model was able to capture from the clinical nursing notes, highlighting the viability of using unstructured clinical data in disease prediction applications. As a part of future work, we intend to validate the proposed *TAGS* model on real-time clinical data and enhance the prediction capabilities further, focusing on the need for time-aware prediction architectures in hospital scenarios. Furthermore, we aim at exploring the power of deep learning architectures in clinical prediction tasks such as disease prediction, length of stay prediction, hospital readmission, and phenotype classification.

Acknowledgments

We gratefully acknowledge the use of facilities at the Department of Information Technology, NITK Surathkal, funded by the Government of India’s DST-SERB Early Career Research Grant (ECR/2017/001056) to Sowmya Kamath S. Any findings, opinions, and recommendations or conclusions expressed in this paper are those of the authors and do not reflect the views of the funding agencies.

References

- [1] Julia Adler-Milstein and Ashish K Jha. Hitech act drove large gains in hospital electronic health record adoption. *Health Affairs*, 36(8):1416–1422, 2017.

- [2] John Angiolillo, S. Trent Rosenbloom, Melissa McPheeters, G. Seibert Tregoning, Russell L. Rothman, and Colin G. Walsh. Maintaining automated measurement of choosing wisely adherence across the icd 9 to 10 transition. *Journal of Biomedical Informatics*, 93:103142, 2019. ISSN 1532-0464. doi: <https://doi.org/10.1016/j.jbi.2019.103142>. URL <http://www.sciencedirect.com/science/article/pii/S1532046419300607>.
- [3] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. Multi-label classification of patient notes a case study on icd code assignment. *arXiv preprint arXiv:1709.09587*, 2017.
- [4] Richard E Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015.
- [5] Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.
- [6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [7] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.
- [8] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [9] Timothy G Buchman, Kenneth L Kubos, Alexander J Seidler, and Michael J Siegforth. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit. *Critical care medicine*, 22(5):750–762, 1994.
- [10] Jacob Calvert, Qingqing Mao, Jana L Hoffman, Melissa Jay, Thomas Desautels, et al. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Annals of Medicine and Surgery*, 11:52–57, 2016.
- [11] Rich Caruana, Shumeet Baluja, and Tom Mitchell. Using the future to” sort out” the present: Rankprop and multitask learning for medical risk evaluation. In *Advances in neural information processing systems*, pages 959–965, 1996.
- [12] Leo Anthony Celi, Sean Galvin, Guido Davidzon, Joon Lee, Daniel Scott, and Roger Mark. A database-driven decision support system: customized mortality prediction. *Journal of personalized medicine*, 2(4):138–148, 2012.

- [13] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.
- [14] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM, 2015.
- [15] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv:1512.03542*, 2015.
- [16] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 371. American Medical Informatics Association, 2016.
- [17] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.
- [18] Elinor CG Chumney, Andrea K Biddle, Kit N Simpson, Morris Weinberger, Kathryn M Magruder, and William N Zelman. The effect of cost construction based on either drg or icd-9 codes or risk group stratification on the resulting cost-effectiveness ratios. *Pharmacoeconomics*, 22(18): 1209–1216, 2004.
- [19] Gilles Clermont, Derek C Angus, Stephen M DiRusso, Martin Griffin, and Walter T Linde-Zwirble. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Critical care medicine*, 29(2):291–296, 2001.
- [20] Sarah A Collins, Kenrick Cato, David Albers, Karen Scott, et al. Relationship between nursing documentation and patients’ mortality. *American Journal of Critical Care*, 22(4):306–313, 2013.
- [21] Gregory F Cooper, Constantin F Aliferis, Richard Ambrosino, John Aronis, Bruce G Buchanan, Richard Caruana, Michael J Fine, Clark Glymour, Geoffrey Gordon, Barbara H Hanusa, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial intelligence in medicine*, 9(2):107–138, 1997.

- 735 [22] Healthcare Cost, Utilization Project (HCUP), et al. Introduction to the
736 hcup national inpatient sample (nis) 2012. *Agency for Healthcare Research
737 and Quality, Rockville*, 2014.
- 738 [23] David R Cox. The regression analysis of binary sequences. *Journal of the
739 Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.
- 740 [24] Filip Dabek and Jesus J Caban. A neural network based model for pre-
741 dicting psychological conditions. In *International Conference on Brain
742 Informatics and Health*, pages 252–261. Springer, 2015.
- 743 [25] Darcy A Davis, Nitesh V Chawla, Nicholas Blumm, Nicholas Christakis,
744 and Albert-László Barabasi. Predicting individual disease risk based on
745 medical history. In *Proceedings of the 17th ACM conference on Informa-
746 tion and knowledge management*, pages 769–778. ACM, 2008.
- 747 [26] Jesse Davis and Mark Goadrich. The relationship between precision-recall
748 and roc curves. In *Proceedings of the 23rd international conference on
749 Machine learning*, pages 233–240. ACM, 2006.
- 750 [27] Sanjay V Desai, Tyler J Law, and Dale M Needham. Long-term compli-
751 cations of critical care. *Critical care medicine*, 39(2):371–379, 2011.
- 752 [28] Sebastien Dubois and Nathanael Romano. Learning effective embeddings
753 from medical notes. 2017.
- 754 [29] Sebastien Dubois, Nathanael Romano, David C Kale, Nigam Shah, and
755 Kenneth Jung. Learning effective representations from clinical notes.
756 *arXiv preprint arXiv:1705.07025*, 2017.
- 757 [30] National Center for Health Statistics et al. Icd-9-cm official guidelines for
758 coding and reporting, 2006.
- 759 [31] Jim Grigsby, Robert Kookan, and John Hershberger. Simulated neural
760 networks to predict outcomes, costs, and length of stay among orthopedic
761 rehabilitation patients. *Archives of physical medicine and rehabilitation*,
762 75(10):1077–1081, 1994.
- 763 [32] Neil A Halpern, Stephen M Pastores, John M Oropello, and Vladimir
764 Kvetan. Critical care medicine in the united states: addressing the inten-
765 sivist shortage and image of the specialty. *Critical care medicine*, 41(12):
766 2754–2761, 2013.
- 767 [33] Nils Yannick Hammerla, James Fisher, Peter Andras, Lynn Rochester,
768 Richard Walker, and Thomas Plötz. Pd disease state assessment in nat-
769 uralistic environments using deep learning. In *AAAI*, pages 1742–1748,
770 2015.

- [34] C William Hanson and Bryan E Marshall. Artificial intelligence applications in the intensive care unit. *Critical care medicine*, 29(2):427–435, 2001.
- [35] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.
- [36] J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2015. *ONC Data Brief*, 35:1–9, 2016.
- [37] G. Hernandez-Ibarburu, D. Perez-Rey, E. Alonso-Oset, R. Alonso-Calvo, K. de Schepper, L. Meloni, and B. Claerhout. ICD-10-CM extension with ICD-9 diagnosis codes to support integrated access to clinical legacy data. *International Journal of Medical Informatics*, 129:189 – 197, 2019. ISSN 1386-5056. doi: <https://doi.org/10.1016/j.ijmedinf.2019.06.010>. URL <http://www.sciencedirect.com/science/article/pii/S1386505619301972>.
- [38] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [39] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.
- [40] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer Methods and Programs in Biomedicine*, 177:141 – 153, 2019. ISSN 0169-2607. doi: <https://doi.org/10.1016/j.cmpb.2019.05.024>. URL <http://www.sciencedirect.com/science/article/pii/S0169260718309945>.
- [41] Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.
- [42] Yohan Jo, Lisa Lee, and Shruti Palaskar. Combining lstm and latent topic modeling for mortality prediction. *arXiv preprint arXiv:1709.02842*, 2017.
- [43] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

- [44] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376, 2017.
- [45] Kaung Khin, Philipp Burckhardt, and Rema Padman. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. *arXiv preprint arXiv:1810.01570*, 2018.
- [46] Sujin Kim, Woojae Kim, and Rae Woong Park. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare informatics research*, 17(4):232–243, 2011.
- [47] William A Knaus, Jack E Zimmerman, Douglas P Wagner, Elizabeth A Draper, and Diane E Lawrence. Apache-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine*, 9(8):591–597, 1981.
- [48] Gokul S Krishnan and S Sowmya Kamath. A supervised learning approach for icu mortality prediction based on unstructured electrocardiogram text reports. In *International Conference on Applications of Natural Language to Information Systems*, pages 126–134. Springer, 2018.
- [49] Gokul S Krishnan and S Sowmya Kamath. Evaluating the quality of word representation models for unstructured clinical text based icu mortality prediction. In *Proceedings of the 20th International Conference on Distributed Computing and Networking*, pages 480–485. ACM, 2019.
- [50] Gokul S Krishnan and Sowmya Kamath. A novel ga-elm model for patient-specific mortality prediction over large-scale lab event data. *Applied Soft Computing*, 2019.
- [51] Leah S Larkey and W Bruce Croft. Automatic assignment of icd9 codes to discharge summaries. Technical report, Technical report, University of Massachusetts at Amherst, Amherst, MA, 1995.
- [52] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.
- [53] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.
- [54] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

- [55] Joon Lee, Daniel J Scott, Mauricio Villarroel, Gari D Clifford, Mohammed Saeed, and Roger G Mark. Open-access mimic-ii database for intensive care research. Institute of Electrical and Electronics Engineers, 2011.
- [56] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.
- [57] Yen-Fu Luo and Anna Rumshisky. Interpretable topic features for post-icu mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 827. American Medical Informatics Association, 2016.
- [58] Yuan Luo. Recurrent neural networks for classifying relations in clinical notes. *Journal of biomedical informatics*, 72:85–95, 2017.
- [59] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.
- [60] Bert A Mobley, Renee Leasure, and Lynda Davidson. Artificial neural network predictions of lengths of stay on a post-coronary care unit. *Heart & Lung: The Journal of Acute and Critical Care*, 24(3):251–256, 1995.
- [61] Alvaro Monge and Charles Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. 1997.
- [62] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [63] Anika Oellrich, Nigel Collier, Tudor Groza, Dietrich Rebholz-Schuhmann, Nigam Shah, Olivier Bodenreider, Mary Regina Boland, et al. The digital revolution in phenotyping. *Briefings in bioinformatics*, 17(5):819–830, 2015.
- [64] Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.
- [65] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Grisel, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [66] Romain Pirracchio. Mortality prediction in the icu based on mimic-ii results from the super icu learner algorithm (sicula) project. In *Secondary Analysis of Electronic Health Records*, pages 295–313. Springer, 2016.

- [67] Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. Mortality prediction in intensive care units with the super icu learner algorithm (sacula): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52, 2015.
- [68] Aaditya Prakash, Siyuan Zhao, Sadid A Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Condensed memory networks for clinical diagnostic inferencing. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [69] Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational recurrent adversarial deep domain adaptation. 2016.
- [70] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 2018.
- [71] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018.
- [72] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016.
- [73] SR Rassek, M Lorenzi, L Lee, S Devji, M McBride, and K Goddard. Reclassification of icd-9 codes into meaningful categories for oncology survivorship research. *Journal of cancer epidemiology*, 2010, 2010.
- [74] Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine Learning for Healthcare Conference*, pages 73–100, 2016.
- [75] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.
- [76] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of machine learning research*, 5(Jan):101–141, 2004.
- [77] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

- [78] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
- [79] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.
- [80] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.
- [81] Suchi Saria and Anna Goldenberg. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, 30(4):70–75, 2015.
- [82] Yutaka Sasaki et al. The truth of the f-measure. *Teach Tutor mater*, 1(5):1–5, 2007.
- [83] Elizabeth L Stone. Clinical decision support systems in the emergency department: Opportunities to improve triage accuracy. *Journal of Emergency Nursing*, 45(2):220–222, 2019.
- [84] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.
- [85] Kathleen B To and Lena M Napolitano. Common complications in the critically ill patient. *Surgical Clinics*, 92(6):1519–1557, 2012.
- [86] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009.
- [87] Jack V Tu and Michael RJ Guerriere. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. *Computers and biomedical research*, 26(3):220–229, 1993.
- [88] Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- [89] J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and LG Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, 1996.

- [90] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM, 2009.
- [91] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, pages 1–16, 2018.
- [92] Ian ER Waudby-Smith, Nam Tran, Joel A Dubin, and Joon Lee. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PloS one*, 13(6):e0198687, 2018.
- [93] Peter Wiemer-Hastings, K Wiemer-Hastings, and A Graesser. Latent semantic analysis. In *Proceedings of the 16th international joint conference on Artificial intelligence*, pages 1–14. Citeseer, 2004.
- [94] Christine M Wollschlager and Arnold R Conrad. Common complications in critically ill patients. *Disease-a-month*, 34(5):225–293, 1988.
- [95] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.
- [96] Safoora Yousefi, Congzheng Song, Nelson Nauata, and Lee Cooper. Learning genomic representations to predict clinical outcomes in cancer. *arXiv preprint arXiv:1609.08663*, 2016.
- [97] Min Zeng, Min Li, Zhihui Fei, Ying Yu, Yi Pan, and Jianxin Wang. Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*, 324:43–50, 2019. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2018.04.081>. URL <http://www.sciencedirect.com/science/article/pii/S0925231218306246>. Deep Learning for Biological/Clinical Data.
- [98] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.
- [99] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.
- [100] Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for todays critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.