FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes

Tushaar Gangavarapu*, *Student Member, IEEE,* Gokul S Krishnan, *Student Member, IEEE,* Sowmya Kamath S, *Senior Member, IEEE,* and Jayakumar Jeganathan, *MD*

Abstract—Accurate risk stratification using patient data is a vital task in channeling prioritized care. Most state-of-the-art models are predominantly reliant on digitized data in the form of structured Electronic Health Records (EHRs). Those models overlook the valuable patient-specific information embedded in unstructured clinical notes, which is the prevalent medium employed by caregivers to record patients' disease timeline. The availability of such patient-specific data presents an unprecedented opportunity to build intelligent systems that provide exclusive insights into patients' disease physiology. Moreover, very few works have attempted to benchmark the performance of deep neural architectures against the state-of-the-art models on publicly available datasets. This paper presents significant observations from our benchmarking experiments on the applicability of deep learning models for the clinical task of ICD-9 code group prediction. We present *FarSight*, a long-term aggregation mechanism intended to recognize the onset of the disease with the earliest detected symptoms. Vector space and topic modeling approaches are utilized to capture the semantic information in the patient representations. Experiments on MIMIC-III database underscored the superior performance of the proposed models built on unstructured data when compared to structured EHR based state-of-the-art model, achieving an improvement of 19.34% in AUPRC and 5.41% in AUROC.

Index Terms—Clinical decision support systems, disease prediction, healthcare analytics, ICD-9 code group prediction, precision medicine.

1 Introduction

Accurate disease prediction and quantification of patients' health condition at the earliest stages of diagnosis is central to clinical decision-making and channeling prompt care to critical patients [1]. Until recently, the healthcare industry was restricted by a conservative treatment approach resulting in limited patient-centric diagnostic capabilities [2]. With the advent of technological advancements and the extensive drive towards digitization, the utilization of the abundantly available heterogeneous clinical data to enhance the quality of life and leveraging such data for evidence-based medicine has gained momentum. For instance, in hospitals, Intensive Care Units (ICUs) are critical-care environments that depend on regular monitoring of various parameters pertaining to the condition of critically ill patients, thus generating large amounts of data. Such data could be vital in the development of Clinical Decision Support Systems (CDSSs) with enhanced predictive capabilities, essential to promote patient-centric and evidence-based treatments, in turn reducing morbidity and mortality rates, and facilitating improved risk assessment. Structured data in the form of Electronic Health Records (EHRs) are manually coded and contain valuable healthcare information, including symptoms, procedures, medications, diagnostic codes, and lab results. Modeling the data available via EHRs using machine and deep learning for survival analysis, mortality prediction, causal effect inference, physiologic decline detection, and others has sparked widespread interest [3].

Despite the substantial role of structured EHRs in enabling precision medicine based practices, their adoption in developing countries is minimal. For clinical decision-making, healthcare personnel in such countries still rely on human evaluation of the unstructured nursing notes. CDSSs used in hospital scenarios are built on structured EHR data, which are readily amenable to standard statistical analysis. However, unstructured clinical text and medical images contain valuable information concerning a patient's state. In particular, clinical nursing notes contain extensively documented subjective assessments and concerns regarding a patient's clinical condition (see Fig. 1). Such notes contain valuable assessments and intuitions of nurses and visiting doctors who continuously monitor the patient. Mining and modeling such data can help discover novel and hidden patterns and relationships needed for effective clinical decision-making. Recent research in the field of health informatics has shown that the information captured in unstructured nursing notes includes caregivers' observations and intuitions that do not fit into the accompanying recorded structured data [3]. Despite the patient-centric richness and abundance of such unstructured healthcare data, the majority of it remains unexplored. Primary challenges in modeling such raw and informally-written clinical text for the prediction of clinical outcomes include: 1) *longitudinality*, multiple measurements or repeated events are available for a subject, 2) *heterogeneity*, the attributes and events often vary conspicuously from one patient to another, 3) *voluminosity*, multiple detailed assessments are maintained for every patient, and 4) *complex temporal and linguistic structure*, rich medical jargon and non-standard abbreviations (e.g., *pt*, *hx*, and *s/p*) are abundant in nursing notes (a sample note is shown in Fig. 2). Thus, the ability to effectively extract and

*Corresponding author

- *T. Gangavarapu is with Automated Quality Assistance (AQuA) Machine Learning Research, Kindle Content Experience at Amazon.com, Inc. E-mail: tusgan@amazon.com (T. Gangavarapu)*
- *G.S. Krishnan and S.S. Kamath are with the Healthcare Analytics and Language Engineering (HALE) Lab, Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, Karnataka, India.*
- *J. Jeganathan is with the Department of Medicine, Kasturba Medical College, Manipal University, MAHE, Mangalore 575001, Karnataka, India.*

Manuscript received 08 July, 2019; revised 07 February, 2020; accepted 18 February, 2020. This work was supported by research under the Early Career Research Grant (ECR/2017/001056), awarded by the Government of India's DST-SERB to Sowmya Kamath S.

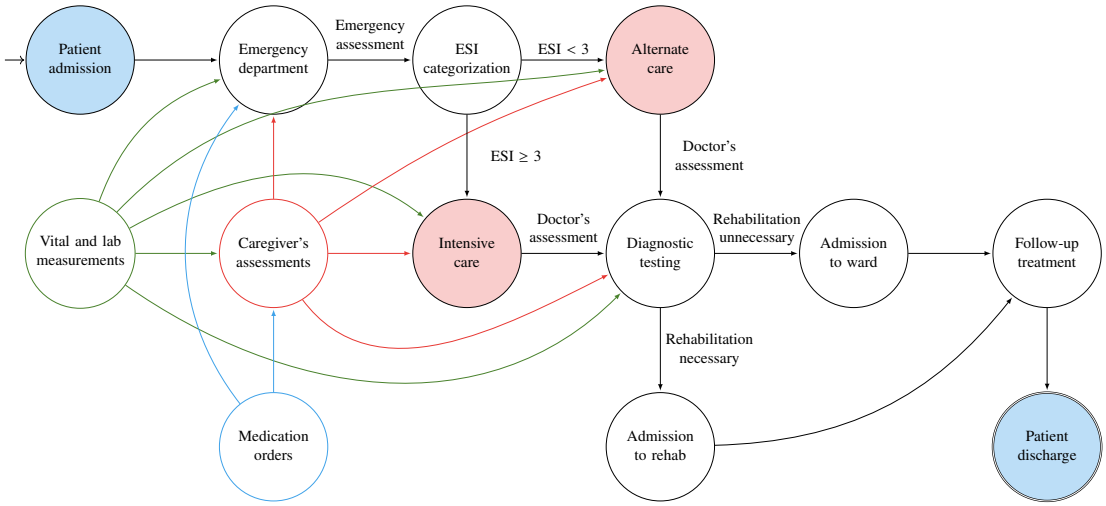


Fig. 1: State transitions of a patient trajectory pertaining to a single hospital admission (across multiple episodes). Note that the caregiver's notes contain patient-specific information concerning several clinical assessments made throughout the stay, including vital and lab measurements and medication orders. A patient can develop a complication at any time, and we aim at detecting the onset of the disease with the earliest recorded symptoms. (ESI is Emergency Severity Index.)

> Pt 62 yo F having dinner with friends when she experienced a sudden onset of retro-sternal chest pain. Pain was desc as heaviness and radiated into lft shoulder. Accomp by SOB, nausia, and an occipital headache. Friends and family here to visit.
>
> Pat O/A - A: Clear. B: Eupnoeic resp. Intensity R=L. Nil adventitious breath sounds. Symm chest excursion. C: Normocardic normotens. Strong reg rad pulse. Centrally and peripherally well profuse. D: Alert, oriented, and coop. PERTL (3 mm). E: Skin warm and dry to touch.
>
> Pain currently (4/10) desc as an aching sensation located L anterior chest and radiating to lft thumb. Plan: ECG, IV access, analgesia, bloods. Recheck ABGs, lytes. Suction prn. Med for pain. Cont vent wean. Note: Wife req son not be informed.

Fig. 2: Sample de-identified nursing note from critical care. Note the inconsistency, absence of grammatical structure, informal word usage, and extensive medical jargon.

consolidate the rich patient-specific information embedded in the nursing notes determines the efficacy of a CDSS. Additionally, there is also a need for multi-label assignment, from a large set of potential labels, owing to the diverse nature of the disease symptoms.

Over the years, owing to the public availability of de-identified large healthcare databases such as Medical Information Mart for Intensive Care (MIMIC-III) [4], an escalation in healthcare data mining and modeling to determine diagnostic measures needed to augment healthcare policies and effectively assess the severity-of-illness was observed. Seminal works [5], [6] on forecasting the length-of-stay and predicting mortality reported promising results with the application of machine learning models to structured critical patient data. Recent advances concerning practical progress in clinical decision-making, and prediction of prominent events and outcomes (e.g., phenotyping, mortality prediction, and ICD-9 code group prediction) using machine and deep learning have been extensively benchmarked on MIMIC databases. Pirracchio [7] reported an improved performance over several traditional severity scoring systems employed in hospitals, in predicting hospital mortality using a super learner algorithm, which was an ensemble of various machine learning models. Johnson *et al*. [8] compared several state-of-the-art works against Logistic Regression (LR) and gradient boosting, using an extracted feature set from the MIMIC-III database for the clinical ICU mortality prediction task. Recently, Harutyunyan *et al*. [9] used multitask RNNs to empirically validate four clinical prediction tasks on the MIMIC-III database. To tag patient notes of the MIMIC-II and III databases by identifying the label-relevant sentences, Baumel *et al*. [10] developed the hierarchical attention bidirectional Gated Recurrent Unit (GRU). Purushotham *et al*. [1] benchmarked a suite of five clinical prediction tasks, including ICD-9 code group prediction on the MIMIC-III database and compared their performance with state-of-the-art works and severity scoring systems.

ICD-9 codes are a taxonomy of diagnostic codes that are widely used by medical personnel, including doctors, health insurance companies, and public health agencies, to classify diseases and a wide range of symptoms, causes of injury, disorders, infections, and other medical conditions. Such categorization is paramount in cost-effective analyses, epidemiology studies, and designing healthcare policies. Segregating full-code predictions from category-level (group) predictions is often desirable, owing to the high granularity of the ICD-9 diagnostic codes. Each diagnostic code group comprises a set of similar diseases, and almost every medical condition can be categorized into a unique code group. High predictability of ICD-9 codes is facilitated by accurate ICD-9 code group prediction. Since the patients are grouped by diagnoses, ICD-9 code groups report on symptoms, severity, and resource utilization across agencies, thus facilitating research, billing, and tracking. Besides, disease-specific staging systems could capture the symptoms, severity, and resource utilization within a specific code group.

Most state-of-the-art works [7], [8] utilize machine learning models built on digitized clinical data and numerical assessments through structured EHRs to facilitate the prediction of significant clinical events and outcomes. Although some of these contemporary efforts [1], [9] attempted to benchmark their performance using deep learning models, they neglected the rich patient-specific information available in the informally-written nursing notes. Owing to practical constraints, ICUs often suffer from the limited availability of equipment and trained medical staff. Furthermore, due to the wide variety and complexity-levels of

ICU complications, there is often a lack of accurate knowledge of the etiology of such complications, resulting in an inability to assess patient mortality risk accurately. Effective modeling of unstructured nursing notes to facilitate early detection of high-risk patients and provide prioritized care to prevent further complications, in turn curtailing the mortality and morbidity rates, is essential. However, structured EHR data based state-of-the-art model [1] reports modest performance in ICD-9 code group prediction. Hence, there is a need for an effective strategy that facilitates accurate ICD-9 code group prediction, in turn enhancing the ICD-9 code predictability.

In this paper, we attempt to model the rich patient information embedded in clinical nursing notes using vector space (Doc2Vec) and topic modeling (Nonnegative Matrix Factorization (NMF)), for deriving optimal patient-specific data representations. *FarSight*, an aggregation mechanism intended to detect the onset of the disease with the earliest recorded symptoms, infections, and disorders, forms the core of our work. Furthermore, deep neural architectures including Multi-Layer Perceptron (MLP), Convolutional Neural Network (ConvNet), Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Convolutional LSTM (Conv-LSTM), and Segment-level GRU (Seg-GRU) are benchmarked for the code group prediction task, and the proposed *FarSight*-aggregated unstructured modeling is evaluated against naïve note aggregation strategy and structured EHR based state-of-the-art model using standard evaluation metrics. The key contributions of our work are mainly three-fold:

- Design of *FarSight*, a long-term aggregation mechanism that employs future lookup to detect disease onset with the earliest recorded symptoms, to enable prioritized care and prevent further complications.
- Leveraging effective vector space and topic modeling approaches to derive optimal data representations from the unstructured clinical text, essential in accurate ICD-9 code group prediction. Our experimental results corroborate the efficacy of the proposed strategy when compared to state-of-the-art models built on structured patient data.
- Designing a technique that utilizes voluminous nursing notes for accurate risk stratification, thus eliminating the dependency on the availability of structured EHRs. This eliminates a significant roadblock in the development of CDSSs for hospitals in developing nations with low structured EHR adoption rates.

The remainder of this paper is organized as follows: In Section 2, we present a discussion on the existing works in the domain of our research. Section 3 elucidates the MIMIC-III database and the proposed methods designed as a part of *FarSight*, for patient-specific clinical feature extraction. Data modeling, deep neural architectures utilized in ICD-9 code group prediction, experimental validation, and performance benchmarking are presented in Section 4. Finally, Section 5 concludes this work with insights into future research avenues.

2 Related Work

Several attempts have been made to exploit the heterogeneity and richness of the healthcare data in EHRs. Systems that provide healthcare services are being actively developed to aid the identification of high-risk individuals, management of hospital resources, and planning of personalized treatment (e.g., MatrixFlow [11] and Intelligent Care Delivery Analytics [12]). In this section, we provide a brief overview of the existing machine and deep learning models in predicting the clinical outcomes and then discuss the existing works built on benchmarking healthcare databases.

Early works [13], [14] report that the machine learning models, especially feed-forward neural networks, obtain good results on medical risk assessment and mortality prediction. Furthermore, in modeling the mortality risk among hospitalized patients, Celi *et al*. [15] showed that feed-forward neural networks almost always outperform several severity-of-illness scores and LR. Recent advances in deep learning have led to the development of novel neural architectures that showed promising results in a variety of clinical prediction tasks, including length-of-stay prediction, inpatient mortality prediction, diagnoses on general EHR data, and diagnostic code group prediction [16]. Che *et al*. [17] developed a scalable deep neural framework for disease diagnosis that uses prior knowledge from medical ontologies to learn clinically relevant features. Dabek and Caban [18] employed a feed-forward neural network to improve the predictability of various psychological conditions such as depression, behavioral disorders, anxiety, and post-traumatic stress disorder. Khin [19] developed a Bi-LSTM model with deep contextualized word embeddings and variational dropouts that achieved superior performance and faster convergence in de-identifying nursing notes. These previous works exemplify the power of leveraging machine and deep neural architectures in healthcare applications.

Over the years, many works have addressed the problem of modeling disease progression, both within a hospital episode and for chronic illness. A cluster of 45 clinical, physiological, and ICU treatment variables was used by Cohen *et al*. [20] to identify complex metabolic states and facilitate patient monitoring. Zhou *et al*. [21] utilized lab test and cognitive scores, and genetic and demographic data to propose a disease progression model based on fused group lasso formulation. Utilizing the heterogeneous and incomplete patient medical records, Wang *et al*. [22] built disease progression models on clinical findings and comorbidities. Choi *et al*. [23] adopted a context-sensitive multivariate Hawkes process to model the temporal progression of patients, and infer disease relationship network for the prediction of patient-specific diseases. Some works utilized clinical time-series data to facilitate the multi-label prediction of diagnostic codes using feed-forward neural networks [24], LSTM networks [25], and temporal ConvNets [26] to capture the comorbidities in the hidden layers implicitly. Recent works [27], [28] leveraged the power of deep neural architectures to model the disease and clinical time-series data. These works establish the need for early patient-specific disease prediction in the development of an efficient CDSS.

To examine the structural, linguistic, and topical differences among unstructured medical narratives, Feldman *et al*. [29] mined the radiology, physician, clinical nursing, and ECG narratives. The authors only presented a foundation to effectively mine clinical nursing notes, which is extended in this work. Zalewski *et al*. [30] proposed a framework to combine various patient heath state modalities for the stratification of medical risk. Their approach was aimed at tackling the high-dimensionality and sparsity of the nursing notes (from the MIMIC-II database) through the use of the Hierarchical Dirichlet techniques. However, the authors employed an LR predictor to facilitate the mortality rate estimation, and did not evaluate their performance against the deep neural architectures and recent works. The ease of benchmarking machine and deep learning models for accurate prediction of clinical events and

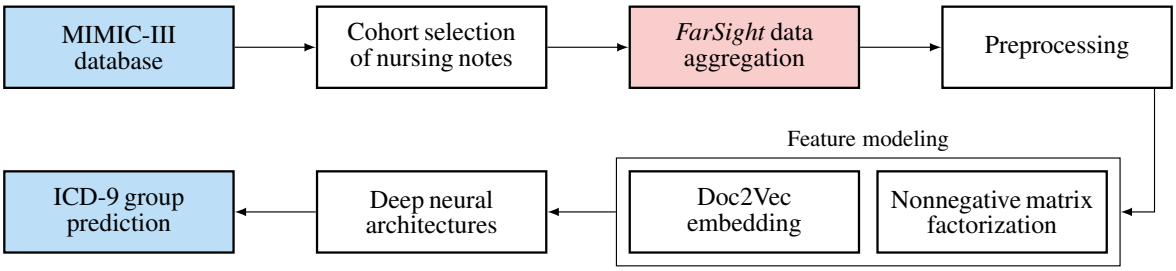


Fig. 3: NLP pipeline used in the prediction of the ICD-9 code group.

outcomes (e.g., mortality and diagnostic code group) is facilitated by the public availability of de-identified healthcare databases such as MIMIC-II and MIMIC-III. To predict the patient-specific mortality in ICUs, Pirracchio [7] used the MIMIC-II database and showed that the super learner algorithm, an ensemble of several machine learning models, outperformed traditional severity-of-illness scoring systems. Although the super learner algorithm performed better than traditional prognostic scoring systems, it was not benchmarked against the recent machine and deep learning models. The challenges in reproducing the results reported by 28 recent and related publications on the publicly available MIMIC-III database was studied by Johnson *et al*. [8]. They extracted a simple set of features and compared the performance reported in the studies against LR and gradient boosting models using the extracted features. Furthermore, to facilitate fair comparison among the proposed methods and account for the large heterogeneity in the studies, the authors emphasized the need for improvement in the way of reporting the performance.

More recently, a comprehensive deep neural approach using multi-task Recurrent Neural Networks (RNNs) was developed by Harutyunyan *et al*. [9] to validate four clinical prediction benchmarking tasks on MIMIC-III empirically. While their work reported promising results in clinical prediction, the authors only compared their model with LR and LSTM models, and failed to benchmark against state-of-the-art machine learning models (especially, super learner) and prognostic scoring systems. Purushotham *et al*. [1] performed consistent and exhaustive benchmarking experiments on several clinical prediction tasks, including mortality prediction, length-of-stay prediction, and ICD-9 code group prediction using MIMIC-III. They benchmarked their work against several severity-of-illness scores and machine learning models. However, their work neglects the rich patient-specific information available in the clinical nursing notes. For the clinical task of patient-specific mortality prediction, Krishnan and Kamath [31] proposed a novel hybrid metaheuristic-based feature modeling approach to process large-scale lab event data—their approach outperformed several severity-of-illness scores and machine learning models. Nevertheless, their study utilizes large-scale structured lab event data to facilitate the clinical prediction task.

Huang *et al*. [32] modeled the unstructured discharge summaries of the MIMIC-III database using state-of-the-art deep neural architectures for predicting the (top−10) ICD-9 code categories. Zeng *et al*. [33] developed a deep transfer framework to improve the diagnostic coding process through a transfer of domain knowledge from medical subject headings. While these recent works are focused on modeling unstructured text for the prediction of clinical outcomes, they ignore the valuable patient-specific information present in informally-written nursing notes. Furthermore, modeling clinician's notes facilitates devising healthcare policies and effective clinical decision support, in addition to reliable billing, as opposed to discharge summaries, which only facilitate accurate billing. Stone [34] explored several opportunities in assisting medical personnel in high-pressure distractive situations with limited patient history of sustained trauma, intended on improving the triaging accuracy of the CDSSs. Our work utilizes the rich patient-centric information to stratify medical risk, thus extending the efforts of the author by aiding the underlying CDSS with minimized risk of clinical deterioration, increased triaging accuracy, and optimized patient outcomes.

Our work explicitly explores a significantly underutilized health-related resource, i.e., unstructured clinical nursing notes, for the development of intelligent CDSSs. Such an initiative would be especially advantageous for hospitals in developing countries where the adoption rate of structured EHRs is relatively low. Our objective is to advance the efforts of the state-of-the-art studies by modeling the valuable information present in such notes, which is often lost in the transcription of nursing notes into structured EHRs. Furthermore, our work facilitates an exhaustive comparative study to assess the performance of the proposed data modeling approaches with a variety of deep neural architectures for the clinical task of ICD-9 code group prediction.

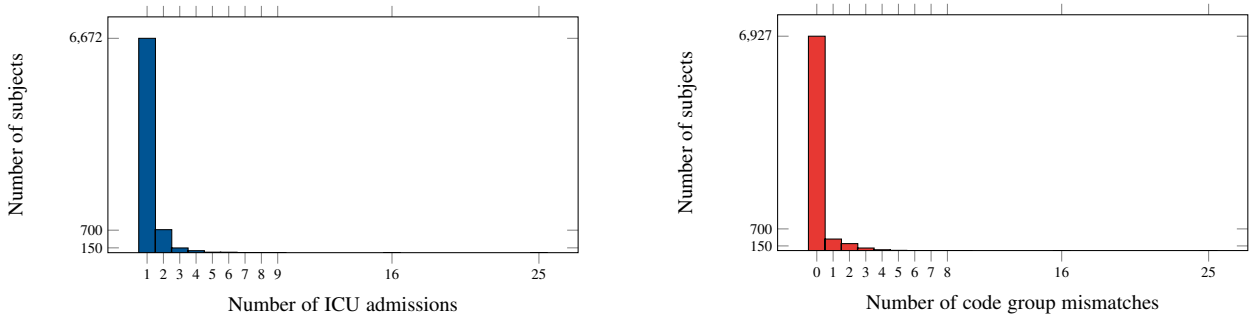
3 Materials and Methods

In this section, we describe the specifics of the techniques designed for preprocessing and extraction of features for the multi-label task of ICD-9 code group prediction from the unstructured clinical notes available in the MIMIC-III dataset. A detailed overview of the Natural Language Processing (NLP) pipeline deployed to facilitate the clinical diagnostic code group prediction task is depicted in Fig. 3. The various subprocesses and models are discussed in subsequent sections.

3.1 Dataset Specifics and Cohort Selection

The MIMIC-III (v1.4) database is a publicly available, sizeable critical care database developed and maintained by the Massachusetts Institute of Technology Laboratory for Computational Physiology. It integrates comprehensive and diverse de-identified health-related data associated with approximately $60,000$ admissions of critical care patients at the Beth Israel Deaconess Medical Center, USA, between 2001 and 2012. The database includes crucial information including demographics, laboratory test results, vital sign measurements (~ one data point per hour), procedures, medications, imaging reports, nursing (caregiver) notes, and in- and out-of-hospital mortality. Furthermore, it supports a diverse range of analytic studies spanning clinical decision-rule improvement, epidemiology, and electronic tool development.

The MIMIC-III (v1.4) database contains $223,556$ nursing notes among $2,083,180$ note events, corresponding to $7,704$ distinct patients. The detailed statistics of the nursing note text



(a) Distribution of the admissions to the ICU.

(b) Distribution of the code group mismatches.

Fig. 4: Statistics of the data extracted from the MIMIC-III database.

TABLE 1: Statistics of the nursing note text corpus.

Parameter	Total
Total clinical nursing notes	223,556
Total sentences in the nursing notes	5,244,541
Total words in the nursing notes	79,988,065
Total unique words in the nursing notes	715,821

corpus are tabulated in Table 1. During the data preparation phase, we considered certain inclusion criteria to select the MIMIC-III subjects for our study. First, using the age at the time of admission to the ICU, the records of neonates (age below 15) were identified and discarded, similar to the cohort criteria adopted by the state-of-the-art studies [1], [8]. Furthermore, to maintain consistency in benchmarking with respect to the related works [1], [8], [35], and to avoid possible information loss during analysis, only the first admission to the ICU for each MIMIC-III subject was considered, and all the later admissions were discarded. The number of ICD-9 code group mismatches[1] from ICU patients' first admission to their later admissions is summarized in Fig. 4b. Note that the diagnostic code groups in the first admission of more than 94% of the nursing notes overlap with those occurring in the later ICU admissions (see Fig. 4b). Thus, retaining only the first ICU admission allows for faster risk prediction using the earliest detected conditions, with reduced computational complexity and information loss. In this study, we aim at facilitating the prediction of ICD-9 code groups that are recorded only in the first ICU admission (with an average of 176.49 episodes per patient) of the MIMIC-III subjects.

3.2 Data Extraction

The MIMIC-III database contains 26 relational tables, of which, the data of our final patient cohort was extracted from four tables: *noteevents*, *admissions*, *patients*, and *diagnoses_icd*. The *noteevents* table contains unstructured nursing notes, electrocardiogram reports, echo reports, and radiology reports for both outpatient and inpatient stays. Information pertaining to patients' ICU admissions is recorded in the *admissions* table and was used to obtain the patients' time of admission to the ICU. The date-of-birth of each patient was extracted from the *patients* table, which contains the charted data for all the MIMIC-III subjects. The *diagnoses_icd* table comprises ICD-9 diagnostic codes of the MIMIC-III subjects. These tables provide the most relevant data and clinical diagnostic features and hence, are selected for the prediction task of ICD-9 code group prediction. For instance, a patient p (born on T_{DoB}), admitted to the hospital (with an admission number I_{hadm}) at time T_{adm}, with age $T_{\text{adm}} - T_{\text{DoB}}$ (must be > 15 (consistent with the criteria in Section 3.1)) has multiple nursing notes (corresponding to multiple episodes (I_{note}s), extracted using I_{hadm}) corresponding to multiple ICD-9 code groups. Fig. 4 shows the statistics of the data extracted from the MIMIC-III database. As per the defined cohort selection criteria presented in Section 3.1, the dataset extracted from the selected tables contained nursing notes of 7,638 MIMIC-III subjects with the median age of 66 years (Quartile Q_1–Q_3: 52 – 78 years).

3.3 Data Cleansing, Aggregation, and Preprocessing

Several erroneous entries exist in the data extracted from MIMIC-III due to various factors such as missing values, noise, incorrect or duplicate entries, outliers, and clerical errors. First, we identified and filtered out the nursing notes with clerical errors and erroneous entries using the *iserror* attribute of the *noteevents* table. Second, we segregated and deduplicated identical patient records. The resultant patient cohort obtained after handling the erroneous entries contained nursing notes corresponding to 6,532 patients (140,792 clinical notes)—the data in these nursing notes were aggregated using the *FarSight* approach to detect the onset of the disease with the earliest recorded symptoms.

3.3.1 *FarSight: Long-Term Aggregation by Future Lookup*

As the need for critical care facilities like ICUs grows, the limited availability of resources including specialized monitoring equipment and trained clinical staff, serves as the bottleneck. In addition, a lack of precise knowledge concerning the etiology of ICU complications can lead to delayed and imprecise recognition of patients at high-risk, thus hindering preemptive treatment options. As a result, the requisite care is often delivered only after the development of a particular complication. Therefore, detection of disease onset when the earliest recorded infections or symptoms are observed is of utmost importance, as it can significantly reduce mortality and morbidity rates. Towards this objective, we present *FarSight*, a long-term aggregation mechanism, which facilitates the aggregation of the patient data using a future lookup on all the later detected symptoms and diseases[2].

Let $\mathcal{P}$ be the set of all patients, indexed by p. For each patient, we have a sequence of clinical nursing notes, $\Phi^{(p)} = \{(\eta_n^{(p)}, \mathcal{I}_n^{(p)})\}_{n=1}^{N(p)}$, with each nursing note $\eta_n^{(p)}$ and the corresponding

1. A patient is said to have n code mismatches if n codes in the set of codes from later admissions are not present in the first admission code set.

2. In this context, 'later detected diseases' at time T are the diseases recorded in the medical records after time T.

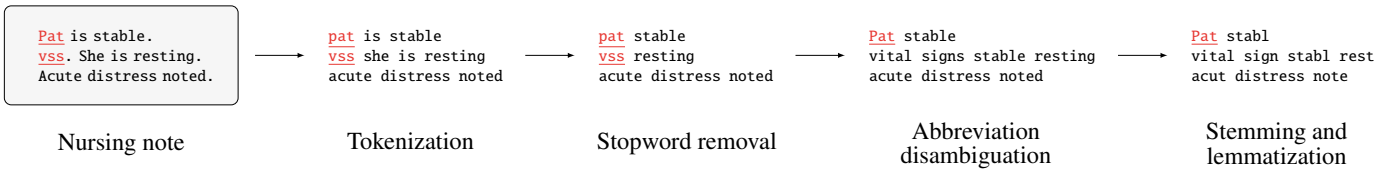


Fig. 5: An example indicating the outcomes of data preprocessing, including tokenization, stopword removal, clinical abbreviation normalization, and stemming/lemmatization process.

ICD-9 diagnostic code (group) $\mathcal{I}_n^{(p)}$ indexed by n, and with $N(p)$ number of notes (of total N nursing notes) for a patient p. Furthermore, the nursing notes of a patient are ordered from oldest to the most recent. Now, the aggregation of the ICD-9 code groups across the nursing notes of a patient is performed using *FarSight*, through a future lookup of the diseases in the long run (dependent on the number of nursing notes recorded for that specific patient (p), concerning several episodes during single hospital admission), resulting in $\Phi^{(p)} = \{(\eta_n^{(p)}, \mathcal{I}^{(p)})\}_{n=1}^{N(p)}$, where $\mathcal{I}^{(p)} = \{\mathcal{I}_n^{(p)}\}_{n=1}^{N(p)}$. Note that, while the aggregation of diagnostic code groups seems to be incremental in nature, the objective is to predict the diseases and complications that are most likely to be observed in the subsequent episodes of a patient's current hospital admission—*FarSight* facilitates such prediction through aggregation of diagnostic code groups across all the episodes recorded for a patient, thus performing long-term aggregation through future lookup. Ultimately, our goal is to learn a generalizable function ($\mathcal{G}$) that estimates the probability of classifying a given clinical nursing note $\eta_n^{(p)}$ into a set of ICD-9 diagnostic codes:

$$\mathcal{G}(\Phi^{(p)}) \approx Pr(\mathcal{I}^{(p)} \mid \eta_n^{(p)}) \tag{1}$$

It is to be noted that, the proposed *FarSight* mechanism facilitates multi-label classification by aggregating the diagnostic code groups across a patient's multiple medical records, rather than aggregating the raw medical text in the nursing notes. Such an aggregation facilitates risk assessment at the initial stages of the disease, with the earliest detected infections and symptoms, and with a reliable accuracy level.

Consider the nursing notes $(\{\eta_n^{(p)}\}_{n=1}^{N(p)})$ of a patient (p) ordered chronologically; assuming $N(p)$ to be three, we have three medical records of the patient p corresponding to three (distinct) diagnostic code groups $(\{\mathcal{I}_n^{(p)}\}_{n=1}^{N(p)=3})$. By employing the *FarSight* aggregation mechanism, we map each of the three medical records to all the ICD-9 code groups observed in the patient p's nursing notes, i.e., $\{\mathcal{I}_n^{(p)}\}_{n=1}^{N(p)=3}$. Simply put, each $\eta_n^{(p)}$ corresponds to $\{\mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}, \ldots, \mathcal{I}_{N(p)}^{(p)}\}$. It is important to stress that, *FarSight* aggregation is only effective when the disease symptoms are progressive and related (e.g., *sore throat* $\longrightarrow$ *cold* $\longrightarrow$ *fever* vs. *sore throat* $\longrightarrow$ *leukaemia*). Since this study specifically considers the first ICU admission of a MIMIC-III subject (see Section 3.1), *FarSight* can be employed to stratify risk using the earliest detected symptoms. However, through naïve aggregation of nursing notes using patient identification numbers, we have $\eta_1^{(p)} \oplus \eta_2^{(p)} \oplus \cdots \oplus \eta_{N(p)}^{(p)}$ mapping to $\{\mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}, \ldots, \mathcal{I}_{N(p)}^{(p)}\}$ ($\oplus$ denotes concatenation). Thus, *FarSight*-aggregated data can help train the underlying classifier in identifying all the possible diagnostic groups, by capturing the episode-specific characteristics, i.e., training at the clinical note level. In contrast, patient-based aggregated data aids in training the predictor at the patient level. Furthermore, the diagnostic code groups of $\eta_i^{(p)}$ $(1 < i < N(p))$ predicted using a model trained on *FarSight*-aggregated data $(\{(\eta_n^{(p)}, \mathcal{I}_n^{(p)})\}_{n=1}^{N(p)=3})$ would include (with high probability) $\mathcal{I}_i^{(p)}$, owing to the training at nursing note granularity. However, employing a classifier trained on naïvely aggregated data to predict the diagnostic code groups of $\eta_i^{(p)}$ $(1 < i < N(p))$ might not include $\mathcal{I}_i^{(p)}$, as episode-specific characteristics are lost.

3.3.2 Data Preprocessing

Despite the inherent content-rich nature of the patient-specific information available in the clinical nursing notes, they are raw, sparse, informally-written, complexly structured, and voluminous. Thus, any transformation of raw medical text into a canonical form extends the learnability and generalizability of the underlying deep neural architectures. Such normalization not only allows for the separation of concerns but also helps maintain consistency. To achieve this, we subject all the notes to NLP processing, which included tokenization, stopword removal, and stemming/lemmatization. First, we removed multiple spaces and special characters. Next, we experimented with multiple tokenizers including MedPost[3], Penn bio tagger[4], NLTK[5], Stanford log-linear part-of-speech tagger[6], and GENIA tagger[7], to segment the medical text in the nursing notes into several primary building blocks (tokens). MedPost tokenizer splits the input text at hyphens, slashes, internal periods, and punctuation within numbers (e.g., *IL-20 i.e. 1,000 U/ml* is split as *IL␣-␣20␣i␣.␣e␣.␣1␣,␣000␣U␣/␣ml*), while Penn bio tagger splits the words at slashes (e.g., *0.05 U/ml* is split as *0.05␣U␣/␣ml*), and hence are not employed in this study. Moreover, we observed that the NLTK tokenizer was similar (in the splitting scheme) to Stanford log-linear part-of-speech and GENIA taggers, with respect to DNA sequences (e.g., *CCAAAGCGTAAAAGG*), words with numbers and letters (e.g., *15th*), and hyphenated compound words (e.g., *x-ray*). Thus, we employed the NLTK tokenizer to facilitate the tokenization of nursing text. Utilizing the NLTK English stopword corpus, we removed stopwords from the generated tokens. Furthermore, punctuation marks (except hyphens and slashes) were also removed. References to images (e.g., *MRI_Scan.jpeg*) were removed, and character case folding was performed. Note that, word-length based token removal was not performed to eliminate the loss of important medical information (e.g., *CT*, *DEXA*, *MRI*, and *PET*). Before any further processing, medical concept normalization through disambiguation of abbreviations (into their respective long forms) was facilitated using CARD, an open-source framework for clinical abbreviation recognition and disambiguation [36]. It must be noted that, despite meticulous nursing abbreviation disambiguation, a large number of non-standard abbreviations, typographical errors, and medical jargon result in clinical notes

3. ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz.
4. https://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger.
5. http://www.nltk.org/.
6. https://nlp.stanford.edu/software/tagger.shtml.
7. http://www.nactem.ac.uk/tsujii/GENIA/tagger/.

being heterogeneous and noisy—so, it is vital to develop a robust classifier, while lowering the burden of preprocessing. Lastly, suffix stripping was performed through stemming, followed by lemmatization for the conversion of the stripped tokens into their respective base forms. Additionally, we eliminated the tokens appearing in less than ten nursing notes (e.g., *spot*, *cope*, and *inch*) in order to lower the computational complexity of training (the total number of tokens pre- and post-elimination were 188,742 and 32,687 respectively) and mitigate problems arising due to overfitting. An example indicating the outcomes of the preliminary preprocessing steps is depicted in Fig. 5.

3.4 Clinical Feature Modeling

Let $\Pi = \{\Phi^{(p)}\}_{p=1}^{\mathcal{P}}$ be the set of all the clinical nursing notes in the MIMIC-III corpus. Each nursing note (η_n) constitutes a variable length of tokens, drawn from a sizeable vocabulary $\mathbb{V}$, making Π complex. Furthermore, each patient has a variable number of such nursing notes, adding to the complexity of Π. Thus, it is critically important to obtain a transformation (T) of the unstructured medical text into a machine-processable form (e.g., a fixed length real number vector), i.e., $T : \Pi \longrightarrow \mathbb{R}^d$. The patient information can now be transformed into an easier-to-use form, $\pi = T(\Pi)$, $\pi \in \mathbb{R}^d$. Usually, the aim is to have $d \ll |\mathbb{V}|$ to ensure the tractability of the learning problem; thus, we try to learn:

$$\mathcal{G}(\Phi^{(p)}) = \mathcal{G}(T(\Phi^{(p)})) \approx Pr(\mathcal{I}^{(p)} \mid \eta_n^{(p)}) \tag{2}$$

Despite the promising performance of rule-based and traditional dictionary-based NLP transformations, they require manual effort in adaption and lack automation [35]. Deriving optimized vector representations of the underlying unstructured nursing note corpus is vital to the performance and practicality of the classification models powering a CDSS. In this study, we use vector space and topic modeling to structure the raw medical text and enable an optimal representation of the patient cohort.

3.4.1 Vector Space Modeling of Clinical Notes

Vector space modeling facilitates the representation of each clinical nursing note as a point in a $d-$dimensional vector space ($d \ll |\mathbb{V}|$). Bag-of-Words (BoW) is a traditional transformation that captures the importance of a concept in the given vocabulary, and the weight of each term is computed as the frequency of its occurrence in the respective nursing note. The Term-Weighting (TW) numerical statistic is a prominent transformation of the BoW that captures both the specificity of a clinical concept as well as its relative importance. For a given nursing note $\eta_n^{(p)}$ of a patient p with $N(p)$ nursing notes, the weight $W_m^{(n)}$ of a term $t_m^{(n)}$ (of total $|t^{(n)}|$ terms) occurring $f_m^{(n)}$ times is given as:

$$W_m^{(n)} = \begin{cases} \left(1 + \log_2 f_m^{(n)}\right)\left(\log_2 \frac{N(p)}{|t^{(n)}|}\right), & \text{if } f_m^{(n)} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

However, the models utilizing BoW or associated transformations suffer from issues, including sparsity and high-dimensionality due to the one-hot encoding of every clinical term. Moreover, BoW transformations fail to capture the intuition of semantically similar clinical notes having similar vector representations. For instance, in BoW transformation, two terms with a tightly-coupled semantic relationship (e.g., *tumor* and *cancer*) could be mapped to entries with considerable distance. A solution is seen in the adoption of Doc2Vec or Paragraph Vector (PV) network, which efficiently learns the clinical term representations in a data-driven manner to cope with these shortcomings.

Doc2Vec facilitates a numerical transformation of variable length nursing notes into low-dimensional fixed-length document embeddings. It provides content-related measurements, typically by learning the distributed distributions using a neural network structure with one shallow hidden layer. The basic principle is as follows: in any given corpus, several clinical terms are used in the prediction of the subsequent term. By utilizing a self-learned embedding matrix, these clinical terms are mapped to numeric vectors and are fed to a neural network, whose output is the predicted term. Mini-batch stochastic gradient descent is used in learning the parameters of the neural network and the word embedding matrix. Doc2Vec extends this basic principle by learning document-level or paragraph representations through the use of an additional vector which represents the semantics of the entire document. Since Doc2Vec incorporates semantic textual features, it is influential in several NLP tasks including question type classification and document-level sentiment detection [37].

In this study, we chose the PV Distributed Memory (PV-DM) variant of Doc2Vec over the PV Distributed BoW (PV-DBoW) variant, owing to its ability to preserve the word order of the clinical terms and comparatively better performance [37]. To obtain the Doc2Vec style features from the raw clinical text, we utilized the implementation in the Python Gensim package, with an embedding size of 500 (trained for 25 epochs), determined empirically using grid-search.

3.4.2 Topic Modeling of Clinical Notes

Topic modeling aims at finding a set of topics (collection of terms) from a collection of documents that best represents the underlying corpus. Latent semantic analysis and other traditional methods of information retrieval compute the Singular Value Decomposition (SVD) of the BoW or TW matrix to generate a lower approximation of the matrix—such methods often deal with matrix computations of high complexity. NMF is a popular multivariate analysis approach that aims at factoring a data matrix ($M \in \mathbb{R}^{|\mathbb{V}| \times N}$) by minimizing the reconstruction error, with nonnegativity constraints. This can be viewed as learning an unnormalized probability distribution over the topics [38]. Formally, NMF seeks a factorization model for a given data matrix M and a target rank $\mathcal{T}$ (number of topics) to explain the data matrix (M), where $W \geq 0$, $H \geq 0$, and $\mathcal{T} \leq \min\{|\mathbb{V}|, N\}$ (as shown in (4)). The unnormalized probabilities are learned by

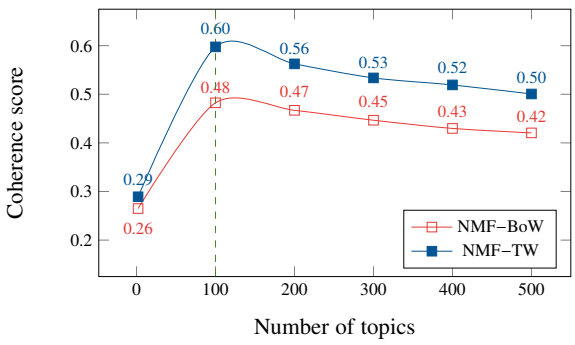


Fig. 6: Comparison of coherence scores to determine the optimal number of NMF topics.

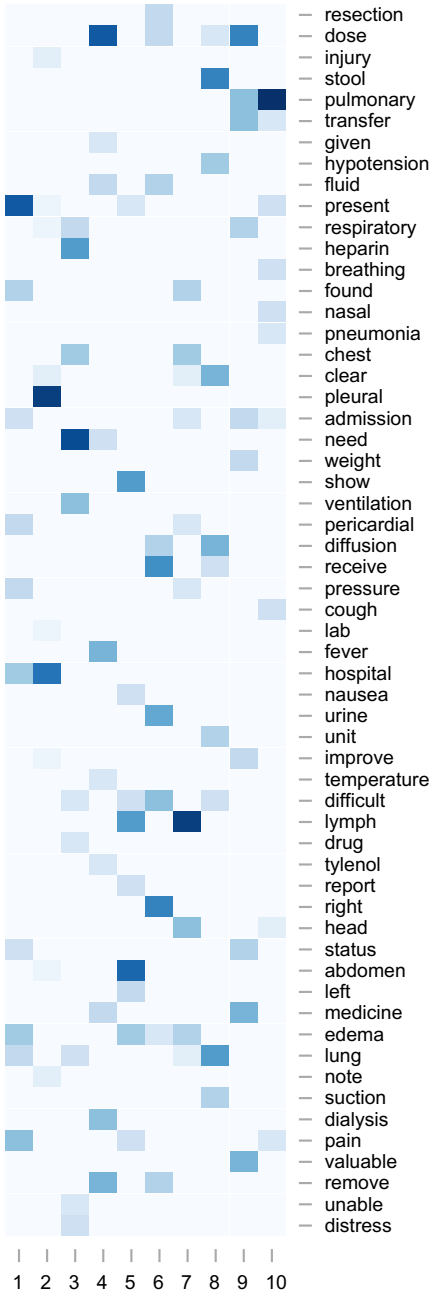


Fig. 7: Correlations between the top ten terms' membership in the top ten NMF clusters.

randomly initializing each set of probabilities and then updating them according to a set of iterative rules defined in (5).

$$M \approx WH^{\mathrm{T}},\ W \in \mathbb{R}^{|\mathbb{V}|\times\mathcal{T}},\ H \in \mathbb{R}^{N\times\mathcal{T}} \tag{4}$$

$$H \longleftarrow H \cdot \frac{W^{\mathrm{T}}M}{W^{\mathrm{T}}WH^{\mathrm{T}}} \qquad W \longleftarrow W \cdot \frac{MH}{WH^{\mathrm{T}}H} \tag{5}$$

At first glance, NMF is an alternative factoring model similar to SVD that considers different constraints (orthogonality) on the latent factors. However, the effectiveness of NMF when modeling real-life nonnegative data (e.g., text, images, and audio spectra) has sparked widespread interest in the fields of signal processing and data analytics [39]. Representing real-life data into nonnegative matrices and factoring them into latent factors yields intriguing results, and thus, NMF is popularly recognized as a workhorse in data analytics.

As is the case with other clustering approaches, determining the optimal number of NMF clusters is a challenging problem. Moreover, learning topics from a multinomial distribution of words from sparse and noisy textual data can often be hard to interpret. Perplexity can be used to address this issue as it measures the generalizability of a model. However, perplexity and human judgment may not always be correlated; often, they are anti-correlated [40]. Semantic Coherence (SC) is a way of evaluating models with a higher guarantee of human interpretability. In our work, we adopt NMF with SC, as it accounts for the semantic similarity between high scoring clinical terms. We employ the C_v variant of the coherence measurement with the Normalized Pointwise Mutual Information Score (NPMI) as the confirmation measure, owing to its more significant correlation with the available human-judged data [41].

Let $\mathcal{T}_i = \{t_1, t_2, \ldots, t_n\}$ be a topic generated from a topic model, represented by its top$-n$ most probable terms (t_ks). A topic depicts greater coherence when the average pairwise similarity among the terms of that topic is high. Given a predefined similarity score ($\mathrm{Sim}(t_k, t_l)$)[8], we compute the SC score using:

$$\mathrm{SC}(\mathrm{Sim}, \mathcal{T}_i) = \frac{\sum_{1\le k\le n-1} \mathrm{Sim}(t_k, t_l)}{\binom{n}{2}} \tag{6}$$

where $t_k, t_l \in \mathcal{T}_i$. The NPMI similarity score is used in finding collocations and associations between the words and is computed as per (7) and (8). To obtain the final conformation score, we average the individual confirmation scores obtained for all the topics ($\mathcal{T}_i$s).

$$\mathrm{NPMI}(t_k, t_l) = \frac{\mathrm{PMI}(t_k, t_l)}{-\log_2(\Pr(t_k, t_l))} \tag{7}$$

$$\mathrm{PMI}(t_k, t_l) = \log_2\left(\frac{\Pr(t_k, t_l)}{\Pr(t_k)\Pr(t_l)}\right) \tag{8}$$

The optimal number of topics in NMF was determined to be 100, by comparing the coherence scores of several NMF models obtained by heuristically varying the number of topics from 2 to 500 (see Fig. 6). Furthermore, for the ease of interpretation of the topics derived from NMF, Fig. 7 depicts a heatmap presenting the correlations among the top ten terms' membership in the top ten clusters. Additionally, from Fig. 7, it can be observed that the NMF topics effectively capture specific medical terms such as *edema*, *pericardial*, *pleural*, *heparin*, *tylenol*, *resection*, *hypotension*, and *pulmonary*, from the unstructured nursing notes. In this study, we built the NMF matrices on both BoW and TW matrices, to enable an exhaustive comparison. Moreover, we model the BoW and TW matrices using NMF without coherence scoring (set to 150 topics, which was determined empirically using grid-search). The implementations available in the Python Gensim package were employed to implement the NMF models.

4 ICD-9 Code Group Prediction

In our work, we focus on ICD-9 code group prediction as a multi-label classification problem, where, each nursing note of every patient is mapped to multiple diagnostic code groups. The ICD-9 codes of a given admission from MIMIC-III are mapped into 19 distinct diagnostic groups[9]. The ICD-9 code range of 760 – 779

8. In our work, we use NPMI as the similarity measure.

9. The code ranges used for mapping can be accessed at http://tdrdata.com/ipd/ipd_SearchForICD9CodesAndDescriptions.aspx.

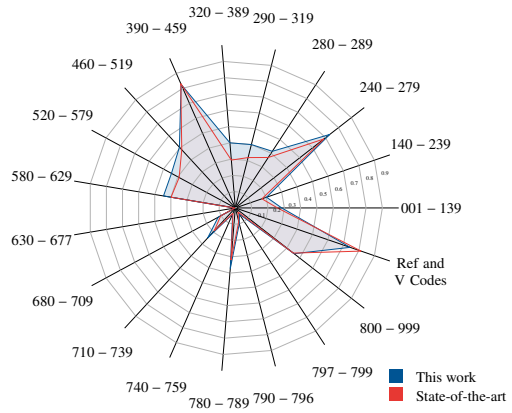


Fig. 8: Comparison with the state-of-the-art model [1] concerning the percentage of patients within an ICD-9 code group.

corresponds to the *conditions originating in the perinatal period*, and is usually assigned to neonates (age < 15), who are excluded from this study as per the defined patient cohort (see Section 3.1). Hence, our dataset does not contain any records in the ICD-9 code range of 760 – 779. Furthermore, our study classifies all the Reference (Ref) and supplemental V-codes into the same code group, to lower the computational complexity of training.

Fig. 8 depicts a radial plot comparing the statistics of our work to that of Purushotham *et al*. [1], concerning the ratio of the number of patients in a particular code group to the total number of patients in the cohort. Despite variations in the data and cohort selection, our work and the state-of-the-art work [1] share similar statistics with respect to ICD-9 code groups (see Fig. 8), thus ensuring a fair comparison of performance.

4.1 Deep Neural Architectures

For the clinical task of multi-label ICD-9 code group prediction, we employed six deep neural architectures: MLP, ConvNet, LSTM, Bi-LSTM, Conv-LSTM, and Seg-GRU (depicted in Fig. 9). We used the implementations available in the Python Keras package with the Tensorflow backend. Grid-search was used to determine the optimal values of the hyperparameters employed in the underlying deep neural models. The deep neural models were trained to minimize a cross-entropy loss (mean squared error prediction loss) function using an Adam optimizer, with a batch size of 128, for eight epochs.

4.1.1 Multi-Layer Perceptron

MLP is a fully connected feed-forward artificial neural network with multiple layers of processing elements (neurons) interacting through weighted connections. MLP offers several advantages, including fault tolerance, generalizability, adaptive learning, and parallelism. Typically, MLP consists of an input layer, one or more hidden layers, and one classification layer at the top to solve the prediction task. The input to the first layer is comprised of a $d-$dimensional embedding (topics) of $\eta_n^{(p)}$, and the output of each layer serves as the input to the subsequent layer. Formally, MLP is a transformation function $g : \mathbb{R}^n \longrightarrow \mathbb{R}^k$, where k is the size of the output vector (19, here). The transformation from layer l ($y^{(l)}$) to the following layer $l+1$ ($y^{(l+1)}$) can be written as:

$$y^{(l+1)} = g^{(l)}(y^{(l)}) = s^{(l)}(W^{(l)} \cdot y^{(l)} + b^{(l)}) \tag{9}$$

where, $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias at layer l, and $s^{(l)}$ is a nonlinear activation function, which is usually tanh, logistic sigmoid, or Rectified Linear Unit (ReLU) function. During training, MLP uses a supervised approach called backpropagation to update and learn optimal values of weights and biases. The gradient of the loss function is calculated using backpropagation, which aids MLP to learn the internal representations, in turn allowing it to learn any arbitrary mappings within the network. The backpropagation algorithm addresses the dependencies between the target classes through a global error function, in the case of multi-label classification. In our study, we use an MLP network with one hidden layer of 75 ReLU processing units and a prediction layer of 19 sigmoid processing units (see Fig. 9a).

4.1.2 Convolutional Neural Architecture

ConvNets are a regularized variation of the deep MLP architecture which are aimed at minimal processing. They utilize layers with convolving filters, which are applied to local features. Due to their transition invariance characteristics and shared-weights architecture, ConvNets are space invariant. They are shown to be effective in a variety of NLP tasks, including semantic parsing, sentence modeling, and search query retrieval [42]. Consider that a clinical nursing note $\eta_n^{(p)}$ is modeled to produce an $n-$dimensional embedding $\mathcal{E}_{1:n}^{(p)} \in \mathbb{R}^n$, where $t_{i:i+j}$ refers to the concatenation of terms $t_i, t_{i+1}, \ldots, t_{i+j}$. A convolution operation involving a filter $f \in \mathbb{R}^{hn}$ is applied to a window of h terms to produce a new feature ($\mathcal{F}_i$) (10), where $s^{(l)}$ and $b^{(l)}$ are the nonlinear activation and bias at layer l. To produce a feature map ($\mathcal{F} \in \mathbb{R}^{n-h+1}$), we now apply this filter to every possible window of terms in the embedding $\{\mathcal{E}_{1:h}^{(p)}, \mathcal{E}_{2:h+1}^{(p)}, \ldots, \mathcal{E}_{n-h+1:n}^{(p)}\}$ (11).

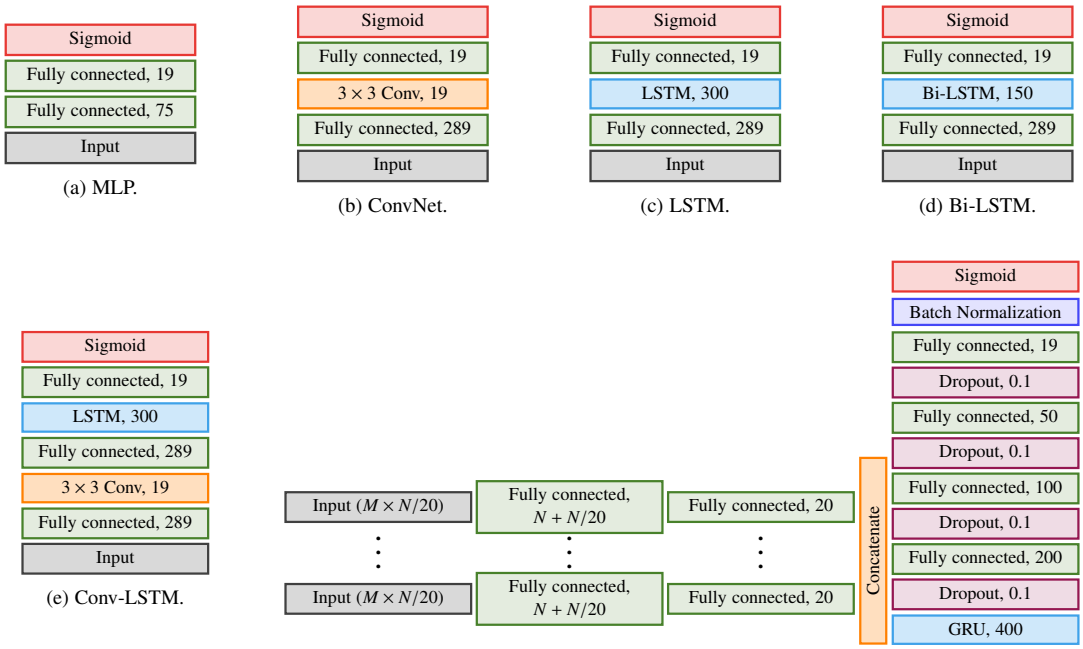
$$\mathcal{F}_i = s^{(l)}(f \cdot \mathcal{E}_{i:i+h-1}^{(p)} + b^{(l)}) \tag{10}$$

$$\mathcal{F} = [\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_{n-h+1}] \tag{11}$$

Here, we extract one feature from one filter, and this process can be extended to obtain multiple features from multiple filters (of varying sizes). The features from the penultimate layer are passed to a fully connected layer using a nonlinear activation function. We employed one fully connected layer of 289 ReLU processing units, and one ConvNet layer with 3×3 convolution window and a feature map size of 19. Finally, the code group prediction is facilitated by a fully connected layer of 19 sigmoid processing units (see Fig. 9b).

4.1.3 Long Short-Term Memory Architectures

LSTM is a special type of RNN that effectively overcomes the gradient vanishing problem and captures long-term dependencies, which is crucial to predict the code groups using nursing notes accurately. To determine the extent to which LSTM memory units must memorize the current state (c_t) and retain the previous state (c_{t-1}), LSTMs employ an adaptive gating mechanism. More specifically, an LSTM memory unit is composed of four gates: the input gate (i), the forget gate (f), the output gate (o), and the candidate value for the cell state (g). An LSTM update at a time step t and layer l can be formulated as shown in (12) through (14),



(a) MLP. (b) ConvNet. (c) LSTM. (d) Bi-LSTM. (e) Conv-LSTM. (f) Seg-GRU.

Fig. 9: Schematic overview of the deep neural architectures employed in this study.

where $\odot$ denotes the Hadamard product, $y_t^{(l)}$ is the output at a time step t, and $W^{(l)} \in \mathbb{R}^{4n \times 2n}$ is the weight matrix at a layer l.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^{(l)} \begin{pmatrix} y_t^{(l-1)} \\ y_{t-1}^{(l)} \end{pmatrix} \tag{12}$$

$$c_t^{(l)} = f \odot c_{t-1}^{(l)} + i \odot g \tag{13}$$

$$y_t^{(l)} = o \odot \tanh(c_t^{(l)}) \tag{14}$$

In nursing notes, the semantic meaning of a term is often influenced by the terms before and after it. Thus, the predictability of diagnostic codes can be enhanced by accessing both past and future input features for a given time—a Bi-LSTM network can facilitate such functionality. In doing so, we can effectively utilize future features (via backward states) and past features (via forward states) for a particular time frame. The precise form of the classification relationship obtained by summarizing future and past term representations can be given by:

$$\overrightarrow{h_t^{(l)}} = s^{(l)}(\overrightarrow{W^{(l)}} \cdot x_t^{(l)} + \overrightarrow{V^{(l)}} \cdot \overrightarrow{h_{t-1}^{(l)}} + \overrightarrow{b}) \tag{15}$$

$$\overleftarrow{h_t^{(l)}} = s^{(l)}(\overleftarrow{W^{(l)}} \cdot x_t^{(l)} + \overleftarrow{V^{(l)}} \cdot \overleftarrow{h_{t+1}^{(l)}} + \overleftarrow{b}) \tag{16}$$

$$y_t^{(l)} = s'^{(l)}(U \cdot h_t^{(l)} + k) = s'^{(l)}(U \cdot [\overrightarrow{h_t^{(l)}}; \overleftarrow{h_t^{(l)}}] + k) \tag{17}$$

where $y_t^{(l)}$ is the output state at a time step t, $h_t^{(l)}$ is the hidden state at a time step t, $s^{(l)}$ and $s'^{(l)}$ are the activation functions, $W^{(l)}$, $V^{(l)}$, and $U^{(l)}$ are the weight matrices, and b is the bias at a layer l.

We set the dimensions of the embedding and LSTM hidden state to 289 (17 time steps with 17 features each) and 300 respectively. A sigmoid activation of the final LSTM output facilitates the multi-label classification (see Fig. 9c). The Bi-LSTM architecture is similar to that of LSTM, except that Bi-LSTM employs two LSTM layers of 150 hidden states each (see Fig. 9d).

4.1.4 *Convolutional Long Short-Term Memory Architecture*

A convolution layer effectively extracts the high-level features from a given precomputed embedding of a clinical nursing note. However, to capture the long-term dependencies in the nursing notes, we need a substantial number of convolutional layers—such dependencies are easily captured and retained by an LSTM network. Thus, a hybrid Conv-LSTM architecture both captures the high-level features and retains the long-term dependencies over time. We used a hybrid Conv-LSTM network with one fully connected layer of 289 ReLU processing units, one ConvNet layer with 3×3 convolution window and a feature map size of 19, followed by another fully connected layer of 289 ReLU processing units, and an LSTM layer with 300 hidden nodes. ICD-9 code group prediction is facilitated by a sigmoid activation of the final LSTM output (see Fig. 9e).

4.1.5 *Segment-level Gated Recurrent Unit*

GRUs are a gating mechanism in RNNs, similar to LSTM networks (with output gate) but with fewer parameters, as it lacks an output gate. Each recurrent unit in a GRU network adaptively captures dependencies of different time scales. A GRU memory unit is composed of two gates: the reset gate (r) and the update gate (z). A precise form of the GRU update can be formulated as:

$$\begin{pmatrix} r \\ z \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \end{pmatrix} W^{(l)} \begin{pmatrix} h_t^{(l-1)} \\ h_{t-1}^{(l)} \end{pmatrix} \tag{18}$$

$$\tilde{h}_t^{(l)} = tanh(V^{(l)} \cdot h_t^{(l-1)} + U^{(l)}(r \odot h_{t-1}^{(l)})) \tag{19}$$

$$h_t^{(l)} = (1 - z) \odot h_{t-1}^{(l)} + z \odot \tilde{h}_t^{(l)} \tag{20}$$

where $\odot$ denotes the Hadamard product, $h_t^{(l)}$ is the hidden state at a time step t, and $W^{(l)}$, $V^{(l)}$, and $U^{(l)}$ are the weight matrices at a layer l. The GRU computes the candidate hidden state $\tilde{h}_t^{(l)}$ and then smoothly extrapolates it (gated by the update gate).

We employ a segment-level GRU, where the input embedding of a nursing note (of size N) is split column-wise into 20 segments, followed by a fully connected layer of $N + N/20$

TABLE 2: Code group prediction using the data in the nursing notes aggregated using *FarSight*.

Data model	Classifier	Performance scores				
		ACC	MCC	F1	AUPRC	AUROC
Doc2Vec $(140,792 \times 500)$	MLP	0.7873 ± 0.0006	0.5587 ± 0.0011	0.7103 ± 0.0027	0.6577 ± 0.0012	0.7795 ± 0.0014
	ConvNet	0.8053 ± 0.0005	0.5938 ± 0.0011	0.7332 ± 0.0006	0.6810 ± 0.0011	**0.7967 ± 0.0004**
	LSTM	0.7986 ± 0.0016	0.5804 ± 0.0025	0.7250 ± 0.0026	0.6705 ± 0.0027	0.7885 ± 0.0016
	Bi-LSTM	0.8018 ± 0.0008	0.5861 ± 0.0018	0.7265 ± 0.0043	0.6758 ± 0.0023	0.7906 ± 0.0026
	Conv-LSTM	**0.8069 ± 0.0023**	**0.5961 ± 0.0040**	**0.7338 ± 0.0022**	**0.6824 ± 0.0039**	0.7962 ± 0.0019
	Seg-GRU	0.7779 ± 0.0020	0.5332 ± 0.0051	0.6794 ± 0.0054	0.6505 ± 0.0030	0.7605 ± 0.0035
NMF−BoW $(140,792 \times 150)$	MLP	0.7829 ± 0.0006	0.5498 ± 0.0009	0.7029 ± 0.0016	0.6530 ± 0.0017	0.7744 ± 0.0007
	ConvNet	0.7965 ± 0.0007	0.5750 ± 0.0013	0.7187 ± 0.0041	0.6688 ± 0.0018	0.7860 ± 0.0020
	LSTM	0.7921 ± 0.0005	0.5652 ± 0.0016	0.7093 ± 0.0030	0.6638 ± 0.0018	0.7794 ± 0.0017
	Bi-LSTM	0.7894 ± 0.0007	0.5596 ± 0.0015	0.7042 ± 0.0020	0.6619 ± 0.0017	0.7758 ± 0.0012
	Conv-LSTM	**0.8048 ± 0.0021**	**0.5897 ± 0.0042**	**0.7240 ± 0.0034**	**0.6806 ± 0.0031**	**0.7911 ± 0.0024**
	Seg-GRU	0.7945 ± 0.0063	0.5666 ± 0.0137	0.7039 ± 0.0114	0.6698 ± 0.0057	0.7772 ± 0.0080
NMF−TW $(140,792 \times 150)$	MLP	0.7953 ± 0.0004	0.5740 ± 0.0010	0.7167 ± 0.0010	0.6696 ± 0.0015	0.7850 ± 0.0005
	ConvNet	0.8174 ± 0.0006	0.6181 ± 0.0008	0.7489 ± 0.0016	0.6948 ± 0.0014	0.8091 ± 0.0014
	LSTM	0.8129 ± 0.0015	0.6062 ± 0.0028	0.7347 ± 0.0020	0.6908 ± 0.0024	0.7992 ± 0.0012
	Bi-LSTM	0.8076 ± 0.0016	0.5952 ± 0.0034	0.7280 ± 0.0037	0.6839 ± 0.0024	0.7936 ± 0.0024
	Conv-LSTM	**0.8282 ± 0.0023**	**0.6368 ± 0.0042**	**0.7562 ± 0.0021**	**0.7089 ± 0.0046**	**0.8157 ± 0.0019**
	Seg-GRU	0.8249 ± 0.0021	0.6273 ± 0.0050	0.7434 ± 0.0057	0.7089 ± 0.0019	0.8073 ± 0.0040
NMF−BoW with SC $(140,792 \times 100)$	MLP	0.7820 ± 0.0004	0.5476 ± 0.0007	0.7011 ± 0.0008	0.6517 ± 0.0013	0.7735 ± 0.0006
	ConvNet	0.7956 ± 0.0002	0.5731 ± 0.0007	0.7174 ± 0.0021	0.6672 ± 0.0017	0.7852 ± 0.0011
	LSTM	0.7905 ± 0.0004	0.5619 ± 0.0003	0.7066 ± 0.0035	0.6623 ± 0.0020	0.7777 ± 0.0019
	Bi-LSTM	0.7889 ± 0.0009	0.5598 ± 0.0005	0.7076 ± 0.0040	0.6598 ± 0.0024	0.7774 ± 0.0023
	Conv-LSTM	**0.8003 ± 0.0015**	**0.5817 ± 0.0033**	**0.7218 ± 0.0038**	**0.6735 ± 0.0024**	**0.7885 ± 0.0027**
	Seg-GRU	0.7918 ± 0.0041	0.5622 ± 0.0087	0.7034 ± 0.0112	0.6659 ± 0.0022	0.7763 ± 0.0070
NMF−TW with SC $(140,792 \times 100)$	MLP	0.7961 ± 0.0003	0.5753 ± 0.0009	0.7175 ± 0.0010	0.6703 ± 0.0016	0.7856 ± 0.0006
	ConvNet	0.8192 ± 0.0006	0.6199 ± 0.0025	0.7466 ± 0.0043	0.6983 ± 0.0013	0.8077 ± 0.0023
	LSTM	0.8142 ± 0.0014	0.6087 ± 0.0034	0.7367 ± 0.0050	0.6918 ± 0.0016	0.8003 ± 0.0030
	Bi-LSTM	0.8096 ± 0.0006	0.5998 ± 0.0012	0.7318 ± 0.0030	0.6860 ± 0.0011	0.7961 ± 0.0019
	Conv-LSTM	**0.8343 ± 0.0031**•	**0.6459 ± 0.0073** • ○	**0.7602 ± 0.0068** • ○	**0.7170 ± 0.0045** • ○	**0.8192 ± 0.0046** • ○
	Seg-GRU	0.8285 ± 0.0028	0.6350 ± 0.0064	0.7502 ± 0.0060	0.7131 ± 0.0034	0.8120 ± 0.0048

If the result is significantly higher (through a two-tailed paired samples Wilcoxon signed-rank test at a significance level of 5%) than the best performing deep neural (machine learning) model on naïvely aggregated data, it is indicated using a • (○).

ReLU processing units, and another fully connected layer of 20 ReLU processing units. The outputs from various segments are then concatenated channel-wise and are flattened. Regularization prevents co-adaptation of the hidden units and is hence necessary. The flattened output is passed through a series of 0.1 dropouts and fully connected ReLU processing units for regularization. Finally, the obtained output is subject to batch normalization to stabilize the network and reduce the covariance shift. A sigmoid activation of the normalized output facilitates the prediction (see Fig. 9f).

4.2 Performance Benchmarking and Discussion

To validate our approach, we performed exhaustive benchmarking experiments on the clinical nursing notes obtained from the MIMIC-III database as per the defined cohort. A significant challenge was the multi-label prediction, where a set of probable ICD-9 code groups were to be predicted for a given clinical nursing note. To assess the predictability of the proposed approaches, we employ a pair-wise comparison of the actual and the predicted diagnostic code group sets, performed via five-fold cross-validation (the means and the standard errors of the mean are presented). Furthermore, to accurately assess the performance of the proposed methods, we employed five standard evaluation metrics including Accuracy (ACC), MCC score, F1 score, Area Under the Precision-Recall Curve (AUPRC), and Area Under the ROC Curve (AUROC). In this study, a pairwise comparison of the predicted and actual diagnostic code groups is presented.

The results of our experiments and the related studies are tabulated in Tables 2, 3, and 4. In Table 2, the performance of the proposed modeling approaches that are built on *FarSight*-aggregated clinical nursing data is summarized. Table 3 shows the performance of all the proposed modeling approaches built on data obtained by naïvely aggregating the patients' nursing notes using their identification numbers. We observe that the NMF−TW with SC approach built on *FarSight*-aggregated data and modeled using Conv-LSTM consistently outperforms other data modeling and classification approaches with respect to all the metrics. Also, the performance of the proposed models drastically increased by 2.47% in ACC, 16.07% in MCC, 13.43% in F1, 16.13% in AUPRC, and 6.50% in AUROC when the data was aggregated using the *FarSight* long-term aggregation mechanism. Furthermore, we compared the actual and predicted (using Conv-LSTM trained on NMF−TW with SC representations) number of clinical notes that received a particular diagnostic code group in Fig. 10. It was observed that the diagnostic code ranges including $001-139$, $280-289$, $320-389$, $460-519$, $630-677$, and $780-789$ had less than 100 mismatches ($< 0.007\%$); $520-579$, $580-629$, and $800-999$ had less than 500 mismatches ($< 0.35\%$) between the actual and predicted ICD-9 code groups across $140,792$ nursing notes. We also remarked that the maximum number of mismatches (over $3,500$) corresponded to Ref and V-codes ($4,078$ (2.90%)), and the code range of $710-739$ ($4,366$ (3.10%)). Note that the statistics presented above were measured as the maximum mismatches across all the cross-validation folds. Table 4 illustrates the ICD-9 code group prediction performance of conventional machine learning models including K-Nearest Neighbors (KNN) with $K = 15$, LR as One-vs-Rest (OvR) with stochastic average gradient solver, Support Vector Machines (SVM) as OvR with

TABLE 3: Code group prediction (with deep learners) using nursing notes aggregated naïvely by patient identification numbers.

Data model	Classifier	Performance scores				
		ACC	MCC	F1	AUPRC	AUROC
Doc2Vec $(6,532 \times 500)$	MLP	0.7898 ± 0.0031	0.5195 ± 0.0088	0.6542 ± 0.0069	0.5914 ± 0.0089	0.7556 ± 0.0030
	ConvNet	0.7729 ± 0.0028	0.4841 ± 0.0049	0.6341 ± 0.0056	0.5679 ± 0.0056	0.7399 ± 0.0036
	LSTM	**0.8018 ± 0.0030**	**0.5427 ± 0.0062**	**0.6731 ± 0.0110**	**0.6098 ± 0.0054**	**0.7634 ± 0.0067**
	Bi-LSTM	0.7964 ± 0.0033	0.5308 ± 0.0083	0.6673 ± 0.0081	0.6003 ± 0.0094	0.7594 ± 0.0055
	Conv-LSTM	0.7989 ± 0.0027	0.5321 ± 0.0044	0.6604 ± 0.0035	0.6050 ± 0.0039	0.7554 ± 0.0030
	Seg-GRU	0.7673 ± 0.0046	0.4558 ± 0.0121	0.5991 ± 0.0109	0.5533 ± 0.0092	0.7179 ± 0.0064
NMF−BoW $(6,532 \times 150)$	MLP	0.7810 ± 0.0026	0.4995 ± 0.0066	0.6179 ± 0.0070	0.5838 ± 0.0067	0.7354 ± 0.0030
	ConvNet	**0.7972 ± 0.0029**	**0.5392 ± 0.0085**	**0.6555 ± 0.0083**	**0.6068 ± 0.0072**	**0.7596 ± 0.0053**
	LSTM	0.7801 ± 0.0042	0.4960 ± 0.0074	0.6225 ± 0.0066	0.5776 ± 0.0085	0.7366 ± 0.0041
	Bi-LSTM	0.7776 ± 0.0042	0.4904 ± 0.0094	0.6189 ± 0.0089	0.5735 ± 0.0070	0.7333 ± 0.0061
	Conv-LSTM	0.7870 ± 0.0033	0.5141 ± 0.0075	0.6363 ± 0.0068	0.5920 ± 0.0083	0.7449 ± 0.0038
	Seg-GRU	0.7893 ± 0.0074	0.5218 ± 0.0117	0.6436 ± 0.0026	0.5973 ± 0.0108	0.7495 ± 0.0025
NMF−TW $(6,532 \times 150)$	MLP	0.7878 ± 0.0042	0.5153 ± 0.0108	0.6321 ± 0.0103	0.5939 ± 0.0093	0.7445 ± 0.0058
	ConvNet	**0.8065 ± 0.0033**	**0.5616 ± 0.0083**	**0.6739 ± 0.0075**	**0.6231 ± 0.0073**	**0.7707 ± 0.0050**
	LSTM	0.7858 ± 0.0026	0.5083 ± 0.0076	0.6327 ± 0.0068	0.5881 ± 0.0091	0.7410 ± 0.0041
	Bi-LSTM	0.7800 ± 0.0044	0.4950 ± 0.0106	0.6249 ± 0.0054	0.5796 ± 0.0107	0.7349 ± 0.0041
	Conv-LSTM	0.7876 ± 0.0014	0.5167 ± 0.0074	0.6440 ± 0.0123	0.5936 ± 0.0064	0.7482 ± 0.0074
	Seg-GRU	0.7946 ± 0.0034	0.5304 ± 0.0103	0.6432 ± 0.0128	0.6044 ± 0.0101	0.7497 ± 0.0077
NMF−BoW with SC $(6,532 \times 100)$	MLP	0.7787 ± 0.0039	0.4910 ± 0.0091	0.6075 ± 0.0087	0.5790 ± 0.0073	0.7295 ± 0.0054
	ConvNet	**0.7956 ± 0.0028**	**0.5358 ± 0.0075**	**0.6550 ± 0.0067**	**0.6046 ± 0.0081**	**0.7599 ± 0.0039**
	LSTM	0.7770 ± 0.0012	0.4885 ± 0.0058	0.6176 ± 0.0114	0.5722 ± 0.0034	0.7336 ± 0.0069
	Bi-LSTM	0.7757 ± 0.0039	0.4832 ± 0.0104	0.6118 ± 0.0116	0.5698 ± 0.0081	0.7292 ± 0.0055
	Conv-LSTM	0.7823 ± 0.0042	0.5041 ± 0.0090	0.6331 ± 0.0076	0.5828 ± 0.0092	0.7436 ± 0.0047
	Seg-GRU	0.7800 ± 0.0042	0.4969 ± 0.0164	0.6267 ± 0.0259	0.5796 ± 0.0095	0.7411 ± 0.0160
NMF−TW with SC $(6,532 \times 100)$	MLP	0.7884 ± 0.0037	0.5162 ± 0.0105	0.6316 ± 0.0090	0.5943 ± 0.0093	0.7441 ± 0.0052
	ConvNet	**0.8062 ± 0.0029**	**0.5608 ± 0.0085**	**0.6718 ± 0.0087**	**0.6229 ± 0.0087**	**0.7695 ± 0.0046**
	LSTM	0.7847 ± 0.0033	0.5049 ± 0.0101	0.6307 ± 0.0131	0.5848 ± 0.0079	0.7404 ± 0.0073
	Bi-LSTM	0.7804 ± 0.0033	0.4949 ± 0.0097	0.6211 ± 0.0121	0.5784 ± 0.0066	0.7335 ± 0.0085
	Conv-LSTM	0.7919 ± 0.0036	0.5246 ± 0.0096	0.6436 ± 0.0095	0.5997 ± 0.0069	0.7489 ± 0.0054
	Seg-GRU	0.7928 ± 0.0049	0.5268 ± 0.0156	0.6446 ± 0.0158	0.6009 ± 0.0120	0.7511 ± 0.0093

TABLE 4: Code group prediction (with machine learners) using nursing notes aggregated naïvely by patient identification numbers.

Data model	Classifier	Performance scores				
		ACC	MCC	F1	AUPRC	AUROC
BoW $(6,532 \times 14,665)$	KNN	0.7741 ± 0.0023	0.4912 ± 0.0025	0.6320 ± 0.0019	0.5454 ± 0.0022	0.7405 ± 0.0019
	LR as OvR	**0.8056 ± 0.0019**	**0.5418 ± 0.0026**	**0.6668 ± 0.0012**	**0.6094 ± 0.0026**	**0.7348 ± 0.0012**
	SVM as OvR	0.7549 ± 0.0015	0.5064 ± 0.0016	0.6148 ± 0.0021	0.5789 ± 0.0018	0.6452 ± 0.0007
	RF ensemble	0.7255 ± 0.0027	0.4067 ± 0.0012	0.5182 ± 0.0025	0.5133 ± 0.0023	0.6670 ± 0.0014
TW $(6,532 \times 14,665)$	KNN	0.7866 ± 0.0012	0.5306 ± 0.0032	0.6697 ± 0.0021	0.5920 ± 0.0025	0.7689 ± 0.0016
	LR as OvR	**0.8143 ± 0.0014**	**0.5845 ± 0.0035**	**0.6874 ± 0.0030**	**0.6378 ± 0.0032**	**0.7804 ± 0.0017**
	SVM as OvR	0.7414 ± 0.0015	0.4007 ± 0.0036	0.5207 ± 0.0028	0.5249 ± 0.0026	0.6801 ± 0.0015
	RF ensemble	0.7653 ± 0.0011	0.4449 ± 0.0031	0.5484 ± 0.0023	0.5517 ± 0.0024	0.6951 ± 0.0013

radial basis function kernel, and Random Forest (RF) ensemble with 100 trees (maximum depth of 2), using nursing notes aggregated naïvely by patient identification numbers. This study does not include BoW or TW modeling on *FarSight*-aggregated data, owing to the high-dimensionality and sparsity of such statistical transformations of the underlying corpus $(140,792 \times 32,687)$. From Fig. 11, it is evident that the proposed deep learners trained on *FarSight*-aggregated data outperform the conventional machine learners and deep learners trained on naïvely aggregated data.

The ability of a model to effectively capture the true and false positives and negatives in risk assessment is of paramount importance, owing to the critical nature of the task itself. AUPRC measures the number of true positives from the set of positive predictions, while AUROC captures the hit and miss rates. AUPRC varies with a change in the ratio of the target classes in the data and hence, is more revealing than AUROC in this context (see Fig. 8) [43]. Precision captures the proportion of the patient records that the proposed model predicted to have a risk that actually had a risk, while recall expresses the ability to find all the patients at risk. These are captured using the F1 score, while the MCC score accounts for the true positives, and false positives and negatives, thus serving as a balanced measure even with class imbalance.

To facilitate the prediction of clinical outcomes, most existing works, including the state-of-the-art model (considered for benchmarking) [1], rely on the structured nature of the EHRs, modeled as feature sets. From Fig. 11, it can be observed that our model built on the unstructured nursing text and modeled using *FarSight*-aggregated data, significantly outperforms the state-of-the-art model by 19.34% in AUPRC and 5.41% in AUROC, and the hierarchical attention GRU model [10] by 35.71% in F1 score. Moreover, most of the existing works benchmarked their performance only on the AUPRC and AUROC metrics, while neglecting to assess the performance of their models with metrics most suited in the cases of imbalanced data, as is the case with most of the real-world data. We argue that the reliability and other critical aspects of the underlying CDSS can be accurately

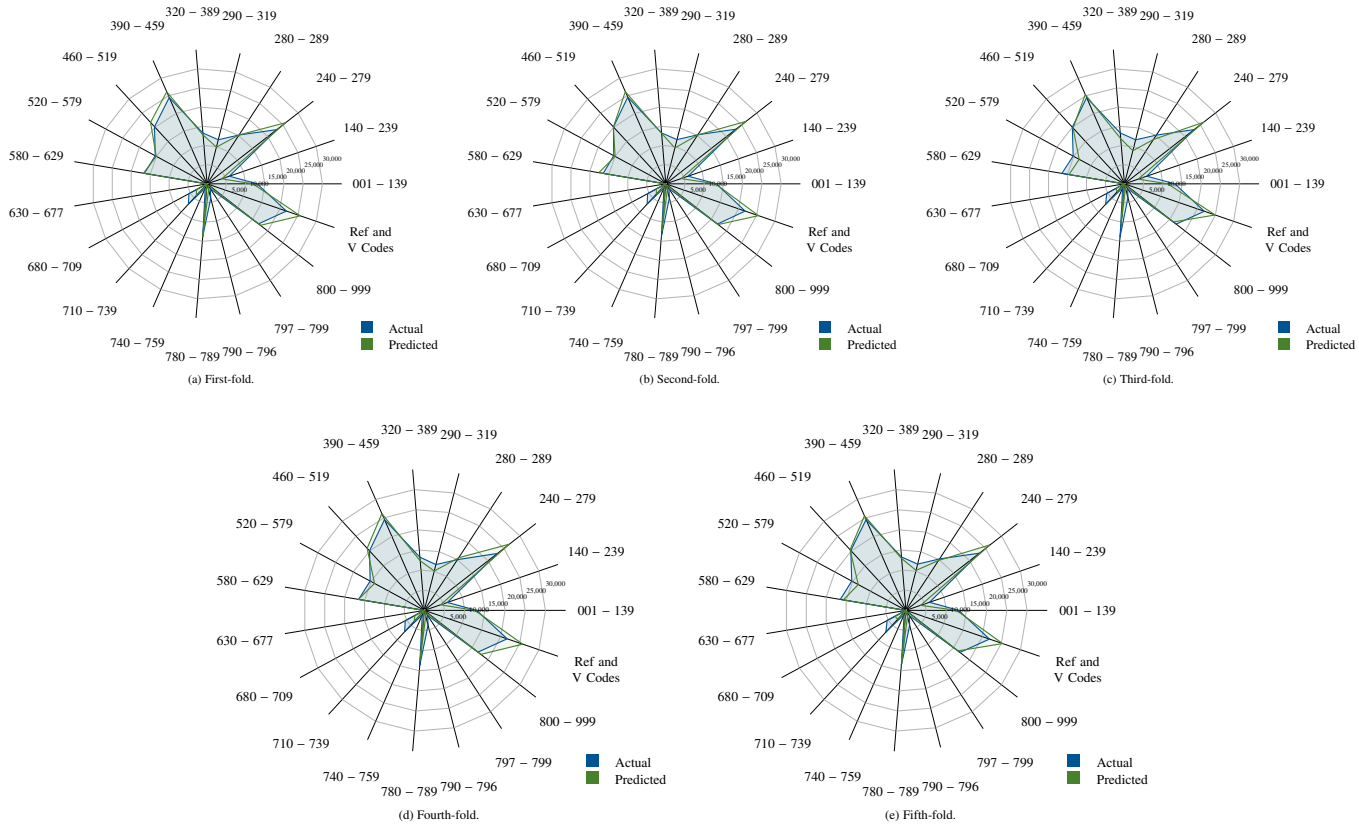


Fig. 10: Comparison of the number of nursing notes concerning actual and predicted (using Conv-LSTM trained on NMF−TW with SC representations) ICD-9 code groups across the cross-validation folds.

and explicitly captured by assessing the performance of a model using targeted metrics like ACC, F1, and MCC scores, which are incorporated in our work. This offers significant insights into the hit and miss ratios, true positive rates, and other capabilities of the underlying CDSS, vital when dealing with real-world clinical data. The NMF−TW with SC model was able to capture discriminative features of the nursing notes needed for the classifier to learn and generalize. Also, the *FarSight* aggregation strategy effectively facilitates accurate risk assessment, well in advance, with an overall accuracy of 83.3%. Thus, a CDSS equipped with the predictive capabilities of *FarSight*-aggregation and NMF−TW with SC modeling could demonstrate evidence-based and patient-centric risk assessments.

In addition to establishing the superiority of the proposed modeling strategy over other existing approaches, we also analyzed the trends in diagnostic code group prediction performance (using Conv-LSTM trained on representations obtained using NMF−TW with SC) with the variations in the percentage of nursing notes used as training data. As graphed in Fig. 12, it can be observed that the prediction performance constantly increases with an increase in the amount of available information for training. It is also interesting to note that the proposed system achieved an overall accuracy of 79%, AUPRC of 0.6606 (9.95% more than the state-of-the-art model), and AUROC of 0.7788 (0.21% more than the state-of-the-art model), using 14,079 nursing notes, which corresponding to just 10% of the training data. These results signify the quality of patient-specific information present in the raw and unstructured nursing notes, and the effectiveness of the proposed aggregation and modeling strategies in extracting and leveraging such information for facilitating accurate ICD-9 code group prediction. Furthermore, these observations corroborate the suitability of *FarSight* for clinical decision support in real-world hospital scenarios, especially in developing countries with limited resources and low structured EHR adoption rates.

4.3 AI-assisted Clinical Decision Support: Interpretability vs. Accuracy

The trustworthiness of computational CDSSs often poses a two-fold dependency that includes predictive accuracy of the underlying models and their potential to justify the suggested recommendations. To this end, even the most potent theory-agnostic neural learning models often prove inadequate in justifying the decisions made, i.e., they are programmed to learn from large amounts of data, while their understanding of the causal structure of the underlying problem remains obscured. Deep neural models are trained to learn the weights on the neurons in the network from given supervised instances, so as to automatically construct a mathematical model to map the input data (here, raw nursing notes) to target labels (i.e., diagnostic code groups). Such neural systems mimic the functioning of neurons in the human brain and can be trained on millions of inputs to facilitate high predictive accuracy. Despite such performance, they are inscrutable to humans; even when a set of highly-weighted features are extracted from the trained model, the relationship between those features and the target variable could be opaque and indirect. Moreover, varying the input settings and permuting any segment of the data can result in the construction of significantly different models. A more critical issue arises from the lack of relation between the associations gleaned from the underlying data and the causal relationships. For example, even when the neural system associates smoking

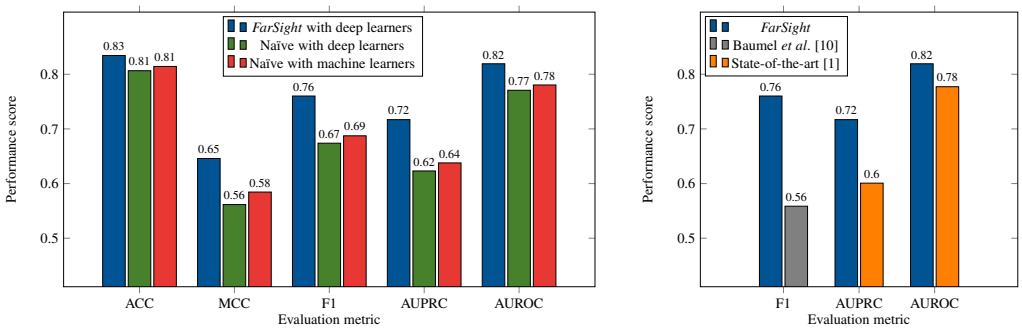


Fig. 11: Comparison of the best performing models (aggregated with and without *FarSight*) and benchmarking works in ICD-9 code group prediction.

with cancer, it is challenging to reason as to how smoking can be causally related to cancer. Therefore, despite logical decision-making facilitated by machine learning systems, they still far short of expectations in terms of accountability.

However, the ability to intervene effectively in medical scenarios is often derived from experience rather than the exploitation of causal relationships—such experience often precedes the ability to understand why such interventions work [44]. Furthermore, what interpretability amounts to, is extremely subjective in the sense that human decisions are interpretable as we can rationalize such decisions; but, such reasoning can be subject to several post-hoc realizations, in which sense, all the machine learning models are also interpretable. However, if interpretability measures the capability of simulating the decision-making model, then the underlying machine learning models can be time stepped through every computation involved in a prediction; however, it is impractical to interpret complex neural models this way, and in a similar sense, human decisions are also equally uninterpretable. Thus, any recommendations where interpretability is prioritized over diagnostic accuracy are counter-productive in clinical systems, as the patients whose diseases go undiagnosed are heavily affected.

The causal relevance might not be related to the associations mined using highly predictive models, thus resulting in the use of neural predictive systems for purposes to which they are not suited. Based on this reasoning, we emphasize on the accuracy of the predicted ICD-9 diagnostic code groups (rather than the interpretability of the decisions made) alongside the diagnoses evaluated by physicians through several baseline standards, including patients' history, examination, and investigations. In most cases, especially in developing countries with low structured EHR adoption rates, physicians' diagnoses do not follow a strict coding scheme. Therefore, the strategy presented in our work could facilitate and aid the physician in documenting the recommended diagnoses in a more consistent format, thus enabling effective clinical decision support, demographic and epidemiology studies, and healthcare insurance policies. Furthermore, the physician depends only on the patients' history, examination, and investigations while arriving at the diagnosis; the richness of information resulting from numerous assessments of continuous monitoring by the nursing staff is usually neglected. Thus, employing the strategy and models presented in this paper, which utilize such valuable information, would complement the physician's effort in arriving at a better and more accurate diagnosis.

4.4 Towards ICD-10-CM Diagnostic Coding Systems

Despite the pervasive use of ICD-9 medical taxonomy in healthcare systems, it is not robust in serving the medical needs of the future. Owing to the limited number of diagnostic codes and the rigid structure of coding, ICD-9 provides limited support to hospital inpatient procedures and patients' medical conditions. Hence, there is a need to transition to ICD-10 Clinical Modification (ICD-10-CM), which enhances the quality of medical data for the development of CDSSs, epidemiological research, processing insurance claims, tracking health conditions, and many others. With a granularity nearly five times higher than that of ICD-9 taxonomy, ICD-10-CM enables significant specificity in identifying disease conditions across several dimensions, including severity, laterality, and complexity. Moreover, ICD-10-CM offers extensive support in classifying poisonings, injuries, and external causes. The ontology also introduces several new concepts missing in ICD-9, such as alcohol level, blood type, and underdosing.

In the ICD-9 coding taxonomy, $3,000$ parent concepts are available, which extend to approximately $14,000$ alphanumeric leaf codes. Adding to the ICD-9 hierarchy, ICD-10-CM enables nearly $30,000$ parent concepts, rolling out to about $70,000$ leaf concepts, with a depth level of up to seven. The increased

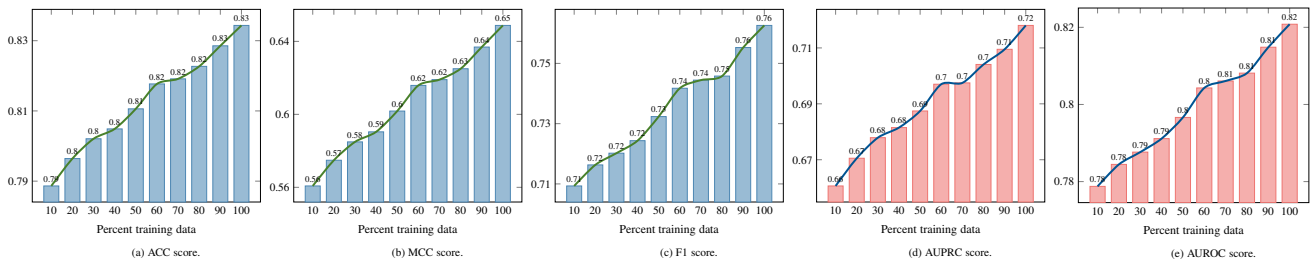


Fig. 12: Plots depicting the effect of variations in the percent of nursing notes used for training on the ICD-9 code group prediction performance facilitated by *FarSight*-aggregated, NMF−TW with SC modeled data, classified using Conv-LSTM neural model.

specificity of ICD-10-CM allows for added flexibility and clarity in disease representations. For instance, `S52.521` denotes *torus fracture of lower end of right radius*, while `S52.521A` extends on `S52.521`, by indicating *the initial encounter for torus fracture of lower end of right radius*. Additionally, ICD-10 ontology presents several combination codes that allow for the reporting of a single code when expressing multiple diagnostic elements. For instance, consider `I26.01`, an ICD-10 code used to indicate *septic pulmonary embolism with acute cor pulmonale*, which in ICD-9 would be denoted using `415.0` and `415.12`. Hence, from a research perspective, it is vital to develop automated and assistive approaches that facilitate adaptation of the existing legacy ICD-9 data quality and analytic algorithms to ICD-10-CM.

While the enormous increase in the number of concepts available in ICD-10-CM is an attempt to solve the limitations in ICD-9, several recent studies have shown that more concepts do not necessarily imply better representation of diseases, especially concerning the questionable importance of certain newly introduced concepts (e.g., `W59.22` indicates *being struck by turtle*, `Y92.253` represents *occurrence of external causes at the opera house*, and others) [45]. Additional concerns include the rising costs of medical coders' training and hospital systems' adaptation to ICD-10-CM, along with the loss of coding accuracy during the transition from ICD-9 to ICD-10. Thus, in addition to designing code-to-code mapping systems, it is vital to develop intelligent mechanisms that extend clinical decision support through accurate ICD-10 code prediction capabilities.

5 Summary

In this study, we employed *FarSight*, a long-term aggregation strategy to detect the onset of the disease with the earliest recorded symptoms, vital in channeling prioritized care, for effective clinical decision support. Our approach is built on the valuable patient-specific information extracted from unstructured nursing notes, by specifically addressing the challenges of longitudinality, heterogeneity, voluminosity, and complexity in structure. Our model leverages vector space and NMF topic modeling to deduce the most representative feature space, which was then used to enable accurate disease prediction using deep neural architectures. The proposed approaches were benchmarked on five standard evaluation metrics to assess their performance when used in multiple contexts—predictability, hit and miss ratios, reliability, and the ability to perform well in the cases of imbalanced data. The proposed NMF−TW with SC model on *FarSight*-aggregated data captured the rich information in the informally-written nursing notes and outperformed the structured EHR-based state-of-the-art model by 19.34% in AUPRC and 5.41% in AUROC. Furthermore, we observed a drastic improvement in the performance of *FarSight*-aggregated unstructured modeling over the naïve note aggregation strategy. Moreover, our proposed model eliminates the dependency on structured EHRs, which is a significant roadblock in developing countries with low structured EHR adoption rates.

Our approach effectively assesses the risk, well in advance; however, we intend to explore certain further enhancements. First, our modeling strategy, while eliminating the need for structured EHR data, models the unstructured nursing text under the assumption that nursing notes record all the clinical assessments. Second, the current modeling strategy is specific to MIMIC-III data, and this study does not account for real-time clinical data. As a part of the future work, we intend to model structured EHR measurements alongside the rich information from the nursing notes. We also aim at extending the proposed strategies to model real-time streaming clinical data, to account for the need for time-aware and reliable CDSS models in real-world scenarios, through the evaluation and analysis of various clinical markers obtained instantaneously. The CDSS would facilitate demographic and epidemiology studies, time-series clinical data modeling, and real-time decision support. Furthermore, the predictive model would be subject to continuous quality improvement through auditing of system performance and reassessment of accuracy in light of varying clinical contexts. Thus, the proposed system modeled as a clinician-oriented interface can assist caregivers in deriving patient-specific and evidence-based real-time assessments, and can be effortlessly adapted to an existing hospital information system. We also propose to extend the learnings and observations of this study to design and develop a more robust CDSS through the use of ICD-10-CM medical taxonomy.

References

[1] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *Journal of biomedical informatics*, 2018.

[2] P. S. Mathew and A. S. Pillai, "Big data solutions in healthcare: Problems and perspectives," in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE, 2015, pp. 1–6.

[3] S. Dubois, N. Romano, D. C. Kale, N. Shah, and K. Jung, "Learning effective representations from clinical notes," *arXiv preprint arXiv:1705.07025*, 2017.

[4] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, 2016.

[5] C. W. Hanson and B. E. Marshall, "Artificial intelligence applications in the intensive care unit," *Critical care medicine*, vol. 29, no. 2, pp. 427–435, 2001.

[6] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, and W. 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*, vol. 29, no. 2, pp. 291–296, 2001.

[7] R. 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*. Springer, 2016, pp. 295–313.

[8] A. E. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: a mortality prediction case study," in *Machine Learning for Healthcare Conference*, 2017, pp. 361–376.

[9] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *arXiv preprint arXiv:1703.07771*, 2017.

[10] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classification of patient notes: case study on icd code assignment," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] A. Perer and J. Sun, "Matrixflow: temporal network visual analytics to track symptom evolution during disease progression," in *AMIA annual symposium proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 716.

[12] D. Gotz, H. Stavropoulos, J. Sun, and F. Wang, "Icda: a platform for intelligent care delivery analytics," in *AMIA annual symposium proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 264.

[13] R. Caruana, S. Baluja, and T. Mitchell, "Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation," in *Advances in neural information processing systems*, 1996, pp. 959–965.

[14] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis *et al.*, "An evaluation of machine-learning methods for predicting pneumonia mortality," *Artificial intelligence in medicine*, vol. 9, no. 2, pp. 107–138, 1997.

[15] L. A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, and R. Mark, "A database-driven decision support system: customized mortality prediction," *Journal of personalized medicine*, vol. 2, no. 4, pp. 138–148, 2012.

[16] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[17] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. ACM, 2015, pp. 507–516.

[18] F. Dabek and J. J. Caban, "A neural network based model for predicting psychological conditions," in *International conference on brain informatics and health*. Springer, 2015, pp. 252–261.

[19] K. Khin, P. Burckhardt, and R. Padman, "A deep learning architecture for de-identification of patient notes: Implementation and evaluation," *arXiv preprint arXiv:1810.01570*, 2018.

[20] M. J. Cohen, A. D. Grossman, D. Morabito, M. M. Knudson, A. J. Butte, and G. T. Manley, "Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis," *Critical Care*, vol. 14, no. 1, p. R10, 2010.

[21] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1095–1103.

[22] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 85–94.

[23] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive hawkes process," in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 721–726.

[24] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PloS one*, vol. 8, no. 6, p. e66341, 2013.

[25] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[26] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Machine Learning for Healthcare Conference*, 2016, pp. 73–100.

[27] N. Y. Hammerla, J. Fisher, P. Andras, L. Rochester, R. Walker, and T. Plötz, "Pd disease state assessment in naturalistic environments using deep learning," in *29th AAAI Conference on Artificial Intelligence*, 2015.

[28] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu, "Variational recurrent adversarial deep domain adaptation," 2016.

[29] K. Feldman, N. Hazekamp, and N. V. Chawla, "Mining the clinical narrative: all text are not equal," in *2016 IEEE international conference on healthcare informatics (ICHI)*. IEEE, 2016, pp. 271–280.

[30] A. Zalewski, W. Long, A. E. Johnson *et al.*, "Estimating patient's health state using latent structure inferred from clinical time series and text," in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 449–452.

[31] G. S. Krishnan and S. Kamath, "A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data," *Applied Soft Computing*, 2019.

[32] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes," *Computer Methods and Programs in Biomedicine*, vol. 177, 2019.

[33] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang, "Automatic ICD-9 coding via deep transfer learning," *Neurocomputing*, vol. 324, pp. 43 – 50, 2019.

[34] E. L. Stone, "Clinical decision support systems in the emergency department: Opportunities to improve triage accuracy," *Journal of Emergency Nursing*, vol. 45, no. 2, pp. 220–222, 2019.

[35] G. S. Krishnan and S. S. 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*. Springer, 2018, pp. 126–134.

[36] Y. Wu, J. C. Denny, S. Trent Rosenbloom, R. A. Miller *et al.*, "A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (card)," *Journal of the American Medical Informatics Association*, vol. 24, pp. 79–86, 2016.

[37] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[38] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 952–961.

[39] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *arXiv preprint arXiv:1803.01257*, 2018.

[40] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

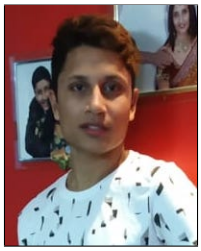
[41] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*. ACM, 2015, pp. 399–408.

[42] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[43] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[44] A. J. London, "Artificial intelligence and black-box medical decisions: Accuracy versus explainability," *Hastings Center Report*, vol. 49, no. 1, pp. 15–21, 2019.

[45] L. Manchikanti, A. D. Kaye, V. Singh, and M. V. Boswell, "The tragedy of the implementation of ICD-10-CM as icd-10: Is the cart before the horse or is there a tragic paradox of misinformation and ignorance," *Pain physician*, vol. 18, no. 4, pp. E485–E495, 2015.



Tushaar Gangavarapu is with the Automated Quality Assistance (AQuA) Machine Learning Research at Amazon.com, Inc., India. He served as the NITK Student Ambassador for the Intel Artificial Intelligence group and CodeNation for over three years. His current research interests are related to Healthcare Informatics, Natural Language Processing, Game-based Learning and Assessment, Learning Sciences, Social Multimedia and Social Network Analysis, Cognitive and Psychological Trait Modeling, and Bio-Inspired Evolutionary Computing. He is a scientific researcher at the Human Centered Computing Group (HCCG) and Healthcare Analytics and Language Engineering (HALE) research lab. He is currently focused on improving readers' content experience through automated quality assistance.

Gokul S Krishnan received his B.Tech degree in Computer Science and Engineering from Cochin University of Science and Technology, Kerala, India; and M.Tech degree in Computer Science and Engineering from VIT University, Vellore, Tamil Nadu, India. He is currently pursuing his Ph.D. at the National Institute of Technology Karnataka, Surathkal, India, specializing in Healthcare Analytics. His research interests include Healthcare Analytics, Data Science, Natural Language Processing, and Web Semantics.

Sowmya Kamath S is with the Department of Information Technology, National Institute of Technology Karnataka, Surathkal. She earned her B.Tech and M.Tech degrees from Manipal University, and Ph.D. degree from NITK Surathkal. Her research interests lie in the areas of Healthcare Analytics, Information Retrieval, Machine Learning, and Natural Language Processing. She heads the Healthcare Analytics and Language Engineering (HALE) Lab at NITK, and her current projects focus on the development of semantics/contextual modeling approaches for large-scale multimodal data like clinical data, for knowledge discovery. She received the Early Career Research Grant for her work in Healthcare Analytics from Govt. of India and is also Senior Member, IEEE.

Jayakumar Jeganathan is with the Department of General Medicine, Kasturba Medical College and Hospital, Manipal Academy of Higher Education (MAHE), Mangalore, Karnataka, India. He earned his MD in Internal Medicine from Kasturba Medical Hospital, Manipal University, India, in 2004. He is a life member of the Association of Physicians of India and Indian Medical Association. His research interests include Infectious Diseases and Diabetes Mellitus. He has several National and International publications to his credit.