Coherence-based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification

Tushaar Gangavarapu* tushaargvsg45@gmail.com
Gokul S Krishnan gsk1692@gmail.com
Sowmya Kamath S sowmyakamath@nitk.edu.in
Healthcare Analytics and Language Engineering (HALE) Lab
Department of Information Technology
National Institute of Technology Karnataka, Surathkal, Mangaluru, India

Abstract

In hospitals, critical care patients are often susceptible to various complications that adversely affect their morbidity and mortality. Digitized patient data from Electronic Health Records (EHRs) can be utilized to facilitate risk stratification accurately and provide prioritized care. Existing clinical decision support systems are heavily reliant on the structured nature of the EHRs. However, the valuable patient-specific data contained in unstructured clinical notes are often manually transcribed into EHRs. The prolific use of extensive medical jargon, heterogeneity, sparsity, rawness, inconsistent abbreviations, and complex structure of the clinical notes poses significant challenges, and also results in a loss of information during the manual conversion process. In this work, we employ two coherence-based topic modeling approaches to model the free-text in the unstructured clinical nursing notes and capture its semantic textual features with the emphasis on human interpretability. Furthermore, we present *FarSight*, a long-term aggregation mechanism intended to detect the onset of disease with the earliest recorded symptoms and infections. We utilize the predictive capabilities of deep neural models for the clinical task of risk stratification through ICD-9 code group prediction. Our experimental validation on MIMIC-III (v1.4) database underlined the efficacy of *FarSight* with coherence-based topic modeling, in extracting discriminative clinical features from the unstructured nursing notes. The proposed approach achieved a superior predictive performance when benchmarked against the structured EHR data based state-of-the-art model, with an improvement of 11.50% in AUPRC and 1.16% in AUROC.

*Corresponding author.

1 Introduction

Until recently, the healthcare industry had an inclination towards conservative approaches for the treatment and diagnosis of patients, resulting in less patient-centric and imprecise assessments (Mathew and Pillai, 2015). Intensive Care Units (ICUs) utilize the most advanced medical resources to treat and monitor critically ill patients. However, such advanced medical interventions in ICUs often make patients vulnerable to several complications (To and Napolitano, 2012). Various infections, including barotrauma, short- and long-term intubation, catheter-associated urinary tract infection, weaning errors, ventilator-associated pneumonia, gastrointestinal tract bleeding, and infections from unrecognized drug interactions, are associated with invasive ICU devices (Wollschlager and Conrad, 1988). The lack of accurate knowledge of the etiology of such complications leads to the inability to accurately stratify risk, due to which, in most cases, adequate care is provided to patients only after the development of a complication (Huddar et al., 2016). With the advent of digitization, advancement in technology, need for evidence-based medicine, increased population, and rising rates of chronic diseases, the utilization of ever-increasing heterogeneous medical data to improve the quality of life has become imperative. Specifically, ICUs are data-rich environments where several parameters of patients are monitored continuously. Such data can be vital to improve the existing Clinical Decision Support Systems (CDSSs), develop new treatments, and predict prominent clinical events and outcomes. Furthermore, such CDSSs could promote evidence-based and patient-centric treatments, resulting in reduced hospital mortality and morbidity rates, and improved risk assessment.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 1012–1022
Hong Kong, China, November 3-4, 2019.

```
Pat is 83 yo F w/PMHx for CLL and hypotens,
who was admited for an elective total hip
arthroplasty for persistent hip pain. NGT
to low cont suct. Family here to visit.

Pat initially sustained a right hip
fracture after a fall in [**2137**], and
had an ORIF performed at the time. Gave med
for pain. Has had right hip pain ever since,
and also has AVN of the right femoral head.

She came in today for elective tot hip repl.
In the OR today, patient had an estimated
1600cc EBL, and received 6u pRBC.
I/Os were 7200cc in (3.7L LR, 1.5L pRBCs).
```

Figure 1: Sample de-identified nursing note from critical care. Observe the absence of grammatical structure, informal word usage, and extensive medical jargon.

Structured medical data in the form of Electronic Health Records (EHRs) contain numerical assessments (e.g., lab results) and are amenable to standard statistical analysis (Huddar et al., 2016). However, unstructured clinical text and images also contain valuable information concerning the state of a patient. In particular, clinical nursing notes maintain objective and subjective assessments of a patient's condition. Such raw notes contain the intuitions and observations of nurses and caregivers who regularly monitor the patient. This valuable patient-specific information present in the clinical nursing notes has the potential to uncover hidden clues about the mental state (e.g., family support and mental fitness) and the health of a patient (Jo et al., 2015). Such information is not found in EHRs or elsewhere (Dubois et al., 2017). However, these notes are informally written, and modeling such notes is challenging due to their high-dimensionality, rawness, sparsity, complex linguistic and temporal nature, inconsistent abbreviations, and occurrence of rich medical jargon (a sample note is shown in Figure 1).

The voluminosity of nursing notes can be observed from the heavy-tailed distribution of the MIMIC-III nursing notes across various patients (see Figure 2), with an average of 176.49 nursing notes per patient. The presentation, analysis, and interpretation of the data present in such notes in a medically appropriate and usable format determine the competence of the underlying CDSS (Wang et al., 2018). Furthermore, there is often a need to assign multiple labels to a patient entry, owing to the diverse and manifold nature of the disease symptoms of the patients (Baumel et al., 2018). Risk stratification as ICD-9[1] code group prediction can help in predicting disease onset and its severity, thus facilitating preventive and prioritized care, and reduction of hospital mortality and morbidity rates.

With the availability of large de-identified healthcare databases such as MIMIC-III[2] (Johnson et al., 2016), modeling patient data using machine and deep learning to predict prominent clinical events and outcomes has sparked widespread interest. Early works (Tu and Guerriere, 1993; Doig et al., 1993; Grigsby et al., 1994; Clermont et al., 2001; Hanson and Marshall, 2001) have reported on the superior performance of machine learning models in forecasting the length-of-stay and mortality, for ICU patients. More recently, Pirracchio (2016) used an ensemble of several machine learning models that offered improved performance in ICU mortality prediction over various

[1]International Classification of Diseases, ninth revision.
[2]Medical Information Mart for Intensive Care.

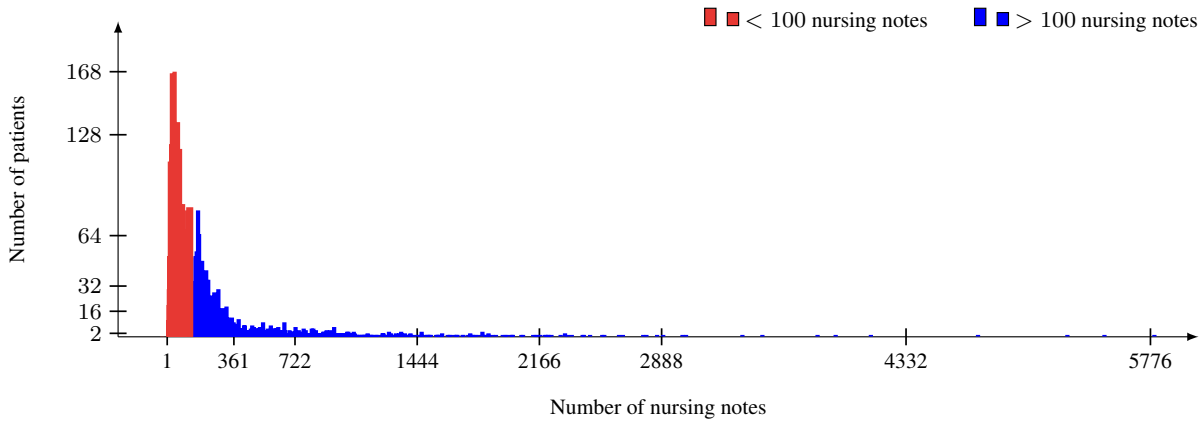


Figure 2: Distribution of the nursing notes across various MIMIC-III subjects.

severity scoring systems. Feldman et al. (2016) mined the clinical nursing, radiology, physician, and ECG narratives to study the linguistic, structural, and topical differences among them. The authors only provided a foundation for mining clinical notes effectively, and in our work, we extend their efforts by effectively modeling the underlying patient representations of the nursing text through effective topic modeling and deep neural learning. Johnson et al. (2017) extracted a set of features from the MIMIC-III database for ICU mortality prediction and compared several state-of-the-art models against gradient boosting and logistic regression. The authors stressed the need for improvement in the way of reporting performance to ensure a fairer comparison. Most of these models utilize machine learning models built on structured EHR data for the prediction of clinical tasks.

Recent works show promising results in modeling patient data using deep learning approaches. Harutyunyan et al. (2017) benchmarked their performance on four clinical prediction tasks on the MIMIC-III database using multitask recurrent neural networks. Zalewski et al. (2017) presented a viable framework to combine several modalities of a patient's health states for risk stratification. Their approach was built on the hierarchical Dirichlet method, aimed at tackling the sparsity and high-dimensionality of the nursing notes extracted from the MIMIC-II database. However, the authors used a logistic regression model to predict the mortality rate of the patients and did not evaluate their performance with the recent works and deep neural architectures. Purushotham et al. (2018) reported a suite of five clinical prediction tasks, including the length-of-stay, mortality, and ICD-9 code group prediction on the MIMIC-III database using deep learning models and benchmarked their performance against the existing state-of-the-art methods and severity scoring systems. However, mining and modeling the valuable patient-specific information in unstructured clinical nursing notes for the development of CDSSs remains mostly uncommon.

In this paper, we discuss an approach to model the rich patient-specific information in the unstructured clinical nursing notes, to aid in the risk stratification as an ICD-9 code group prediction task. ICD-9 codes are a taxonomy of diagnostic codes used for cost-effectiveness analysis, epidemiology studies, and designing healthcare policies. Accurate ICD-9 code group prediction not only promotes better ICD-9 code determination, but also facilitates more reliable risk stratification by reporting on the severity, symptoms, and the use of resources across code groups, thus aiding disease-specific staging systems. In our work, two coherence-based topic modeling approaches, Coherence-based Latent Dirichlet Allocation (C-LDA) and Coherence-based Nonnegative Matrix Factorization (C-NMF) are employed to capture the semantic relationships between the textual features of the clinical notes and derive optimal data representations with a higher guarantee on human interpretability. We employ *FarSight* to aggregate the documented patient data in a way intended to detect the onset of the disease with the earliest recorded symptoms. Furthermore, we benchmark the performance of our proposed topic models using two neural architectures, including Multi-Layer Perceptron (MLP) and Attention-based Long Short Term Memory (A-LSTM). Additionally, we perform a sensitivity analysis to assess the statistical significance of the obtained results.

The remainder of this paper is structured as follows: Section 2 describes the MIMIC-III database, the preprocessing steps, and the topic modeling approaches employed to obtain the optimal data representations from the raw clinical nursing notes. The deep neural architectures employed in the clinical task of ICD-9 code group prediction along with the discussion of the experimental results of our benchmarking are presented in Section 3. Finally, Section 4 summarizes this paper with highlights on future research possibilities.

2 Materials and Methods

In this section, we discuss in detail, the Natural Language Processing (NLP) pipeline designed to facilitate multi-label ICD-9 code group prediction, and the same is depicted in Figure 3.

2.1 Dataset and Cohort Selection

MIMIC-III (v1.4) is a publicly available large healthcare database with comprehensive medical data of over $40,000$ ICU patients. The healthcare database contains $223,556$ nursing notes extracted from $2,083,180$ note events (*noteevents* table), corresponding to $7,704$ distinct patients (*diagnoses_icd* table). Two selection criteria were employed in the cohort selection. Firstly, only

those records corresponding to the patients older than 15 (adults) were retained using the patient's age at the time of admission to the ICU (extracted from *admissions* and *patients* tables). Secondly, only the first admission of a patient to the hospital was considered. Both these steps were followed in accordance with the existing literature (Johnson et al., 2017; Purushotham et al., 2018). The resultant dataset comprises nursing notes of $7,638$ patients with a median age of 66 years (Quartile $Q_1 - Q_3$: $52 - 78$ years).

2.2 Data Cleaning

The data extracted from the MIMIC-III database contained erroneous patient entries due to several factors, including missing values, duplicate or incorrect records, outliers, and noise. The erroneous entries were filtered out using the *iserror* attribute of the *noteevents* table. Then, duplicate patient records were identified and deduplicated. The resultant dataset comprised of nursing notes corresponding to $6,532$ patients, and the data in these records were aggregated using the proposed *FarSight* technique.

2.3 FarSight: Long-Term Aggregation

It is crucial to detect the onset of the disease with the earliest detected symptoms, to provide preventive care and reduce the mortality and morbidity of complications. We propose *FarSight*, which is designed to aggregate the patient data using a future lookup on all the detected diseases in the later medical records concerning that patient. Let $\mathcal{P}$ be the set of all patients, and let a patient p have a sequence of N clinical notes, $\mathcal{S}^{(p)} = \{(\eta_i^{(p)}, \mathcal{I}_i^{(p)})\}_{i=1}^{N}$, with each clinical note $\eta_i^{(p)}$ mapped to an ICD-9 code $\mathcal{I}_i^{(p)}$ indexed in the order from the oldest to the most recent. Now, *FarSight* aggregates the ICD-9 codes across the nursing notes of a patient using a future lookup, resulting in $\mathcal{S}^{(p)} = \{(\eta_i^{(p)}, \mathcal{I}^{(p)})\}_{i=1}^{N}$, where $\mathcal{I}^{(p)} = \{\mathcal{I}_i^{(p)}\}_{i=1}^{N}$. Ultimately, we aim at learning a function $\mathcal{F}$ to estimate the probability of classifying a given nursing note $\eta_j^{(p)}$ into a set of diagnostic code groups: $\mathcal{F}(S^{(p)}) \approx Pr(\mathcal{I}^{(p)} \mid \eta_j^{p})$. Instead of aggregating several patient records, *FarSight* only aggregates the ICD-9 codes across a particular patient's nursing notes to facilitate risk stratification at the initial stages of the disease with the earliest recorded symptoms and infections.

2.4 Data Preprocessing

Data (text) normalization is performed to facilitate the transformation of inconsistent and informally written medical text into a consistent canonical form. Preprocessing includes tokenization, stopword removal, and stemming/lemmatization. Tokenization splits the nursing text into words (tokens). Using the NLTK English stopword corpus, we removed the stopwords from the generated set of tokens. Next, references to images (e.g., *PET_Scan.jpg*) were removed, and character case folding was performed. Word length based token removal was not performed to retain medical abbreviations such as *CT*, *MRI*, *DEXA*, and *PET*. Lastly, stemming was employed to facilitate suffix stripping, followed by lemmatization to convert the stripped tokens into their base forms. The tokens appearing in less than ten clinical notes were eliminated to mitigate overfitting and lower the computational complexity of training.

2.5 Topic Modeling of Clinical Notes

Let the set of all nursing notes be $\mathbb{S} = \{\mathcal{S}^{(p)}\}_{p=1}^{\mathcal{P}}$. Each nursing note η_j constitutes a variable length of words from a large vocabulary $\mathbb{V}$, making $\mathbb{S}$ very complex. Thus, a transformation (T) of the unstructured clinical text to a machine-processable form ($T : \mathbb{S} \rightarrow \mathbb{R}^k$ ($k \ll |\mathbb{V}|$)) is vital to the efficacy and performance of the underlying deep neural architectures.

Topic modeling aims at finding a set of topics

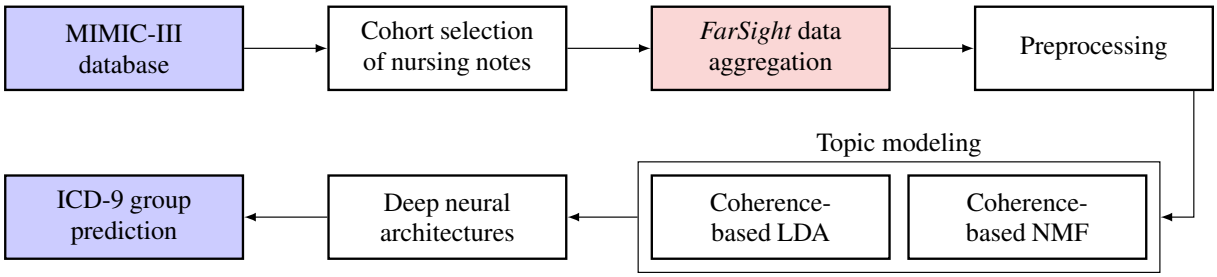


Figure 3: NLP pipeline used in the prediction of the ICD-9 code group.

from a set of clinical notes that best represents the corpus. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a cluster analysis approach based on the three-layer Bayesian framework including documents, topics, and tokens. LDA draws a mixture of topics from the Dirichlet distribution and facilitates a flat and soft probabilistic clustering of tokens into topics and documents into topics. LDA posits that each term and clinical note belong to a set of clinical topics with a certain probability. Nonnegative Matrix Factorization (NMF) (Lee and Seung, 2000) is a matrix factorization approach that decomposes multivariate data into topics. In NMF, each topic is a nonnegative linear combination of the tokens in the vocabulary. NMF iteratively decomposes the data matrix ($N \times |\mathbb{V}|$) into two lower rank matrices with $\mathcal{T}$ topics ($N \times \mathcal{T}$ and $\mathcal{T} \times |\mathbb{V}|$). These topic models capture the context of occurrence and co-occurrence, which is essential for accurate predictability of the underlying deep neural models.

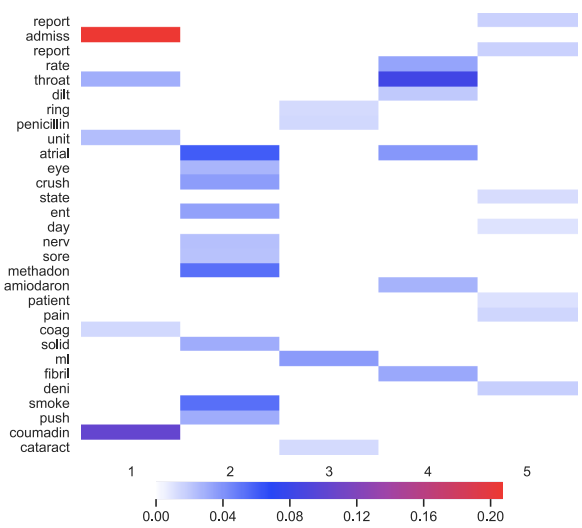
Determining the optimal number of LDA or NMF clusters is a challenging task. To address this issue, we utilize the Topic Coherence (TC) or semantic coherence (Röder et al., 2015) between the topics to derive the optimal number of clusters. Furthermore, when topics are learned from a multinomial distribution over words from noisy and sparse text data, they are less coherent and hard to interpret. TC evaluates topic models with a greater guarantee of human interpretability. This study adopts LDA and NMF with TC (C-LDA and C-NMF) as TC accounts for the semantic similarity between the higher scoring tokens and facilitates the generation of human-understandable topics. We employ the C_v variant of coherence measurement with a Normalized Pointwise Mutual Information (NPMI) score (Bouma, 2009) as the confirmation measure, due to its high correlation with the available human-judged data (Röder et al., 2015). Let $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ be a topic generated from a topic model which is represented using its top-k most probable tokens (t_is). Note that higher values of the average pairwise similarity among the tokens in $\mathcal{T}$ imply greater coherence of the topic. For a predetermined similarity measure $S(t_i, t_j)$ (here NPMI), the coherence score is computed as:

$$\text{Coherence}_S(\mathcal{T}) = \frac{\sum_{1 \le i \le k-1} \sum_{i+1 \le j \le k} S(t_i, t_j)}{\binom{k}{2}} \quad (1)$$

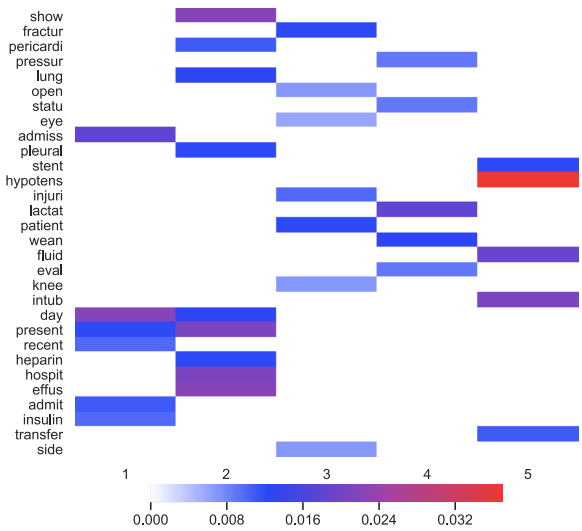
where $t_i, t_j \in \mathcal{T}$. The coherence score comes from external data, i.e., the data not used during training (we employed the full set of English Wikipedia articles), and is intended to regularize the topic models. The NPMI similarity score is an extension of the pointwise mutual information score, and is used in finding associations and collocations between the words (Aletras and Stevenson, 2013). The NPMI score is computed as:

$$\text{NPMI}(t_i, t_j) = \frac{\text{PMI}(t_i, t_j)}{-\log_2(\Pr(t_i, t_j))} \quad (2)$$

$$\text{PMI}(t_i, t_j) = \log_2 \left(\frac{\Pr(t_i, t_j)}{\Pr(t_i)\Pr(t_j)} \right) \quad (3)$$



(a) Coherence-based LDA.



(b) Coherence-based NMF.

Figure 4: Correlations between top terms' membership in top five topic modeling clusters.

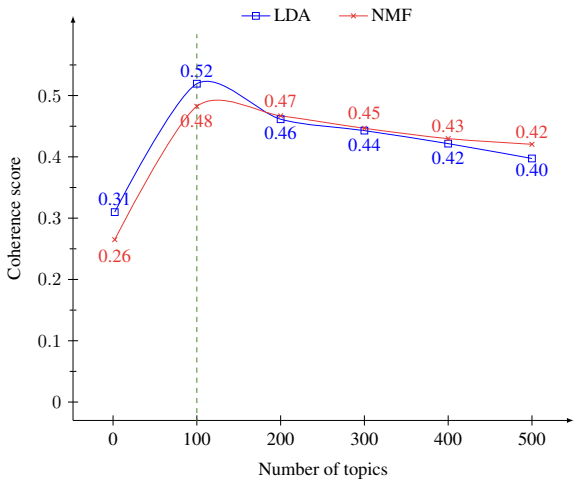


Figure 5: Coherence score comparison to determine the optimal number of topics.

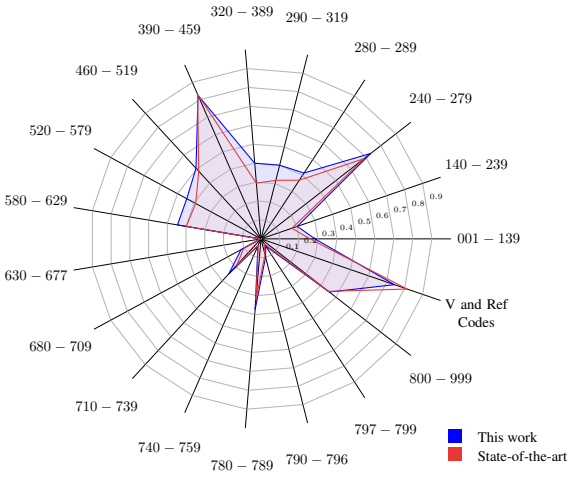


Figure 6: Comparison of ICD-9 code group statistics with the state-of-the-art model (Purushotham et al., 2018).

The individual confirmation measures obtained for all topics ($\mathcal{T}_i$s) are averaged to obtain the final coherence score.

The number of topics for both LDA and NMF models was determined to be 100, by computing the coherence score of several topic models obtained by varying the number of topics. The LDA and NMF matrices were built on a bag-of-words representation of the clinical notes. For the ease of interpretation, a heat map presenting the correlations between top terms' membership in top five C-LDA clusters is presented in Figure 4a, and top five C-NMF clusters is depicted in Figure 4b. From Figure 4, it can be observed that both the C-LDA and C-NMF models effectively capture specific clinical terms, including *penicillin*, *cataract*, *coumadin*, *insulin*, *heparin*, and *pleural* from the raw nursing text. Figure 5 shows the coherence score comparison of LDA and NMF models with the number of topics varying from 2 to 500.

3 ICD-9 Code Group Prediction

ICD-9 codes are a taxonomy of diagnostic codes typically used by healthcare professionals and insurers when discussing medical conditions. This study only focuses on category-level (group) predictions, owing to the high granularity of the diagnostic codes. Each code group comprises a set of similar diseases, and most of the health conditions can be categorized into a unique group. This study focuses on the risk stratification as a multi-label problem, where each nursing note is mapped to multiple ICD-9 code groups. The ICD-9 codes for a given admission are mapped into 19 distinct code groups[3]. Note that the ICD-9 code range of 760 – 779 was left out since it corresponds to the *conditions originating in the perinatal period*, which is usually assigned to newborns, who are excluded from this study as per the defined cohort selection criteria (see Section 2.1). Additionally, to lower the computational cost of training, we merged all the reference and supplemental V-codes into a single code group. Figure 6 presents a spider plot depicting the statistics of the ratio of the number of patients in a particular code group to the total number of patients in the cohort. Although our work and the state-of-the-art (Purushotham et al., 2018) differ in data and cohort selection, it can be observed from Figure 6 that both the works share similar statistics concerning the ICD-9 code groups, thus facilitating a fair comparison of performance.

3.1 Deep Neural Architectures

We used two deep neural architectures, Multilayer Perceptron (MLP) and Attention-based LSTM (A-LSTM), for the multi-label ICD-9 code group prediction task. The deep models were trained to minimize a binary cross-entropy loss function using an Adam optimizer, with a batch size of 128, for eight epochs.

3.1.1 Multi-Layer Perceptron

The MLP is a feed-forward artificial neural network consisting of multiple layers of neurons (nodes) interacting using weighted connections.

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

MLP offers several advantages including adaptive learning, fault tolerance, parallelism, and generalizability. The output of a neuron in every layer serves as an input to the subsequent layer. A neuron in the current layer (l) with the input $I^{(l)}$ is activated in the following layer ($l+1$) as $g^{(l)}(W^{(l)} \cdot I^{(l)} + b^{(l)})$, where $g^{(l)}$ is a non-linear activation such as Rectified Linear Unit (ReLU), tanh, or logistic sigmoid, and $b^{(l)}$ and $W^{(l)}$ are the bias and weight matrix at layer l. MLP uses back-propagation to determine the gradient of the loss function needed to learn an optimal set of weights and biases needed to minimize a loss function. This study employs an MLP network with one hidden layer of 75 nodes, activated using a ReLU function, and one output layer of 19 nodes, activated using a sigmoid function.

3.1.2 Attention-based LSTM

The LSTM effectively captures the long-term dependencies and overcomes the gradient vanishing problem which is crucial in the accurate risk stratification using unstructured nursing notes. LSTMs introduce an adaptive gating mechanism to determine the extent to which the LSTM memory units must retain the previous state (c_{t-1}) and memorize the features in the current state (c_t). Typically, four gates composite an LSTM network including the input gate i, the forget gate f, the output gate o, and the candidate value g for the cell state. The precise form of an LSTM update at a layer l and time step t is computed as:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^{(l)} \begin{pmatrix} h_{t-1}^{(l)} \\ h_t^{(l-1)} \end{pmatrix} \quad (4)$$

$$c_t^{(l)} = f \odot c_{t-1}^{(l)} + i \odot g \quad (5)$$

$$h_t^{(l)} = o \odot \tanh(c_t^{(l)}) \quad (6)$$

where $\odot$ denotes element-wise multiplication, h_t is the output at a time step t, and $W^{(l)}$ is a $[4n \times 2n]$ weight matrix at layer l.

Attentive neural models have been successfully applied to several NLP tasks including sentence summarization, text entailment, and reading comprehension (Bahdanau et al., 2014). This study utilizes the attention mechanism for the clinical task of risk stratification as ICD-9 code group prediction. Let H be the matrix of output vectors $[h_1, h_2, \ldots, h_T]$ produced from LSTM. The representation r_j of a nursing note η_j after T time steps is computed as $H \cdot (\text{softmax}(v^T \cdot \tanh(H)))^T$, where v is a trainable parameter. This study utilizes an attention-based LSTM with dimension size of 289 for the embedding (17 time steps) and 300 for the LSTM hidden state. The multi-label classification is facilitated using a sigmoid activation of the final A-LSTM output.

3.2 Experimental Results and Discussion

To experimentally validate the proposed approach, we performed an exhaustive benchmarking on the clinical nursing notes obtained from the MIMIC-III database. The experiments were performed using a server running Ubuntu OS with 56 cores of Intel Xeon processors, 128 GB RAM, 3 TB hard drive, and two NVIDIA Tesla M40 GPUs. A significant challenge arose due to the manifold nature of diseases, as each patient record was assigned a set of ICD-9 code groups. This study employs a pair-wise comparison of the actual and predicted code group sets. Five standard evaluation metrics including Accuracy (ACC), F1 score, MCC score, Area Under the Precision-Recall Curve (AUPRC), and Area Under the ROC Curve (AUROC) were employed to evaluate the performance of the proposed coherence-based modeling approaches, classified using MLP and A-LSTM. Ten-fold cross-validation was performed to assess the predictability of the proposed models. Table 1 tabulates the performance of the proposed modeling approaches using the proposed *FarSight* approach for data aggregation along with two standard baselines. We observe that the proposed C-LDA model outperforms the C-NMF model in accurately classifying the diagnostic ICD-9 code groups. Additionally, from Table 1, we observe that the proposed C-LDA model outperforms the other standard baselines including LDA and NMF without coherence scores.

AUPRC varies with the change in the ratio of the target classes in the data and hence is more informative than AUROC while evaluating imbalanced data (Saito and Rehmsmeier, 2015). F1 score captures both precision and recall of the prediction, and MCC score takes into account, the true positives, false positives, and false negatives, thus serving as a balanced measure even with class imbalance. Due to the significant class imbalance in the underlying corpus (see Figure 6), AUROC and MCC scores serve as accurate evaluation metrics. The existing works, including the state-of-

the-art model (Purushotham et al., 2018), are built on the structured nature of the EHRs, modeled using numerical feature sets (e.g., lab results) to aid in the prediction of clinical events. From Figure 7, we remark that the proposed approach built on the unstructured medical text and preprocessed using the *FarSight* approach outperformed the state-of-the-art model by 11.50% in AUPRC and 1.16% in AUROC. Furthermore, the existing works do not benchmark their performance on metrics other than AUPRC and AUROC. We urge that the other metrics presented in this study aid in the accurate assessment of the proposed models, essential in determining the reliability of the underlying CDSS. It can also be noted that the *FarSight* approach effectively models the unstructured data to facilitate the detection of the onset of the disease with the earliest recorded symptoms, and such modeling results in an improvement in the clinical decision-making process. We observe that utilizing the proposed approach leads to accurate health risk appraisal well in advance, with an overall accuracy of 80%. Thus, CDSSs built on the predictive capabilities of *FarSight*-aggregated and C-LDA classified modeling could demonstrate effective patient-centric and evidence-based risk assessment, thus ensuring proper channeling of preventive and prioritized care.

3.3 Sensitivity Analysis

The experimental results in Table 1 highlight the efficacy of the proposed models over the state-of-the-art model (see Figure 7) and standard baselines, including LDA and NMF without coherence scores, in modeling the raw patient-specific clinical nursing notes. To analyze the significance of the observed performance further, we performed a statistical sensitivity analysis. Sensitivity analysis

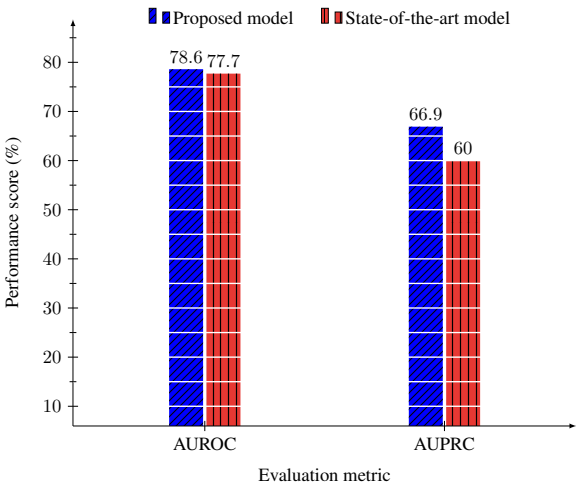


Figure 7: Comparison of the proposed approach with the state-of-the-art model (Purushotham et al., 2018).

(Simar and Wilson, 1998) is a potential approach that facilitates decision-making by measuring the extent to which the optimal solution is sensitive to the change in the input of one or more parameters.

To understand the distribution of the underlying data, we employed the Kolmogorov-Smirnov test for normality (J and Jr, 1951), which revealed that the data was not normally distributed. The performance of an algorithm measured as a result of ten-fold cross-validation forms the treatment population of that approach. Additionally, note that each sample (performance score) in the corresponding treatments of the algorithms under comparison utilize the same k^{th} fold data, and thus the samples are generated as a function of the same input population. Therefore, to perform the sensitivity analysis, we employed a nonparametric paired samples Wilcoxon signed-rank test (Wilcoxon, 1992) at a significance level (α) of 5%. The null hypothesis in Wilcoxon signed-rank test is that the two treatments are drawn from the same distribution,

Data Model	Classifier	Performance score				
		ACC	F1	MCC	AUPRC	AUROC
C-LDA $(140, 792 \times 100)$	MLP	$\mathbf{0.7954 \pm 0.0003}$	0.7175 ± 0.0008	$\mathbf{0.5743 \pm 0.0006}$	$\mathbf{0.6692 \pm 0.0006}$	$\mathbf{0.7857 \pm 0.0004}$
	A-LSTM	0.7932 ± 0.0002	$\mathbf{0.7186 \pm 0.0002}$	0.5712 ± 0.0007	0.6660 ± 0.0007	0.7854 ± 0.0013
C-NMF $(140, 792 \times 100)$	MLP	0.7826 ± 0.0004	0.7011 ± 0.0008	0.5480 ± 0.0007	0.6530 ± 0.0013	0.7735 ± 0.0006
	A-LSTM	0.7811 ± 0.0005	0.6990 ± 0.0040	0.5449 ± 0.0007	0.6510 ± 0.0009	0.7715 ± 0.0026
LDA $(140, 792 \times 100)$	MLP	0.7950 ± 0.0003	0.7168 ± 0.0020	0.5735 ± 0.0012	0.6685 ± 0.0013	0.7848 ± 0.0011
	A-LSTM	0.7930 ± 0.0007	0.7153 ± 0.0034	0.5701 ± 0.0022	0.6655 ± 0.0013	0.7833 ± 0.0020
NMF $(140, 792 \times 100)$	MLP	0.7829 ± 0.0006	0.7029 ± 0.0016	0.5498 ± 0.0009	0.6530 ± 0.0017	0.7744 ± 0.0007
	A-LSTM	0.7815 ± 0.0008	0.6935 ± 0.0052	0.5451 ± 0.0024	0.6535 ± 0.0014	0.7689 ± 0.0031

Table 1: Experimental results for ICD-9 code group prediction using MLP and A-LSTM.

Data Model	Classifier	ACC		F1		MCC		AUPRC		AUROC	
		p	z	p	z	p	z	p	z	p	z
C-LDA $(140,792 \times 100)$	MLP	–		0.005	−2.803	–		–		–	
	A-LSTM	0.005	−2.803	–		0.005	−2.803	0.005	−2.803	0.009	−2.599
C-NMF $(140,792 \times 100)$	MLP	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803
	A-LSTM	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803
LDA $(140,792 \times 100)$	MLP	0.005	−2.803	0.009	−2.599	0.007	−2.701	0.016	−2.395	0.009	−2.599
	A-LSTM	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803
NMF $(140,792 \times 100)$	MLP	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803
	A-LSTM	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803	0.005	−2.803

Table 2: A paired samples Wilcoxon signed-rank test (two-tailed, $p < 0.05$) for the proposed model with the best performance against other modeling strategies.

which is rejected in favor of the alternate hypothesis when the significance level (p-value) resulting from the test is higher than the preset α. Table 2 presents the results of our sensitivity analysis for the proposed model with the best performance against other modeling strategies. From Table 2, it can be observed that the value of p is always lower than the preset α of 0.05. Thus, we conclude that the proposed model with the best performance is statistically significant than the other approaches and baseline methods with respect to all the employed performance evaluation metrics.

4 Concluding Remarks

In this paper, we presented *FarSight*, a technique for detecting the onset of the disease with the earliest recorded symptoms and infections, to provide preventive and prioritized care, in turn aiding in the reduction of the morbidity rate. Two coherence-based topic modeling approaches were employed to capture the semantic information in the nursing notes and derive the optimal data representations with emphasis on the human interpretability of the derived clinical concepts. The obtained data representations were effectively leveraged for diagnostic ICD-9 code group prediction using deep neural architectures. Unlike in the previous works, we benchmarked the performance of our proposed models using several evaluation metrics which are essential in the accurate assessment of the reliability of the models. The proposed model captured the valuable patient-specific information present in the informally written nursing notes and outperformed the structured EHR data based state-of-the-art model with an improvement of 11.50% in terms of AUPRC and 1.16% in terms of AUROC. Furthermore, we also observed that the proposed *FarSight*-aggregated and C-LDA classified model captured the discriminative features of the nursing notes and consistently outperformed several other standard models, including C-NMF, LDA, and NMF. Moreover, our model eliminates the dependency on structured EHRs for the development of CDSSs and is extremely vital in countries with low EHR adoption rates.

Although the proposed approach effectively stratifies the patients' risk and the associated complications, it can be enhanced further, which calls for further research on this topic. First, the proposed approach only models the unstructured nursing text and neglects the structured EHR information (e.g., lab results), which can potentially be utilized to facilitate robust patient profiling. Second, the modeling presented in this study does not account for real-time clinical data. In the future, we intend on exploring the techniques for modeling structured EHR data along with the data modeled from the unstructured clinical nursing notes. We also aim at validating our model on real-time clinical data to enhance its predictability and adaptability, thus focusing on the need for time-aware, dependable architectures in real-world hospital scenarios.

Acknowledgments

This work is funded by the Government of India's DST-SERB Early Career Research Grant (ECR/2017/001056) to Sowmya Kamath S. Any opinions, findings, and recommendations or conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen.

Gilles Clermont, Derek C Angus, Stephen M DiRusso, Martin Griffin, and Walter T Linde-Zwirble. 2001. Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models. *Critical care medicine*, 29(2):291–296.

GS Doig, KJ Inman, WJ Sibbald, CM Martin, and JM Robertson. 1993. Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. *Proceedings. Symposium on Computer Applications in Medical Care*, pages 361–365.

Sebastien Dubois, Nathanael Romano, David C Kale, Nigam Shah, and Kenneth Jung. 2017. Learning Effective Representations from Clinical Notes. *arXiv preprint arXiv:1705.07025*.

Keith Feldman, Nicholas Hazekamp, and Nitesh V Chawla. 2016. Mining the Clinical Narrative: All Text are Not Equal. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 271–280. IEEE.

Jim Grigsby, Robert Kooken, and John Hershberger. 1994. Simulated neural networks to predict outcomes, costs, and length of stay among orthopedic rehabilitation patients. *Archives of physical medicine and rehabilitation*, 75(10):1077–1081.

C William Hanson and Bryan E Marshall. 2001. Artificial intelligence applications in the intensive care unit. *Critical care medicine*, 29(2):427–435.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*.

Vijay Huddar, Bapu Koundinya Desiraju, Vaibhav Rajan, Sakyajit Bhattacharya, Shourya Roy, and Chandan K Reddy. 2016. Predicting Complications in Critical Care Using Heterogeneous Clinical Data. *IEEE Access*, 4:7988–8001.

Frank J and Massey Jr. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American statistical Association*, 46(253):68–78.

Yohan Jo, Natasha Loghmanpour, and Carolyn Penstein Rosé. 2015. Time Series Analysis of Nursing Notes for Mortality Prediction via a State Transition Topic Model. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1171–1180, New York, NY, USA. ACM.

Alistair EW Johnson, Tom J Pollard, and Roger G Mark. 2017. Reproducibility in critical care: a mortality prediction case study. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 361–376, Boston, Massachusetts. PMLR.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for Non-negative Matrix Factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 535–541, Cambridge, MA, USA. MIT Press.

Prabha Susy Mathew and Anitha S Pillai. 2015. Big data solutions in healthcare: Problems and perspectives. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–6. IEEE.

Romain Pirracchio. 2016. *Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project*, pages 295–313. Springer International Publishing, Cham.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. 2018. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA. ACM.

Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):1–21.

Léopold Simar and Paul W Wilson. 1998. Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. *Management science*, 44(1):49–61.

Kathleen B To and Lena M Napolitano. 2012. Common Complications in the Critically Ill Patient. *Surgical Clinics*, 92(6):1519–1557.

Jack V Tu and Michael RJ Guerriere. 1993. Use of a Neural Network as a Predictive Instrument for Length of Stay in the Intensive Care Unit Following Cardiac Surgery. *Computers and Biomedical Research*, 26(3):220–229.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2018. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In Samuel Kotz and Norman L Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer New York, New York, NY.

Christine M Wollschlager and Arnold R Conrad. 1988. Common complications in critically ill patients. *Disease-a-Month*, 34(5):225–293.

Aaron Zalewski, William Long, Alistair EW Johnson, Roger G Mark, and H Lehman Li-wei. 2017. 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)*, pages 449–452. IEEE.