Deep Neural Learning for Automated Diagnostic Code Group Prediction Using Unstructured Nursing Notes

Aditya Jayasimha*†
Automatic Speech Recognition
Samsung Research and Development Institute
Bangalore, Karnataka, India
adityajayasimha@gmail.com

Tushaar Gangavarapu*†‡
Worldwide Deals, Community Shopping
Amazon.com, Inc.
Bangalore, Karnataka, India
tusgan@amazon.com

Sowmya Kamath S
Healthcare Analytics and Language Engineering Lab
Department of Information Technology
NITK Surathkal, Karnataka, India
sowmyakamath@nitk.edu.in

Gokul S Krishnan
Healthcare Analytics and Language Engineering Lab
Department of Information Technology
NITK Surathkal, Karnataka, India
gsk1692@gmail.com

ABSTRACT

Disease prediction, a central problem in clinical care and management, has gained much significance over the last decade. Nursing notes documented by caregivers contain valuable information concerning a patient's state, which can aid in the development of intelligent clinical prediction systems. Moreover, due to the limited adaptation of structured electronic health records in developing countries, the need for disease prediction from such clinical text has garnered substantial interest from the research community. The availability of large, publicly available databases such as MIMIC-III, and advancements in machine and deep learning models with high predictive capabilities have further facilitated research in this direction. In this work, we model the latent knowledge embedded in the unstructured clinical nursing notes, to address the clinical task of disease prediction as a multi-label classification of ICD-9 code groups. We present *EnTAGS*, which facilitates aggregation of the data in the clinical nursing notes of a patient, by modeling them independent of one another. To handle the sparsity and high dimensionality of clinical nursing notes effectively, our proposed *EnTAGS* is built on the topics extracted using Non-negative matrix factorization. Furthermore, we explore the applicability of deep learning models for the clinical task of disease prediction, and assess the reliability of the proposed models using standard evaluation metrics. Our experimental evaluation revealed that the proposed approach consistently exceeded the state-of-the-art prediction model by 1.87% in accuracy, 12.68% in AUPRC, and 11.64% in MCC score.

*Equal contribution to this research.
†A. Jayasimha and T. Gangavarapu completed most of this work at the Healthcare Analytics and Language Engineering Lab, NITK Surathkal, Karnataka, India.
‡Corresponding author.

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371176

CCS CONCEPTS

• **Applied computing** → **Health informatics**; Bioinformatics; • **Computing methodologies** → **Natural language processing**; **Deep learning algorithms**; *Information extraction*; *Cross-validation.*

KEYWORDS

Clinical Decision Support Systems, Deep Learning, Disease Prediction, Healthcare Analytics, Multi-label Classification, Natural Language Processing.

ACM Reference Format:
Aditya Jayasimha, Tushaar Gangavarapu, Sowmya Kamath S, and Gokul S Krishnan. 2020. Deep Neural Learning for Automated Diagnostic Code Group Prediction Using Unstructured Nursing Notes. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3371158.3371176

1 INTRODUCTION

Disease prediction plays a pivotal role, both in clinical care and hospital management. It is a problem that has gained significant interest in the present-day, owing to the recent developments and attention towards AI-assisted precision healthcare [35]. Analyzing recent healthcare reports [17] reveals that more than 83% of the hospitals in the United States have adopted the structured Electronic Medical Record (EMR) systems. The number of patients admitted to hospitals in the United States alone is more than thirty million, which indicates the availability of large amounts of patient data. Such availability has boosted research in several clinical decision-making problems such as length-of-stay prediction, mortality prediction, readmission prediction, phenotyping, and disease prediction, especially by leveraging the predictive power and capabilities of machine and deep learning models [14, 16, 25, 34, 35, 41]. However, a significant challenge is posed in the cases of developing countries, where, the adoption rate of EMR systems is quite low. Most developing countries are yet to define and employ a structured approach for storing and managing patient information, due to which, such information is often only available in the form of informally-written, free-text nursing notes [13, 14].

These free-text documents noted by caregivers record in detail, the clinical status of patients during outpatient care or hospital stay. Medical personnel in such developing countries often resort to a manual evaluation of these raw clinical notes for causal inferences and decision-making [25]. Therefore, nursing notes contain valuable clinical information such as lifestyle, history of illness, symptoms, medications, and treatments, which can be potentially leveraged for intelligent clinical decision-making. Attempts to manually convert such unstructured nursing notes into EMRs often result in a loss of crucial patient information and the relevant subjective assessments that are recorded in them [14, 15]. Hence, it is critical to develop automated methods that can facilitate optimal modeling of unstructured clinical nursing text, for building clinical prediction models. Additionally, the availability of large, public patient databases such as Medical Information Mart for Intensive Care (MIMIC-II [29] and -III [23]) have further boosted the research in such clinical prediction and decision-making tasks. However, mining and modeling such raw clinical notes is challenging, especially due to their rawness, sparsity, high-dimensionality, complex temporal and linguistic structure, rich medical jargons, and abundant abbreviations [21].

International Statistical Classification of Diseases and Related Health Problems (ICD, with the most recent versions, ICD-9 and ICD-10) is the most popular medical coding ontology employed in hospitals, whose alphanumeric codes classify diseases and a wide variety of infections, symptoms, causes of injury, disorders, and others. Such medical coding is typically performed by medical personnel who are trained to understand nursing notes with complex clinical terminology and inconsistency; and map these nursing notes into a set of appropriate codes from a large menu of options (ICD-9 coding scheme has around 13, 000 codes). Hence, medical coding is a time consuming, expensive, and inexact process, which can be overcome by automatic code assignment models. Each ICD-9 code represents a unique disease, and similar diseases are grouped into a diagnostic code group [14, 15, 35]. ICD-9 code group prediction is a clinical prediction problem that aims at multi-label classification of patient records into one or more categories of diseases [7]. Such code group prediction facilitates efficient risk assessment and timely response, enabling medical professionals to better manage medical interventions at the point of care [10]. It is interesting to note that diagnostic code group prediction facilitates more than just accurate billing [6]; it is actively employed in retrospective epidemiology studies [30, 39, 42] and healthcare research including predictive modeling [3–5, 9, 15, 36].

To facilitate intelligent clinical decision-making, several recent works have incorporated the predictive capabilities of machine learning models and deep neural architectures. In 2016, Pirracchio [34] presented the super learner model, which was an ensemble of various machine learning models; the model outperformed several traditional severity scores such as sepsis-related organ failure assessment [40], acute physiologic assessment and chronic health evaluation [24], and simplified acute physiology score [27]. The author underscored the relevance of machine learning models over traditional prognostic scoring systems. However, the proposed super learner model was not benchmarked against the recent machine and deep learning models. Johnson *et al.* [22] replicated the results of 28 related and recent works benchmarked on the MIMIC-III database for the task of mortality prediction. Their work stressed on the significant challenges in reproducing the reported results, owing to the large variations in the patient cohorts and characteristics considered by different studies. Furthermore, they highlighted the need for improvising ways of performance reporting and evaluation of clinical prediction models, in order to take into account the diversity of various approaches, so as to ensure a fair comparison of the reported performances. Harutyunyan *et al.* [16] employed multi-task deep learning architectures to facilitate the prediction of four diverse clinical prediction tasks, and benchmarked their results on the MIMIC-III database. They used Long Short Term Memory (LSTM) and logistic regression models, but did not benchmark their performance against existing severity scoring systems or machine learning models, specifically the super learner.

Purushotham *et al.* [35] also utilized deep neural models to enable three clinical prediction tasks, namely, length-of-stay prediction, mortality prediction, and ICD-9 code group prediction. They utilized feature sets derived from the MIMIC-III database, and benchmarked their results against a number of severity scoring systems and machine learning models. Huang *et al.* [18] used discharge summaries to predict the top-10 ICD-9 code categories, by modeling the discharge summaries using term weighting, Word2Vec, and word sequencing with an embedding matrix, and applying state-of-the-art deep learning classifiers including Convolutional Neural Network (CNN), LSTM, and Gated Recurrent Unit (GRU), to achieve the prediction as a multi-label classification task. Zeng *et al.* [43] also utilized discharge summaries for ICD-9 code assessment, by modeling the discharge summaries using word embeddings and applying a sequential CNN classifier to achieve the prediction as a multi-label classification task. They employed transfer learning from the domain knowledge of medical subject headings [31]. Additionally, they evaluated their proposed model using evaluation metrics such as micro-average precision, micro-average recall, and micro-average F-measure. More recently, Gangavarapu *et al.* [14] designed Term weighting of unstructured notes AGgregated using fuzzy Similarity (TAGS) model, a fuzzy similarity matching method for aggregating the raw nursing notes of the MIMIC-III database. TAGS aimed at removing anomalous and redundant patient records, thus reducing the cognitive burden and improving the predictability of the underlying models. They utilized vector space (TAGS and Doc2Vec) and topic modeling (Latent Dirichlet Allocation (LDA)) approaches to effectively model the nursing notes. Their work employed a wide range of deep neural models for the task of ICD-9 code group prediction, and benchmarked their promising performance against state-of-the-art methods.

In this work, we present the Enhanced TAGS (*EnTAGS*) strategy, aimed at advancing the research problems addressed in the existing works, through the aggregation of nursing note data by independently modeling the individual clinical notes. We propose that our designed *EnTAGS* data modeling approach is aimed at modeling each record independently, while aggregating the diagnostic code groups observed across all the clinical notes corresponding to a patient. We employ Non-negative Matrix Factorization (NMF) to obtain optimized patient representations from the aggregated information, upon which the deep prediction models are built. We

experiment and benchmark the applicability of deep learning models in both vanilla and hybrid versions, for the clinical task of ICD-9 code group prediction. Furthermore, to enable exhaustive comparison, we employ a prolific baseline approach derived from [14], that adopts a naive strategy of aggregating all the nursing notes of a patient by their identification numbers. We juxtapose the findings of our work with both the state-of-the-art TAGS model [14] and the employed baseline approach, to corroborate the efficacy and reliability (measured using standard evaluation metrics) of the proposed strategy and models in automated diagnostic coding of clinical notes. The rest of the paper is organized as follows: the proposed methodology is discussed in detail, in Section 2. The experiments conducted and the observations derived in comparison to the state-of-the-art methods are delivered in Section 3, followed by concluding remarks in Section 4.

2 PROPOSED METHODOLOGY

Figure 1 depicts the pipeline designed to facilitate ICD-9 code group prediction from clinical nursing notes using the proposed *EnTAGS* approach. In the following subsections, we elucidate on the stages in the pipeline proposed to facilitate diagnostic code group prediction. For experimental validation and benchmarking of the proposed strategy, we utilized the MIMIC-III database [23], made available by the Massachusetts Institute of Technology Lab for Computational Physiology, which comprises of diverse health data of more than 40, 000 Intensive Care Unit (ICU) patients.

2.1 Patient Cohort Selection

The MIMIC-III relational database is a freely accessible large critical care database that consists of tables corresponding to deidentified healthcare data associated with over 40, 000 ICU patients. In the database, there are 2, 083, 180 note events (*noteevents* table), from which 223, 556 are nursing notes corresponding to 7, 704 distinct ICU patients, whose diagnostic codes are recorded in the *diagnoses_icd* table. The statistics of the nursing note text corpus utilized in this work are summarized in Table 1. In this study, we consider two preliminary criteria aimed at selecting potential subjects from the MIMIC-III database. First, all the records of patients with age below 15 (extracted using the *patients* and *admissions* tables) are discarded from the study. Second, in order to facilitate disease prediction with the earliest recorded symptoms, only the nursing notes corresponding to the first admission of a patient to a hospital were considered. These two criteria were defined in-line with the conditions considered in the existing benchmarking models [14, 15, 22, 25, 35]. The dataset obtained after cohort selection consisted of nursing notes corresponding to 7, 638 patients, with a median age of 66 years.

Table 1: Statistics of the nursing note corpus of MIMIC-III.

Parameter	Total
Clinical nursing notes	223, 556
Sentences in the clinical nursing notes	5, 244, 541
Words in the clinical nursing notes	79, 988, 065
Unique words in the clinical nursing notes	715, 821

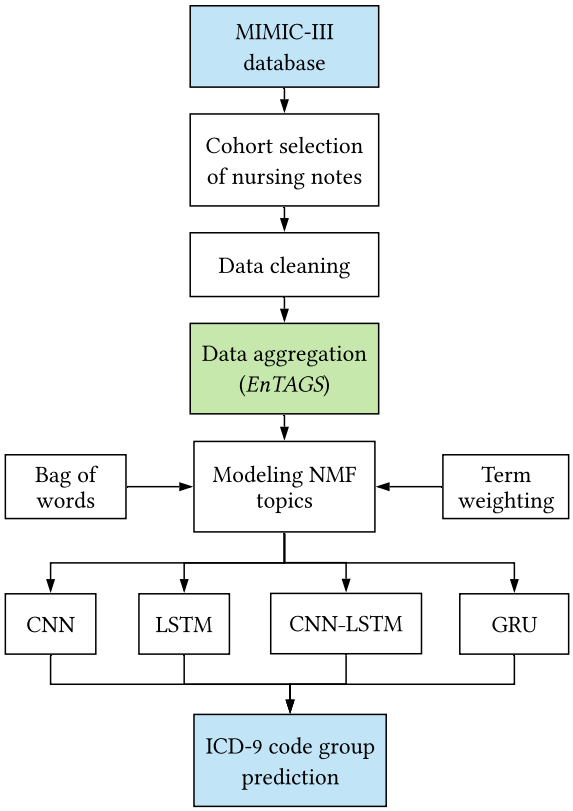


Figure 1: The proposed workflow employed in facilitating ICD-9 code group prediction from clinical nursing notes.

2.2 Data Cleaning and Aggregation

Following the patient cohort selection, we cleaned the obtained dataset by removing faulty entries introduced due to a variety of factors such as inconsistency, incorrect mapping of diagnostic codes, duplicate entries, and others. First, all the entries with a value '1' set for the *iserror* attribute (signifies a faulty record) were identified are removed. Next, all the duplicate patient records were discarded. Post this processing, the resulting dataset comprised of nursing notes corresponding to 6, 532 patients. To obtain a canonical form, the raw clinical text in the nursing notes was cleaned and normalized by removing symbols and special characters, and trimming extra spaces between the words (tokens). File references to images (e.g., *CT_Scan.jpg*) present in the nursing notes were removed, and case folding was performed. In our study, we aim at building a robust classifier to tackle the inconsistency in the nursing text (e.g., *pat*, *pt*, and *patient*), and tokens of all sizes were retained to represent important patient information such as *mg*, *CT*, and others. Following this, to convert the tokens into their base forms, we performed stemming and lemmatization. Additionally, we performed stopword removal to eliminate high-frequency general words such as *while*, *were*, *being*, and others, using the NLTK stopword corpus [8].

Table 2: Comparison of the baseline (with TW−NMF) model against Doc2Vec and LDA approaches employed in the state-of-the-art work [14].

Metric	Baseline with TW−NMF	Doc2Vec from [14]	LDA from [14]
ACC	**0.8065 ± 0.0033**	0.8005 ± 0.0017	0.8034 ± 0.0016
AUPRC	**0.6231 ± 0.0073**	0.6076 ± 0.0033	0.6181 ± 0.0011
AUROC	**0.7707 ± 0.0050**	0.7600 ± 0.0010	0.7649 ± 0.0011
CE	**18.5063 ± 0.1488**	18.6485 ± 0.0539	18.6243 ± 0.0679
F1	**0.6739 ± 0.0075**	0.6655 ± 0.0022	0.6643 ± 0.0013
LRL	**0.4244 ± 0.0095**	0.4388 ± 0.0019	0.4361 ± 0.0018
MCC	**0.5616 ± 0.0083**	0.5386 ± 0.0032	0.5542 ± 0.0022

In this study, we adopted a strategy of aggregating voluminous clinical data that diverges from that employed in the state-of-the-art models. Unlike in TAGS [14] and the other benchmark methods [25, 35], we do not aggregate multiple nursing notes of a patient, recorded across several monitoring episodes of an admission. We consider each record independent of the other, while merging the diagnostic code groups across all the nursing notes maintained for that patient. Such a choice of not independently modeling the diagnostic code groups corresponding to nursing notes was made to handle the large number of near-duplicate clinical notes (notes with little variation in the text but corresponding to the same transcription) that were mapped to different diagnostic code groups. This approach of independently modeling the nursing notes of the patients is termed as *EnTAGS*. Thus, employing *EnTAGS* would effectively handle near-duplicate clinical notes, in turn aiding the predictability of the underlying models. Despite the vitality of *EnTAGS* in enhancing clinical decision-making, the consideration of each patient record mapping to all the corresponding diseases has a major shortcoming of causing false alarms in the prediction. However, we still explore *EnTAGS* as a potential alternative to the existing modeling strategies because: 1) independent modeling of patients' nursing notes considerably reduces false negatives and 2) such aggregation is suitable in scenarios where failing to predict an existing disease has much more severe consequences than inaccurately predicting a disease that may not be present, but has a significant possibility of occurring in the future.

Furthermore, utilizing the proposed *EnTAGS* facilitates the underlying deep neural classifiers to learn to classify each nursing note with respect to the given diagnostic code labels, while the state-of-the-art methods [14, 35] aggregate all the nursing notes recorded for a patient (to facilitate prediction), and they do not employ any specific backtracking mechanisms to focus on which specific portions of the aggregated nursing note maps to which possible diagnostic code group. Hence, the models trained using naively aggregated patient data learn predictions over the entire aggregations rather than individual nursing notes. In contrast, the proposed *EnTAGS* strategy facilitates the mapping of the individual nursing note to all possible disease labels. We argue that such independent modeling can better map similar (as those observed while training) individual nursing notes at prediction time to accurate disease labels. For instance, consider the data of a patient with four in-patient evaluations, corresponding to four distinct nursing notes. Now, suppose that only the first nursing note is employed in diagnostic code group prediction, the *EnTAGS* approach would be able to capture the episode-specific characteristics in this nursing note, better than the methods that are trained on aggregated nursing notes. Additionally, since the ICD-9 code groups across all the nursing notes of a patient are merged, the underlying model would be able to predict the diseases corresponding to the first episode, as well as those observed in the later episodes (corresponding to second, third, and fourth clinical notes)—since the model would have been trained to predict all the disease labels in an admission, for a given nursing note. Additionally, in hospitals, the *EnTAGS* strategy could be potentially employed as a forecasting tool intended on predicting hospital readmission, since it can aid the deep classifier in learning to predict all the possible diagnostic code groups for a given nursing note. Finally, the proposed *EnTAGS* approach utilizes one-hot representations of the diagnostic code groups, which can be extended to include fuzzification of diagnostic code groups, with higher importance (weighting scheme) for the current diagnostic code groups and reducing importance for present and past disease code groups, based on suitable distributions of the underlying patient data.

2.3 Feature Modeling of Nursing Notes

Each nursing note in the preprocessed corpus consists of a varying number of tokens, and to handle the complexity, high-dimensionality, and sparsity of the data, these free-text nursing notes have to be transformed into a machine-processable form. Traditional methods such as Bag-of-Words (BoW) and Term Weighting (TW) present the statistical distributions of the underlying corpus, and therefore suffer from high-dimensionality and sparsity—they also neglect the crucial semantic information in the clinical notes. Thus, it is necessary to utilize optimal patient representations that capture the semantic information in the clinical text, while overcoming the problems of high-dimensionality and sparsity, to realise the capabilities of deep neural learning for effective prediction.

In our work, we utilize NMF [28], a topic modeling approach, for modeling and deriving clinical concepts from the nursing notes. NMF is a group of algorithms utilized in multivariate analysis and

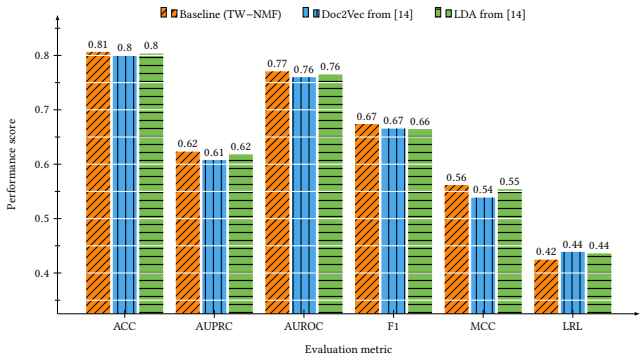


Figure 2: Comparison of the baseline approach (modeled using NMF) with the models in the state-of-the-art work [14].

linear algebra, where a matrix $\mathbb{S}$ is factorized into two matrices $\mathbb{W}$ and $\mathbb{H}$, such that all the three matrices contain non-negative elements. NMF is utilized in a variety of applications in signal and data analytics [2, 11, 19, 20, 28, 32, 33], one of which involves modeling of topics from free-text. Given a BoW or TW matrix of the nursing notes, say $\mathbb{S}$, with dimensions $N \times W$, where N is the number of nursing notes and W is the vocabulary size of the underlying corpus, NMF decomposes $\mathbb{S}$ into two matrices: $\mathbb{A}$ and $\mathbb{B}$ with dimensions $N \times T$ and $T \times W$ respectively. Here, T is the number of topics, which was heuristically determined to be 100, and is used throughout this study. To summarize, the NMF decomposition of a BoW or TW matrix would yield components that are considered as the "clinical concepts" extracted from the notes, which then facilitate the decomposition of nursing notes into a weighted sum of topics. In this study, we utilized the implementations in the Python Gensim package [37, 38]. NMF can be applied on the BoW or TW statistical modeling of the corpus, and in this study, we experiment with both these distributions.

Following a detailed comparison of the baseline model built on NMF with other vector space and topic modeling techniques benchmarked by [14] (observe from Table 2 and Figure 2 that the baseline modeled using NMF outperforms other modeling strategies), NMF was chosen to model the topics in the nursing notes. Since our ultimate goal is to develop time-aware intelligent prediction systems, it is important to remark that a direct comparison of the baseline with the TAGS model [14] is irrelevant, as the TAGS approach models each nursing note using 14, 650 dimensions, while the baseline model facilitates 100-dimensional embeddings (99.32% lower). The TAGS strategy considers every word in the vocabulary, while our models built on NMF utilize the clinical information of a nursing note condensed into a 100-dimensional vector. Additionally, the time taken to train the deep neural models on TAGS-aggregated information would be much higher than that with our proposed strategy. However, since *EnTAGS* considers each nursing note independent of the other, the number of nursing notes input for feature modeling and training is much higher than that with TAGS. It can be inferred that the training of large amounts of nursing notes with each note embedded as a 14, 650-dimensional vector is computationally expensive and impractical both in terms of memory requirements and the time required for feature modeling and training. Therefore, the TW modeling approach employed in TAGS method cannot be replicated with the proposed *EnTAGS* strategy. Potential alternatives to embed the free-text include Doc2Vec and LDA methods experimented in [14], as their dimensions (500 and 100 respectively) and training times are comparable. We observed that the NMF topic modeling outperformed the Doc2Vec and LDA models of the state-of-the-art by 0.75% and 0.39% respectively (in terms of Accuracy (ACC)), and hence we chose NMF to model the features of the *EnTAGS*-aggregated patient information, instead of Doc2Vec or LDA.

2.4 ICD-9 Code Group Prediction

ICD-9 codes are a taxonomy of diagnostic codes employed in the classification of diseases, based on a large number of symptoms, disorders, infections, and others. ICD medical ontology is widely used by the healthcare professionals. Every health condition is mapped

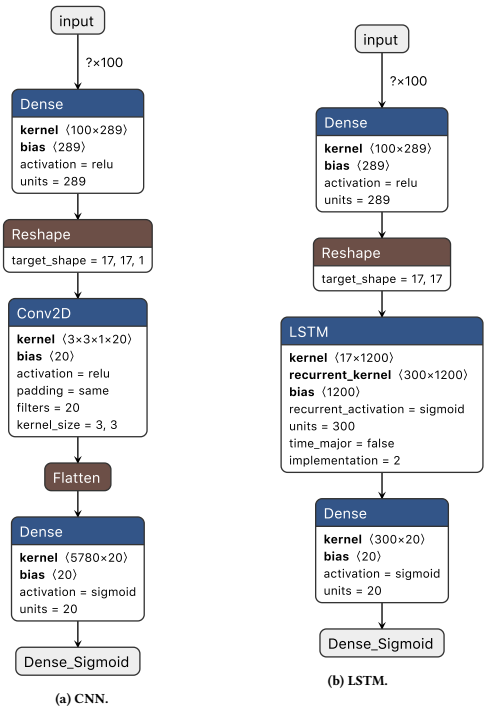
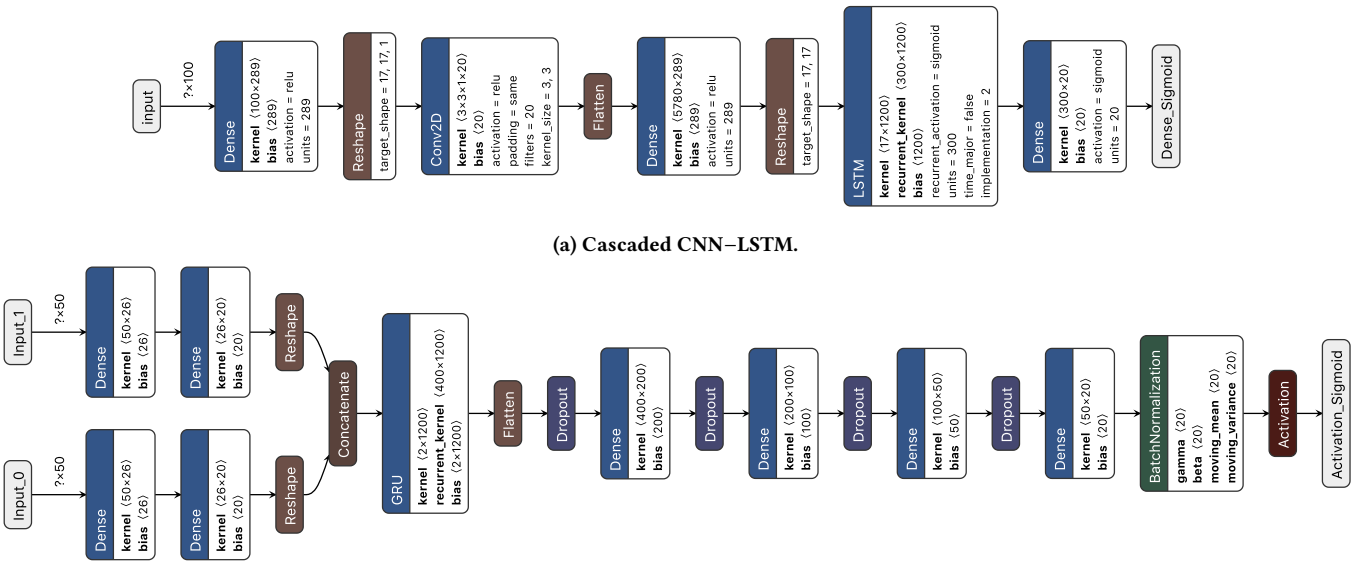


Figure 3: A schematic overview of the vanilla neural architectures employed in this study.

to a unique diagnostic code, and thus the number of such diagnostic codes is very large in number. Consequently, assigning ICD-9 codes to patient records is a tedious task, and previous research works have underscored the importance of automated prediction models for ICD-9 code group prediction, in place of individual ICD-9 code prediction [14, 26]. Since accurate ICD-9 code group prediction can improve the predictability of the corresponding ICD-9 codes by narrowing down the co-domain of the classification mapping, diagnostic code group prediction serves as a preliminary step to ICD-9 code prediction. Therefore, there has been a significant interest in developing automatic code group assignment models, and it has remained a long-standing research challenge [7, 12, 14–16, 35]. Each code group[1] consists of a number of similar diseases. In our work, we aim at predicting the ICD-9 code groups as a multi-label classification problem, where each record in the modeled nursing note corpus is mapped to one or more code groups. The *supplemental V codes* and *reference codes* are classified into the same code group, so as to reduce the computational complexity of training resulting from the large presence of these codes in the dataset. In total, there are 20 distinct ICD-9 code groups, making the problem a 20-class, multi-label classification problem.

We employed four different deep learning architectures to facilitate the ICD-9 code group prediction task. These neural models include two vanilla and two hybrid versions: CNN, LSTM, cascaded CNN–LSTM, and partitioned GRU. The implementations in the Python Keras package with Tensorflow backend [1] were used in

[1]The code ranges used for mapping can be accessed online at: http://tdrdata.com/ipd/ipd_SearchForICD9CodesAndDescriptions.aspx.



(a) Cascaded CNN–LSTM.

(b) Partitioned GRU (two segments).

Figure 4: A schematic overview of the hybrid neural architectures employed in this study.

this study. All the deep neural architectures aimed at minimizing a binary cross-entropy loss function. Additionally, we utilized the Adam optimizer to train the neural models. Batch size was set to 128 for all the models, and the vanilla versions and cascaded CNN–LSTM were trained for eight epochs, while the partitioned GRU model was trained for 27 epochs. To obtain the optimal set of hyperparameters, we employed grid search.

2.4.1 Convolutional Neural Network. CNNs are deep feed-forward neural networks which reduce processing (while training) through parameter sharing—such sharing drastically reduces the number of hyperparameters to be learned and optimized by the network. CNNs can capture the local information (in the form of features) in the nursing notes effectively through the use of a number of filters. In our work, we employed a vanilla CNN model that comprises one dense layer of 289 nodes with ReLU activation function, one convolutional layer with a convolution window size of 3×3 and a feature map size of 20. The final output from the convolution layer is subjected to a sigmoid activation that yields the multi-label prediction. The vanilla CNN architecture employed in this study is presented in Figure 3a.

2.4.2 Long Short Term Memory. LSTMs are a special type of recurrent neural networks that solve the vanishing gradient problem typically observed in the traditional recurrent networks. They capture long-term dependencies very effectively, which plays a crucial role in ICD-9 code group prediction. Each nursing note is a continuous piece of text, and since LSTMs treat the nursing note as a time sequence, they can effectively capture dependencies between various parts of the nursing note and map them to the corresponding ICD-9 code groups. In our work, the utilized LSTM architecture involved a dense layer of 289 nodes with ReLU activation function. We then reshaped the obtained 289 features into 17 time-steps with 17 features each, which was then passed on as the input to an LSTM layer with 300 recurrent units. The final output obtained from the LSTM cell was subject to a sigmoid activation, thus yielding the multi-label prediction. The LSTM model architecture utilized in this work is depicted in Figure 3b.

2.4.3 Cascaded CNN–LSTM. CNNs capture local dependencies in the nursing notes effectively, while LSTMs are efficient in handling long-term dependencies. Therefore, a cascaded CNN–LSTM hybrid neural architecture would be able to capture both these types of interactions in an effective manner. The cascaded CNN–LSTM model architecture employed in our work is shown using Figure 4a. In the neural model, we employed a dense layer of 289 nodes with ReLU activation function, followed by a convolution layer with 3×3 convolution window size and 20 feature map size. The final convolved output was flattened, and the obtained output was passed onto a dense layer of 289 nodes with ReLU activation function. The resulting 289 features were reshaped into 17 time-steps with 17 features in each time-step, which was then fed as the input to an LSTM layer with 300 recurrent units. The resulting output from the LSTM cell was then subject to a sigmoid activation, to facilitate the multi-label prediction.

2.4.4 Partitioned Gated Recurrent Unit. GRUs are an alternative to LSTMs, with slight changes in their implementation. Similar to LSTMs, GRUs also capture long-term dependencies effectively. In this study, with GRUs, we employ a deeper neural architecture than that in the previously presented neural architectures, termed as the "partitioned GRU" neural model. First, we segment the input into 20 distinct non-overlapping partitions. Each input partition is then subject to a dense layer with nodes as many as half the number of records in the partition, which is then followed by a dropout of 0.1 and a dense layer of 20 nodes, again followed by

Table 3: The benchmarking results of the proposed *EnTAGS* approach using four deep neural architectures.

	Classifier	ACC	AUPRC	AUROC	CE	F1	LRL	MCC
BoW–NMF	CNN	0.7965 ± 0.0007	0.6688 ± 0.0018	0.7860 ± 0.0020	18.5327 ± 0.1412	0.7187 ± 0.0041	0.3898 ± 0.0045	0.5750 ± 0.0013
	LSTM	0.7921 ± 0.0005	0.6638 ± 0.0018	0.7794 ± 0.0017	18.6699 ± 0.0822	0.7093 ± 0.0030	0.4024 ± 0.0040	0.5652 ± 0.0016
	CNN–LSTM	0.8048 ± 0.0021	0.6806 ± 0.0031	0.7911 ± 0.0024	18.3760 ± 0.0919	0.7240 ± 0.0034	0.3825 ± 0.0046	0.5897 ± 0.0042
	GRU	0.7945 ± 0.0063	0.6698 ± 0.0057	0.7772 ± 0.0080	18.9530 ± 0.3058	0.7039 ± 0.0114	0.4081 ± 0.0143	0.5666 ± 0.0137
TW–NMF	CNN	0.8174 ± 0.0006	0.6948 ± 0.0014	0.8091 ± 0.0014	17.9663 ± 0.0562	0.7489 ± 0.0016	0.3499 ± 0.0032	0.6181 ± 0.0008
	LSTM	0.8129 ± 0.0015	0.6908 ± 0.0024	0.7992 ± 0.0012	18.2588 ± 0.0398	0.7347 ± 0.0020	0.3694 ± 0.0021	0.6062 ± 0.0028
	CNN–LSTM	**0.8282 ± 0.0023**	**0.7089 ± 0.0046**	**0.8157 ± 0.0019**	**17.6853 ± 0.0566**	**0.7562 ± 0.0021**	**0.3392 ± 0.0036**	**0.6368 ± 0.0042**
	GRU	0.82486 ± 0.0021	0.7089 ± 0.0019	0.8073 ± 0.0040	18.3412 ± 0.2126	0.7434 ± 0.0057	0.3569 ± 0.0079	0.6273 ± 0.0050

a dropout of 0.1. The resulting 20 outputs corresponding to the 20 partitions are then concatenated and flattened, thus obtaining a 400-dimensional vector corresponding to a nursing note. The intuition behind such partitioning is to capture local interactions in various segments of the nursing notes. The obtained 400 features are reshaped into 20 time-steps with 20 features each, which is then fed as the input to a GRU layer with 400 recurrent units. The output from the GRU layer is flattened, which is then followed by four dense layers of 200, 100, 50, and 20 nodes respectively—we employed a dropout of 0.1 after each dense layer, except after the final dense layer (of 20 nodes). After the final dense layer, batch normalization was applied. The purpose of having dropout and batch normalization was to avoid overfitting on the training data, and to normalize the mean and standard deviation of the neural network weights while training such a deep model. Each one of the dense layers in the entire partitioned GRU network was followed by the ReLU activation function, and the final output was subject to a sigmoid activation, thus yielding the multi-label prediction. A sample architecture with two partitions (rather than 20 partitions as employed in our work) is shown in Figure 4b.

3 RESULTS AND DISCUSSION

For the experimental validation of the proposed models and the strategy employed in this study, we utilized a high-end server running Ubuntu OS with 56 cores of Intel Xeon processors, 128 GB RAM, 3 TB hard drive, and two NVIDIA Tesla M40 GPUs. Moreover, to facilitate the benchmarking of the proposed strategy and models against the state-of-the-art work [14], we employed seven standard evaluation metrics: accuracy, Area Under the Precision-Recall Curve (AUPRC), Area Under the ROC Curve (AUROC), Coverage Error (CE), F1 score, Label Ranking Loss (LRL), and Matthews Correlation Coefficient (MCC) score—the choice of these seven evaluation metrics was in accordance with those employed in the state-of-the-art work [14]. Apart from accuracy, evaluation metrics including AUPRC and MCC score play a pivotal role in the accurate assessment of the reliability of the proposed prediction systems, owing to the fact that the underlying patient data is class imbalanced. From the obtained results, we observe that the performance obtained from *EnTAGS*-aggregated, TW–NMF-modeled patient information, classified using the cascaded CNN–LSTM deep neural model, outperforms all the other modeling strategies with respect to each of the seven evaluation metrics. In this study, we report promising results of our proposed *EnTAGS* modeling strategy, with an accuracy of 82.82%, AUPRC of 70.89%, and MCC score of 63.68%. This performance obtained as a result of *EnTAGS* aggregation and NMF modeling reflect the fact that the underlying deep neural networks are able to learn optimal representations over individual nursing notes in an effective way and are able to leverage those representations to facilitate efficient predictions of ICD-9 code groups. Table 3 tabulates the results of the proposed *EnTAGS* modeling approach, classified using the proposed (four) deep neural classifiers on the clinical nursing notes modeled using NMF built on BoW and TW statistical representations of the underlying data.

It is interesting to note from Table 3 that, despite the use of NMF-modeled (permutation invariant) data with deep neural classifiers (CNNs and recurrent networks) that exploit the local structure and dependencies in the data, we observe improved performance across all the evaluation metrics. Such boosted performance can be attributed to the fact that NMF facilitates disentanglement of the hidden structure of the underlying data by learning features that exhibit sparse part-based representations of the data (decomposing data into parts). Additionally, since NMF forces the encoding of the data to be non-negative, the sparse part-based representations are indeed additive representations of data, which can be exploited to improve the performance neural network. Furthermore, non-negative decomposition of voluminous free-text BoW or

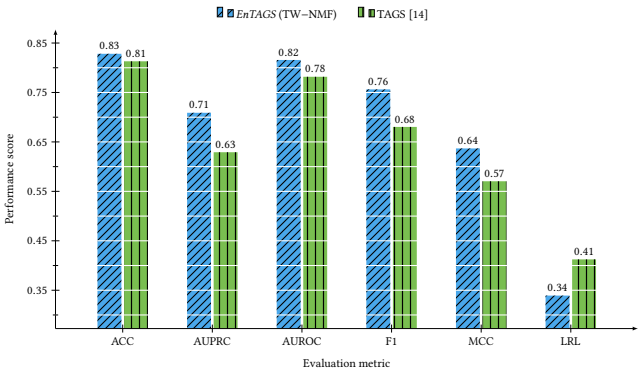


Figure 5: Comparison of the best performing neural model (CNN–LSTM on *EnTAGS*-aggregated NMF-modeled patient data) with the state-of-the-art TAGS model [14].

Table 4: Comparison of the best performing model (CNN–LSTM on *EnTAGS* with TW–NMF) with the state-of-the-art TAGS model [14] and the best performing baseline (CNN on NMF) model.

Evaluation metric	*EnTAGS* with TW–NMF	TAGS [14]	Baseline with TW–NMF	*EnTAGS* Improvement	
				TAGS [14]	Baseline
ACC	**0.8282 ± 0.0023**	0.8130 ± 0.0005	0.8065 ± 0.0033	1.87%	2.70%
AUPRC	**0.7089 ± 0.0046**	0.6291 ± 0.0027	0.6231 ± 0.0073	12.68%	13.77%
AUROC	**0.8157 ± 0.0019**	0.7817 ± 0.0023	0.7707 ± 0.0050	4.35%	5.84%
CE	**17.6853 ± 0.0566**	18.1300 ± 0.1088	18.5063 ± 0.1488	2.45%	4.44%
F1	**0.7562 ± 0.0021**	0.6803 ± 0.0024	0.6739 ± 0.0075	11.16%	12.21%
LRL	**0.3392 ± 0.0036**	0.4124 ± 0.0047	0.4244 ± 0.0095	17.75%	20.08%
MCC	**0.6368 ± 0.0042**	0.5704 ± 0.0020	0.5616 ± 0.0083	11.64%	13.40%

TW representations can provide flexible and interpretable features that can facilitate classification. It is important to understand that patient information captured in nursing notes is often heterogeneous in nature, obtained from multiple sources containing a wide variety of distinctive qualities. Hence, it is by identifying the occurrences of such distinctive qualities that a neural network can facilitate accurate prediction of the associated code groups. The usefulness of NMF in addressing the difficulties of the code group prediction task has been confirmed, as it outperforms the modeling approaches resulting from state-of-the-art embedding systems (see Figure 2). Moreover, owing to the high-dimensionality of the nursing note data, it is easier to train efficient neural models from NMF representations than from raw text. Therefore, we argue that feature learning through NMF modeling is particularity well-suited to train deep neural models, as such techniques have the advantage of adapting to the underlying data and the task at hand.

To enable exhaustive comparison, we juxtapose the results obtained using our models and strategy against those employed in the state-of-the-art work (TAGS model) [14]. The benchmarking of the proposed *EnTAGS* strategy with TAGS and the baseline model is shown in Table 4 and Figure 5. As can be seen from the tabulated results, the proposed *EnTAGS* model outperforms TAGS and the baseline with respect to all the evaluation metrics. We observed an improvement of 12.68% (11.64%) in AUPRC (MCC) over the TAGS model and 13.77% (13.40%) in AUPRC (MCC) over the baseline model. The percentage improvement of the proposed strategy over TAGS and baseline models across other evaluation metrics is tabulated in Table 4. Figure 5 presents a graphical representation corroborating the superior performance of the proposed *EnTAGS* approach against the state-of-the-art method[2].

4 SUMMARY

The clinical task of diagnostic code group prediction is an active research area, and utilizing clinical nursing notes presents an unprecedented opportunity to facilitate the task as a multi-label classification problem. In this study, an efficient data aggregation strategy, *EnTAGS*, which extends the efforts of the state-of-the-art was presented. The proposed *EnTAGS* aggregation strategy aimed at modeling the nursing notes of a patient independent of one another, while aggregating the diagnostic code groups recorded across all the notes of that patient. Moreover, we utilized NMF-based feature modeling (with BoW and TW distributions), which was aimed at capturing the semantic information present in the clinical nursing notes of the patient. We analyzed that feature learning through NMF is particularity well-suited for code group prediction, owing to the heterogeneity of the patient data and the ability of NMF to adapt to such data and the task at hand. We utilized *EnTAGS*-aggregated, NMF-modeled representations of the patient data to train four deep neural architectures, CNN, LSTM, cascaded CNN–LSTM, and partitioned GRU, for the task of ICD-9 code group prediction as a multi-label classification problem. From our experimental validation, we observed that the NMF topic modeling on the TW statistical representation of the nursing notes, when trained using the cascaded CNN–LSTM classifier, resulted in superior performance in comparison to other modeling strategies—we obtained promising results of 82.82% in accuracy, 70.89% in AUPRC, and 63.68% in MCC score; and outperformed the state-of-the-art model by 1.87% in accuracy, 12.68% in AUPRC, and 11.64% in MCC score.

In the current study, ICD-9 code groups of a patient are encoded as one-hot representations. As a part of the future work, we aim at focusing on extending the proposed *EnTAGS* strategy to account for some sense of fuzzy learning and disease predictability, through fuzzification of diagnostic code groups, with higher importance (weighting scheme) for the current diagnostic code groups (corresponding to the latest patient record) and reducing importance (non-zero varying weights) for present and past disease code groups (corresponding to the earlier or later records of the patient), based on suitable distributions of the underlying patient data. We also indent on extending the strategy and models presented in this study to other clinical prediction tasks such as length-of-stay prediction, hospital readmission prediction, (in- and out-patient) mortality prediction, phenotyping, and diagnostic code prediction.

[2]Due to scale variations of CE compared to other metrics, we have not graphed CE in Figure 5; however, the results concerning CE are shown in the Tables 3 and 4.

ACKNOWLEDGMENTS

We acknowledge the use of the facilities at the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, funded by Government of India's DST-SERB Early Career Research Grant (ECR/2017/001056) to Sowmya Kamath S.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*. 280–288.

[3] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. 2018. Improving palliative care with deep learning. *BMC medical informatics and decision making* 18, 4 (2018), 122.

[4] Tian Bai, Ashis Kumar Chanda, Brian L Egleston, and Slobodan Vucetic. 2017. Joint learning of representations of medical concepts and words from ehr data. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 764–769.

[5] Tian Bai, Ashis Kumar Chanda, Brian L Egleston, and Slobodan Vucetic. 2018. EHR phenotyping via jointly embedding medical concepts and words into a unified vector space. *BMC medical informatics and decision making* 18, 4 (2018), 123.

[6] Tian Bai and Slobodan Vucetic. 2019. Improving Medical Code Prediction from Clinical Text via Incorporating Online Knowledge Sources. In *The World Wide Web Conference*. ACM, 72–82.

[7] 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*.

[8] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[9] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*. 301–318.

[10] Sarah A Collins, Kenrick Cato, David Albers, Karen Scott, Peter D Stetson, Suzanne Bakken, and David K Vawdrey. 2013. Relationship between nursing documentation and patients' mortality. *American Journal of Critical Care* 22, 4 (2013), 306–313.

[11] Maurice D Craig. 1994. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing* 32, 3 (1994), 542–552.

[12] Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, and Albert-László Barabasi. 2008. Predicting Individual Disease Risk Based on Medical History. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 769–778. https://doi.org/10.1145/1458082.1458185

[13] Sebastien Dubois, Nathanael Romano, David C Kale, Nigam Shah, and Kenneth Jung. 2017. Effective Representations of Clinical Notes. *arXiv preprint arXiv:1705.07025* (2017).

[14] Tushaar Gangavarapu, Aditya Jayasimha, Gokul S. Krishnan, and Sowmya Kamath S. 2019. TAGS: Towards Automated Classification of Unstructured Clinical Nursing Notes. In *Natural Language Processing and Information Systems*. Springer International Publishing, Cham, 195–207.

[15] Tushaar Gangavarapu, Gokul S Krishnan, and Sowmya Kamath S. 2019. Coherence-Based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 1012–1022. https://doi.org/10.18653/v1/K19-1095

[16] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771* (2017).

[17] J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. 2016. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2015. *ONC data brief* 35 (2016), 1–9.

[18] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. 2019. An empir ical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine* 177 (2019), 141 – 153. https://doi.org/10.1016/j.cmpb.2019.05.024

[19] Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. 2016. Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems*. 1786–1794.

[20] Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. 2013. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing* 62, 1 (2013), 211–224.

[21] Yohan Jo, Lisa Lee, and Shruti Palaskar. 2017. Combining LSTM and latent topic modeling for mortality prediction. *arXiv preprint arXiv:1709.02842* (2017).

[22] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. 2017. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*. 361–376.

[23] 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 (2016), 160035.

[24] William A Knaus, Jack E Zimmerman, Douglas P Wagner, Elizabeth A Draper, and Diane E Lawrence. 1981. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine* 9, 8 (1981), 591–597.

[25] Gokul S Krishnan and S Sowmya Kamath. 2018. 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, 126–134.

[26] Leah S Larkey and W Bruce Croft. 1995. *Automatic assignment of icd9 codes to discharge summaries*. Technical Report. Technical report, University of Massachusetts at Amherst, Amherst, MA.

[27] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. 1993. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Jama* 270, 24 (1993), 2957–2963.

[28] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788. https://doi.org/10.1038/44565

[29] Joon Lee, Daniel J Scott, Mauricio Villarroel, Gari D Clifford, Mohammed Saeed, and Roger G Mark. 2011. Open-access MIMIC-II database for intensive care research. Institute of Electrical and Electronics Engineers.

[30] Nathan Levitan, A Dowlati, SC Remick, HI Tahsildar, LD Sivinski, R Beyth, and AA Rimm. 1999. Rates of initial and recurrent thromboembolic disease among patients with malignancy versus those without malignancy. *Risk analysis using Medicare claims data. Medicine (Baltimore)* 78, 5 (1999), 285–291.

[31] Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88, 3 (2000), 265.

[32] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. 2017. On mixed memberships and symmetric nonnegative matrix factorizations. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2324–2333.

[33] Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 2 (1994), 111–126.

[34] Romain Pirracchio. 2016. 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, 295–313.

[35] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. 2018. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics* 83 (2018), 112–134.

[36] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1, 1 (2018), 18.

[37] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[38] R Rehurek and P Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[39] Donald H Taylor Jr, Truls Østbye, Kenneth M Langa, David Weir, and Brenda L Plassman. 2009. The accuracy of Medicare claims as an epidemiological tool: the case of dementia revisited. *Journal of Alzheimer's Disease* 17, 4 (2009), 807–815.

[40] J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and LG Thijs. 1996. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive care medicine* 22, 7 (1996), 707–710.

[41] Ian ER Waudby-Smith, Nam Tran, Joel A Dubin, and Joon Lee. 2018. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PloS one* 13, 6 (2018), e0198687.

[42] Wolfgang C Winkelmayer, Sebastian Schneeweiss, Helen Mogun, Amanda R Patrick, Jerry Avorn, and Daniel H Solomon. 2005. Identification of individuals with CKD from Medicare claims data: a validation study. *American Journal of Kidney Diseases* 46, 2 (2005), 225–232.

[43] Min Zeng, Min Li, Zhihui Fei, Ying Yu, Yi Pan, and Jianxin Wang. 2019. Automatic ICD-9 coding via deep transfer learning. *Neurocomputing* 324 (2019), 43 – 50. https://doi.org/10.1016/j.neucom.2018.04.081 Deep Learning for Biological/Clinical Data.