NLP Based Clinical Data Analysis for Assessing Readmissions of Patients with COPD

Priyanka V. Medhe[1] **Dinesh D. Puri**[2]
Department of Computer Science and Engineering,
SSBT's College of Engineering and Technology,
North Maharashtra University, Jalgaon

Abstract— **Natural language processing is a computer science field, which focuses on interactions between computers and human (natural) languages. The human languages are ambiguous unlike computer languages, which make its analysis and processing difficult. Most of the data present these days is in unstructured form such as Accident reports, Patient discharge summary, Criminal records etc, which makes it hard for computers to understand for further use and analysis. This unstructured text needs to be converted into structured form by clearly defining the sentence boundaries, word boundaries and context dependent character boundaries for further analysis. With the passage of recent federal legislation many medical institutions are now responsible for reaching target hospital readmission rates. Chronic diseases account for many hospital readmissions and Chronic Obstructive Pulmonary Disease has been recently added to the list of diseases for which the United States government penalizes hospitals incurring excessive readmissions. Though there have been efforts to statistically predict those most in danger of readmission, few have focused primarily on unstructured clinical notes. Here a framework is created for analyzing clinical notes and predicting readmissions of patients. Key steps include many algorithms within the field of data mining and machine learning, so a framework for component selection is created to select the best components. NLP is applied followed by some of processing techniques like, tokenization, stop words removal, stemming, pruning, semantic analysis, POS Tagger etc.**

Keywords— **Chronic Obstructive Pulmonary Disease, Natural Language Processing, Readmissions, Clinical Notes Pre-processing, Prediction Modelling**

I. INTRODUCTION

The Institute of Medicine's report on medical errors demonstrates that adverse events in hospitalized patients are common [1]. A study of 30,121 randomly selected records of hospitalized patients admitted to acute-care hospitals in New York State in 1984 [2] showed that 3.7% had adverse events; of those, 2.6% caused permanent disability, 13.6% caused death, and 28% were negligent. A second study of 15,000 discharges from hospitals in Utah and Colorado in 1992 [3] showed that 2.9% had adverse events; of those, 2.2% caused permanent disability, 6.6% caused death, and 27–32% were negligent. Several studies have attempted to clarify the epidemiology of adverse events [4, 5].

The American Recovery and Reinvestment Act (ARRA) of 2009 [1] emphasized the adoption of health information technology through the Health Information Technology for Economic and Clinical Health Act (HITECH Act) [2]. Two prime components related to this act are Introduction of penalties for hospitals for patient readmission within 30, 60 and 90 day period for specific diagnoses; and Introduction of the concept of Clinical Decision Support Systems (CDSS) in Electronic Health Records through "Meaningful Use" (MU) compliance [3]. Currently, the MU compliance requires a very basic implementation of rule based decision support systems which could be introduced by an office-practice physician based on the combination of demographics, lab results, medications, allergy, and past medical history.

The HITECH Act stipulates that healthcare providers demonstrate the meaningful use of health IT. As part of this act, CMS identified "hospital readmissions for COPD" as a costly problem that needs to be addressed in the United States as a whole [4]. The scope of the problem is very large and cost data is available through CMS. CMS has started penalizing hospitals for excessive 30-day COPD readmissions. As a result, there is an increased amount of pressure on hospitals to adopt the CDSS to identify the candidates for hospital readmission and avoid such readmissions by a series of efforts, such as closely coordinated transition of care. Unfortunately, it is not possible to provide such an extensive level of care for every patient due to the amount of resources needed, shortage in medical staff, and the expenses involved in such care coordination [4]–[6]. Therefore, it is critical to accurately identify candidates for hospital readmission and then avoid such readmission through the use of resources. Further, since patient-hospitalization represents such a large portion of healthcare expenses, health plans, Accountable Care Organizations (ACO), and Managed Services Organizations (MSO) are also targeting hospital readmission in order to improve their profitability. Though predictive modeling for many diseases has seen a large body of research [7]–[10], COPD predictive modeling remains scarce.

The main motivation for this research is the availability of an enormous amount of data that could effectively aid in medical research. These data are mostly available as free text collected through research applications. Processing of these data will provide information that would aid in the research subject recruitment process. This could be achieved by filtering the criteria from the free text to be used in the database queries. Patient data in hospitals includes a significant amount of unstructured data such as physician notes, discharge summaries, and x-ray radiology reports. Since free text is an important part of patient records, including it in predictive analysis is equally important. Despite the inherent value of the clinical information present in the document, a manual review of free text records is very time-consuming process. Therefore, there is interest in developing a Natural Language Processing (NLP) based approach to extract such information from patient records. However, this is not a simple task due to the ambiguity and variations in language used for describing and evaluating any specific patient condition.

User specific use of terminology, abbreviations, and acronyms are often used for describing patient condition. Every physician has a unique style and terminologies for defining a patient problem, encounter or a situation. Due to the variation and complexity in such unstructured information, an architecture which can standardize the information by converting this unstructured data into structured form is required.

For addressing patient safety successfully it requires detecting medical events effectively. As the number of patients seen at medical centres, recognizing events automatically from data which is already available electronically would greatly facilitate patient safety work. Chronic Obstructive pulmonary disease is 3rd amongst all the diseases with highest number of hospital readmissions. Based on the data from 2003-2004, 1 out of 5 Medicare beneficiaries were readmitted for COPD with in a 30-day period with primary diagnoses being COPD. Some problems due to which readmissions occur are improper out-patient management, Poor quality of care and poorly coordinated transition of care. There are many other factors which play an important role in COPD readmission which are more demographics based such as patients residing in low-income areas (7.8 %) are 22 percent more likely to be readmitted then the patients in highest income areas (6.4%) and the readmission rate for black patients' is 30% higher than patients' in any other groups.

II. Background

The Chronic Obstructive pulmonary Disease readmission problem can be reduced by analyzing the discharge summaries and lab reports etc. of the patients who are admitted in a hospital for index diagnoses of COPD. Once, the text analysis on discharge summaries is done, the analyzed text can be divided two categories: Primary factors and Secondary factors. The list of this Primary and Secondary factors is then analyzed with the help of a predictive model and then the probability of readmission will be deduced. So, the background for this thesis can be divided into two categories broadly: Prediction models and Text analysis. The related work includes cTAKES and UIMA. Below is the diagram showing the relationship between Prediction models, Text Analysis, cTAKES and UIMA.

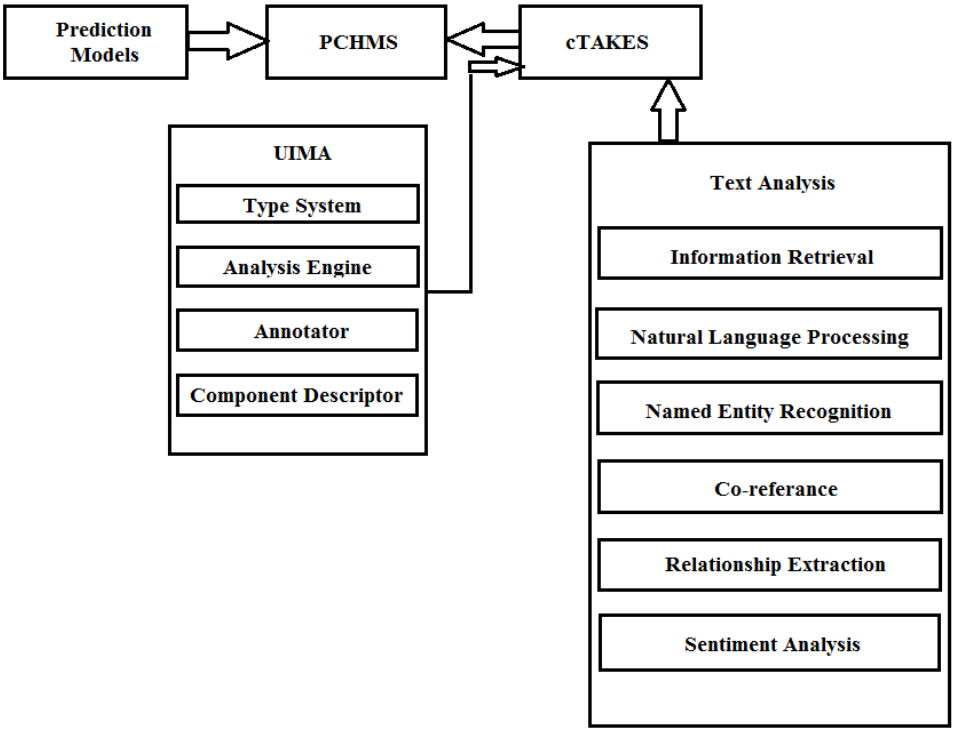

Fig. 1 Literature Survey

A. cTAKES

cTAKES (Clinical Text Analysis and Knowledge Extraction System) is a text analysis system develop by Mayo Clinic. It uses Apache's UIMA (Unstructured Information management Architecture) for converting the unstructured text from patient discharge summaries to structured form. The PCHMS uses cTAKES for all the text analysis and knowledge extraction [11]. It uses Prediction models for assessing the final probability of readmission for a COPD patient.

B. Prediction Models

The process of forming a model to predict the best probability of an outcome is known as Predictive modelling. Almost all the regression models can be used for predicting. The predictive models can be divided in three main categories: Parametric Predictive modelling, Non-parametric predictive modelling and semi-parametric predictive modelling. The parametric predictive models assume one or more parameters which might affect the overall distribution of outcomes and in turn their probability. The non-parametric predictive models use no parameters for determining the distribution of outcomes and their probability. The semi-parametric predictive models include features of both.

C. Text Analysis

The text-based materials are very important source of valuable information and knowledge. There are varying pieces of text which can be treated as Information source such as: Discharge Summaries for Health Care, Accident reports for Road Safety etc. All the sources provide experiment results/summaries as free text which is easily readable by human, but complex for computers to understand. Some of the text can be very complex such as Protein –Protein relationship etc. which needs more complex system for analysis. In order for a system to be most accurate in analysis it should be able to mimic the way the human brain works. Human brain divided every piece of information into set of small bits and then analyse from there. In the same way, the text analysis system divides the process into several steps and each step extracts a certain piece of information is captured. Below are the set of most important components for a text analysis system: Information Retrieval [12], Natural Language Processing [13], Named Entity Recognition [14], Co-reference [15], Relationship Extraction [16] and Sentiment Analysis.

D. UIMA

The data in today's world is in unstructured form like in discharge summaries, handwritten notes, research results/reports etc. This unstructured information is very complex because this information is written with no standard structure being followed. The text may contain spelling mistakes, the author's attitude and the main complexity arises when every person has their own terminologies for defining a problem or situation. Due to the increase in such complex unstructured information, we needed an architecture which can convert this unstructured data into structured form in a standardized manner. The new architecture needs to extract the information which can be later related to concepts and events [17]. The Unstructured Information Management Architecture (UIMA) is divided into four main parts: Type System, Analysis Engine, Annotator and CAS. UIMA is also used by: NLM's MetaMap, YTEX and Detect – HA.

III. RELATED WORK

A systematic review was performed in 2011 by Kansagara et al. which compares data, methodology, and results [18]. The review confirmed that readmission prediction is a difficult problem and recent models do not necessarily perform better than research a decade prior.

A framework was created by researchers at Deakin University to analyse many chronic disease readmissions [19]. The system works by creating schemas to be used when capturing data for a patient. In the case of COPD, a COPD specific template is used. Disease specific models are then built. 1,816 patients were analysed. The model is able to predict 30-day readmission rate in COPD patients with an AUC=0.67 and was an improvement upon co-morbidity baseline methods that are often used for readmission analysis.

Another system which is specific to COPD patients was created by Fan et al. [20]. This system was not however used in the analysis of hospital readmissions. Instead, patients were analysed for COPD exacerbations within the period of one year. Baseline methods for comparison used a model consisting of basic features such as demographics and questionnaire information. An improved model was presented which also included the features spirometry, PaO2, dyspnea, prior exacerbations and co-morbidity. The AUC for this model was 0.68.

Work by Wasfy et al. attempts to use the unstructured data contained in the Electronic Health Record (EHR) to predict 30 day readmission in percutaneous coronary intervention patients [10]. The primary method of NLP in this study was the use of regular expressions to extract specific queries from the clinical note. Although using regular expression based queries can be useful, automatic discovery of new features which may be useful is not possible. The AUC for this research was 0.69. The data distribution was changed to approximately 0.33 readmitted and 0.67 not readmitted.

Recent research by Duggal et al. uses Apache cTAKES to annotate the unstructured EHR [21]. This research specifically looks at the 30-day readmission rate of diabetes patients in an Indian hospital. The data contains 0.129 readmission rate and 9,381 instances. Several machine learning algorithms were compared. Naive Bayes, Logistic Regression, Random Forest, Adaboost, and Neural Networks. The highest AUC for this diabetes study was 0.688 using Random Forests. The results are typical of readmission analysis.

IV. PROPOSED SOLUTION

The proposed system consists of four subsystems: (1) Clinical notes preprocessing (2) Feature Analysis (3) Classification and (4) Performance evaluation. Following assumptions can be made regarding the proposed solution:

- There exists a repository of rich clinical information that we hypothesize contains useful data about patient safety.
- The data are obscured, however, due to the way they are recorded. Much is in narrative form and therefore not amenable to traditional statistical analysis, and even when coded, the data are stored in complex, nested structures that may be difficult to use.
- A set of informatics tools exists such as natural language processing, machine learning, etc. which are capable of extracting the useful patient safety information from the repository in an automated or partially automated fashion.

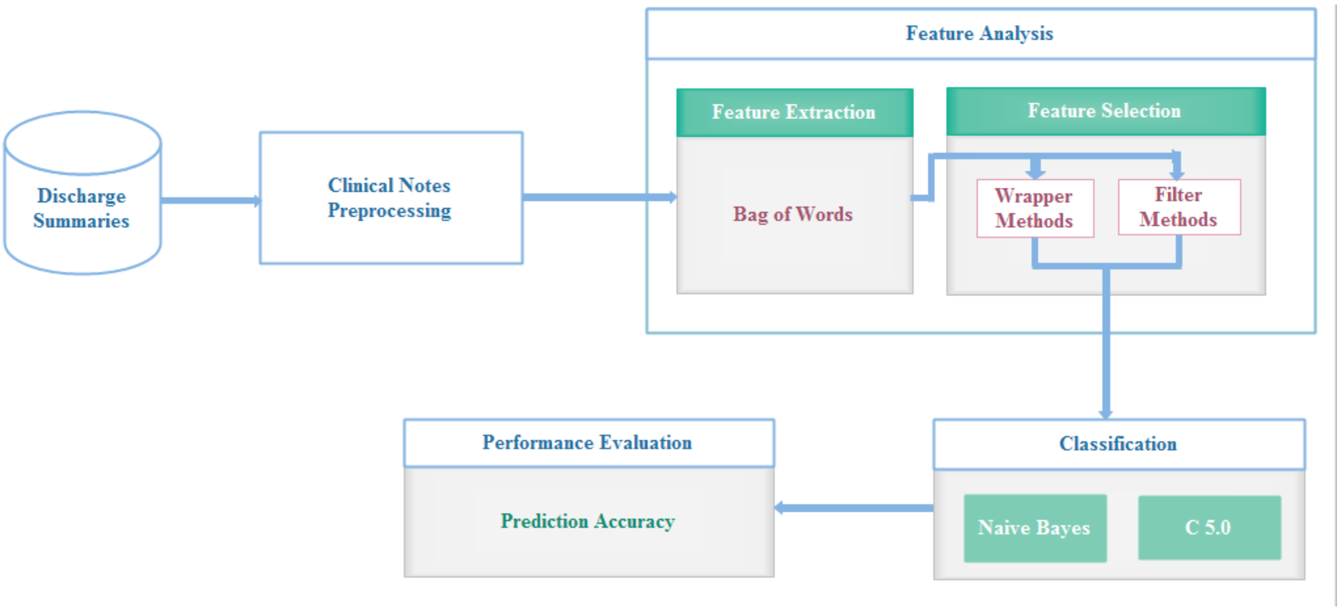

Fig. 2 Architecture of Proposed System

A. *Clinical Notes Pre-processing*

In clinical notes pre-processing the first step is to pre-process the documents, i.e. converting the documents into appropriate needs of data schemes and the second step is to analyse the available data from first step and divide it into clusters. This process is carried out by clustering algorithm. The CNP consist of following components:

- The POS tagger
- Stop words Removal
- Stemming
- Collecting synonyms & hypernym
- Calculate Frequency (Weighting)

B. *Feature Analysis*

In order to build a predictive model, features must be extracted from unprocessed data. A feature is an individual measurable property of the phenomenon being observed. This process can either be manual or algorithmic. Incorporation of domain knowledge can increase the quality of these features and in turn increase the quality of the predictive model. This proposed model has two distinct phases of feature analysis, feature extraction and feature selection.

1) Feature Extraction: Given a piece of natural language, several methods exist for extracting features. Here in this system Bag-of-Words method is used for extracting features. The bag-of-words representation method treats each word in the corpus as a feature. Each document is an instance and each word is either present or not-present in the instance. This method is simple and can be used with most any natural language document. No additional domain knowledge is required to prepare the data. Though simplistic, bag-of-words often produces good results with minimal feature engineering and can be combined with other techniques such as tf-idf to give unequal weighting based on document frequency.

2) Feature Selection: A method used to reduce the number of features is known as feature selection. Ideally, removing features which offer little or no information to the classification algorithm is desired. Feature selection can be broadly categorized into three groups: (1) Filter (2) Wrapper and (3) Embedded. Filter methods use statistical tests to rank features by relevance. They are typically quick to compute compared to other methods but may not find an optimal set of features. Wrapper methods test all possible combinations of features with a fixed classification algorithm and use a performance metric such as accuracy to find the highest score. It may be possible to find the most useful features using wrapper methods, but this method is computationally expensive and will often lead to over fitting. Embedded methods have feature selection built into the classification algorithm. The C5.0 decision tree algorithm is an example of embedded feature selection as it uses Information Gain Ratio to select which features to use in building tree nodes. Filter and wrapper feature selection methods are evaluated in this research and four methods are analyzed to determine which is most useful for final inclusion in the framework.

C. *Classification*

1) Naïve Bayes: Naïve Bayes (NB) is a simple probabilistic classifier that is based upon Bayes' theorem. The classifier assumes independence between features. Though many times this independence assumption is not true, in practice NB still works well. NB uses little memory and can classify new instances quickly. Early methods for e-mail spam detection used NB due to this speed. NB is known to work well in text classification contexts and was chosen for this quality.

2) C5.0: C5.0 is based on the information gain ratio that is evaluated by entropy. To select the test features at each node in the tree, the information gain ratio measure is used. Such a measure is referred to as a feature (attribute) selection measure. The attribute with the highest information gain ratio is chosen as the test feature for the current node. Let D be a set consisting of (D1… Dj) data instance. The class label attribute consist of m distinct values defining m distinct classes, Ci(for I = 1,…,m) and Dj be the number of samples of D in class Ci. The expected information needed to classify a given sample as follows;

$$\text{Splitinfo}_A(D) = -\sum(|Dj|/|D|) * \log((|Dj|/|D|))$$

$$\text{Gain ratio}(A)=\text{Gain}(A) / \text{Splitinfo}_A(D)$$

Where,

$$\text{Gain} = \text{Info}(D)-\text{Info}_A(D)$$

$$\text{Info}(D) = -\sum Pi \log2(Pi)$$

And,

$$\text{Info}_A(D) = -\sum(|Dj|/|D|) * \text{Info}(Dj)$$

$$\text{Info}_A(D)=- \Sigma (|Dj|/|D|)* \text{Info}(Dj)$$

Where Pi = probability of distinct class Ci,D =data Set, A=Sub attribute from attribute, (|Dj|/|D|)=act as weight of jth partition. In other words, Gain (A) is the expected reduction in entropy caused by knowing the value of feature A.

D. Performance Evaluation

An analysis of the AUC of classifiers will be performed. The best performing algorithms will be determined. The statistical methods will be iteratively compared using NB classifier and C5.0 while varying the features selected.

V. Conclusions

This readmission analysis system represents a natural language approach to patient readmission prediction. This approach offers the advantage that separate data collection is not required for readmission prediction since clinical notes are already collected by medical institutions. Additionally, unstructured data requires no data format conversions to be evaluated by an external system. Structured systems using RDBMS typically require many data conversion steps to reach an expected data format. Thus, this system presents easy integration into existing EHR systems. With the increase in EHR systems, clinical notes will become increasingly important and NLP techniques will need to be considered when creating decision support systems. The results may show the importance of feature selection and model creation time to the implementation of practical systems.

Acknowledgment

I express my deep sense of gratitude to the Prof. Dr. K. S. Wani (Principal of SSBT's College of Engineering and Technology), and Prof. Dr. Girish K. Patnaik(HOD of Department of Computer Engineering) and Prof. Mr. Dinesh D. Puri for their cooperation and guidance in completion of this work.

References

[1] "The American Recovery and Reinvestment Act of 2009 Report." [Online]. Available: http:// www.gpo.gov/fdsys/pkg/BILLS-111hr1enr/pdf/BILLS-111hr1enr.pdf. [Accessed: 23-Aug-2016].

[2] "HITECH Act Enforcement Interim Final Rule." [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/administrative/enforcementrule/hitechenforcementifr.html.[Accessed: 23-Aug-2016]

[3] "Electronic Health Records (EHR) Incentive Programs." [Online]. Available: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html. [Accessed: 23-Aug-2016].

[4] D. Goodman, E. Fisher, and C. Chang, "The Revolving Door: A Report on US Hospital Readmissions," *Princeton, NJ Robert Wood Johnson Found*, 2013.

[5] P. Jain, *Prognostic COPD healthcare management system*, no. May. FLORIDA ATLANTIC UNIVERSITY, 2014.

[6] R. Behara, A. Agarwal, F. Fatteh, and B. Furht, "Predicting Hospital Readmission Risk for COPD Using EHR Information," in *Handbook of Medical and Healthcare Technologies*, Springer, 2013, pp. 297–308.

[7] R. Behara, A. Agarwal, V. Rao, and C. Baechle, "Predicting the Occurrence of Diabetes using Analytics," in *Models and Applications in the Decision Sciences: Best Papers from the 2015 Annual Conference*, 1st ed., Pearson Press, 2016, pp. 187–193.

[8] R. Behara, A. Agarwal, V. Rao, and C. Baechle, "Predictive Analytics for Chronic Diabetes Care," in *2015 Annual Meeting of the Decision Sciences Institute Proceedings*, 2015.

[9] R. Wallmann, J. Llorca, I. Gómez-Acebo, Á. C. Ortega, F. R. Roldan, and T. Dierssen-Sotos, "Prediction of 30-day cardiac-related-emergency-readmissions using simple administrative hospital data," *Int. J. Cardiol.*, vol. 164, no. 2, pp. 193–200, 2013.

[10] J. H. Wasfy, G. Singal, C. O'Brien, D. M. Blumenthal, K. F. Kennedy, J. B. Strom, J. A. Spertus, L. Mauri, S. L. T. Normand, and R. W. Yeh, "Enhancing the Prediction of 30-Day Readmission after Percutaneous Coronary Intervention Using Data Extracted by Querying of the Electronic Health Record," *Circ. Cardiovasc. Qual. Outcomes*, vol. 8, no. 5, pp. 477–485, 2015

[11] G. K. Savova, J. J. Masanz, P. V Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 5, pp. 507–513, 2010.

[12] Sneiderman C, Demner-Fushman D, Fiszman M, Ide N, Rindflesch T:Knowledge-based methods to help clinicians find answers in MEDLINE. J Am Med Inform Assoc 2007, 14:772-780.

[13] Ah-Hwee Tan. Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases KDAD'99, page 65-70. (1999).

[14] J. J. Jung. Online named entity recognition method for microtexts in social networking services: A case study of twitter. Expert Systems with Applications, Volume 39, Issue 9, July 2012, Pages 8066-8070.

[15] Soon W, Lim D, Ng H (2001) A machine learning approach to coreference resolution of noun phrases. J Comput Linguist 27(4):521–544.

[16] Ce Gao,Yixu Song, Peifa Jia. A Fast Events Relationship Extraction Method Based on Semi-CRFs. Knowledge Acquisition and Modeling, 2009.

[17] Michael Thomas Egner, Markus Lorch and Edd Biddle. UIMA GRID:Distributed Large-scale Text Analysis, 2007, IEEE.

[18] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, “CLINICIAN ’ S CORNER Risk Prediction Models for Hospital Readmission A Systematic Review,” *Jama*, vol. 306, no. 15, pp. 1688–1698, 2011.

[19] T. Tran, W. Luo, D. Phung, S. Gupta, S. Rana, R. Kennedy, A. Larkins, and S. Venkatesh, “A framework for feature extraction from hospital medical data with applications in risk prediction.,” *BMC Bioinformatics*, vol. 15, no. 1, p. 6596, 2014.

[20] V. S. Fan, S. D. Ramsey, B. J. Make, and F. J. Martinez, “Physiologic variables and functional status independently predict COPD hospitalizations and emergency department visits in patients with severe COPD,” *COPD J. Chronic Obstr. Pulm. Dis.*, vol. 4, no. 1, pp. 29–39, 2007.

[21] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, “Predictive risk modelling for early hospital readmission of patients with diabetes in India,” *Int. J. Diabetes Dev. Ctries.*, 2016.