Received August 29, 2016, accepted October 3, 2016, date of publication October 19, 2016, date of current version November 28, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2618775

Predicting Complications in Critical Care Using Heterogeneous Clinical Data

VIJAY HUDDAR[1], BAPU KOUNDINYA DESIRAJU[2], VAIBHAV RAJAN[1], SAKYAJIT BHATTACHARYA[1], SHOURYA ROY[1], AND CHANDAN K. REDDY[3], (Senior Member, IEEE)

[1] Xerox Research Centre India, Bangalore 560103, India
[2] Institute of Genomics and Integrative Biology, New Delhi 110025, India
[3] Department of Computer Science, Virginia Tech, Arlington, VA 22203, USA

Corresponding author: V. Rajan (vaibhav.rajan@xerox.com)

This work was supported by the U.S. National Science Foundation under Grant IIS-1231742 and Grant IIS-1527827.

ABSTRACT Patients in hospitals, particularly in critical care, are susceptible to many complications affecting morbidity and mortality. Digitized clinical data in electronic medical records can be effectively used to develop machine learning models to identify patients at risk of complications early and provide prioritized care to prevent complications. However, clinical data from heterogeneous sources within hospitals pose significant modeling challenges. In particular, unstructured clinical notes are a valuable source of information containing regular assessments of the patient's condition but contain inconsistent abbreviations and lack the structure of formal documents. Our contributions in this paper are twofold. First, we present a new preprocessing technique for extracting features from informal clinical notes that can be used in a classification model to identify patients at risk of developing complications. Second, we explore the use of collective matrix factorization, a multi-view learning technique, to model heterogeneous clinical data—text-based features in combination with other measurements, such as clinical investigations, comorbidites, and demographic data. We present a detailed case study on postoperative respiratory failure using more than 700 patient records from the MIMIC II database. Our experiments demonstrate the efficacy of our preprocessing technique in extracting discriminatory features from clinical notes as well as the benefits of multi-view learning to combine clinical measurements with text data for predicting complications.

INDEX TERMS Clinical notes, topic models, heterogeneous data, multi–view learning, collective matrix factorization, postoperative respiratory failure.

I. INTRODUCTION

An Intensive Care Unit (ICU) monitors and treats the most critically ill patients in a hospital. ICUs typically have high staff–to–patient ratio and the most advanced medical resources compared to other units. Critically ill patients are often transferred to ICUs from Emergency or Surgery for continuous monitoring and support.

Unfortunately, advanced medical interventions in critical care themselves make the patients vulnerable to several complications [1]. Several infections, such as ventilator associated pneumonia, central line associated bloodstream infection and catheter associated urinary tract infection, are associated with invasive devices used in ICUs. Infections due to previous procedures undertaken are also possible, such as surgical site infections [1]. Ventilatory support provided to many ICU patients are associated with complications like short-term and long-term intubation, barotrauma, gastrointestinal tract bleeding, and weaning errors [2]. Unrecognized drug interactions can also cause problems, including acute renal failure. In addition, ICU patients are susceptible to nutritional complications, acid base problems, and psychological disturbances [2]. Furthermore, ICU survivors are often known to have long-term physical, neuropsychiatric, and quality of life impairments [3].

As the need for ICUs have grown worldwide [4], more ICUs have been created but the availability of resources, both clinical staff and monitoring equipment, remain limited due to many practical constraints. Accurate knowledge of the etiology of ICU complications is often lacking, leading to the inability of accurate identification of high–risk patients and prevention of complications. As a result, in many cases, current medical intervention is reactive and adequate care is provided to patients *after* a complication develops.

The availability of digitized clinical data through Electronic Medical Records (EMR) in hospitals is increasing throughout the world. In particular, ICUs are data–rich environments where multiple parameters of patients are continuously monitored at higher rates and volume compared to other hospital units. This data presents an unprecedented opportunity to study and gain deeper understanding of diseases, develop new treatments and improve healthcare systems. There is, naturally, tremendous interest in identifying high–risk individuals early, much before the onset of disease, to provide preventive care and reduce the morbidity and mortality of complications.

Along with potential opportunities, the data also poses substantial challenges. Patient information is distributed in several heterogeneous sources such as laboratory tests, continuous waveforms, nursing notes, radiology images and reports, genomic data and so on. In addition there are problems associated with privacy, integration of distinct databases, inconsistently used clinical concepts across databases, artifacts due to noise, measurement errors and missing data [5]. The analysis, interpretation, and presentation of this data in a clinically relevant and usable format is one of the main challenges faced by clinical decision support systems [6].

Structured data available as numerical measurements (e.g. lab tests) are easily amenable to standard statistical analysis. However, unstructured and high–dimensional data such as free–text and images also contain valuable information about a patient's condition. Clinical notes, in particular, contain both subjective and objective assessments of a patient's condition. They are recorded by the nursing staff and visiting doctors who monitor the patient regularly (see Figure 3 for an example). However, they are informally written, contain a large number of inconsistently used abbreviations and lack the linguistic structure of more formal documents like radiology reports and discharge summaries. Hence standard tools of linguistic analysis are not effective on such notes.

In this paper we investigate the use of such informally written nursing notes for predicting complications arising in critical care. We also explore multi–view learning based techniques to effectively combine features obtained from nursing notes and other clinical measurements for predictive modeling. We present a detailed case study on predicting postoperative acute respiratory failure and demonstrate the efficacy of our text–based feature extraction and multi–view learning approach for predictive modeling.

The rest of the paper is organized as follows. We begin with a brief overview of related work in predictive analytics in critical care. In section III we describe the heterogeneous sources of data available in hospitals and critical care. Our predictive system is detailed in section IV which describes our text preprocessing techniques for extracting information from nursing notes as well as our multi–view learning based system for predictive modeling. We present a detailed case study on an ICU complication – postoperative respiratory failure – in section V. Empirical results on predicting postoperative respiratory failure using our predictive system are presented in section VI. We conclude in section VII.

II. RELATED WORK

There have been many attempts at exploiting the richness and heterogeneity of clinical data in electronic medical records. A comprehensive introduction can be found in [7]. Systems to provide services such as identifying patients at risk for complications, personalized treatment planning, and hospital resource management are being actively developed. Some well known examples include ICDA, a platform for intelligent care delivery analytics [8] and MatrixFlow, for analysis of disease progression using clinical event sequences [9]. However, there many unresolved challenges related to healthcare data collection, management and modeling. Johnson et al. [5] present a detailed review on such challenges in critical care.

A. TIME SERIES MODELS

A large body of literature can be found on modeling clinical data with applications in physiological signal analysis and disease risk prediction. Modeling the temporality of multivariate clinical data that is noisy, sparse and unevenly sampled presents significant challenges. For example, Mao et al. [10] evaluate various features extracted from clinical time series signals for classification models used in deterioration warning in ICU patients. Wiens et al. [11] study time-series classification techniques to identify patients at risk of hospital acquired infections. A nonparametric Bayesian approach for feature construction from longitudinal health data is presented in [12]. Dürichen et al. [13] propose multi-task Gaussian processes for multivariate physiological time-series analysis. Multi-task Gaussian processes have also been used by Ghassemi et al. [14] to model noisy, heterogeneous and unevenly sampled temporal clinical data, including physiological signals and clinical notes. Batal et al. [15] explore a different temporal pattern mining approach to design classification models for multivariate temporal clinical data. A significant fraction of clinical data is not continuous–valued but categorical such as billing codes, diagnosis and procedure codes, demographic variables etc. There has not been much progress in modeling temporal categorical variables that is also sparse and irregular. Recent progress in this direction is through an interesting use of renewal processes [16]. Another recent work in modeling heterogeneous dynamic clinical data is based on Generalized Linear Dynamic Models [17].

B. DISEASE PROGRESSION MODELS

A related application is modeling the progression of diseases which is useful both within a hospital episode as well as for more chronic illnesses. For example, Wang et al. [18] builds disease progression models using clinical findings and comorbidities using incomplete and heterogeneous patient records. Cohen et al. [19] use clustering of 45 physiological, clinical and treatment variables from ICUs to identify complex metabolic states of patients and use these clusters to track patient states over time. Zhou et al. [20] propose a fused group lasso formulation for disease progression models using lab tests, cognitive scores, genetic and demographic data. A multivariate context-sensitive Hawkes process is used to simultaneously infer disease relationship network and model temporal progression of patients by Choi et al. [21] that can be used to predict patient-specific future diseases.

C. COMPUTATIONAL PHENOTYPING

The related area of computational phenotyping aims to computationally discover meaningful representations of phenotypes, primarily from EMR data, that are clinically useful. The most commonly investigated phenotypes are those that provide clinically useful descriptions of various diseases. These inferred phenotypes can then be used in designing predictive models for diseases. Pivovarov et al. [22] present a graphical model for large-scale discovery of disease models using heterogeneous clinical data (notes, laboratory tests, medications and diagnosis codes). Marlin et al. [23] investigate 13 clinical variables – vitals and lab measurements like heart rate, respiration rate, glucose etc. – in the context of mortality prediction and discover phenotypes, based on cluster patterns, for mortality risk. A similar study in the context of mortality risk, is by Che et al. [24] who specifically uncover phenotypes related to septic shock and circulatory diseases. Kale et al. [25] also use physiological streams to discover phenotypes in critically ill neonates. Ho et al. [26] study phenotypes based on chronic disease indicators and procedure and diagnosis codes. Their phenotypes are represented as tensors where each mode denotes a medical event. Diagnosis codes are also used by Zhou et al. [27] for phenotyping renal disease and congestive heart failure. Schulam et al. [28] studied phenotypes for scleroderma (an autoimmune disease) with the aim of discovering disease subtypes using physiological data. In all these works phenotype representations are given by patterns, based on specific medical features, obtained by sequential mining or temporal abstractions. A different graph–based representation is proposed by Liu et al. [29] that uses diagnosis and medication codes from EMRs.

D. MINING CLINICAL NOTES

The value of clinical notes for providing deeper insights into clinical practice and understanding diseases is well recognized. Finlayson et al. [30] have used over 20 million notes spanning 19 years of data to build a graph of inter-related clinical concepts which in turn can be used to mine association rules, risks of various clinical events as well as patterns of comorbidity, drug-drug, and drug-disease interactions. Clinical notes have been used to model disease symptom and medication relationships ([31]) and to study symptom–symptom interactions ([32]) which in turn have been used to identify previously unknown potential symptoms of diseases. Deeper knowledge of risks and symptoms can directly improve risk models for various diseases. Text mining has also been used for pharmacovigilance for adverse drug event detection and assessment [33]. Hripcsak et al. [34] demonstrate that text analytics can detect clinical conditions in chest X-rays with a consistency that is indistinguishable from that of physicians reviewing the same reports. Adverse event identification due to central venous catheters have used nursing notes in ICU [35]. A recent study by Murff et al. [36] explores the use of text analytics to predict several postoperative complications. They conclude that NLP analysis of EMR data to identify postoperative complications have higher sensitivity and lower specificity compared with patient safety indicators based on discharge coding. In section V we discuss previous studies on postoperative respiratory failure, some of which use text sources for their analysis. A few recent studies have investigated the use of topic modeling for analyzing free–text clinical notes [14], [37], [38]. In all these studies machine learning based models, that effectively use the information in clinical notes, are found to outperform traditional scoring systems (like SAPS-I) in identifying patients with high mortality risk.

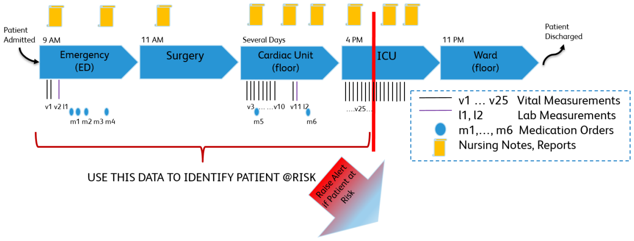


FIGURE 1. Schematic of patient trajectory in a single hospital episode. Various clinical measurements are made throughout the duration of the stay. See Table 1 for a brief description of the measurements. A patient can develop a complication at any time and our aim is to use all the past measurements to identify patients at risk of developing a given complication.

III. CLINICAL DATA IN CRITICAL CARE: CHALLENGES

Figure 1 shows a schematic of a patient in a hospital. A patient is typically moved from one unit to another and in each unit clinical data about the patient is recorded - either digitally in electronic medical records (EMR) or on paper. Table 1 briefly describes the various data sources available in hospitals (including critical care). Standardized codes are available to encode the various data sources, for example LOINC codes for lab tests, RXNORM codes for medication orders and ICD codes for diagnoses and procedures [7].

Two significant challenges in modeling clinical data to determine a patient's clinical state are posed by the heterogeneity and temporality of the data.

A. HETEROGENEITY

Table 1 provides a glimpse of the heterogeneity of clinical data that may be present in a single hospital episode. We have, among others, numerical measurements, text data and image data as well as demographic data. In addition there may be genetic data and data acquired from various wearable

TABLE 1. Various sources of clinical data available in electronic medical records and their characteristics.

Data	Type	Characteristics	Examples
Vitals	Numerical, Temporal	Typically measured every second/minute within ICU and every few hours outside ICU	Blood Pressure, Respiration Rate, Heart Rate
Lab Tests	Numerical	Typically measured a few times, investigation depends on patient's condition and diagnoses	Blood Glucose, Uric Acid
Medication Orders	Numerical, Temporal	Physician orders of prescribed medications	Insulin, Aspirin
Procedures	Numerical, Temporal	Medical/surgical procedures performed on the patient	Craniotomy, endoscopy
Diagnoses	Numerical, Temporal	Diagnoses of past and current conditions	Sepsis, Diabetes
Nursing Notes	Text, Temporal	Assessment of patient's condition including subjective observations	See Figure 3
Radiology	Image, Text	Radiology images accompanied by reports from the radiologists	X-Ray, CT Scan
Demographic	Numerical, Static	Demographic details of the patients	Age, gender, ethnicity

sensors before the hospital episode that may also be used to model a patient's clinical states. Studies have also tried to use data available in social media where patients might provide information about their condition [39]. Features extracted from diverse data sources differ in their datatypes, noise characteristics and signal–to–noise ratio. For example, vitals and lab measurements are typically continuous–valued, procedure and diagnosis codes are encoded using standardized categories and comorbidities are usually in binary format (indicating presence/absence). Features extracted from text and images are very high–dimensional and require suitable preprocessing before modeling. Exploring and modeling the dependencies between such heterogeneous data sources is the aim of statistical techniques such as multi–view learning.

B. TEMPORALITY

There is a notion of temporality in all data where multiple measurements are made during a patient's stay. For example, repeated measurements of blood pressure, temperature or multiple nursing notes or radiology investigations. Some lab tests could also be done repeatedly to assess the patient's condition during a single hospital episode. The sampling rate of these measurements can vary from being highly irregular – only a few times during a patient's stay – to being high frequency data – continuous measurements (every second) through bedside monitors in ICU. Thus traditional time-series modeling approaches that assume regular sampling may not always be applicable on such data. Several different methods have been proposed to model or obtain useful features from such data, when all the measurements are continuous–valued: (1) Using statistical summaries within pre-determined windows to obtain features (e.g. [11]) (2) Using temporal pattern mining techniques (e.g. [15]) (3) Using multi-task Gaussian processes to model unevenly sampled temporal data (e.g. [14]). However many measurements are encoded in binary or categorical data types, for example diagnosis codes or standard risk scores and may also be temporal due to repeated measurements. Not many techniques are available to model such data (a recent work uses renewal processes [16]).

In addition to heterogeneity and temporality, there are other challenges due to sparsity, noise and missing data. Missing clinical data requires special care while modeling – unlike in many other contexts, data is not always missing–at–random and may be missing by choice. For example, each patient undergoes specific lab tests that is determined by the doctor after assessing his/her clinical condition. Thus a lab test appears missing because the doctor does not find it necessary to conduct, given the patient's condition. This can easily be overlooked while analyzing data from patients with diverse clinical conditions. In addition there are problems due to privacy concerns and inconsistent definitions of clinical concepts across various compartmentalized data archives within hospitals. These issues are well described in [5]. Another challenge lies in the effective use of domain knowledge in predictive models. A wealth of data is present in biomedical literature and on the web. The use of such knowledge can potentially enhance the predictive accuracy of models. Examples of such techniques include priors from ontologies in deep learning [24] and the use of hierarchical priors in topic modeling [40].

The source of our data is MIMIC II [41], a publicly available database, part of Physionet [42], containing physiological signals and clinical data of more than 2300 patients in Critical Care. The descriptions of the data in the following are strictly only for this dataset. However many of the data characteristics are common across hospitals. In this paper we use only data available in the critical care (from MIMIC II) but data from the hospital stay before ICU entry can also be used with the same techniques that are described here.

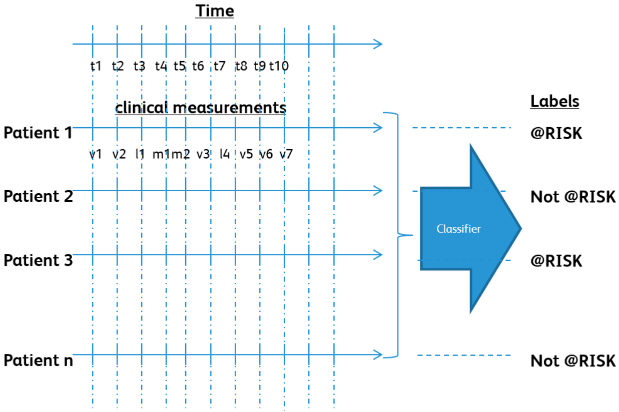


FIGURE 2. Classification model for identifying patients at risk of a given complication. Clinical measurements are used to obtain features that are used to train a classifier. The trained classifier is used to identify patients at risk of developing the complication.

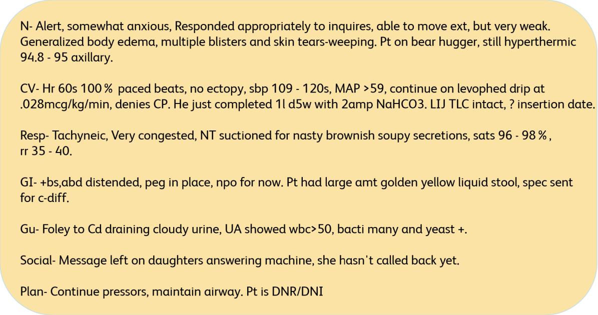
IV. FEATURE DESIGN FOR A PREDICTIVE SYSTEM

We build a predictive system using a classification model. Figure 2 shows a schematic of a classifier used to identify patients at risk of a complication. Historical data of patients is used to train a binary classifier after previous cases of the specific complication are identified and suitably labeled. During prediction the classifier uses the input data to output a label. If the label indicates the patient to be at risk, the clinical staff can be alerted for appropriate intervention.

The first step towards building such a classifier is to preprocess the data and extract features that are used to train the classifier. In the following subsections we describe our preprocessing and feature extraction steps from each of the data sources. While extracting features from numerical data such as lab tests and medication orders are straightforward and have been studied previously [7], extracting features from text data in nursing notes poses additional challenges. We discuss these in detail and describe new methods of extracting features from nursing notes. Finally we also discuss ways of combining the features from heterogeneous sources.

A. TEXT PREPROCESSING AND FEATURE EXTRACTION

Text sources of data include nursing notes and investigation reports. Nursing notes are informally written notes that record the condition of patients. It contains both subjective and objective evaluation of the patients. They are typically written every 3–4 hours in critical care. Figure 3 shows excerpts from an anonymized nursing note from the MIMIC II database.



N- Alert, somewhat anxious, Responded appropriately to inquires, able to move ext, but very weak. Generalized body edema, multiple blisters and skin tears-weeping. Pt on bear hugger, still hyperthermic 94.8 - 95 axillary.

CV- Hr 60s 100 % paced beats, no ectopy, sbp 109 - 120s, MAP >59, continue on levophed drip at .028mcg/kg/min, denies CP. He just completed 1l d5w with 2amp NaHCO3. LIJ TLC intact, ? insertion date.

Resp- Tachyneic, Very congested, NT suctioned for nasty brownish soupy secretions, sats 96 - 98 %, rr 35 - 40.

GI- +bs,abd distended, peg in place, npo for now. Pt had large amt golden yellow liquid stool, spec sent for c-diff.

Gu- Foley to Cd draining cloudy urine, UA showed wbc>50, bacti many and yeast +.

Social- Message left on daughters answering machine, she hasn't called back yet.

Plan- Continue pressors, maintain airway. Pt is DNR/DNI

FIGURE 3. Sample de-identified nursing note from critical care. Note the informal usage and absence of grammatical structure in sentences. The descriptions are organized into various sections that begin with headings: CV (Cardio-Vascular), RESP (Respiratory), GU (Genito-Urinary), GI (Gastro-Intestinal) etc., loosely based on different human physiological systems.

A discharge summary is a report written at the end of a patient's stay in the hospital. It typically includes details of the patient, the healthcare professionals involved during the stay, diagnoses, investigations and complications during the stay, past medical conditions as well as present and future treatment plans. Discharge summaries are intentionally excluded from our analysis since they are written at the end of the patient's stay and cannot be used in a real-time prediction system within the ICU. This also makes the problem harder since discharge summaries contain comprehensive information of patients' past and current medical history which nursing notes lack. Discharge summaries are formal documents and systems analyzing them (using linguistic techniques) rely on their grammatical structure. In comparison, nursing notes are informally written and contain nonstandard and inconsistently used abbreviations.

Observing the nursing notes, we notice that the data is not completely unstructured but is structured into various headings such as "CV", "RESP" etc. See Figure 3 for an example. These headings are neither consistent nor unique; for example, "CV" is also written as "CARDIO" and "CARD" in some notes.

The key idea of our preprocessing method lies in realizing that the importance of a word or phrase in the text, in the context of a complication, is relative to the heading within which it resides. The significance of the same word (e.g. 'blood') differs when it is under the heading "CV" (cardiovascular) than when it is under "GI" (genito-urinary). Hence, we extract features for each heading separately and assign an importance value to each word based on its frequency of occurrence in the training data. Only a fixed percentage of the extracted words are used in further preprocessing of the text. The complete sequence of steps performed is listed below. We denote by *text observation* all the text data for a single patient concatenated together which includes nursing notes and investigative reports of a patient but excludes the discharge summary.

- Extract all the headings from all text observations using predefined rules that identify headings. For example, a word in the beginning of a sentence, followed by colon is considered a heading.
- Eliminate headings and regroup data. Since headings are not consistently provided in the text, several different headings could in reality refer to the same word (example 'CARDIO', 'CV' and 'CARD' all refer to the same heading). We use topic modeling to automatically form clusters that represent different sections of the notes. This is described in detail in the following subsection. After this step all the sections of the notes have consistent headings (i.e. sections under 'CARDIO', 'CV' and 'CARD' would have the same heading) that are different from the original headings.
- Words within the same heading are processed together for each text observation. Stemming, stop word removal and punctuation removal are performed to obtain a list of stem words under each heading (for each text observation).
- Let $n_w(C, H)$ be the number of text observations from class C wherein the word w occurs under heading H. The importance of a word is computed as $I_w(H) = n_w(A, H) - n_w(B, H)$ for classes A and B. Thus words that are more frequent in class A are positive and those for class B are negative and the importance value is an approximate measure of the word's discriminatory power.
- For each heading H, we sort the words with respect to their importance values $I_w(H)$, select the top and bottom 5% (thus selecting from both the most negative and most positive values), and discard the rest.

Within each heading, each of these words forms a feature and the number of occurrences of the word within a text observation is the feature value. A patient's data consists of a feature vector containing all the feature values (for all the headings).

With these preprocessing steps we obtain 7228 features from the text data. Note that this algorithm is an extension of the method presented in [43] wherein clustering based on headings is not automated and requires manual identification of synonymous headings. Here we achieve automation through the use of LDA for clustering.

1) CLUSTERING NOTES WITH TOPIC MODELS

The aim of clustering the notes is to obtain consistently named headings which are then used for feature extraction as described above.

We segment each clinical note utilizing the heading–based structure found within each note. A heading is identified by the set of words between the beginning of a new line and a colon (':'). The set of words between two headings is considered to be a section. We found negligibly few colons within a section and so this rule turned out to be sufficiently accurate. All these words along with the section's heading are considered to be a single document for topic modeling.

During training, notes of all the patients are concatenated and segmented to form individual documents using the segmentation rule described above. The set of all words in all the documents (excluding the words removed during preprocessing) forms our vocabulary. A 50–topic Latent Dirichlet Allocation based topic model [44] is fitted on this document corpus. Each topic is given a unique but arbitrary name (A, B, C etc.). For each document in the training set, the model gives us 50 topic probabilities. The name of the topic with the highest probability is chosen as the heading for each section.

For notes in the test data, we segment the notes using the same rule into different sections. For each section we obtain a new heading by obtaining the topic probabilities from the trained LDA model and choosing the name of the topic with the highest probability.

B. FEATURES FROM HETEROGENEOUS DATA

1) MULTI-VIEW LEARNING

Measurements from heterogeneous sources for the same subject, also called *views*, are commonly presented as co–occurring samples. For example, blood tests, medication orders and nursing notes present multiple views for the same subject in our study. Multi–view learning assumes that the measurement process for each view differs and hence the noise model and the signal–to–noise ratio in each view also differs.

A naïve way to learn from such multiple views is to concatenate the extracted features from each source and use the concatenation as a single data set. This approach does not identify and exclude views or features within views that are unrelated to the outcome variable (the complication to be predicted). Several studies have illustrated the benefits of principled techniques of exploring multi–view data over simple concatenation [45]. These techniques exploit the correlations within and between views to improve predictive learning. A number of these techniques are based on Canonical Correlation Analysis (CCA) [46] that finds linear relationships between multidimensional variables and can be viewed as a way of guiding feature selection and dimensionality reduction towards underlying semantics [45]. Classical CCA has been extended to a probabilistic formulation [47], to a Bayesian formulation [48] and a multi–view generalization of Bayesian CCA is given in [49].

Klami et al. [50] propose a multi–view learning approach that generalizes CCA for arbitrary collection of matrices. It allows each of the views to have a separate low-rank structure that is independent of the other views, as well as structures that are shared only by a subset of them. It also supports multi–view learning on continuous, binary and count observations and is efficient for sparse matrices involving missing data. Unlike many other methods, their Bayesian solution requires no parameter tuning. Thus this method offers several advantages particularly for clinical datasets that have multiple views, contain continuous, binary and count observations, and often have sparse and missing observations.

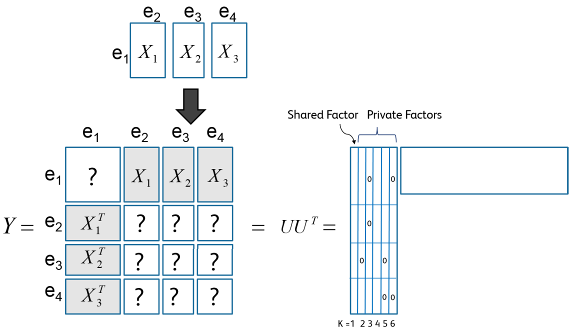


FIGURE 4. **Illustration of CMF for multi–view learning [50].** X_1, X_2, X_3 **are different views representing relationships between entity sets** e_1, e_2, e_3, e_4**. The aim of CMF is to obtain a low dimensional representation (**U**) of the combined data. See text for more details.**

We briefly describe the collective matrix factorization (CMF) approach of [50]. A matrix is a relationship between two *entity sets*. Entity e_i in Figure 4 represents the entity set for the rows and columns of corresponding matrix. For example, if X_1 represents the blood tests for a set of patients, e_1 represents the set of patients, e_2 represents the set of variables related to blood tests and matrix X_1 represents the relationship of e_1 and e_2. Similarly X_2 represents the relationship between e_1 and e_3 (e.g. features from nursing notes).

The aim of CMF is to approximate each matrix with a rank–K factorization along with additional row and column bias terms. Assume M matrices $X_m = [x_{ij}^{(m)}]$ representing E sets of entities, each with cardinality d_e. Let r_m and c_m denote the entity sets corresponding to the rows and columns of matrices. The element corresponding to the row i and column j of the m^{th} matrix can be written as [50]:

$$x_{ij}^{(m)} = \sum_{k=1}^{K} u_{ik}^{(r_m)} u_{jk}^{(c_m)} + b_i^{(m,r)} + b_j^{(m,c)} + \varepsilon_{ij}^{(m)}, \qquad (1)$$

where $\mathbf{U}_e = [u_{ik}^{(e)}] \in \mathbb{R}^{d_e \times K}$ is the low–rank matrix related to the entity set e, $b_i^{(m,r)}$ and $b_j^{(m,c)}$ are the bias terms for the mth matrix, and $\varepsilon_{ij}^{(m)}$ is element–wise independent noise.

The same model is also obtained by collecting all the matrices into a larger symmetric matrix Y of dimension $d = \sum_{e=1}^{E} d_e$. Note that blocks not corresponding to any observed relationship between entities are left blank (this includes all blocks along the diagonal). The CMF model, ignoring noise terms, is given by $Y = UU^T + \epsilon$ where $U \in \mathbb{R}^{d \times K}$ is the concatenation of all the U_e matrices (column-wise). Although, in principle, any symmetric matrix factorization technique, that can handle missing values, can be used to obtain this factorization, the variational Bayesian approximation presented in [50] adds sparsity priors to infer the structured noise in each entity set that is independent of other entities. Thus CMF obtains *private factors* for each entity set as well as *shared factors* that are shared across entity sets.

Matrix U contains row–wise concatenated rank-K representation of all the entity sets. In particular, the block corresponding to entity e_1 in matrix U is the joint representation of all matrices for the entity set e_1. In our example, this corresponds to the patients and can be used as a combined representation of patients shared across all views. Thus, CMF not only helps in reducing the dimensionality of the combined dataset (similar to the standard PCA when there is a single view) but also transfers information between the views in a way that augments learning from the combined representation.

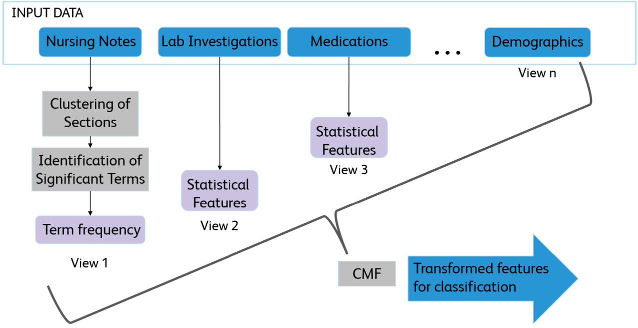


FIGURE 5. Our overall approach using CMF to obtain features for classification. Features obtained from each of the data sources, form the different views. These views are combined using CMF to obtain a low dimensional representation used as features for classification.

2) OVERALL APPROACH

Our overall approach is shown in Figure 5. We extract term frequencies of statistically significant terms within each heading, after clustering the headings, as described in the previous section. For other numerical measurements, we extract the following statistical features from the training sample: sample–based mean, standard deviation, range, skewness and kurtosis. Static demographic features like gender, ethnicity etc. are used directly as features.

CMF, as described above, is used to combine these different views and obtain transformed features. For our dataset entity e_1 is the set of patients and entities $e_2, e_3, e_4, \ldots$ are feature matrices from different views (see Figure 4). These views for our dataset are shown in table 3. Thus matrix X_1 has text features, matrices X_2, X_3, X_4, X_5 have statistical features from chart events, IO events, Labs and Medications respectively, X_6 has comorbidities and X_7 has demographic features. Note that the datatype of each matrix is different – statistical features are continuous-valued, term frequencies from text are ordinal and comorbidities are binary. Categorical demographic features are also converted to binary through one–hot encoding. The input to CMF are the matrices $X_1, \ldots, X_7$. The number of components K is empirically chosen. As described above, the output matrix U of CMF (in Figure 4) has blocks corresponding to each of the entity sets $(e_1, e_2, \ldots)$. The block corresponding to e_1 has a K–dimensional representation of patients that can be used as the transformed feature set for classification.

C. ONLINE MODEL UPDATES AND PREDICTION

A simple batch-wise approach can be used for online monitoring and prediction. The first model is trained on historical data of patients. As more data of all patients arrives, the model can be updated on a regular basis and subsequent predictions can be made using the updated model. The period after which such updates are done depends on the frequency of measurements. With respect to nursing notes in critical care, we found that updating the model every 24 hours gives accurate predictions.

V. POSTOPERATIVE ACUTE RESPIRATORY FAILURE

Acute Respiratory Failure (ARF) occurs when the respiratory system fails in oxygenation and/or CO_2 elimination. It is the end point of respiratory complications (such as pneumonia, atelectasis etc.) and is the most common among them.

ARF occurs postoperatively in about 3% of all surgical cases and death within 30 days occurs in nearly 26% of the cases [51]. Incidence of ARF in general surgical patients ranges from 0.2 to 3.4% [52]–[54] and may be as high as 7% in complicated surgeries such as cardiac, open aortic and upper abdominal surgeries [55]. ARF is often fatal having a mortality rate of $> 25\%$ and management is difficult owing to multiple etiologies. Surprisingly, there has been no change in the incidence and mortality rates over the last 10 years [51]. Khuri et al. [56] show that ARF is an independent predictor of mortality and Dimick et al. [57] have studied the large cost and length of stay associated with ARF. A predictive model for ARF can hence also be utilized in predictive systems for mortality, cost and length of stay.

Respiratory Failure can be classified into two types and a variety of physiological factors are involved in the pathogenesis. Type 1 respiratory failure or hypoxic respiratory failure occurs when partial pressure of oxygen in blood (PaO2) $<$ 60 mm Hg [58]. This is the most common form of respiratory failure and is due to impairment of diffusion of oxygen across the alveoli. Most common causes of diffusion impairment are pulmonary edema or collapse of the alveoli (atelectasis) leading to ventilation-perfusion mismatch. Type II respiratory failure or hypercapneic respiratory failure occurs when partial pressure of carbon dioxide in blood (PaCO2) $>$ 50 mm Hg [58]. It occurs because of the inability of lungs to exhale carbon dioxide. Most common causes of this type of failure are neurological or muscular dysfunction leading to decrease in ventilation. Trauma leading to severe injuries to brain or chest wall is the other factor. Thus, respiratory failure is defined as PaO2 $<$ 60 mm Hg or PaCO2 $>$ 50 mm Hg. Another definition used for ARF is the inability to extubate (i.e. breathe without mechanical ventilation) for more than 48 hours [59]. There is no general agreement on the time after the surgery within which ARF is considered to be a postoperative complication [60]. Usually within 3-7 days after surgery is considered best [52], [61]–[64] although, some authors consider time frames upto 30 days after surgery [51], [65].

A. RESPIRATORY FAILURE: INCIDENCE AND RISK FACTORS

Many risk factors associated with ARF have been identified in an attempt to evaluate patients preoperatively. They are of two types, patient–related and procedure–related. Smetana et al. [60] present a systematic literature survey and meta-analysis to identify the preoperative risk factors for the postoperative respiratory complications. All these risk factors, in decreasing order of the pooled estimates of odds ratios, are listed in Table 2.

TABLE 2. Risk factors for postoperative acute respiratory failure listed in decreasing order of pooled estimates of odds ratios found in previous studies. See text for more details.

Patient–related	Procedure–related
Age, ASA class, Abnormal chest radiograph, CHF, Arrhythmia, Functional dependence, COPD, Weight Loss, Comorbidities, Tobacco Use, Impaired sensorium, Corticosteroid use, Alcohol Use	Surgical Site, Emergency surgery, Prolonged surgery General anesthesia, Transfusion

Although several risk factors have been identified, prediction of ARF remains a challenge. Six previous studies in the literature have explored predictive models for ARF. Several predictors have been identified by these studies and are mainly categorized into preoperative and intraoperative predictors (See [58, Table 1]). We now briefly describe these studies.

Ramachandran et al. [62] used data of 222,094 patients undergoing non-emergent, non-cardiac surgery from the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database. The NSQIP database defines unplanned intubation as placement of an endotracheal tube and mechanical or assisted ventilation because of the onset of respiratory or cardiac failure manifested by severe respiratory distress, hypoxia, hypercapnia, or respiratory acidosis within 30 days of an operation. The incidence of unanticipated early postoperative intubation was reported to be 0.83–0.9%. Independent predictors of unanticipated early postoperative intubation included current ethanol use, current smoking, dyspnea, chronic obstructive pulmonary disease, diabetes mellitus needing insulin therapy, active congestive heart failure, hypertension requiring medication, abnormal liver function, cancer, prolonged hospitalization, recent weight loss, body mass index less than 18.5 or greater than 40 kg/m^2, medium-risk surgery, high-risk surgery, very high-risk surgery, and sepsis ([62, Table 2]).

Gupta et al. [51] also use the NSQIP dataset to study ARF. Their definition of ARF is postoperative mechanical ventilation of greater than 48 hours or unplanned intubation within 30 days of surgery. They found five predictors: type of surgery, emergency case, dependent functional status, preoperative sepsis, and higher American Society of Anesthesiologists (ASA) class ([51, Table 4]). On a validation dataset they report AUC of 0.897. Hua et al. [65] also use NSQIP data, to identify preoperative risk factors of ARF. They fit a Cox Proportional Hazards model and hazard ratios are reported for the predictors ([65, Table 2]). Main predictors from this study are patient age, ASA class, the presence of preoperative sepsis and total operative time.

There are two studies which analyze the use of electronic anaesthesia records and discharge summaries. Breueckman et al. [64] define their outcome as re-intubation in the hospital after primary extubation in the operating room, leading to unplanned mechanical ventilation within the first three postoperative days. Independent predictors for re-intubation reported in this study are: ASA Score of 3 or more, emergency surgery, high-risk surgical service, history of congestive heart failure, and chronic pulmonary disease. Weights of 3, 3, 2, 2, and 1 are assigned to these predictors, respectively, based on their beta coefficient in the predictive model. The score yields a calculated AUC of 0.81, whereas each point increment is associated with a 1.7-fold (odds ratio: 1.72 [95% CI: 1.551.91]) increase in the odds for re-intubation in the training dataset. Using a validation dataset (of 16,884 cases), the score had an AUC of 0.80 and similar estimated probabilities for re-intubation.

Blum et al. [63] use data from an institutional database of surgical, ARF and death registries to identify risk factors for ARF. Preoperative risk factors for ARF development include ASA status 3-5 emergent surgery, renal failure, chronic obstructive pulmonary disease, number of anesthetics during the admission, and male sex. After matching, intraoperative risk factors include drive pressure, fraction of inspired oxygen, crystalloid administration, and erythrocyte transfusion.

However, all these existing studies have major limitations. It is important to note that these studies are very heterogeneous. For example, the definition of ARF varies across the studies: 4 out 6 studies [51], [62], [64], [65] consider ARF to be prolonged mechanical ventilation or unplanned intubation whereas other two consider PaO2 and PaCo2 values to define ARF [61], [63]. The timeframe considered is also different with 4 studies considering less than 7 days after surgery [61]–[64] and other two [51], [65] considering less than 30 days after surgery. An important point to note is that all these studies are retrospective studies including data from discharge summaries etc. which will not be available during the stay of patient in the hospital and hence cannot be used in a real-time predictive system. Another limitation is that all these models require extensive manual evaluation by trained physicians and is time consuming.

To our knowledge no previous work has investigated the use of nursing notes and investigative reports during the patient's stay to predict ARF and we take the first step in this direction. We are also unaware of any previous work that combines text sources and other clinical data for real–time model building and prediction in ICUs.

VI. EXPERIMENTS

Our experiments are designed to answer the following questions:

- Do nursing notes in critical care provide sufficient discriminatory signal to distinguish between ARF and non-ARF cases?
- Can data from nursing notes be combined effectively with the variety of patient measurements present to design a classifier?
- Can a real-time online system effectively identify patients at risk of ARF?

A. EXPERIMENT SETTINGS

1) DATASET

Postoperative Acute Respiratory Failure cases are identified using ICD9 code 518.5. We obtain clinical data of 91 ARF patients from the MIMIC II database. We also obtain data of 684 patients who underwent various kinds of surgeries. Thus we have a dataset of 775 patient records: 91 from ARF class and 684 from Non-ARF class.

For ARF cases, we only use data before the diagnosis of ARF. The diagnosis timestamp is not present in MIMIC II and so we use the following criteria to infer the time. We use the lab measurements to check if any of the following three conditions are true: (1) PaO2 $<$ 60 mm Hg (2) PaCO2 $>$ 50 mm Hg (3) PaO2/FiO2 $<$ 300, and if so, we use the timestamp of the lab measurement as the diagnosis timestamp. Further, we manually examine all the nursing notes to ensure that ARF is not diagnosed earlier than the inferred timestamp.

2) CLINICAL MEASUREMENTS

Along with nursing notes that are periodically recorded in the ICUs, we also use the measurements shown in Table 3. Only those measurements that are present in more than 20% of the patients in our dataset are chosen. For temporal numerical measurements – Chart events, IO events, Blood Tests and Medications – we extract the following statistical features for train and test samples: mean, standard deviation, range, skewness and kurtosis. Static features – demographics and comorbidities – are used directly as features. Comorbidities are encoded as binary variables, all demographics except age are categorical variables which are converted to binary variables through one–hot encoding. For CMF, we consider each set of measurements as given in Table 3 as separate views.

TABLE 3. Different views of a patient: clinical measurements, in addition to nursing notes, used to obtain features for classification.

Views	Measurements
Chart Events (Ch)	Central Venous Pressure, Arterial Blood Pressure, Heart Rate, Respiratory Rate, Oxygen Saturation, Weight
IO Events (IO)	Saline, Urine Output
Labs (L)	Blood PH, Glucose, Blood Urea Nitrogen, Platelet Count, Hematocrit, Chloride, Calcium, Magnesium, Sodium, Potassium
Medications (M)	Propofol, Insulin
Comorbidities (Co)	Presence/Absence of 30 diseases/conditions: congestive heart failure, cardiac arrhythmias, valvular disease, pulmonary circulation, peripheral vascular, hypertension, parlysis, other neurological, chronic pulm, diabetes uncomplicated, diabetes complicated, hypothyrodism, renal failure, liver disease, peptic ulcer, aids, lymphoma, metastatic cancer, solid tumor, rheumatoid arthritis, coagulopathy, obesity, weight loss, fluid electrolyte, blood loss anemia, deficiency anemias, alcohol abuse, drug abuse, psychoses, depression
Demographics (D)	Age, Gender, Ethnicity, Marital Status, Religion

3) EVALUATION METRIC

We use the Area Under the ROC Curve (AUC) as our evaluation metric. All results show average AUC and standard deviation over 5–fold cross validation.

4) BASELINES

To evaluate our text preprocessing method for feature extraction, we compare with the standard method used in text analysis, of Term Frequency – Inverse Document Frequency (TF-IDF). Stemming, stop word removal and punctuation removal are performed before obtaining the TF-IDF vectors. We also compare our method with the method of Ghassemi et al. [38] that uses topics obtained from Latent Dirichlet Allocation as features for mortality prediction using text data.

5) CLASSIFICATION

These features are evaluated using five classifiers: Logistic Regression with L2 regularization (LR), Support Vector Machines (SVM), Decision Tree (DT), AdaBoost (AB) and Random Forest (RF). We use Scikit-learn implementations of these classifiers [66]. Default settings are used in each case, except for Random Forest where we use 1000 trees.

Features obtained from various clinical sources are combined through CMF and evaluated using the same five classifiers. The R package from CRAN for CMF is used to obtain the transformed features.

Since there is class imbalance of nearly 1:7.5 in the dataset, we use SMOTE [67] to inflate the minority class during training. When only text data is used, we use principal component analysis (PCA) to reduce the dimensionality of the feature matrix. The number of components is chosen to be 250 determined empirically through cross–validation to obtain the best predictive accuracy. When CMF is used along with other clinical data, PCA is not required since CMF itself obtains a transformed feature set assuming a low rank factorization.

B. RESULTS

1) OFFLINE CLASSIFICATION WITH TEXT-BASED FEATURES

Table 4 shows classification results with five different classifiers using three different feature sets all of which are obtained from the nursing notes.

TABLE 4. Classification results: Average AUC (with standard deviation) over five folds. Best result for each classifier in bold.

Classifier	AUC		
	Ours	TF-IDF	Ghassemi et al.
LR	**0.828** (0.032)	0.803 (0.044)	0.802 (0.041)
SVM	**0.873** (0.068)	0.828 (0.053)	0.798 (0.036)
DT	**0.623** (0.056)	0.58 (0.043)	0.573 (0.051)
AB	0.783 (0.023)	**0.784** (0.05)	0.741 (0.07)
RF	**0.844** (0.036)	0.825 (0.054)	0.818 (0.059)

Logistic regression and Random Forest obtain the best classification results. With four out of the five classifiers, our method outperforms the standard TF-IDF approach. With all the five classifiers, our method outperforms the method of Ghassemi et al. Overall, the best result of AUC 0.844 is obtained with Random Forest on the features obtained using our method. Random Forest is also the least sensitive to data dimensionality and the AUC remains above 0.82 when PCA is not used.

2) DISCRIMINATORY FEATURES

We study the most discriminatory terms found using the Gini importance scores in the trained Random Forest. Note that for this experiment, we use the entire feature matrix without PCA–based dimensionality reduction. Terms within the following headings were found to be the most discriminatory: Neurology (O), Genito-Urinary and Gastro-Intestinal (AN), Cardiovascular (AW) and Respiratory (W). Under these headings, the most discriminatory terms are words that indicate general health status like ‘alert’, ‘stable’, ‘pulse’, ‘wean’, ‘well’, ‘good’ and ‘pain’. Class frequencies of twenty of these words, shown in Figure 6 are high in Non ARF patients as expected.

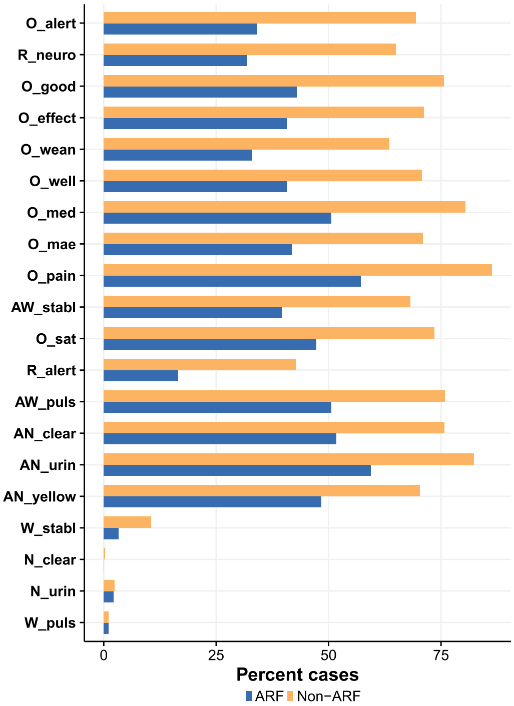


FIGURE 6. Proportions of most discriminatory text features (format: heading_term) in both classes: ARF and Non-ARF. These terms suggest general improvement in health condition and are present in significantly higher proportion in Non-ARF patient notes.

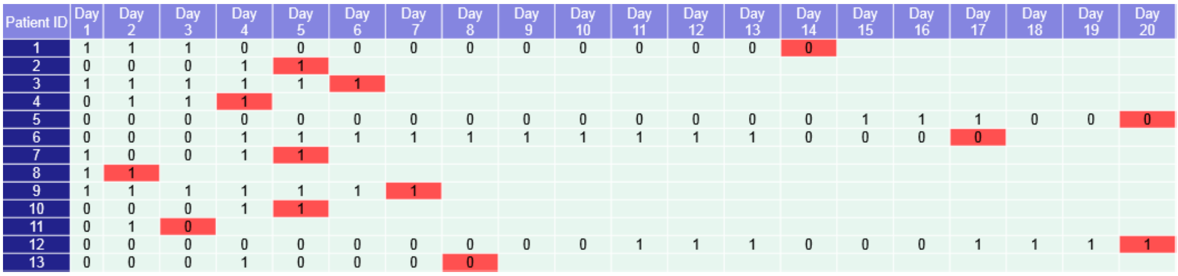
O_alert:
O: neuro awake alert , O: neuro pt alert, O: nuero pt cont awake alert, O: neuro alert oriented, R: neuro alert oriented
O_well:
O: pt moves well, O: ap agitation due dts fairly well controlled alcohol gtt, O: tolerating clear liquids well
O_pain:
O: comfort denies pain, O: requesting pain meds, O: neuro following commands rarely opens eyes pain moving rt side
AW_stable:
AW: atrial sensing ventricular pacing stable hemodynamics, AW: cv hr low s briefly s after lopressor dose bp stable
AN_clear, AN_yellow, AN_urin:
AN: gu patent foley clear yellow urine, AN: gu status huo bdline qs amber urine, AN: gu indwelling foley intact draining clear yellow urine

FIGURE 7. Nursing notes excerpts corresponding to significant features (see Figure 6). Terms are prepended with headings inferred from LDA.

TABLE 5. Classification results using features from each view individually – from Table 3 and from text based features (repeated from Table 4) – and (last column) using CMF that combines all views: Average AUC over five folds.

	AUC							
	Ch	IO	L	M	Co	D	Text	CMF
LR	0.554	0.502	0.503	0.564	0.592	0.533	0.828	**0.881**
SVM	0.545	0.527	0.509	0.537	0.556	0.567	0.873	0.582
DT	0.513	0.569	0.504	0.507	0.549	0.521	0.623	0.619
AB	0.552	0.594	0.546	0.518	0.585	0.584	0.783	0.831
RF	0.584	0.512	0.51	0.523	0.504	0.512	0.844	0.848

Figure 7 shows excerpts from the nursing notes containing these words. This demonstrates the value of nursing notes that record personal observations of the clinical staff. Patients who show visible improvement in their condition are less likely to develop ARF and possibly other complications as well.



Patient ID	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0						
2	0	0	0	1	1															
3	1	1	1	1	1	1														
4	0	1	1	1																
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
6	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0			
7	1	0	0	1	1															
8	1	1																		
9	1	1	1	1	1	1	1													
10	0	0	0	1	1															
11	0	1	0																	
12	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	1
13	0	0	0	1	0	0	0	0												

FIGURE 8. Online (day-wise) prediction of ARF using nursing notes for 13 patients. Prediction (0: not at risk of ARF, 1: at risk of ARF) made at the end of each day using all previous nursing notes. Red cells mark the day when clinical diagnosis of ARF is made. In these 13 cases, our classifier predicts the possibility of ARF many days before diagnosis.

3) OFFLINE CLASSIFICATION WITH COMBINED FEATURES

In addition to the text features obtained from nursing notes we use features from other clinical data listed in Table 3. Table 5 shows the classification results (first six columns) when each of these views is individually used. Note that PCA is applied only to the feature matrix from text data and not for features obtained from other clinical measurements. We observe that none of these views can yield significant predictive accuracy.

We also reproduce the results obtained by using text features alone, using our preprocessing method, from Table 4. With Logistic Regression, Random Forest and Adaboost, CMF that combines all features yields higher predictive accuracy than that obtained using text features alone. These results are for $K = 1000$ in CMF chosen through cross-validation to obtain the best predictive accuracy. Logistic Regression with CMF on the combined features obtains the best AUC of 0.881.

4) ONLINE CLASSIFICATION

We evaluate online prediction using only our text based features. After training the model on the training data, we contruct features from all the notes written during a single day for each patient and predict the risk of ARF. Figure 8 shows a schematic of the prediction for 13 patients who developed ARF. At the end of each day, the classifier predicts 1 (at risk for ARF) or 0 (not at risk for ARF) using the notes of that day. Different patients are diagnosed with ARF on different days – marked bv red cells. We observe that in all these 13 cases, the classifier predicts 1 and hence infers the imminent risk of ARF much before the clinical diagnosis. In total there are 91 ARF patients with mean ICU length of stay: 28.5 days (standard deviation: 23.75) and median ICU length of stay: 20 days (ICU length of stay statistics for all 775 patients are: mean 13.5 days, standard deviation 16.75 and median 8 days).

Overall classification results over five folds (for 775 patients) are shown in Table 6. Note that while prediction is done each day, we report the accuracy on a per–patient basis, i.e. accuracy measures the proportion of test set patients correctly identified as ARF (on any day on or before clinical diagnosis) or Non-ARF cases. *Prediction Time*, defined only for correctly identified ARF cases, is the number of hours between our classifier's identification of ARF and the clinical diagnosis. We observe that Logistic Regression and Random Forest obtain the best accuracy values and can accurately identify more than 80% of the patients accurately on an average more than 3 days in advance. Early identification of high–risk patients can lead to prioritized care and potentially, prevention of the complication.

TABLE 6. Online Classification results: average per–patient accuracy over five folds and, for correctly identified ARF cases, average prediction time (number of hours before diagnosis).

Classifier	Accuracy (%)	Prediction Time (Hours)
LR	88.58	87.28
SVM	74.79	138.83
DT	40.53	114.21
AB	76.46	158.58
RF	81.62	129.4

VII. CONCLUDING DISCUSSION

In this paper we investigate the use of unstructured clinical notes for predicting complications in critical care. We present a new preprocessing technique for extracting features from such clinical notes. We also explore the use of Collective Matrix Factorization (CMF) based multi–view learning to model heterogeneous clinical data. Using postoperative respiratory failure in critical case as an example, we present empirical results to illustrate the efficacy of these techniques.

Similar to previous studies, such as [14], [37], and [38], we find that informally written nursing notes are extremely valuable sources of information for predictive modeling. Our new preprocessing technique exploits the presence of sections within the nursing notes. However identification of these sections is problematic due to inconsistent naming of the section headings. We solve this problem by inferring clusters on the notes (through topic modeling). Statistically significant pairs of section heading and terms within each section are then used as features for classification. *Our approach is found to outperform previous best techniques for feature extraction – TF-IDF and topics from LDA [38]* – that do not exploit the structure found in these otherwise unstructured notes.

We find that nursing notes contain several general indicators of the improvement in a patient's condition that distinguish patients who do not develop complications from those who do. Text based features can also be used in an online manner, using for example all the notes available in consecutive 24-hour windows, to identify the risk of complications.

We find that risk of complications can be identified several days in advance using such a system, with an overall accuracy of above 80%.

Heterogeneity in clinical data poses significant challenges for predictive modeling. Apart from nursing notes, many other measurements are available for patients in critical care. These include, among others, lab measurements, medication orders, comorbidity information and demographic details. These can be considered different *views* of the patient and often have different datatypes and noise characteristics. CMF offers a principled approach to combining diverse datatypes and perform multi–view learning that can provide a low dimensional representation of the combined data. *We observe that while each of the views does not have high discriminatory value in classification, with the use of CMF the predictive value of the combined data is higher than that of only text–based features.* A limitation of the CMF–based approach is the loss of interpretability with respect to the significance of a feature in the final prediction. This is also the case with PCA–transformed features in case of a single–view feature matrix. The importance of the original feature in the final prediction cannot be determined since the transformation produces new features that are a combination of all the input features. We only use statistical summaries of the temporal data as features. Use of several recent techniques for modeling sparse, irregularly sampled temporal clinical data (such as [13] and [16]) in combination with multi–view learning might be a useful direction to explore.

REFERENCES

[1] K. B. To and L. M. Napolitano, "Common complications in the critically ill patient," *Surgical Clinics North Amer.*, vol. 92, no. 6, pp. 1519–1557, 2012.

[2] C. M. Wollschlager and A. R. Conrad, "Common complications in critically ill patients," *Disease-a-Month*, vol. 34, no. 5, pp. 225–293, 1988.

[3] S. V. Desai, T. J. Law, and D. M. Needham, "Long-term complications of critical care," *Critical Care Med.*, vol. 39, no. 2, pp. 371–379, 2011.

[4] N. A. Halpern, S. M. Pastores, J. M. Oropello, and V. Kvetan, "Critical care medicine in the United States: Addressing the intensivist shortage and image of the specialty," *Critical Care Med.*, vol. 41, no. 12, pp. 2754–2761, 2013.

[5] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care," *Proc. IEEE*, vol. 104, no. 2, pp. 444–466, Feb. 2016.

[6] O. Badawi *et al.*, "Making big data useful for health care: A summary of the inaugural MIT critical data conference," *JMIR Med. Informat.*, vol. 2, no. 2, p. e22, 2014.

[7] C. K. Reddy and C. C. Aggarwal, *Healthcare Data Analytics*, vol. 36. Boca Raton, FL, USA: CRC Press, 2015.

[8] D. Gotz, H. Stavropoulos, J. Sun, and F. Wang, "ICDA: A platform for intelligent care delivery analytics," in *Proc. AMIA Annu. Symp.*, 2012, pp. 264–273.

[9] A. Perer and J. Sun, "Matrixflow: Temporal network visual analytics to track symptom evolution during disease progression," in *Proc. AMIA Annu. Symp.*, 2012, pp. 716–725.

[10] Y. Mao, W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, "An integrated data mining approach to real-time clinical monitoring and deterioration warning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. 2012, pp. 1140–1148.

[11] J. Wiens, E. Horvitz, and J. V. Guttag, "Patient risk stratification for hospital-associated C. Diff as a time-series classification task," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 467–475.

[12] S. Saria, D. Koller, and A. Penn, "Learning individual and population level traits from clinical temporal data," in *Neural Inf. Process. Syst. (NIPS), Predictive Models Personalized Med. Workshop*, 2010.

[13] R. Dürichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, "Multitask Gaussian processes for multivariate physiological time-series analysis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 314–322, Jan. 2015.

[14] M. Ghassemi *et al.*, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 446–453.

[15] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht, "A pattern mining approach for classifying multivariate temporal data," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, 2011, pp. 358–365.

[16] T. A. Lasko, "Efficient inference of Gaussian-process-modulated renewal processes with application to medical event data," in *Proc. Uncertainty Artif. Intell.*, 2014, p. 469–476.

[17] K. L. C. Barajas and R. Akella, "Dynamically modeling patient's health state from electronic medical records: A time series approach," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 69–78.

[18] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 85–94.

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

[20] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1095–1103.

[21] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive hawkes process," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2015, pp. 721–726.

[22] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad, "Learning probabilistic phenotypes from heterogeneous EHR data," *J. Biomed. Informat.*, vol. 58, pp. 156–165, Dec. 2015.

[23] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proc. 2nd ACM SIGHIT Int. Health Informat. Symp.*, 2012, pp. 389–398.

[24] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 507–516.

[25] D. C. Kale, Z. Che, Y. Liu, and R. Wetzel, "Computational discovery of physiomes in critically ill children using deep learning," in *Proc. Workshop DMMI AMIA*, 2014, pp. 1–2.

[26] J. C. Ho, J. Ghosh, and J. Sun, "Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 115–124.

[27] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 135–144.

[28] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2956–2964.

[29] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 705–714.

[30] S. G. Finlayson, P. LePendu, and N. H. Shah, "Building the graph of medicine from millions of clinical narratives," *Sci. Data*, vol. 1, 2014, Art. no. 140032, doi:10.1038/sdata.2014.32.

[31] Y. Ling, Y. An, and X. Hu, "A matching framework for modeling symptom and medication relationships from clinical notes," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, 2014, pp. 515–520.

[32] P. Sondhi, J. Sun, H. Tong, and C. Zhai, "Sympgraph: A framework for mining clinical notes through symptom relation graphs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1167–1175.

[33] R. Harpaz *et al.*, "Text mining for adverse drug events: The promise, challenges, and state of the art," *Drug Safety*, vol. 37, no. 10, pp. 777–790, 2014.

[34] G. Hripcsak, C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson, and P. D. Clayton, "Unlocking clinical data from narrative reports: A study of natural language processing," *Ann. Internal Med.*, vol. 122, no. 9, pp. 681–688, 1995.

[35] J. F. E. Penz, A. B. Wilcox, and J. F. Hurdle, "Automated identification of adverse events related to central venous catheters," *J. Biomed. Informat.*, vol. 40, no. 2, pp. 174–182, 2007.

[36] H. J. Murff *et al.*, "Automated identification of postoperative complications within an electronic medical record using natural language processing," *JAMA*, vol. 306, no. 8, pp. 848–855, 2011.

[37] L.-W. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of ICU patients using topic models inferred from unstructured progress notes," in *Proc. AMIA Annu. Symp.*, 2012, pp. 505–511.

[38] M. Ghassemi *et al.*, "Unfolding physiological state: Mortality modelling in intensive care units," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 75–84.

[39] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc. ICWSM*, 2013, p. 2.

[40] F. Doshi-Velez, B. C. Wallace, R. Adams, "Graph-sparse LDA: A topic model with structured sparsity," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2575–2581.

[41] M. Saeed, C. Lieu, G. Raber, and R. G. Mark, "Mimic II: A massive temporal ICU patient database to support research in intelligent patient monitoring," in *Proc. Comput. Cardiol.*, 2002, pp. 641–644.

[42] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[43] V. Huddar, V. Rajan, S. Bhattacharya, and S. Roy, "Predicting postoperative acute respiratory failure in critical care using nursing notes and physiological signals," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2014, pp. 2702–2705.

[44] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[45] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[46] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[47] C. Archambeau, N. Delannay, and M. Verleysen, "Robust probabilistic projections," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 33–40.

[48] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 965–1003, 2013.

[49] S. Virtanen, Y. Jia, A. Klami, and T. Darrell, "Factorized multi-modal topic model," in *Proc. 28th Conf. Annu. Conf. Uncertainty Artif. Intell. (UAI-12)*, 2012, pp. 843–851.

[50] A. Klami, G. Bouchard, and A. Tripathi, "Group-sparse embeddings in collective matrix factorization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[51] H. Gupta *et al.*, "Development and validation of a risk calculator predicting postoperative respiratory failure," *Chest J.*, vol. 140, no. 5, pp. 1207–1215, 2011.

[52] A. M. Arozullah, J. Daley, W. G. Henderson, and S. F. Khuri, "Multifactorial risk index for predicting postoperative respiratory failure in men after major noncardiac surgery," *Ann. Surgery*, vol. 232, no. 2, pp. 242–253, 2000.

[53] R. G. Johnson, A. M. Arozullah, L. Neumayer, W. G. Henderson, P. Hosokawa, and S. F. Khuri, "Multivariable predictors of postoperative respiratory failure after general and vascular surgery: Results from the patient safety in surgery study," *J. Amer. College Surgeons*, vol. 204, no. 6, pp. 1188–1198, 2007.

[54] J. Canet *et al.*, "Prediction of postoperative pulmonary complications in a population-based surgical cohort," *The J. Amer. Soc. Anesthesiologists*, vol. 113, no. 6, pp. 1338–1350, 2010.

[55] O. Gajic *et al.*, "Early identification of patients at risk of acute lung injury: Evaluation of lung injury prediction score in a multicenter cohort study," *Amer. J. Respiratory Critical Care Med.*, vol. 183, no. 4, pp. 462–470, 2011.

[56] S. F. Khuri *et al.*, "Determinants of long-term survival after major surgery and the adverse effect of postoperative complications," *Ann. Surgery*, vol. 242, no. 3, p. 326–341, 2005.

[57] J. B. Dimick, S. L. Chen, P. A. Taheri, W. G. Henderson, S. F. Khuri, and D. A. Campbell Jr., "Hospital costs associated with surgical complications: A report from the private-sector national surgical quality improvement program," *J. Amer. College Surgeons*, vol. 199, no. 4, pp. 531–537, 2004.

[58] G. Sachdev and L. M. Napolitano, "Postoperative pulmonary complications: Pneumonia and acute respiratory failure," *Surgical Clinics North Amer.*, vol. 92, no. 2, pp. 321–344, 2012.

[59] J. Canet and L. Gallart, "Postoperative respiratory failure: Pathogenesis, prediction, and prevention," *Current Opinion Critical Care*, vol. 20, no. 1, pp. 56–62, 2014.

[60] G. W. Smetana, V. A. Lawrence, and J. E. Cornell, "Preoperative pulmonary risk stratification for noncardiothoracic surgery: Systematic review for the American college of physicians," *Ann. Internal Med.*, vol. 144, no. 8, pp. 581–595, 2006.

[61] D. J. Kor *et al.*, "Derivation and diagnostic accuracy of the surgical lung injury prediction model," *J. Amer. Soc. Anesthesiologists*, vol. 115, no. 1, pp. 117–128, 2011.

[62] S. K. Ramachandran, O. O. Nafiu, A. Ghaferi, K. K. Tremper, A. Shanks, and S. Kheterpal, "Independent predictors and outcomes of unanticipated early postoperative tracheal intubation after nonemergent, noncardiac surgery," *J. Amer. Soc. Anesthesiologists*, vol. 115, no. 1, pp. 44–53, 2011.

[63] J. M. Blum *et al.*, "Preoperative and intraoperative predictors of postoperative acute respiratory distress syndrome in a general surgical population," *J. Amer. Soc. Anesthesiologists*, vol. 118, no. 1, pp. 19–29, 2013.

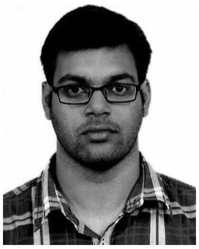
[64] B. Brueckmann *et al.*, "Development and validation of a score for prediction of postoperative respiratory complications," *The J. Amer. Soc. Anesthesiologists*, vol. 118, no. 6, pp. 1276–1285, 2013.

[65] M. Hua, J. Brady, and G. Li, "A scoring system to predict unplanned intubation in patients having undergone major surgical procedures," *Anesthesia Analgesia*, vol. 115, no. 1, pp. 88–94, 2012.

[66] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[67] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

VIJAY HUDDAR received the B.E. degree in computer science from the Gogte Institute of Technology, Belgaum, Karnataka, in 2010, and the M.Tech. degree in computer science (IT) from the International Institute of Information Technology, Bengaluru, in 2014. Prior to the master's degree, he was a Systems Engineer with Infosys Ltd. for a period of two years. He has been a Research Engineer with the Machine Learning and Statistics Group, Xerox Research Centre India (XRCI), since 2014. His interests include machine learning, algorithm design, statistics, and analytics related to healthcare. He holds six filed U.S. patents and three publications in peer-reviewed international conferences for his research in XRCI.



BAPU KOUNDINYA DESIRAJU received the M.B.B.S. degree from the Andhra Medical College, Visakhapatnam. He is currently pursuing the Ph.D. degree in computational biology with the CSIR-Institute of Genomics and Integrative Biology, New Delhi. His long-term vision is to build integrated models of health and disease using data ranging from cellular, physiological, medical records, environmental, geographical, and social networks and application of these models to help improve healthcare. He is currently a Physician turned Data Scientist. His passion is to understand human physiology and disease better with the help of data science tools. He has three publications in peer-reviewed international journals.

VAIBHAV RAJAN received the B.E. degree in computer science from BITS, Pilani, India, and the M.S. and Ph.D. degrees in computer science from the Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland. He was with Hewlett-Packard Labs. He is currently a Senior Research Scientist with the Xerox Research Centre India, where he leads a Project on Clinical Decision Support Analytics. He holds three granted and 27 filed patents and has authored over 30 peer-reviewed articles in leading conferences and journals. His current research interests include algorithm design and statistical inference, particularly in the areas of healthcare and bioinformatics. He was a recipient of the ERS IASC Young Researchers Award 2014 given by European Regional Section of the International Association for Statistical Computing.

SAKYAJIT BHATTACHARYA received the master's degree in statistics from the Indian Statistical Institute, the second master's degree from Michigan State University, USA, and the Ph.D. degree in statistics from University College Dublin, Ireland. His Ph.D. dissertation deals with imputation and derivation of tools to detect outliers in a data set. He was a Post-Doctoral Fellow with the University of Guelph, Canada. He is currently a Research Scientist at Xerox Research Centre India (XRCI). He joined XRCI as Post-Doctoral Researcher in December 2012 and works in the area of analytics. His research interests are in regression, time series analysis, and cluster analysis through mixture model based approach. He was a recipient of the ERS IASC Young Researchers Award 2014 given by European Regional Section of the International Association for Statistical Computing.

SHOURYA ROY is currently the Leader of the Text and Graph Analytics Group, Xerox Research Centre India. Over the last 15 years, his research focus has been on computational linguistics, machine learning, and data mining and their applications to different domains such as customer care, education, and healthcare. He was one of the leaders starting the "Noisy Text Analytics" research area and organized a successful workshop series workshop on analytics for noisy unstructured text data from 2007 to 2012. He holds 12 patents and many other patent applications are in different stages of filing process. He has authored over 50 research papers in leading conferences and journals.

CHANDAN K. REDDY (M'01–SM'13) received the Ph.D. degree from Cornell University and the M.S. from Michigan State University. He is currently an Associate Professor with the Department of Computer Science, Virginia Tech. He has authored over 70 peer-reviewed articles in leading conferences and journals including SIGKDD, WSDM, ICDM, SDM, CIKM, TKDE, DMKD, TVCG, and PAMI. His primary research interests are Data Mining and Machine Learning with applications to Healthcare Analytics, Bioinformatics, and Social Network Analysis. His research is funded by the National Science Foundation, the National Institutes of Health, the Department of Transportation, and the Susan G. Komen for the Cure Foundation. He was a recipient of the Best Application Paper Award at ACM SIGKDD Conference in 2010 and the Best Poster Award at the IEEE VAST Conference in 2014, and was a finalist of the INFORMS Franz Edelman Award Competition in 2011. He is a senior member of the IEEE and life member of the ACM.