Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 11–24
DOI: 10.15439/2023F3385
ISSN 2300-5963 ACSIS, Vol. 35

Deciphering Clinical Narratives – Augmented Intelligence for Decision Making in Healthcare Sector

Lipika Dey
TCS Research
India
lipika.dey@tcs.com

Sudeshna Jana
TCS Research
India
sudeshna.jana@tcs.com

Tirthankar Dasgupta
TCS Research
India
dasgupta.tirthankar@tcs.com

Tanay Gupta
TCS Research
India
gupta.tanay@tcs.com

***Abstract*—Clinical notes that describe details about diseases, symptoms, treatments, and observed reactions of patients to them, are valuable resources to generate insights about the effectiveness of treatments. Their role in designing better clinical decision making systems is being increasingly acknowledged. However, the availability of clinical notes is still an issue due to privacy violation concerns. Hence most of the work done are on small datasets and neither the power of machine learning is fully utilized, nor is it possible to validate the models properly. With the availability of the Medical Information Mart for Intensive Care (MIMIC-III v1.4) dataset for researchers though, the problem has been somewhat eased. In this paper we have presented an overview of our earlier work on designing deep neural models for prediction of outcomes and hospital stay for patients using MIMIC data. We have also presented new work on patient stratification and explanation generation for patient cohorts. This is early work targeted towards studying trajectories for treatment for different cohorts of patients, which can ultimately lead to discovery of low-risk models for individual patients to ensure better outcomes.**

***Index Terms*—Clinical Notes, BioNER, Clustering, Anomaly Detection, Autoencoder, Shapley Value**

I. Introduction – Clinical Narratives and Their Uses

COMMUNICATION within healthcare systems is dominated by textual narratives. This includes a diverse array of documents generated by different sources for various purposes. These texts can be broadly classified as follows:

(a). Clinical notes, especially nursing notes attached to Electronic Health Records (EHR) of patients admitted to the hospital for treatment, contain valuable information about the patients, symptoms, diagnoses, treatments, chronic and past ailments, drug prescriptions, and adverse drug effects, if any, for the patient. Collections of such texts can be effectively mined to gather insights to improve healthcare for other patients [1]. If not explicitly, these texts also provide insights about a physician's possible reasons for following a path of treatment. Nursing notes are time-stamped and therefore can provide a view into the trajectories of recovery. Clinical notes are also highly sensitive in nature since they contain personal identification details. Hence, though large volumes of clinical data are generated across the world, restricted access to the data due to security and privacy concerns is a major bottleneck for researchers. The publicly available Medical Information Mart for Intensive Care (MIMIC) database [2] has eased out this problem to some extent. This data has been anonymized and specially made available for research purposes only. This database contains daily records of over $40,000$ patients with details about their illness, symptoms, medical history, results of diagnostic tests, treatments, nursing observations, discharge summaries along with some patient demographic details like age, gender etc. along with the number of days spent in ICU and other wards for each admission.

(b). Pathological and radiology reports - These are semi-structured narratives resulting from various targeted examinations like findings from radiology images, blood tests, etc. Researchers in the area of computer vision have been long engaged in developing predictive systems from the images for automated detection of diseased cells, tissues, or organs. Recent developments in the area of multi-modal analytics have spurred interest in using textual descriptions along with images for better predictive models.

(c). Bio-medical literature or technical publications that report scientific advances in the area of life sciences and healthcare, include documents like journal articles, case studies, systematic reviews, and clinical guidelines published by regulatory bodies. These documents are important sources of information for those who work in the areas of drug discovery, designing new treatment protocols, etc. Text mining of bio-medical scientific literature is an old established area. The aim is to come up with systems that can help in the easy assimilation of knowledge from this vast incremental repository using sophisticated information extraction and reasoning methodologies. A comprehensive review of text mining techniques for extraction information from bio-medical texts is presented in [3].

(d). Social Media texts - Patient-generated text like tweets or blog posts play a critical role in gathering insights about individual and collective experiences about a drug, treatment, clinical trials, or care facility. Social media text analysis has played a critical role in detecting and assimilating adverse drug effects, especially in obtaining new knowledge about

preconditions or co-occurring conditions that cause adverse effects of a drug. With the increasing popularity of patient support groups like PatientsLikeMe, social media content analysis is gaining new heights. There are dedicated groups for different diseases offering hitherto undiscovered insights about rare diseases to the entire community [4].

The repository of clinical texts is huge, analysis of which can yield deep insights for clinical decision-making, drug discovery, and healthcare management systems. While text mining from reports, literature, and social media has been actively pursued for some time now, the mining of clinical notes from EHRs is a fairly new area, primarily due to the non-availability of such texts earlier. Even to this day, the volume of clinical records of patients available for study purposes is fairly low. This is a severe impediment to the development of machine learning systems, as these are known to require large volumes of data. Nevertheless, with increasing focus on personalized and optimized healthcare management, the analysis of clinical texts is gaining importance. Analytical applications of clinical documents can be broadly classified as -

- Descriptive analytics: which is targeted at knowledge discovery from the scientific literature. This is a fairly mature area of research actively pursued by researchers working in the areas of natural language processing and text mining. They often aim to discover information about new entities and relations reported in the literature. Dalianis provides a detailed review of clinical text mining and its applications along with the challenges of analyzing such data in the book [1]. This book also presents a comprehensive study of clinical text mining in non-English languages.
- Diagnostic analytics: this area digs deeper into clinical texts to unearth causal explanations about events.
- Predictive analytics: this area seeks to employ predictive models to predict possibilities of repeat occurrence of known events. One of the major consumers of predictive models designed using past patient data are hospital management authorities. Hospital admission notes have turned out to be very useful for the purpose, as these can be used to predict ICU length of stays [5], hospital readmissions, procedure requirements, etc. ICU stay prediction is an important problem for hospitals, since ICU facilities are expensive to set up, and their optimal use and availability are imperative to ensure better outcomes through proper resource planning [6]–[8]. Obtaining advance information about the possible length of ICU stay or duration of hospitalization are also useful for patients and their families, as it helps in better expectation management and planning from their side also. Predicting individual patient outcomes is gaining importance as clinical decision making is increasingly focusing on individualized care. Given the wide variability among individuals however, a step forward in this direction is to move towards understanding patient cohorts - that is group of patients who are more similar to each other, than to rest of the patients suffering from the same disease. This is known as patient stratification. Patient stratification is gaining interest from medical as well as machine learning researchers, as it holds the promise to deliver better outcomes to the exceptions. These are early days in this area, and most studies have been conducted on very focused dataset.
- Prescriptive analytics: this is an extension of predictive analytics, where the intent is to prescribe the best possible actions from among a set of possible actions, to achieve a desired outcome under a given state. Prescriptive analytics can be viewed as a natural follow-up of patient stratification. It ideally requires simulation of future possibilities to arrive at the best possible decision for an individual, by studying the current state, possible interventions and their effects using a simulation and then arriving at a feasible conclusion.

In this paper, we shall be primarily presenting our work done in the areas of predictive analytics and patient stratification using clinical text like hospital admission notes, with an end goal of providing decision intelligence for better care management and increased visibility into patient cohorts respectively. The rest of the paper is organized as follows. Section II provides an overview of the MIMIC dataset. Section III presents related work done in the area of predictive analytics, along with our earlier work in the area. It presents an overview of different types of predictions done with clinical data including text by different groups, and also an overview of our earlier work done for predicting ICU length of stay and procedure requirements for patients based on the first day's admission notes. Section III-A provides a comparative study of the performance of all the methods. This is followed by patient stratification work. Section IV presents an overview of related work in the area of patient stratification by other researchers. From section V onwards we present our work done in the area of patient stratification using clinical texts, which to the best of our knowledge has not been attempted before. We have presented deep learning based methods to generate explainable clusters from clinical notes of patients admitted with a specific disease. The explanations generated for the clusters provide more insights into the co-morbidities of each cohort present within the group. The cohort-treatment associations are also obtained. We present results of experiments done with pneumonia patients of the MIMIC dataset in section VII. Detailed insights into the cohorts obtained are presented in the form of cluster-wise patient statistics in terms of age and hospital stay, along with cluster-specific association of symptoms, treatments, and final recovery information obtained from the discharge summaries. We believe the insights can pave the way for analyzing the effectiveness of the treatment trajectories and thereby customizing them in future, to improve treatment effectiveness and reduce mortality, if possible. We conclude with a lot of future possibilities in section VIII.

II. The Dataset

As our primary data source, we have used the MIMIC-III v1.4 database [2], which contains the details of over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012. This database has pre-existing Institutional Review Board (IRB) approval, and researchers can access the data after successfully completing the training course "Data or Specimens Only Research" provided by the Collaborative Institutional Training Initiative (CITI).

The MIMIC-III database contains details of 46,520 distinct patients with 58,976 hospital admissions. This database includes both structured and unstructured clinical events documented for patients during hospital admissions. Importantly, the database adheres to stringent anonymization protocols, meticulously safeguarding patient privacy. Moreover, to ensure heightened privacy protection, precise dates and times of events have been intentionally obscured. Instructions for accessing this dataset are available in the website https://mimic.mit.edu/docs/gettingstarted/.

III. Building Predictive Models for Patient Care with Clinical Notes

In this section, we present an overview of work done in the area of predictive analytics with electronic health records.

In [9], a neural network based model is used to predict the length of remaining hospital stay for a patient at the time of exit from the ICU unit. In this study, authors used several medical attributes like patients' demography, CPT events, services, procedures, diagnosis, etc. of 31,018 patients from the MIMIC database. In another study [13], Harutyunyan et al. proposed a channel-wise LSTM model using multitask training for predicting mortality along with a forecast of the remaining time to be spent in ICU made at each hour of stay. Predictions were generated from 17 clinical variables like Capillary refill rate, blood pressure, fraction inspired oxygen, Glasgow coma scale, glucose level, heart rate, etc. of patients from the mimic database. Afterward, in 2020 [12], a deep learning architecture based on the combination of temporal convolution and pointwise convolution was proposed to predict the length of ICU stay. This work used the eICU critical care dataset [18], which contained records of 118,534 unique patients, and predictions were based on structured features like patients' gender, age, hour of admission, height, weight, ethnicity, Unit Stay, Physician Speciality, etc. In [10], a study was presented on the prediction of length of stay in ICU and mortality, using several machine learning algorithms on a set of patients from the MIMIC database based on their vital signs like heart rate, blood pressure, temperature, respiratory rate, and patient's demography like age, gender, height, weight, etc. In 2021 [11], Su et al. developed several machine learning models for predicting mortality, severity, and length of stay for a set of 2224 Sepsis patients who were admitted to the ICU of Peking Union Medical College Hospital. In their predictive models, authors used patients' clinical parameters such as age, $P(v\text{-}a)CO_2$ /$C(a\text{-}v)O_2$, SO_2, oxygenation index, white blood cell count, oxygen concentration, temperature, etc. from the first 6h in the ICU.

It is worth mentioning here that none of the above-mentioned works used textual data for prediction. Most of them have used only structured clinical parameters for predicting various clinical events. The richness of clinical notes has not been fully exploited for prediction. In particular, nursing notes play a crucial role in capturing essential patient information that extends beyond the physiological metrics recorded by laboratory tests or radiology reports. These notes encompass a wide range of details, including symptoms, overall health condition, administered medications, performed procedures, and devised treatment strategies. Moreover, they occasionally encompass insights into a patient's response to care and treatment, often described through behavioral observations meticulously documented by the caregiving professional.

Figure 1 shows a sample nursing note with different portions of text color-coded, to highlight the different categories of information that a note may contain. Use of linguistic expressions like *"severe multilobar pnx"*, *"worsening multifocal pnx"*, *"No abdominal pain, no further bleeding"* provide an added dimension of human assessment, that cannot be captured through numbers only, but can be important while distinguishing between two similar patients who are possibly responding differently to the treatment. The notes are very comprehensive in nature. With the database containing almost as many notes for each patient as the number of days of admission, this offers quite a rich collection to work with. Additionally, within the EHR system, a discharge summary is also a crucial component of the patient's medical records that provides a concise overview of a patient's hospital stay, their medical condition, treatments received, and instructions for follow-up care upon their discharge from the hospital. Often the information is repeated across these sources. The redundancy helps in data verification, especially since the data can be quite noisy. Other fields which are more structured in nature like age, gender, admission diagnosis, medications, mortality, etc. are also used for predictive modeling.

Recently, in 2021 Aken et al. [15], have introduced a model, called CORe, on top of BioBERT for predicting multiple clinical outcomes along with the duration of ICU stay. The authors devised a distinct note extracted from the discharge summary, leveraging it within their predictive framework.

In 2022 [16], we proposed a model which utilized nursing notes of the first day of ICU along with clinical parameters from laboratory tests, to predict whether a patient would need an ICU stay of short or long duration, where the partition of short or long stays was decided by median length of stay recorded in data. Further in [17], this work was extended to additionally predict the need for critical procedures such as bypass surgery, stenting, tracheotomy, and cholecystectomy - which were the most commonly occurring ones in the dataset, along with an ICU stay. The objective was to predict these procedures based on the first day's nursing notes in which these were not explicitly mentioned. We also proposed using a framework called "Local Interpretable Model-agnostic

TABLE I
PERFORMANCE ANALYSIS OF DIFFERENT MODELS FOR ICU LENGTH OF STAY PREDICTION.

Earlier Works	Dataset	Features used	ICU stay classes	Methods	Best Result
[9]	31,018 patients from MIMIC database	patients' demography, CPT events, services, procedures, diagnosis, etc.	ICU stay$\leq$5 days, class 0; ICU stay$>$5 days, class 1	neural network based model	80% accuracy
[10]	44,000 ICU stays from MIMIC database	patient's vital signs - heart rate, BP, temp, resp rate, age, gender, height, weight, etc.	ICU stay$\leq$2.64 days, class 0; ICU stay$>$2.64 days, class 1	Machine learning algorithms	65% accuracy using random forest algorithm
[11]	2224 Sepsis patients from Peking Union Medical College Hospital Intensive Care Medical Information System and Database (PICMISD)	patient's age, P(v-a)CO_2 /C(a-v)O_2, SO_2, oxygenation index, WBC count, oxygen concentration, bpm, temp, etc.	ICU stay (LOS) ($>$6 days, $\leq$ 6 days)	logistic regression, random forest, and XGBoost model	sensitivity = 0.79, specificity = 0.66, F1 score = 0.69, AUC = 0.76 using Random forest
[12]	eICU critical care dataset	patient's gender, age, hour of admission, height, weight, ethnicity, unit stay, Physician Speciality, etc.	classifying in 10 classes - one for ICU stays shorter than a day, seven day-long buckets for each day of the first week, one for stays over one week but less than two, and one for stays over two weeks	combination of temporal convolution and pointwise convolution	Kappa score = 0.58
[13]	MIMIC database 42276 ICU stays	17 structured clinical variables - capillary refill rate, blood pressure, fraction inspired oxygen, Glasgow coma scale, glucose, heart rate, etc from first 24 hours of admission	classifying in 10 classes - one for ICU stays shorter than a day, seven day-long buckets for each day of the first week, one for stays over one week but less than two, and one for stays over two weeks	LSTM-based neural network models	AUC-ROC = 0.84
[14]	22,353 patients from MIMIC database	Clinical Notes	Remaining ICU stay time is discretized into 10 classes {0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7, 7-8, 8-14, 14+}	multi-model neural network	Kappa score = 0.453
[15]	38,013 admissions from MIMIC database	created notes from discharge summaries	four categories - Under 3 days, 3 to 7 days, 1 week to 2 weeks and more than 2 weeks	pre-trained CORe model on top of BioBERT	AUC-ROC = 72.53%
[16]	**22,789 admissions from MIMIC database**	**nursing notes from first 24h along with 20 vital signs and lab measurements available in first 24h of ICU stay**	**"Short" if ICU LOS$<$4days and "Long" if ICU LOS$\geq$4days**	**trans-former based deep neural network model**	**Accuracy = 79.20%, AUC-ROC = 87.33% Kappa score = 0.594**
[17]	**28,659 admissions from MIMIC database**	**nursing notes from first 24h along with 20 vital signs and lab measurements available in first 24h of ICU stay**	**"Short" if ICU LOS$<$3days and "Long" if ICU LOS$\geq$3days**	**multimodal multiobjective transformer network**	**Accuracy = 84%**

Explanations" (LIME) to obtain explanations about the prediction outcome. LIME creates several subsets from the original data containing only a part of the original attributes. It then computes the influence of the attributes on the classification, based on the presence or absence of certain features in the selected text. To capture the details and the various nuances of a nursing note, we proposed using transformer-based models for representing them. Embeddings for the nursing notes were generated using BlueBERT [19] and Clinical BioBERT [20]. Along with this, for each admission, we computed four types of Severity of Illness (SOI) scores: APACHE II, SAPS II, SOFA, and OASIS based on data collected within 24 hours of ICU admission. This way, the proposed model made use of both structured as well as unstructured data. The novelty of the

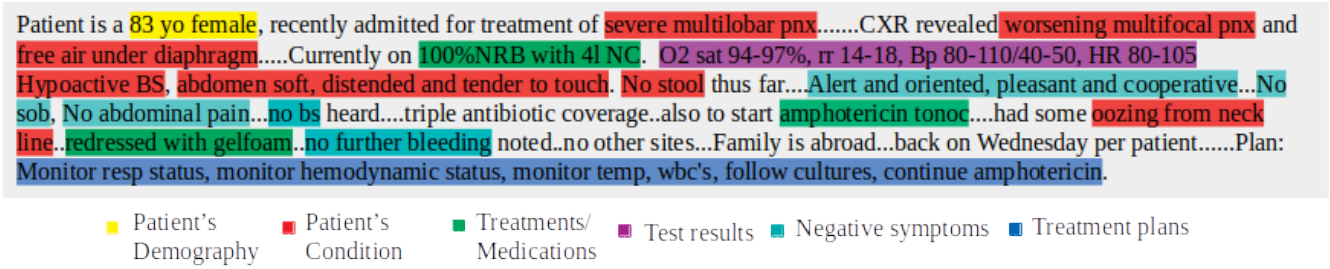


Fig. 1. A Sample nursing note collected from MIMIC-III. Colored components are the extracted events and entities.

proposed model was its design as a multi-objective prediction model, where critical procedures also were predicted along with the duration of ICU stay. Different kinds of networks were experimented with, namely the Convolutional Neural Network (CNN) [21] and Long Short Term Memory (LSTM) [22] for prediction. Additionally, it was observed that the use of a Term Frequency - Inverted Document Frequency (TF-IDF) vector, impacts the performance of the model significantly. It might be due to the fact that the TF-IDF vector can capture the common and distinct features across the classes very effectively. Figure 2 presents the prediction architecture that performed the best along with the proposed representation scheme.

The input representation for a patient is the vector generated by concatenating the outputs of the BiLSTM layer as a result of embedding the corresponding first admission note, the TF-IDF feature vector for clinical entities extracted from the note, and the SOI scores of the patient. The concatenated vector embedding is simultaneously fed into two task-specific fully connected layers, one for predicting ICU length-of-stay and another for predicting the possibilities of each of the four surgical interventions – bypass surgery, stenting, tracheotomy, and cholecystectomy. For the length-of-stay classification, we have used the softmax activation function with binary cross-entropy loss L_{LOS}. For the intervention prediction tasks, since these are not mutually exclusive outcomes, we have trained the prediction layer using the sigmoid activation function with binary cross-entropy loss functions $L_{intervention}$. Finally, we have defined a joint loss function using a linear combination of the loss functions for the two tasks as:

$$L_{Joint} = \lambda * L_{loss} + (1 - \lambda) * L_{intervension} \quad (1)$$

, where λ controls the contribution of losses of the individual tasks in the overall joint loss.

A. Evaluation of Predictive Models for ICU LOS prediction

As discussed earlier, one major aspect that distinguishes the models from each other is the set of predictor variables used. The choice of prediction method is often guided by the choice of the predictor features. The evaluation metrics used by the studies are also not always the same. This makes comparison of methods a little tricky. However, we have presented a compilation of the models, the metrics, and the reported performances of these models in Table I. The last two rows of the table I present the performance of our proposed models over a set of ICU patients from the MIMIC database. The prediction accuracies for bypass surgery, stenting, tracheotomy, and cholecystectomy were found to be $89\%, 83\%, 55\% and 54\%$ respectively. The performance of the last two categories were not so good due to lack of enough data in the set. Interestingly however, while only 3% of the total tracheotomy procedures done later were mentioned in first day's notes, our model could predict 70% of them in the first day. However, the false positive rates were high for the proposed model. This can be reduced with more data. Overall, the significant gains for acquiring valuable insights into procedure requirements on the first day itself are quite significant. Consequently, such an approach offers prospects for enhanced planning and decision-making. It was observed that a total of 15691 unique diseases are recorded in the MIMIC database as key reasons for which patients were admitted to ICU. It was found that in the dataset we used, the topmost category, which constituted about 4% of the entire set were patients of Pneumonia, followed by around 2% each of Sepsis, Coronary Artery Disease, Congestive Heart Failure, and Gastrointestinal bleeding. This somewhat explains the observation of the cardiac procedures as most frequent followed by tracheotomy and cholecystectomy.

As seen in Table I, the performances of predictive systems are improving over time. While these systems are good for hospital management, particularly in efficiently managing resources for heterogeneous sets of patients, when it comes to providing better patient experience, the trajectory of clinical decision-making is moving towards providing more personalized care management. While these are early days of designing personalized care management systems, existing literature strongly suggests that rather than working with very large groups, the stratification of patients into homogeneous subgroups or clusters is likely to play a major role in enabling personalized treatments. In the next section, after describing novel techniques for patient clustering, we will demonstrate the effectiveness of the methods utilizing patient data afflicted with Pneumonia, which was the predominant disease present in the dataset.

IV. Patient Stratification - a review of related works

Most of the predictive models that have been proposed earlier in literature, have worked on large heterogeneous patient

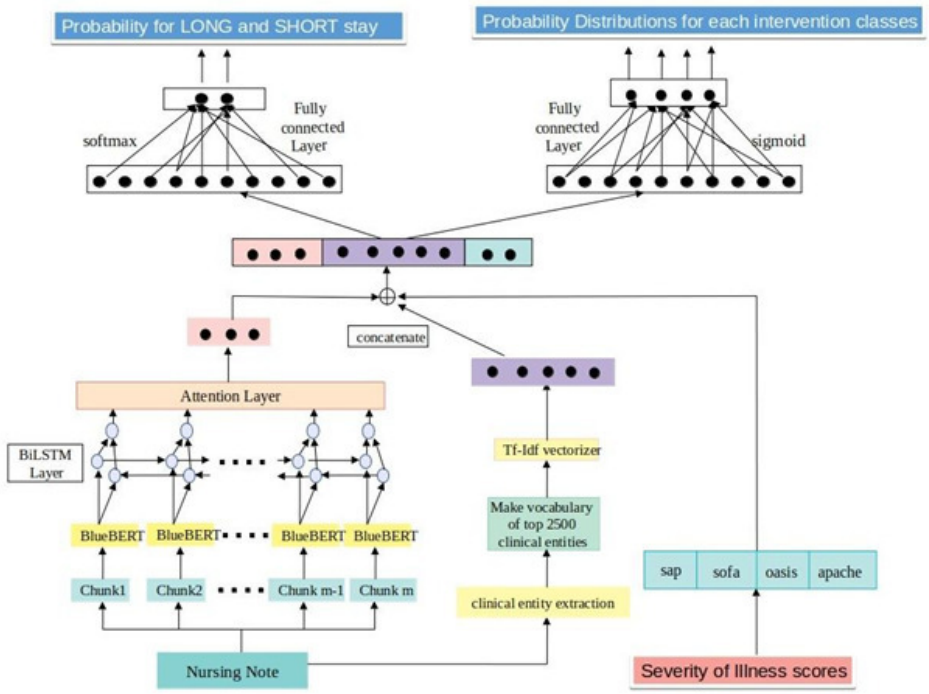


Fig. 2. Overview of proposed multimodal multitask framework for predicting the ICU length of stay and necessity of the interventions. Process the nursing notes in chunks by the BlueBERT model and add a BiLSTM-attention layer on the top. We also extract 2500 medical entities from these notes and make a TF-IDF representation. Then the note representations from the BlueBERT-BiLSTM -Attention network, TF-IDF representation, and four severity of illness scores are concatenated and two task-specific fully connected layers are applied to obtain the final predictions.

populations, which were not able to discover obtain much insights into risk factors for individual patients and hence not very suitable for personalized decision making. Patient stratification techniques that can identify homogeneous groups and thereafter group-specific risk factors, is therefore proposed as a better way to gain insights about outcome differences.

In recent years, several data clustering approaches have been proposed for patient stratification and subsequent analysis of the clusters using different cluster validity indices [23], [24]. In 2021, Alexander et al. [25] investigated the clinical heterogeneity of Alzheimer's disease patients using electronic health records (EHR). In 2022, Angelini et al. [26] proposed an explainable clustering method to identify dominant osteoarthritis endotypes using different biochemical markers, to design tailored treatments and drive drug development. In a recent paper [27], Bhavani et al. have applied hierarchical clustering (DTW-HC) and partitioning around medoids (DTW-PAM) from the first 8 hours of hospitalization records, to identify sub-phenotypes of infected patients. Chen et al. In [28] have presented a model for prediction and risk stratification of kidney outcomes in IgA Nephropath, which is a common disease worldwide. The intent is to predict long-term outcomes and stratifying risk for clinical decision-making. They have also stated that these kinds of work can be useful for designing future clinical trials. The model used was gradient tree boosting implemented in the eXtreme Gradient Boosting (XGBoost) system. The dataset itself was quite small. Kanwal et al. in [29] present stratification of patients suffering from Non-Fatty Liver Disease (NAFLD), which is largely asymptomatic, and success of treatment depends on optimal timing and accurate assessment of fibrosis risk. The work describes the NAFLD Clinical Care Pathway that was developed to assist clinicians in diagnosing and managing NAFLD with clinically significant fibrosis (stage F2–F4) based on the best available evidence using a statistical approach. In [30] Seinen et al. have presented a detailed review of prognostic prediction models that use unstructured clinical text.

V. Patient Stratification Using Nursing Notes - Obtaining Insights about Patient Cohorts and Exceptions

We now present methods for generating explainable patient cohorts based on their condition at admission, by clustering first day's nursing notes of patients. Rather than using the entire dataset of all patients, our focus will now be on individual diseases. We propose the use of auto-encoders to represent patient health conditions, which is a different approach from using the transformer-based representations presented earlier. Further, we propose the use of SHAP values for explaining the clusters. We also show that the use of autoencoders within a disease category improves the accuracy of prediction of the duration of hospitalization.

While analyzing the prediction results using the LIME framework described in section III, we found that clinical

notes, one factor that could affect the performance of prediction could be highly variable in style and content. While some caregivers record only the symptoms that are present on a given day, some others meticulously note down the absence of common symptoms, adverse reactions, psychological state of patients, appetite, etc. The use of non-standard terminology and abbreviations are known to be quite common. Using widely variable terms to describe the same condition is very common in clinical texts. For example - *icterus* and *jaundice* refer to the same disease. Similarly, the terms *brain abscess, intracerebral abscess, cerebral abscess* all refer to the same state. Over the years, bio-medical dictionaries like UMLS [31] have been prepared to document these. Though the BlueBERT embeddings that we used earlier, could capture linguistic nuances like difference between *severe pain* and *mild pain*, these could not always capture the similarities or differences between two notes based on the medical terms used. Therefore, before getting into the stratification work, where the similarity would play a significant role, we implemented an additional processing layer, wherein every clinical note underwent initial processing through the Biomedical dictionaries for standardization of terms. Using this step it was now possible to also distinguish different types of entities like symptoms, diseases, drugs, etc. which could help in grouping patients better. For example, people suffering from the same disease and undergoing the same treatment, but who showed different reactions to it, could now be put into different groups. This led to the idea of using a different embedding altogether which we shall now explain.

The details of the processing pipeline using the biomedical dictionaries are presented below.

Entity extraction: We have used Scispacy [32] and MetaMap [33] for extracting health conditions from patients' clinical notes.

Scispacy: ScispaCy is one of the most robust model for processing biomedical, scientific, and clinical texts on several NLP tasks such as part-of-speech tagging, dependency parsing, named entity recognition, etc. In our work, we have used the pre-trained scispaCy model *en_ner_bc5cdr_md*, which was trained on the BC5CDR corpus for recognizing disease names mentioned in a clinical note.

MetaMap: Nowadays, another entity extraction tool MetaMap is widely used for identifying medical entities. This was developed by the National Library of Medicine (NLM) to map biomedical text to concepts in the Unified Medical Language System (UMLS) [31]. We have processed clinical notes using MetaMap and extracted eight medical entities such as "Sign or Symptom", "Disease or Syndrome", "Acquired Abnormality", "Anatomical Abnormality", "Congenital Abnormality", "Injury or Poisoning", "Mental Process", and "Mental or Behavioral Dysfunction".

Detecting Negations: The Negex algorithm [34] which detects the presence of negative modifiers like "no", "not", etc. is then applied to detect negative mentions of entities in a text. The original list was enhanced to accommodate commonly appearing negation concepts such as "deny", "refuse", "absent", "decline" etc. that occur frequently in clinical notes. For example, given a sentence "*The patient has shortness of breath but denies any chest pain*", two symptoms identified should be *shortness of breath, neg chest pain*. These negative symptoms have a significant contribution to describing individual patients.

Entity Standardization: All the extracted entities are then standardized using the UMLS Metathesaurus [31], [35]. This is important since a single medical condition like "Hypertension" may be referred to as "High blood pressure", "Arterial hypertension" or "Hypertensive disorder" by different professionals. UMLS contains an exhaustive list of such situations and assigns a "Concept Unique Identifier (CUI)" to each. However, we have observed that several entities still did not have an exact match with any UMLS concept. These entities were mapped using an approximate string-matching algorithm [36], that found the closest UMLS concept based on Levenshtein distance measure [37]. For entities that could not be mapped to any UMLS concept, unique identifiers were created to ensure that no health condition was ignored. Examples of such entities from the MIMIC dataset include terms like "airway swelling", "overdistention of lung" etc. To avoid confusion, we refer to these also as CUIs.

Each clinical note can now be represented in terms of the CUIs present in it, either to indicate the presence or absence of a symptom. Consequently, a patient's status at a particular point in time can also be expressed in terms of these CUIs.

Let the collective list of CUIs describing the diseases and symptoms for a particular study be denoted by *health status (H)*. Given a patient p, the health condition at time t is denoted by a vector of $h_i \in H$, where the value of h_i is set to 1 if h_i mentioned in the corresponding clinical notes, -1 if it is mentioned negatively, and 0 if h_i is not mentioned. It may be noted that the CUIs associated with a patient are expected to change over time as treatment progresses. Consequently, the patient may be represented by different vectors over the same space as t changes.

Patient Medication Information: Besides health conditions, each patient also has medications that are prescribed based on these conditions. Considering a unified set of medicines M, at a given time t, each patient p is also associated with a binary vector $< m_i >, i = 1..|M|$, where $m_i = 1$, if medicine m_i is ongoing for p at time t, otherwise 0. This binary vector also changes over time.

A. Creating Dense Representation of Patients in Terms of Health Conditions using Autoencoders

While the number of unique diseases and symptoms obtained from any patient database is very high since all people do not exhibit all symptoms or diseases, the above vectors are high-dimensional and sparse. An autoencoder-based transformation is applied to obtain a dense representation in a lower-dimensional space [38], [39]. In an autoencoder (AE) model, the "encoder" network creates a compressed representation of the input data by capturing the essential characteristics and underlying patterns, while the "decoder"

network learns to reconstruct the original input data from the compressed representation while minimizing the loss of information [40]. We next show how dense representations are used for prediction as well as clustering purposes.

B. Autoencoder-based Prediction of Duration of Hospitalization

Before getting into the details of clustering, we validated the representation using it in a predictive framework similar to the ones described earlier first. Figure 3 presents a CNN-based architecture that is used for the prediction of hospital / ICU stay duration using the autoencoded representation of the first day's clinical notes. The network performance was tested with a dataset of 2106 "Pneumonia" patients who had undergone one admission for the disease. The set comprises individuals both with and without ICU admission. Since the number of total patients in this category is still quite small for a deep neural architecture, we have used all the data for predicting long or short hospital stay, rather than ICU stay. The median stay for this set of patients was 9 days. We conducted two experiments to check the performance. For the first experiment, the cut-off between short and long stays was assumed to be 7 days, while for the second experiment, it was assumed to be 9 days. While the accuracy is found to be 83% for 7 days, it is 86% for 9 days, which is the median. Thus median appears to be a good estimator for deciding long or short stays. It was found that the system performed more errors for patients with short stays, which were erroneously classified as long. Some of these patients, had deceased after a short stay. Though not exactly comparable, the prediction performance is found to be much higher than all reported works presented earlier, which may be due to the focused dataset or the representation, or both. Whether this kind of performance can be observed for all other diseases also, needs to be further explored, but this would also need more data for each of the corresponding categories.

VI. Patient Stratification using autoencoders - Generating explainable patient clusters

The autoencoded vector representations are clustered using the k-means clustering algorithm [41], with Euclidean distance as the distance metric to identify similarity among patients. For a given value of k, a set of k cluster centers are chosen randomly, and then each data point is assigned to the cluster that is found by iteratively minimizing the within-cluster distance among the points. In order to determine the right value of k, we have made use of *silhoutte coefficient* [42]. The silhouette coefficient of each point measures how similar it is to other points within the same cluster in comparison to points in other clusters. The average silhouette coefficient computed from all the points provides a measure of the cohesiveness of each cluster along with their separation or distinctiveness from each other. Starting with 2, the value of k is iteratively increased as long as the silhouette score also increases with it. The ideal value of k is the one that yields the highest average silhouette score, beyond which the score starts to decrease steadily.

In order to generate human-interpretable explanations for the clusters, we used Shapley values [43], which can measure the contribution of each feature of each individual towards the final outcome, while preserving the sum of contributions of all. We wanted the explanations to be in terms of diseases and symptoms, including the dominant symptoms in a cluster, as well as the distinguishing aspects between the clusters. The CUI-based representation was used for the purpose. Treating the cluster labels as the target outcomes, a Random Forest classifier [44] was trained to predict the target labels, using the CUI vector-based representation of the patients. The trained model was then analyzed using SHAP TreeExplainer [43], [45], [46], to gain insights into the decision-making process. This method provides not only the contribution of each symptom to a particular label but also the SHAP values for each patient, thereby helping with the interpretation of why a patient has been assigned to a particular cluster. They also help in the interpretation of misclassifications by the model, if any.

A similar approach was followed to derive Shapley values for patients and clusters in terms of medicines. In this case, the Random Forest classifier was trained to predict the target cluster labels using the medicine vectors and then analyzed with the SHAP TreeExplainer. It helps in identifying the most significant medicines for each cluster, thereby providing insights about the common and distinct medicines administered to patients belonging to different clusters.

A. Identifying Anomalies within Clusters

The obtained SHAP values for patients' health conditions are now used to compute the anomaly score for each individual. Anomalous individuals - within a cluster who may demonstrate distinctive characteristics in terms of certain symptoms or response patterns, different from other members, should have higher scores. For a cluster C, the health condition anomaly score of a patient p is denoted by $\alpha_C(p)$ and computed as follows:

$$\alpha_C(p) = \sum_{h \in H} |\omega_h(p, C) * (m_h(C) - v_h(p))|,$$

where H denotes the set of all CUIs, $v_h(p)$ is either 1, 0, or -1 based on whether the symptom h is present for p or not, as described earlier, $m_h(C)$ is the median value of the symptom h in C and $\omega_h(p, C)$ denotes the SHAP value for symptom h for patient p with respect to C. $\alpha_C(p)$ is normalized for each cluster to keep the scores within 0 to 1. The higher anomaly score indicates that the patient is more distinctive from his/her neighbors.

Since the duration of sickness also varies largely among patients, a normalized anomaly score is also computed in terms of duration. For a patient p with duration of sickness d, the

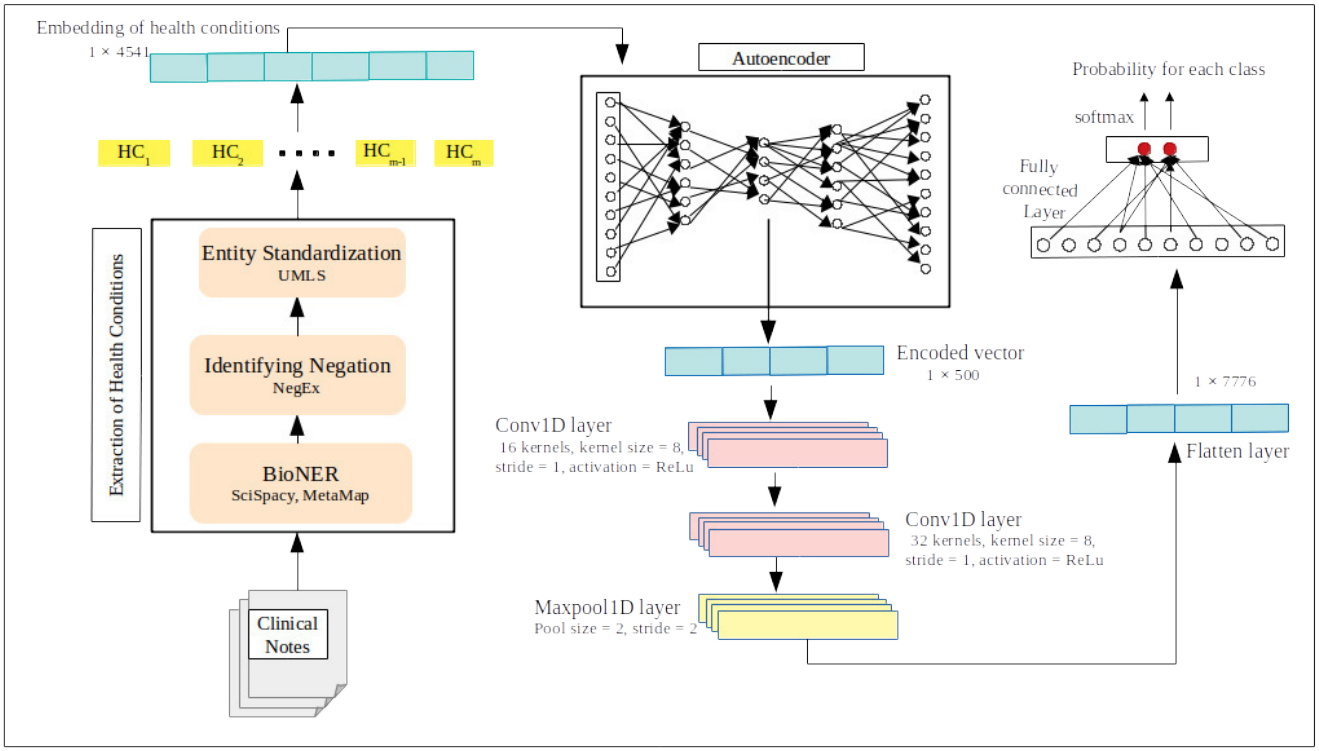


Fig. 3. Overview of the proposed framework for processing clinical notes, generating representations with an autoencoder, and subsequently applying CNN to predict hospitalization duration.

duration anomaly score denoted by $\gamma_C(p)$ with respect to other members of C is computed as follows :

$$\gamma_C(p) = |d - m_C| / \max_{i \in C} \gamma_C(i),$$

where m_C is the mode value of duration of disease for cluster C.

Absolute anomaly score $\chi(p)$ is computed using the following formula:

$$\chi_C(p) = 0.5 * \sqrt{\alpha_C(p)^2 + \gamma_C(p)^2},$$

This distributes the scores within the first quadrant of a unit circle, centered at the origin. The anomaly score is high for the set of points that are far away from the origin, the higher with the x and the y axes values providing insights about the symptom anomalies and duration anomaly respectively.

B. Assessing Causal effects of Significant Medications within Clusters

We further propose the use of causal analysis of clusters to analyze the effect of medications on patients. Causal analysis attempts to identify the effect of different treatments on groups of patients, based on observed outcomes. For patient stratification, the duration of sickness may be considered as an observed outcome, while medications or procedures are the treatment variables.

Causal analysis was done using the DoWhy package [47]. Given the diseases and symptoms as common cause variables, the duration of sickness as the outcome, and the medications as treatments administered, DoWhy generates an initial causal graph then estimands are identified using the graph. The final causal estimates are obtained using Propensity Score Weighting [48] as the estimator, and refutation as the validation technique. The causal estimates are validated using two different methods namely "adding random common cause" and "data subset validation" [49]. While the first method estimates the effect of a treatment by adding random independent variables, the second one does the same taking subsets of data. Either way, causal estimates are assumed to be good if the results don't show high perturbations, indicated by the $p\ value$.

VII. Results of Patient Stratification for Pneumonia patients in MIMIC dataset

In this section, we present results for analyzing data of 2106 patients from the MIMIC-III v1.4 database [2], who were all admitted to the hospital and diagnosed as suffering from "Pneumonia", using the above methods. The intent was to:

(a) Obtain explainable clusters of patients based on their health conditions recorded in the first set of clinical notes on admission.

(b) Identify anomalies within each cluster and also generate explanations for the anomalous behaviors.

(c) Identify the significant medicines for each cluster using the SHAP values.

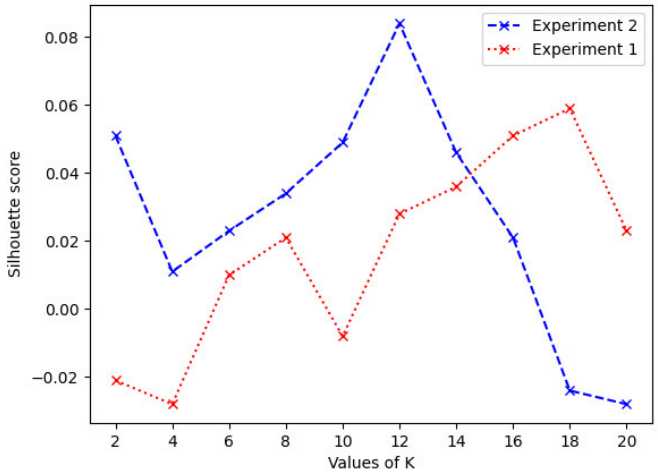


Fig. 4. Average silhouette scores for k = 2 to 20 for Experiment 1 (red) and Experiment 2 (blue).

(d) Study the causal effect of significant medicines on the duration of disease for the clusters.

As mentioned earlier, each of these patients had multiple clinical notes attached to them, approximately one per day of admission. The notes varied quite a bit in content. While some contained only incremental information, some were detailed and overlapped with earlier ones. A total of $23,737$ notes were obtained. After pre-processing all the notes as described in section V, a comprehensive list of 4800 unique health conditions was identified. Initially, the entire set of $23,737$ records was used to generate 500-dimensional autoencoded representations for the patients. However, the clustering results were not satisfactory. Hence, after some experimentation, this was changed to use only the first day's notes. The number of unique health status reduced to 2803. The 500 dimensional auto-encoders were then generated using the vectors for the first-day notes only. Figure 4 presents the average silhouette scores for k values ranging from 2 to 20 for both experiments. The best score was achieved for k equal to 12, using the second method. Figure 5 presents the distribution of these 12 clusters plotted using tSNE [50]. It shows that the clusters are fairly distinct and well-separated. Analysis of the SHAP summaries also revealed that the clusters were distinct and well-segregated from each other.

Since the trade names of drugs varied a lot, though their compositions were same, therefore the drugs corresponding to the notes were obtained from the database and mapped to their drug classes using the pre-trained *gpt-3.5-turbo* model [51]. However, the drug summaries generated by SHAP were not found to be very distinct from each other. This can be because the same drug might be prescribed for two different health conditions, or two different drugs might have been administered for the same health condition by different physicians. This needs to be analyzed in depth further, which remains a future task.

Figure 6 presents the SHAP summaries in terms of major health conditions present and absent, and the drugs identified for a few clusters. It shows that cluster 0 predominantly consists of patients suffering from *Endometriosis* and *Diabetes* along with pneumonia, and do not exhibit *Paroxysmal familial ventricular fibrillation*. Similarly, patients belonging to cluster 4 are found to suffer from *Lung consolidation* and do not have *Paroxysmal familial ventricular fibrillation*. Cluster 6 predominantly comprises patients with *Atrial Premature Complexes* and does not exhibit *Lung Consolidation* or *Paroxysmal familial ventricular fibrillation*. It may be observed that the most common co-morbidity was some or the other form of cardiac disease. The SHAP summary for drugs for clusters 0, 4, and 6, show that the most significant medicine administered to cluster 0 patients is *antidiabetic hormones*, which is obviously to handle diabetes, for patients of cluster is *proton pump inhibitor* and *hypoglycemic agent*, and the most prevalent medication for patients of cluster 6 is found to be *Vasopressor* which is for regulating blood pressure.

Anomaly scores revealed that the most anomalous patient in cluster 0 is a 67-year-old patient with a hospital stay of 62 days, deviating significantly from the cluster's mode value of 7 days. This anomaly can be attributed to the coexistence of Endometriosis, Paroxysmal familial ventricular fibrillation, and Acquired abnormality of atrium, a combination not observed in other members of this group. For cluster 4, while most patients were aged between 60 to 80, the most anomalous patient was a 33-year-old person who spent 81 days in the hospital, against the cluster mode value of 6 days. This person had all significant common symptoms of the cluster along with *Paroxysmal familial ventricular fibrillation*. The top two most anomalous persons of cluster 6 are aged 72 and 52 years, with 58 and 46 days of stays against cluster mode of 8 days. The anomalous symptom for the first person is *renal cyst*, while for the second patient, they are *lung consolidation* and *liver failure*. While most patients in this cluster are aged between 60 to 80, a third anomalous patient is aged 33 and suffered from multiple comorbidities but was admitted for only 3 days. Thus, it can be seen that the anomaly score is able to identify patients who are demographically outliers, even though age was not taken into account for computing the score.

Consequently, causal analysis was done for the drugs that were found to be significant for these clusters. For cluster 0, the effect of *antidiabetic hormone* is -0.459. The negative values indicate that these medicines contributed towards reduced duration of stay. Also, we have observed that this medicine is given 83% of short-stayed patients. On the other hand, the effect of *proton pump inhibitor* is 2.746 and this medicine is given to 89% of long-stayed patients. For cluster 4, the causal effect values of medicine *beta-blocker* was -2.7, for *Analgesic anti-platelet* it was -2.02, and for *Bronchodilator* it was -1.51, contributed towards in decreasing in the duration of hospital stay. *Opioid analgesic* had the highest positive value, indicating that this did not play a role in reducing the duration. For cluster 6, in which 84% of patients suffer from *arterial premature complexes*, medicines *Corticosteroid (-3.17), Anticonvulsant/ neuropathic pain agent (-3.02), Opioid*

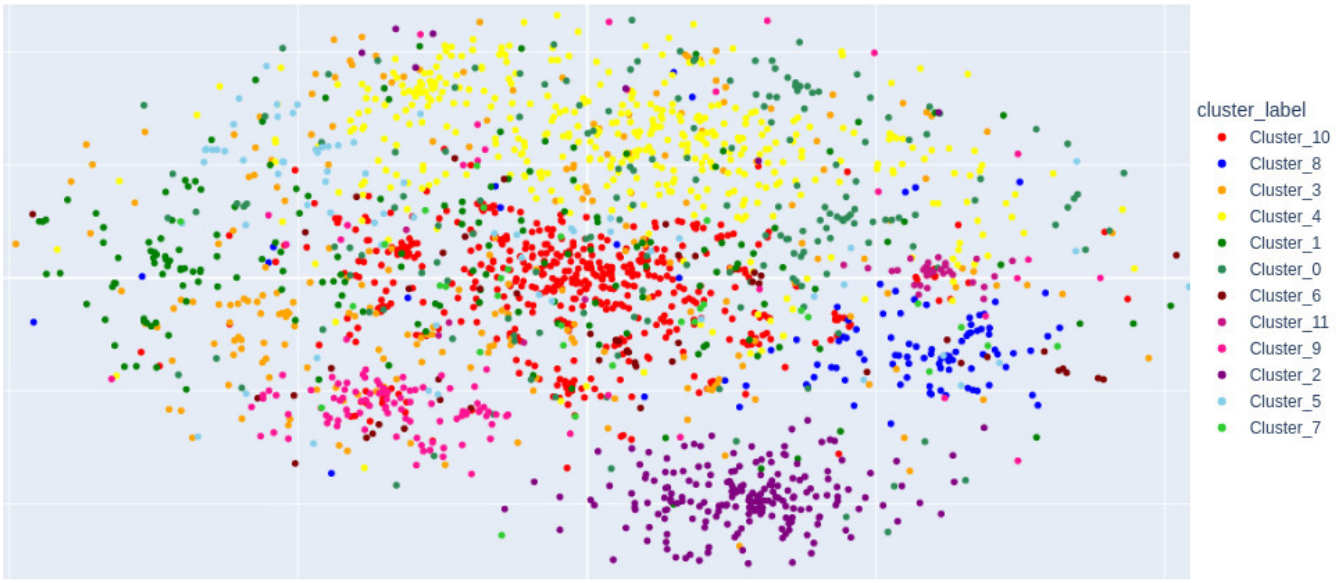


Fig. 5. t-SNE visualization of 12 clusters.

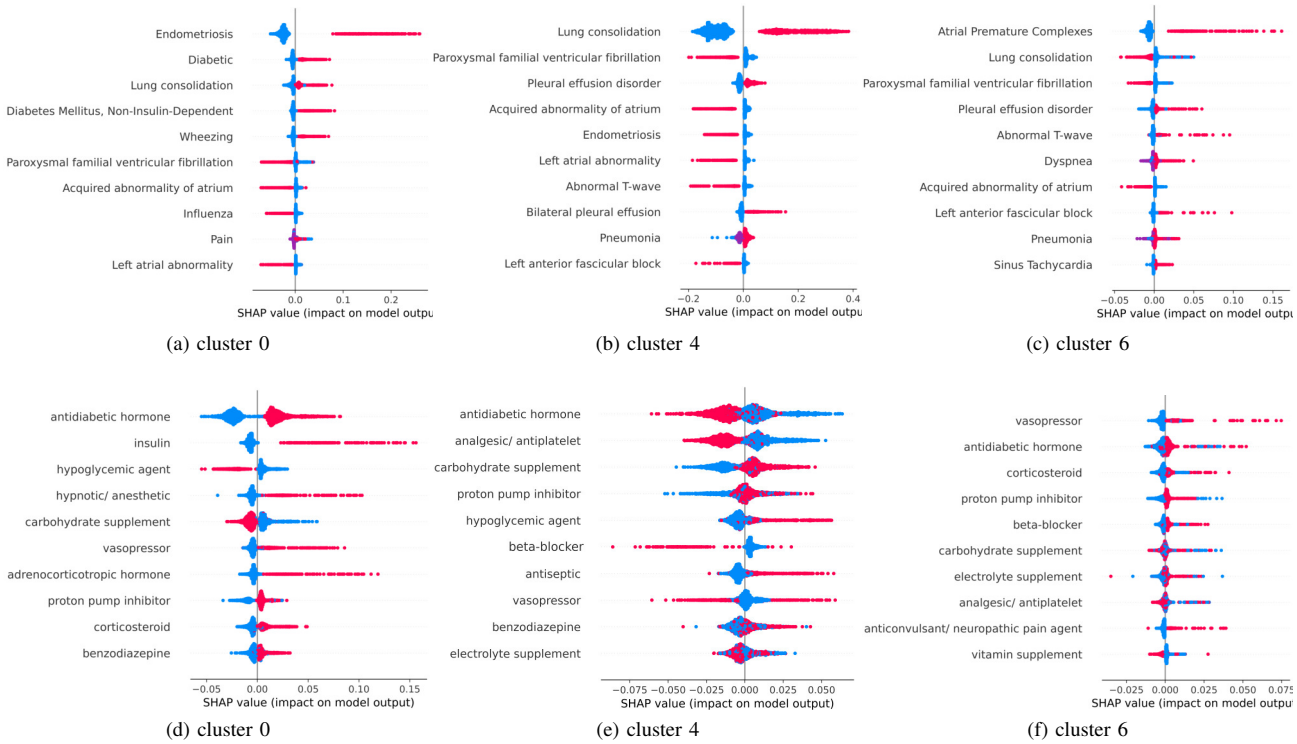
analgesic (-2.4), Vasopressor (-1.02) are found to causally reduce the duration of stay. Further analysis for cluster 6 reveals that 58% of patients with less than 6 days of hospital stay, 52% of those with 6 - 10 days, and 50% of those with higher than 10 days, were administered with *Corticosteroid* on the first day. The manual inspection also reveals that *Corticsteroid* was not administered to the top three anomalous patients, whose details were presented earlier, on the first day. Likewise, we conducted similar analyses for the remaining clusters. Though the exact implications of the results are best analyzed by healthcare experts, our analysis reveals that the results obtained from anomaly detection, their explanations, and causal analysis are all consistent with each other.

We also extracted the final recovery status of the patients from the discharge summaries. Based on the descriptions, we identified three major states at the time of discharge - (a). deceased patients, (b). patients whose vital signs were stable, could ambulate independently, and were coherent, (c). patients who were lethargic but mentally alert and ambulated with assistance. Figure 7 shows the distribution of the different clusters of patients identified earlier (based on their initial states) across these three categories. The top three categories of deceased patients are from clusters 7, 0 and 11, who suffered from comorbidities like ventricular hypertrophy, endometriosis or showed severe signs of edema. The results indicate that there is a need to look deeper into these cohorts, especially into the symptoms presented by the deceased patients. Comparative analysis of lengths of hospital admission for all patient clusters and deceased patients of each cluster are shown in Figure 8. It can be seen that the medians of the deceased patients for each cluster are almost same as those of other patients, which is around 10. While majority of the patients survived, a small percentage could not, and future work would be to delve deeper into the reasons for these differing outcomes. Superficial analysis at this point reveals that these patients had higher number of co-morbidities at admission time, some uncommon ones like *Septic embolus, Kidney Failure, AIDS-Associated Nephropathy, sepsis* etc. Data and the trajectory of treatment of these patients can be analyzed further to see what could have been done to ensure a positive outcome.

Figure 9 presents cluster-wise classification accuracy of predicting LOS for test patients, obtained by the classifier mentioned in section V-B. It is observed that there were no classification error for patients of cluster 6, which also incidentally had the least variation in terms of length of stay for patients as shown in 8. In future we would like to work on prediction of outcomes for patients within individual clusters. We are also exploring the role of SHAP values in explaining the prediction outcomes since though the LIME framework used earlier provides measure of associations between words and classes, it does not provide any explanation about why a particular class was assigned to an individual based on a collection of features.

VIII. Conclusion

Clinical notes are the backbone of healthcare systems. Well-written documents can provide valuable insights about diseases, patients and treatment effectiveness. They can provide a wealth of information about the similarities and diversity of situations across different situations. Deciphering the notes themselves however is a difficult problem due to the inherent variations in terms of style and content, which result from individual and organizational preferences. Obtaining these notes for analytical tasks is also a difficult problem. Though their utility is well-acknowledged, still these are not available in volumes that can help in designing machine-learning models for analyzing them. The key concerns are those of security



(a) cluster 0 (b) cluster 4 (c) cluster 6

(d) cluster 0 (e) cluster 4 (f) cluster 6

Fig. 6. SHAP values for Cluster 0, 4, and 6 of health conditions (top) and drugs (bottom). Red indicates high significance and blue low. Right side indicates presence and left side indicates absence of a feature.

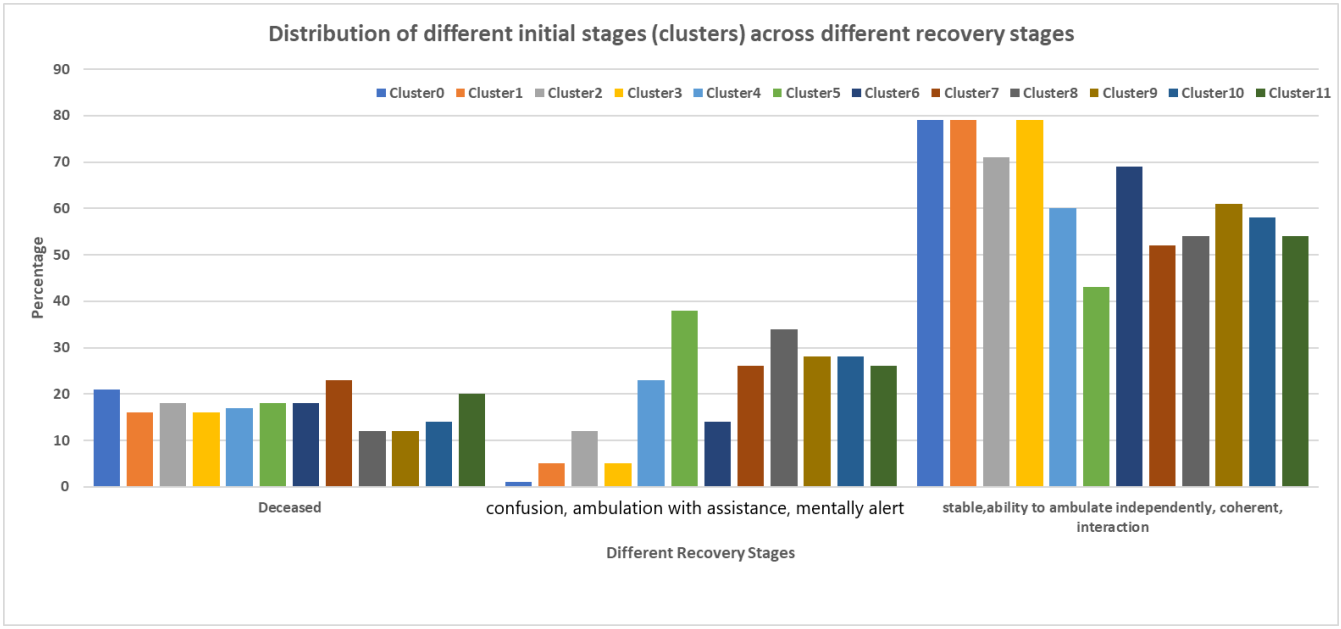


Fig. 7. Distribution of patients of twelve clusters derived from initial stages across recovery classes.

and privacy violations. Various groups however have started reporting their work on proprietary data. While this does establish the legitimacy of the problem, it is often not easy to reproduce the results in another setting or conduct a comparative analysis of the results obtained. The MIMIC dataset available for researchers alleviates some of the problems. It is a fairly large dataset with substantial number of clinical notes associated with details about diseases, treatment and outcomes.

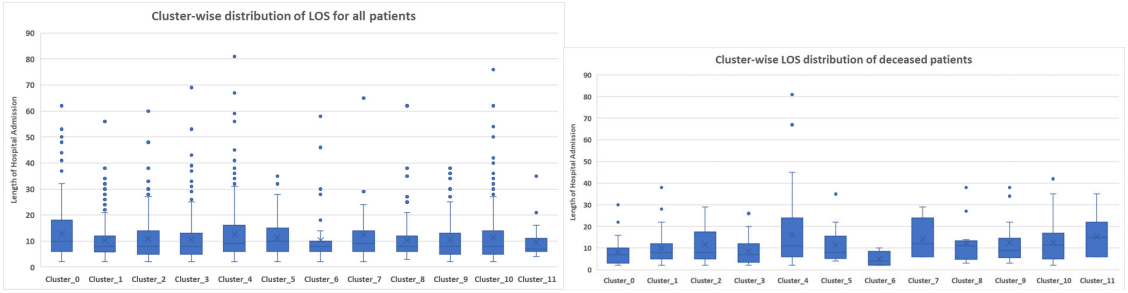


Fig. 8. (a) Clusterwise distribution of length of stay for all patients (b) Clusterwise distribution of length of stay for deceased patients.

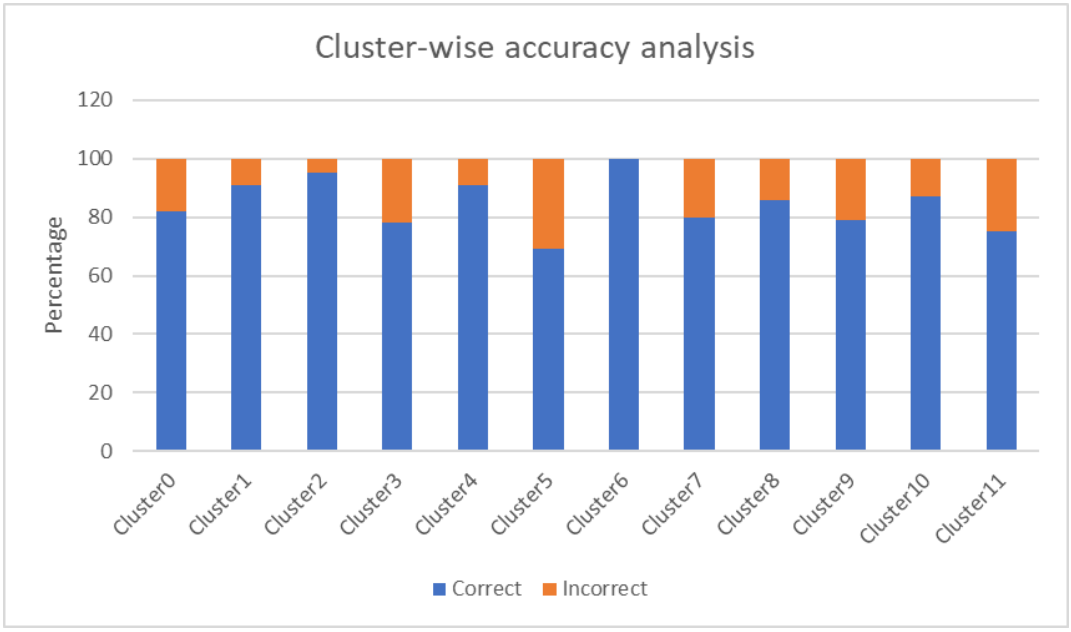


Fig. 9. Clusterwise accuracy of length of stay classification.

The use of clinical notes in predicting length of hospital admission, readmission possibilities, treatments as well as expected clinical outcomes has been prevalent for quite some time. Presently, the use of unstructured clinical notes is on the rise for development of prognostic prediction models. The focus is on developing explainable models. It is expected that robust and trustworthy prediction models will change the course of clinical practice as treatment procedures will move from majority focused designs to more customized designs.

In this paper, we have presented some initial work that we have started for explainable patient stratification. We have shown that deep learning based representations can effectively capture the richness of clinical notes and thereby be used to provide valuable insights about patient cohorts as well as exceptions within them. In future, we intend to extend our work in generating complete trajectories using the proposed representations. These trajectories in conjunction with the outcomes can help in risk assessment for patients, and thereby help in steering towards low-risk trajectories, especially for patients who are outliers.

REFERENCES

[1] H. Dalianis, *Clinical text mining: Secondary use of electronic patient records*. Springer Nature, 2018.
[2] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
[3] B. Percha, "Modern clinical text mining: a guide and review," *Annual review of biomedical data science*, vol. 4, pp. 165–187, 2021.
[4] T. L. Rodziewicz and J. E. Hipskind, "Medical error prevention," *StatPearls. Treasure Island (FL): StatPearls Publishing*, 2020.
[5] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin, "A systematic review of the prediction of hospital length of stay: Towards a unified framework," *PLOS Digital Health*, vol. 1, no. 4, p. e0000017, 2022.
[6] K.-C. Chang, M.-C. Tseng, H.-H. Weng, Y.-H. Lin, C.-W. Liou, and T.-Y. Tan, "Prediction of length of stay of first-ever ischemic stroke," *Stroke*, vol. 33, no. 11, pp. 2670–2674, 2002.
[7] A. Lim, "Statistic methods for modeling incidence of infectious diseases mortality and length of stay in hospital fot patients dying in southern thailand," Ph.D. dissertation, Prince of Songkla University, Pattani Campus, 2009.
[8] D. A. Huntley, D. W. Cho, J. Christman, and J. G. Csernansky, "Predicting length of stay in an acute psychiatric hospital," *Psychiatric services*, vol. 49, no. 8, pp. 1049–1053, 1998.
[9] T. Gentimis, A. Ala'J, A. Durante, K. Cook, and R. Steele, "Predicting hospital length of stay using neural networks on mimic iii data," in *2017*

IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE, 2017, pp. 1194–1201.

[10] K. Alghatani, N. Ammar, A. Rezgui, A. Shaban-Nejad *et al.*, "Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation," *JMIR Medical Informatics*, vol. 9, no. 5, p. e21347, 2021.

[11] L. Su, Z. Xu, F. Chang, Y. Ma, S. Liu, H. Jiang, H. Wang, D. Li, H. Chen, X. Zhou *et al.*, "Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models," *Frontiers in Medicine*, vol. 8, p. 883, 2021.

[12] E. Rocheteau, P. Liò, and S. Hyland, "Predicting length of stay in the intensive care unit with temporal pointwise convolutional networks," *arXiv preprint arXiv:2006.16109*, 2020.

[13] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[14] S. Khadanga, K. Aggarwal, S. Joty, and J. Srivastava, "Using clinical notes with time series data for icu management," *arXiv preprint arXiv:1909.09702*, 2019.

[15] B. van Aken, J.-M. Papaioannou, M. Mayrdorfer, K. Budde, F. A. Gers, and A. Löser, "Clinical outcome prediction from admission notes using self-supervised knowledge integration," *arXiv preprint arXiv:2102.04110*, 2021.

[16] S. Jana, T. Dasgupta, and L. Dey, "Using nursing notes to predict length of stay in icu for critically ill patients," in *Multimodal AI in healthcare: A paradigm shift in health intelligence*. Springer, 2022, pp. 387–398.

[17] ——, "Predicting medical events and icu requirements using a multimodal multiobjective transformer network," *Experimental Biology and Medicine*, vol. 247, no. 22, pp. 1988–2002, 2022.

[18] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[19] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.

[20] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] X. Li and K.-C. Wong, "Evolutionary multiobjective clustering and its applications to patient stratification," *IEEE transactions on cybernetics*, vol. 49, no. 5, pp. 1680–1693, 2018.

[24] R. W. Grant, J. McCloskey, M. Hatfield, C. Uratsu, J. D. Ralston, E. Bayliss, and C. J. Kennedy, "Use of latent class analysis and k-means clustering to identify complex patient profiles," *JAMA network open*, vol. 3, no. 12, pp. e2 029 068–e2 029 068, 2020.

[25] N. Alexander, D. C. Alexander, F. Barkhof, and S. Denaxas, "Identifying and evaluating clinical subtypes of alzheimer's disease in care electronic health records using unsupervised machine learning," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–13, 2021.

[26] F. Angelini, P. Widera, A. Mobasheri, J. Blair, A. Struglics, M. Uebelhoer, Y. Henrotin, A. C. Marijnissen, M. Kloppenburg, F. J. Blanco *et al.*, "Osteoarthritis endotype discovery via clustering of biochemical marker data," *Annals of the Rheumatic Diseases*, vol. 81, no. 5, pp. 666–675, 2022.

[27] S. V. Bhavani, L. Xiong, A. Pius, M. Semler, E. T. Qian, P. A. Verhoef, C. Robichaux, C. M. Coopersmith, and M. M. Churpek, "Comparison of time series clustering methods for identifying novel subphenotypes of patients with infection," *Journal of the American Medical Informatics Association*, vol. 30, no. 6, pp. 1158–1166, 2023.

[28] T. Chen, X. Li, Y. Li, E. Xia, Y. Qin, S. Liang, F. Xu, D. Liang, C. Zeng, and Z. Liu, "Prediction and risk stratification of kidney outcomes in iga nephropathy," *American journal of kidney diseases*, vol. 74, no. 3, pp. 300–309, 2019.

[29] F. Kanwal, J. H. Shubrook, L. A. Adams, K. Pfotenhauer, V. W.-S. Wong, E. Wright, M. F. Abdelmalek, S. A. Harrison, R. Loomba, C. S. Mantzoros *et al.*, "Clinical care pathway for the risk stratification and management of patients with nonalcoholic fatty liver disease," *Gastroenterology*, vol. 161, no. 5, pp. 1657–1669, 2021.

[30] T. M. Seinen, E. A. Fridgeirsson, S. Ioannou, D. Jeannetot, L. H. John, J. A. Kors, A. F. Markus, V. Pera, A. Rekkas, R. D. Williams *et al.*, "Use of unstructured text in prognostic clinical prediction models: a systematic review," *Journal of the American Medical Informatics Association*, vol. 29, no. 7, pp. 1292–1302, 2022.

[31] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[32] M. Neumann, D. King, I. Beltagy, and W. Ammar, "Scispacy: Fast and robust models for biomedical natural language processing. arxiv 2019," *arXiv preprint arXiv:1902.07669*.

[33] A. R. Aronson, "Metamap: Mapping text to the umls metathesaurus," *Bethesda, MD: NLM, NIH, DHHS*, vol. 1, p. 26, 2006.

[34] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of biomedical informatics*, vol. 34, no. 5, pp. 301–310, 2001.

[35] B. McInnes, Y. Liu, T. Pedersen, G. Melton, and S. Pakhomov, "Umls:: similarity: Measuring the relatedness and similarity of biomedical concepts." Association for Computational Linguistics, 2013.

[36] P. A. Hall and G. R. Dowling, "Approximate string matching," *ACM computing surveys (CSUR)*, vol. 12, no. 4, pp. 381–402, 1980.

[37] R. Haldar and D. Mukhopadhyay, "Levenshtein distance technique in dictionary lookup methods: An improved approach," *arXiv preprint arXiv:1101.1232*, 2011.

[38] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[39] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 490–497.

[40] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *arXiv preprint arXiv:2003.05991*, 2020.

[41] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[42] T. M. Kodinariya, P. R. Makwana *et al.*, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[43] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using shapley values," in *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*. Springer, 2020, pp. 17–38.

[44] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Ensemble machine learning: Methods and applications*, pp. 157–175, 2012.

[45] I. Ekanayake, D. Meddage, and U. Rathnayake, "A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using shapley additive explanations (shap)," *Case Studies in Construction Materials*, vol. 16, p. e01059, 2022.

[46] S. Cohen, E. Ruppin, and G. Dror, "Feature selection based on the shapley value," *other words*, vol. 1, p. 98Eqr, 2005.

[47] A. Sharma and E. Kiciman, "Dowhy: An end-to-end library for causal inference," *arXiv preprint arXiv:2011.04216*, 2020.

[48] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, no. 3, pp. 399–424, 2011.

[49] A. Sharma, V. Syrgkanis, C. Zhang, and E. Kıcıman, "Dowhy: Addressing challenges in expressing and validating causal assumptions," *arXiv preprint arXiv:2108.13518*, 2021.

[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[51] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen *et al.*, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023.