Patient Case Similarity (Review 01)

Sanchita Goswami[1], Darshan M S[1], Gaurav H[1], Srivatsa K S[1],Umme Kulsum[1] ,Dr.Manjunath KV[2]

[1] UG Student, Dept of CSE, Presidency University, Bengaluru, India

[2] Associate Professor, Dept of CSE, Presidency University, Bengaluru, India

Abstract - Evaluating patient similarity is a critical task in healthcare informatics, enabling various applications such as cohort studies and treatment effectiveness research. This task is often based on Electronic Health Records (EHRs), which are inherently heterogeneous, longitudinal, and sparse, presenting challenges in data representation and interpretability. Existing approaches often lose temporal information by relying on aggregated vector-based representations of patient data. In this paper, we propose an advanced patient similarity evaluation framework that leverages both static and dynamic patient data while preserving temporal characteristics. Our multi-model framework integrates deep learning methods, including Bidirectional Encoder Representations from Transformers (BERT) for contextual data, convolutional neural networks (CNN) for semantic feature extraction, and a long-short-term-memory (LSTM)-based autoencoder to capture temporal dependencies and reduce dimensionality. By fusing clinical narrative data with temporal EHR data, the proposed model generates more interpretable and accurate patient similarity measures. Empirical evaluations on real-world clinical datasets demonstrate significant improvements in classification accuracy and patient outcome predictions over traditional methods, showcasing the framework's potential for enhancing precision medicine.

***Key Words*:** patient; patient similarity network; precision medicine; big data; personalized healthcare; patient-centred framework; deep learning; electronic health records; transformers; BERT; autoencoder

LITERATURE SURVEY:

I. "Multi-model and multi-slice ensemble learning architecture based on 2Dconvolutional neural networks for Alzheimer's disease diagnosis"

This paper proposes a novel approach for Alzheimer's Disease (AD) diagnosis that combines information from multiple imaging modalities (T1-weighted MRI, T2-weighted MRI, and amyloid PET) and multiple slices of each brain image. The architecture utilizes 2D convolutional neural networks (CNNs) to extract features from brain images and ensemble learning to combine the predictions from multiple models. By leveraging multi-modal and multi-slice information, the model aims to capture a more comprehensive understanding of AD and improve diagnostic accuracy.

II. "A deep learning-based multi-model ensemble method for cancer prediction"

This paper presents a multi-modal ensemble method for cancer prediction that combines information from various data sources, including clinical data, imaging data, and genomic data. The method utilizes deep learning techniques to extract features from each modality and ensemble learning to integrate the predictions from multiple models. By leveraging the complementary information from different modalities, the model aims to improve prediction accuracy and provide valuable insights for clinical decision-making.

III. "A NOVEL MULTI-MODEL PATIENT SIMILARITY NETWORK DRIVEN BY FEDERATED DATA QUALITY AND RESOURCE PROFILING"

This paper introduces a patient similarity network (PSN) framework that leverages federated learning and multi-modal data to improve patient care. The PSN combines information from various sources such as

electronic health records (EHRs), clinical notes, and imaging data to identify patients with similar characteristics and treatment histories. By leveraging federated learning, the PSN can be trained collaboratively across multiple institutions without sharing sensitive patient data, ensuring privacy and security. The framework aims to optimize resource allocation and provide personalized care by identifying patients who may benefit from similar treatments or interventions.

IV. "Toward Precision Medicine in Intensive"

This paper explores the potential of precision medicine in intensive care settings. It discusses the challenges and opportunities associated with applying personalized approaches to critically ill patients. The paper highlights the importance of patient similarity analysis in identifying patients with similar characteristics and predicting outcomes.

V. "Patient similarity for precision medicine: A systematic review"

This paper provides a comprehensive review of existing methods for patient similarity analysis in the context of precision medicine. The authors discuss various methods for measuring similarity, including demographic information, medical history, lab results, and medications. They also highlight the challenges associated with patient similarity analysis, such as data quality, standardization, and privacy concerns.

VI. "Patient Similarity Emerging Concepts in Systems and Precision Medicine"

This paper explores the concept of patient similarity in detail, defining it as a multifaceted construct that encompasses various factors such as demographics, clinical characteristics, genetic makeup, lifestyle habits, and environmental exposures. The paper discusses the applications of patient similarity in personalized medicine, clinical decision support, drug discovery, and disease prediction. It also highlights the challenges associated with patient similarity, such as data quality, privacy concerns, and ethical considerations.

VII. "A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine"

This paper introduces a patient similarity network (PSN) framework that leverages multi-modal deep learning to identify patients with similar characteristics. The PSN combines information from various data sources, including clinical notes, vital signs, and lab results, to capture complex relationships between patient features. The framework aims to improve personalized treatment recommendations, risk stratification, and clinical decision support.

VIII. "Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding"

This paper proposes a deep learning approach for patient similarity measurement using Medical Concept Embedding (MCE). MCE represents medical concepts as vectors in a high-dimensional space, allowing the model to capture semantic relationships between concepts. By leveraging MCE, the model can learn complex patterns in patient data and identify meaningful similarities between patients.

IX. "Comparing clinical information with claims data: Some similarities and differences"

This paper investigates the relationship between clinical information and claims data for patient care similarity. The authors compare the two data sources in terms of their ability to capture patient characteristics and medical histories. They find that while claims data can be a useful source of information, it has limitations in terms of detail and subjectivity. The paper highlights the importance of using both clinical information and claims data to obtain a comprehensive understanding of patient care.

X. "A method for inferring medical diagnoses from patient similarities"

This paper introduces a novel method for predicting medical diagnoses based on patient care similarity. The authors propose a patient care similarity metric that considers various factors such as demographics, medical history, symptoms, and treatments. By comparing a new patient to existing patients with known diagnoses, the system can infer the most likely diagnosis based on the similarity between their medical care patterns. The paper demonstrates the effectiveness of the proposed method in improving the accuracy and efficiency of medical decision-making.

Objectives:

The primary objective of this project is to develop a robust system capable of identifying similar patient cases based on their medical reports. This task involves applying machine learning algorithms to analyze patient data and extract meaningful patterns that can be used to group patients with similar characteristics.

- **Improved Patient Outcomes:** By identifying similar patients, healthcare providers can tailor treatment plans to individual needs, leading to better outcomes.
- **Treatment and Drug Recommendations:** Understanding patient similarities can aid in recommending appropriate treatments or drugs for new patients.
- **Prediction of Clinical Outcomes:** Analyzing similar cases can help predict the potential outcomes of a specific treatment or intervention.
- **Clinical Decision Support:** Providing recommendations based on the experiences of similar patients can support clinicians in making informed decisions.
- **Research:** Identifying similar patient cases can facilitate research on specific diseases or treatments.

By successfully completing this project, we aim to contribute to the advancement of patient care by providing a valuable tool for identifying similar patient cases and improving clinical decision-making.

Proposed Model:

The model proposed in this work consists of four key components that aim to capture patient similarity based on Electronic Health Records (EHRs) data. These components include contextual embedding of medical concepts, temporal patient representation, unsupervised patient similarity, and supervised patient similarity using convolutional neural networks (CNNs). The model emphasizes both the temporal characteristics of medical events and the learning of medical concept embeddings, improving the precision of patient similarity measurements.

Datasets:

Our model is trained on a real world longitudinal EHRs database of 218,680 patients for the course of over four years. According to the reasons presented at the beginning of this section, we select four patient cohorts from the EHRs data, namely, Chronic Obstructive Pulmonary Disease (COPD), Diabetes, Heart Failure, and Obesity.

Table I provides a summary of the patient cohorts used in our experiment. Each cohort consists of a set of case patients who are confirmed with one of the four diseases according to their medical diagnosis, and each patient comes with a set of medical events including diagnosis and medications. In each patient encounter, we use the International Classification of Disease-Version 9 (ICD-9) codes to denote the diagnosis of diseases that a patient suffers from. All the clinical events about medications are pre-processed to normalize the descriptions based on brand names and clinical dosages.

Cohorts	# Patients	# Events
COPD	2,000	247,043
Diabetes	2,000	259,074
Obesity	2,000	211,496
Heart Failure	1,135	165,254
Total	7,135	882,867

Table I: Summary of EHRs datasets for patients clustering.

We construct datasets with medical events collected from patients who were confirmed of having the disease by medical experts. We develop the criteria that any patients presented in the datasets has at least forty events. The requirement is set to ensure that each test case has minimum events of clinical history that could be used in reasonable analytics tasks in healthcare. Also, to enable distinctly cluster without overlapping among cohorts, we remove patients who suffers from more than one disease in the cohort list. Finally, there are 8,000 remaining patients and 6,064 distinct clinical events. Medical event appearing in more than 90% of patients or present in fewer than five patients are removed from the datasets to avoid biases and noise in the learning process.

In the following experiments, we use two datasets:

DATASET-I uses the complete patients' events while DATASET-II reserves historical events except those labelled as cohort identifiers. On DATASET-I, we split the dataset into training and test sets with same number of patients, and other patients left for validation. As for DATASET-II, we construct the data sets in accordance with DATASET-I. A few of patients are filtered out because of the limited number of their medical events. Table II summaries the two datasets.

Data	# Patients	# Events
DATASET-I TRAIN	3,211	396,072
TEST	3,210	399,804
DEV	714	86,991
DATASET-II TRAIN	3,083	373,145
TEST	3,080	377,287
DEV	685	81,392

Table II: Summary of modelling datasets.

A. Contextual Embedding of Medical Concepts

- **Objective**: The goal is to learn **contextual embeddings** of medical concepts from patient EHRs. This method provides a better representation compared to conventional one-hot encoding.
- **Approach**: Inspired by NLP methods like **Word2Vec**, the system treats medical events as "words" and the surrounding medical events (those happening before and after) as the "context" around each medical event. By encoding medical events into vector spaces, these embeddings reflect medical contextual correlations.
- **Key Considerations**: The model adapts window lengths to different medical events (chronic conditions may have larger contexts compared to acute ones) and incorporates timestamps to differentiate between events happening days versus years apart.

B. Temporal Patient Representation

- **Representation Format**: Each patient p is represented as a matrix X with dimension d × Np, where d is the fix embedding dimension and Np is the total number of visit patient p has.. Each visit is represented by summing all the medical vectors from that visit, preserving temporal order.
- **Challenge**: Comparing patients with different numbers of visits NpN_pNp introduces variability in matrix dimensions, which requires specialized techniques for similarity computation

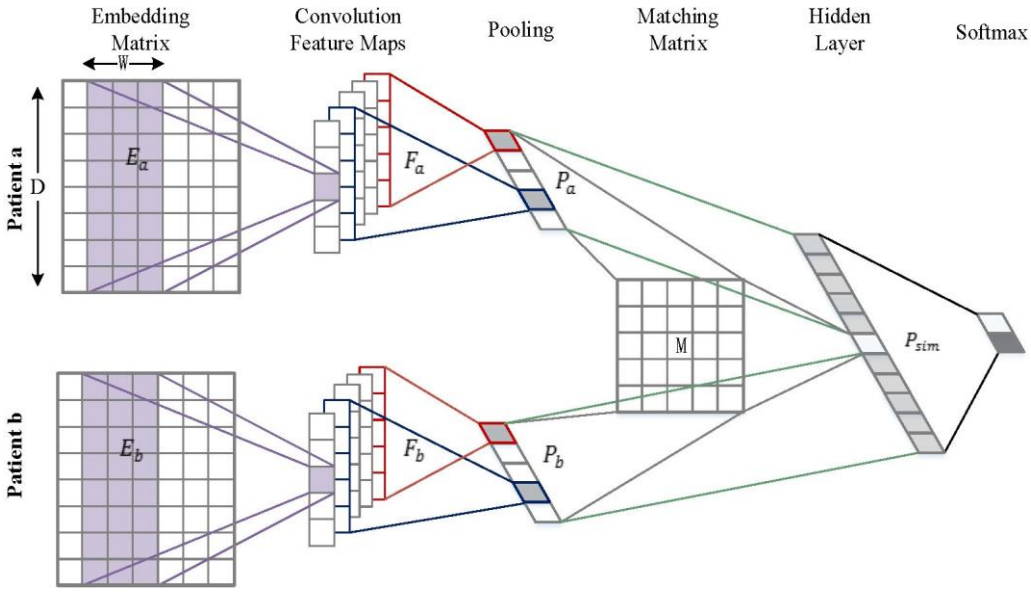

Figure 1: The overall framework of supervised patient similarity matching. To train the singular neural network, embedding matrices of pairs of patients E_a,E_b passed through convolutional filters are mapped into feature maps. We build the deep embedding patients representations P_a,P_b for patients by pooling patients feature maps into the intermediate vectors. With the rich feature vectors, we learn a symmetrical similarity matrix M for measuring the distance between patient a and b.

C. Unsupervised Patient Similarity

- **RV Coefficient**: Used to compute similarity between two patients based on their temporal representations. Given matrix representations X ∈R^n×k and Y ∈R^m×k the **RV coefficient** measures linear relations between the two matrices.

$$\mathrm{RV}(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{X}\mathbf{X}^{`}\mathbf{Y}\mathbf{Y}^{`})}{\sqrt{\mathrm{tr}(\mathbf{X}\mathbf{X}^{`})^2 \mathrm{tr}(\mathbf{Y}\mathbf{Y}^{`})^2}}$$

- **dCov Coefficient**: Measures **non-linear relations** between patient records using **distance covariance** (dCov). This method captures more complex relationships between patient data, ensuring flexibility in assessing similarity

D. Supervised Patient Similarity (CNN-based)

- **Model Architecture**: Inspired by **semantic matching** in NLP, a **Convolutional Neural Network (CNN)** architecture is proposed to compute patient similarity with supervision. CNN filters extract relevant medical concept sequences from patient EHR data, allowing the model to capture patterns across patient records.
- **Convolution and Pooling**: Event sequences are processed through convolutional filters, and **max pooling** is applied to reduce dimensionality while retaining important features.
- **Matching Matrix**: The similarity between patient representations xa and xb, is computed using a **matching matrix** MMM, optimized during training: $\mathrm{sim}(\mathbf{x}_a,\mathbf{x}_b) = \mathbf{x}^T{}_a \mathbf{M}\mathbf{x}^T{}_b$
- **Optimization**: The model uses supervised learning techniques, such as **square loss** and **AdaGrad**, to minimize prediction error and improve similarity accuracy. Regularization techniques like **dropout** are employed to prevent overfitting.

E. Optimization and Regularization

- **Loss Function**: Square loss is used for regression tasks, calculating the difference between predicted similarity scores and ground truth.
- **Training**: The parameters of the model, including embeddings and network layers, are jointly optimized using **Backpropagation** and **Stochastic Gradient Descent (SGD)**, specifically leveraging **AdaGrad** for efficient updates.
- **Regularization**: **Dropout** is applied in the penultimate layer to reduce overfitting, ensuring robustness in the model's predictions.

Drawbacks:

1. Complexity of Temporal Representation

- **Challenge**: The model's reliance on temporal patient representation, where each patient's visit history is encoded as a matrix, introduces significant complexity. The variation in the number of visits across patients means that comparing these matrices requires non-trivial methods like the RV coefficient and dCov coefficient.

- **Drawback**: The complexity may lead to difficulties in scalability when processing very large datasets with millions of patients and numerous medical events, resulting in high computational costs.

2. Limited Interpretability

- **Challenge**: Deep learning models, especially those using CNNs and complex embeddings, often act as "black boxes," making it difficult to interpret how the model arrived at specific similarity scores.
- **Drawback**: In clinical settings, interpretability is crucial. Physicians and healthcare professionals need clear, interpretable insights from the model to make informed decisions. The lack of transparency in the embedding and similarity calculation could limit trust in the system.

3. Difficulty Handling Rare Conditions

- **Challenge**: The model relies on frequency-based window sizing in the contextual embedding step, which assumes that more frequent events have broader contexts (chronic conditions). However, rare medical conditions may not appear frequently enough to provide meaningful embeddings.
- **Drawback**: This frequency-based approach may struggle with rare diseases or uncommon medical events, leading to inaccurate similarity scores for patients with less common conditions.

4. Data Sparsity in EHRs

- **Challenge**: Electronic Health Records (EHRs) are inherently sparse because patients only interact with the healthcare system periodically, and medical events are not consistently recorded for all conditions.
- **Drawback**: The sparsity of EHR data can reduce the quality of the learned embeddings, especially when medical histories are incomplete. This could negatively impact the accuracy of patient similarity measures, particularly for patients with irregular or infrequent healthcare visits.

5. Need for Extensive Labelling in Supervised Learning

- **Challenge**: The supervised patient similarity model requires labelled data to train the CNNs and optimize the matching matrix.
- **Drawback**: In many healthcare settings, obtaining sufficient labelled data is a challenge, as creating labelled datasets for medical conditions requires expert input and is time-consuming. This could limit the applicability of the supervised approach in real-world settings where labelled data is scarce.

6. Sensitivity to Hyperparameters

- **Challenge**: The performance of the model depends on a range of hyperparameters such as the window size in contextual embedding, the number of convolution filters in CNNs, and the choice of loss functions.
- **Drawback**: Fine-tuning these hyperparameters can be a time-consuming and resource-intensive process. In some cases, suboptimal choices of hyperparameters can lead to poor performance, especially for complex and diverse patient populations.

7. Generalization to Different Populations

- **Challenge**: The proposed model is trained and tested on specific datasets, often reflecting the population from which the data was collected.
- **Drawback**: The model may not generalize well to different populations with varying medical practices, demographics, or disease distributions. For example, a model trained on EHR data from a hospital in the U.S. may not perform well on data from a hospital in Europe or Asia due to differences in healthcare systems and patient characteristics.

REFERENCES

[1] Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis (2021)
[2] A deep learning-based multi-model ensemble method for cancer prediction (2017)
[3] A NOVEL MULTI-MODEL PATIENT SIMILARITY NETWORK DRIVEN BY FEDERATED DATA QUALITY AND RESOURCE PROFILING (2023)
[4] Toward Precision Medicine in Intensive (2019)
[5] Patient similarity for precision medicine: A systematic review (2018)
[6] Patient Similarity Emerging Concepts in Systems and Precision Medicine (2016)
[7] A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine (2022)
[8] Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding (2016)
[9] Comparing clinical information with claims data: Some similarities and differences (1991)
[10] A method for inferring medical diagnoses from patient similarities (2013)

Timeline of the Project (Gantt Chart)