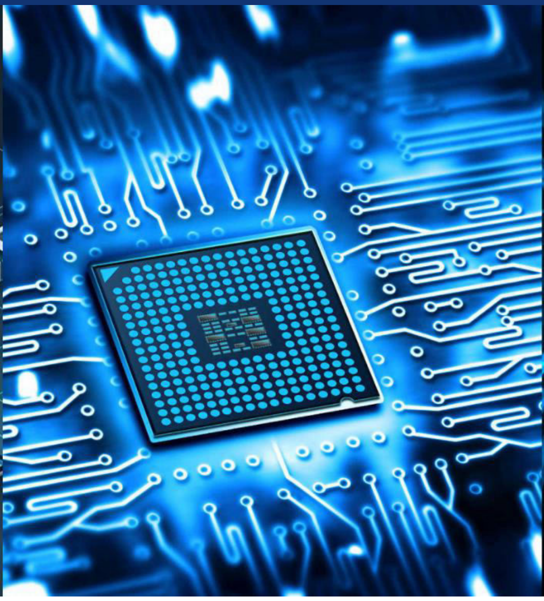
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.625

Volume 13, Issue 1, January 2025

www.ijircce.com ijircce@gmail.com +91-9940572462 +91 63819 07438 www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Patient Case Similarity for Predictive Healthcare Analytics: A Study using Electronic Health Records (EHR) and Machine Learning

Chandan R[1], Tharun Kumar[2], Shree Chakra[3], Affan[4], Himansu Sekhar Rout[5] [0009-0004-4373-8117] *

Assistant Professor, Department of CSE, Presidency University, Itgalpura, Bangalore, India

UG Student [CSE], Department of CSE, Presidency University, Itgalpura, Bangalore, India

ABSTRACT: In the evolving field of predictive healthcare, ana- lyzing patient case similarity has emerged as a key approach for tailoring patient care and improving outcomes. This study presents a methodology for identifying and evaluating patient similarity based on Electronic Health Records (EHR) data, leveraging machine learning to predict health outcomes and guide treatment decisions. Using a cancer diagnosis dataset and a custom-built model, we analyze historical patient data to find similar cases, providing clinicians with insights into potential disease progression and personalized treatment pathways. This paper discusses the pre-processing and modelling steps, along with backend integration for clinical use. Our findings underscore the effectiveness of similarity-based predictions in enhancing healthcare delivery, particularly in high-stakes or emergency contexts, by offering rapid, data-driven insights.

KEYWORDS: EHR (Electronic Health Records), Similarity Matrix, Classification.

I.INTRODUCTION

The healthcare industry faces increasing demands for inno- vative solutions that not only manage growing patient volumes but also address the complexities of personalized care. In this environment, predicting patient health outcomes with high accuracy is crucial, especially in cases where timely intervention can significantly influence prognosis. Traditional methods often struggle to capture the nuanced similarities among patients with complex conditions, limiting their effec- tiveness in providing personalized treatment recommendations. This research aims to overcome these challenges by focus- ing on patient case similarity, a method that uses historical patient data to identify individuals with comparable medical histories. By analyzing this similarity, clinicians can gain valu- able insights into potential disease progressions and treatment outcomes for a current patient, based on the experiences of similar past cases. This study specifically examines cancer patients' EHR data, employing machine learning techniques to compute patient similarity scores. These scores help to classify patients and predict health trajectories, improving the precision and responsiveness of healthcare delivery.

Through a carefully structured methodology, we utilize a cancer diagnosis dataset to evaluate the viability of case similarity models in practical healthcare settings. Our approach combines data preprocessing, feature selection, and machine learning to construct a predictive model that can be deployed in real clinical environments. By providing a detailed case similarity analysis, we aim to demonstrate the advantages of this approach in fostering data-driven, personalized healthcare solutions. This paper contributes to the growing body of literature on predictive healthcare analytics, offering insights into the potential of machine learning to transform patient care.

II. LITERATURE REVIEW

The concept of patient similarity analysis within predictive healthcare analytics has gained traction over the past decade, particularly as Electronic Health Records (EHR) become more accessible for large-scale research. Patient similarity mod- els are essential for identifying individuals with comparable medical histories, symptoms, and outcomes, which can be instrumental in personalized medicine, disease progression forecasting, and treatment optimization.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Sharafoddini et al. (2017) pioneered work on prediction models based on patient similarity using EHR data, emphasizing the importance of data preprocessing and feature selection in accurately grouping patients with similar conditions. Their study highlighted that patient similarity could enhance diag- nostic precision and contribute to effective treatment planning by uncovering patterns within large datasets. Additionally, Chan et al. (2010) demonstrated the application of machine learning algorithms to calculate patient similarity scores, fo- cusing on optimizing predictive accuracy in medical contexts. Their approach underscored the potential of patient similarity models to support clinical decision-making by providing a systematic way to analyze historical data.

Further advancements in machine learning have enabled more refined models capable of handling complex, multi-dimensional health data. For example, recent studies have leveraged Natural Language Processing (NLP) for symptom analysis, extracting valuable information from unstructured EHR data fields such as patient notes and descriptions of symptoms. These methods have expanded the potential of patient similarity models, making them applicable across var- ious clinical settings. However, despite these advancements, challenges remain, including data quality issues, interpretabil- ity of similarity metrics, and the need for robust evaluation frameworks to ensure that similarity scores translate into meaningful clinical insights.

This study builds upon previous work by applying machine learning to a cancer diagnosis dataset, aiming to refine patient similarity analysis and demonstrate its utility in a practical healthcare context. By focusing on high-stakes cases such as cancer diagnosis, we seek to highlight the value of patient similarity in predicting outcomes and assisting healthcare providers in making well-informed, data-driven decisions.

III. DATASET AND METHODOLOGY

A. DATASET

The dataset used for this study, cancer diagnosis data from kaggle, provides detailed information on cancer patients, including demographic variables, diagnosis information, treat- ment histories, and outcomes. The dataset serves as the foun- dation for patient similarity analysis, allowing us to identify patterns that might indicate similar health trajectories among patients.

Data Overview: The cancer diagnosis dataset contains var- ious fields critical for similarity analysis, such as demographic information, medical records, and symptom descriptions.

Data Preprocessing: To ensure the dataset is suitable for machine learning analysis, several preprocessing steps are nec- essary, including handling missing data, normalization, feature engineering, and text processing for symptom descriptions.

Exploratory Data Analysis (EDA): An initial exploration of the dataset reveals trends such as age distribution, common cancer stages, and outcome patterns, which inform the modeling approach and similarity analysis.

B. PROPOSED METHODOLOGY

The proposed methodology utilizes patient similarity scor- ing to predict health outcomes for cancer patients based on past Electronic Health Record (EHR) data.
Data Acquisition and Preprocessing: Data cleaning, scaling, and NLP processing for symptom descriptions. Feature Selection: Key features selected include demographic variables diagnosis information, and treatment data. Model Develop- ment: The primary model uses classification such as K-nearest neighbors (KNN), and cosine similarity for similarity scoring. Backend Development: A Flask-based backend provides real- time access to similarity scores and predictions. Validation and Testing: The model’s accuracy was validated with cross- validation metrics including accuracy, F1 score, and AUC- ROC curve.

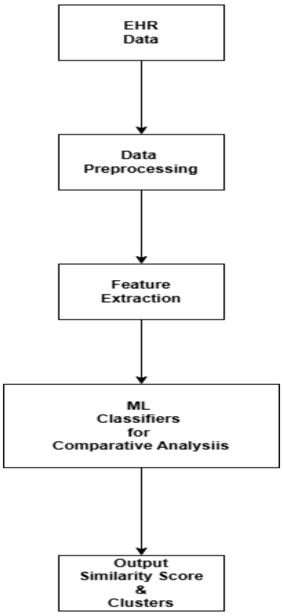

EHR Data
Data Preprocessing
Feature Extraction
ML Classifiers for Comparative Analysiis
Output Similarity Score & Clusters

Fig. 1. Proposed Model

Scaling Techniques: Scaling techniques like Standard Scaler and MinMax Scaler are essential in preparing data for machine learning models, especially those sensitive to feature magnitudes, such as distance-based algorithms (e.g., k-nearest neighbors). Standard Scaler transforms data to have a mean of 0 and a standard deviation of 1.

calculate similarity score: This function calculates the similarity score between two patient vectors using Euclidean distance. By computing the norm between two vectors, it measures how similar or different they are in terms of their feature values. This score forms the basis of comparing patients in terms of their symptoms, history, or other relevant medical data.

generate similarity matrix: This function generates a similarity matrix for a dataset by computing pairwise similarity scores between each pair of patients. The matrix has dimensions n by n, where n is the number of patients, and each entry i, j represents the similarity score between patient i and patient j. To optimize computation, the function calculates each score once, reflecting it in both i, j and j, i.

evaluate similarity accuracy: This function evaluates the accuracy of the similarity matrix by calculating the F1 score, which measures the precision and recall of similarity predictions against true labels. It applies a threshold to classify similarity, converts the continuous similarity matrix into binary predictions, and compares these predictions with the true similarity values provided by ytrue.

classify new patient: This function uses KMeans clustering to classify a new patient based on their feature data. It takes the patient's data, scales it, and assigns it to the nearest cluster using a pre-trained KMeans model. This helps in identifying which patient group or profile the new patient fits best.

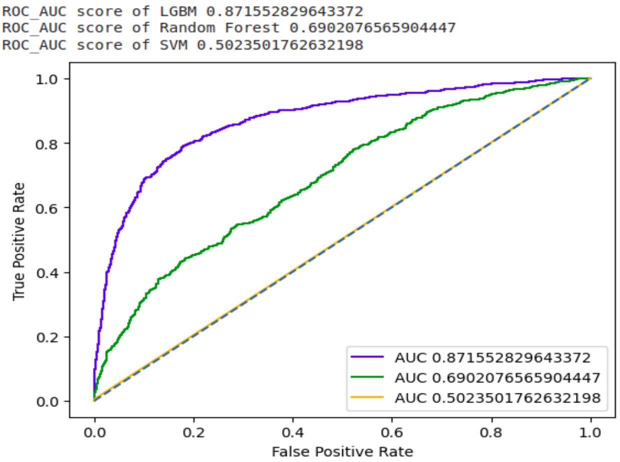
researcher interface: This function simulates a researcher interface where researchers can query the similarity of specific patients to others. For a given patient index, it retrieves the top five similar cases from the similarity matrix, allowing researchers to study patterns or relationships based on similarity scores.

doctor interface: This function simulates a doctor interface that classifies new patients based on their data and retrieves similar cases within the same cluster. It assigns the new patient to a cluster and then finds up to five similar cases within that cluster, helping doctors identify potentially relevant cases for diagnosis or treatment guidance. Additional function for Similarity Matrix: The function then generates and prints the similarity matrices for both training and test data, as well as the similarity RMSE on the training set. The researcher interface and doctor interface functions are also demonstrated with example data, showing the practical application of similarity queries and patient classification.

IV. RESULT AND DISCUSSION

Evaluation Metric	Lightgbm	Random forest	SVM
Accuracy	0.8162	0.6387	0.6147
Precision	0.81	0.81	0.61
Recall	0.69	0.59	0.61
F1-score	0.74	0.77	0.60
ROC-AUC	0.87	0.69	0.50
Execution time(sec)	174	985	51.3

TABLE I EXPERIMENT RESULT TABLE



```
ROC_AUC score of LGBM 0.871552829643372
ROC_AUC score of Random Forest 0.6902076565904447
ROC_AUC score of SVM 0.5023501762632198
```

Fig. 2. ROC-AUC of LGBM, Random Forest, and SVM

TABLE II EXPERIMENT RESULT TABLE

Evaluation Metric	ResNet-50	Basic CNN
Accuracy	0.97	0.94
Loss	0.09	0.13
Validation Accuracy	0.81	0.84
Validation Loss	0.51	0.41

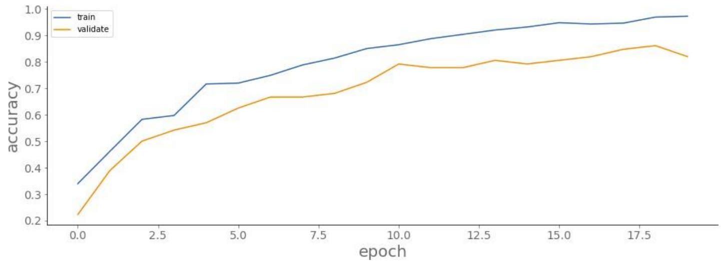


Fig. 3. Resnet-50 Training and Validation Accuracy

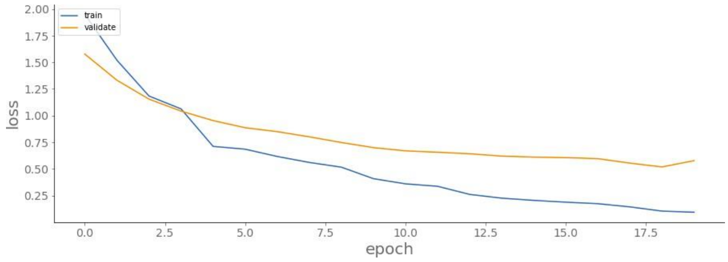


Fig. 4. Resnet-50 Training and Validation Loss

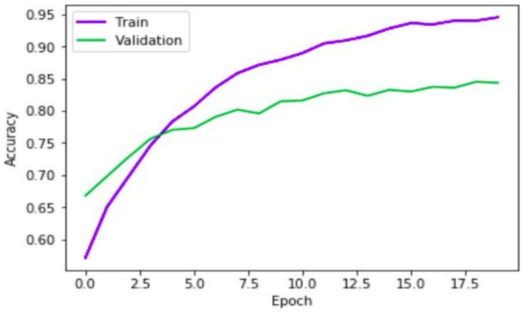


Fig. 5. CNN Training and Validation Accuracy

V.CONCLUSION

This study explored the use of electronic health records (ehr) and machine learning techniques to determine patient case similarity for predictive healthcare analytics. By leveraging structured and unstructured data within ehr systems, the research demonstrated the potential of similarity-based approaches to improve patient care through more personalized treatment plans and accurate risk prediction.

REFERENCES

[1] Heba Mohsen et al, "Classification using Deep Learning Neural Networks for Brain Tumors", Future Computing and Informatics, pp 1- four (2017).

[2] Stefan Bauer et al, "Multiscale Modeling for Image Analysis of Brain Tumor Studies", IEEE Transactions on Biomedical Engineering, fifty nine(1): (2012).

[3] Atiq Islam et al, "Multi-fractal Texture Estimation for Detection and Segmentation of Brain Tumors", IEEE, (2013).

[4] Meiyan Huang et al, "Brain Tumor Segmentation Based on Local Independent Projectionbased Classification", IEEE Transactions on Biomedical Engineering, IEEE, (2013).

[5] AndacHamamci et al, "Tumor-Cut: Segmentation of Brain Tumors on Contrast Enhanced MR Images for Radiosurgery Applications", IEEE Transactions on Medical Imaging, 31(3): (2012).

[6] Bjoern H. Menze et al, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", IEEE Transactions on Medical Imaging, (2014).

[7] Jin Liu et al, "A Survey of MRI-Based Brain Tumor Segmentation Methods", TSINGHUA Science and Technology, 19(6) (2011).

[8] Shamsul Huda et al, "A Hybrid Feature Selection with Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis", IEEE Access, 4: (2017).

[9] R. Karuppathal and V. Palanisamy, "Fuzzy based automatic detection and category technique for MRI-mind tumor", ARPN Journal of Engineering and Applied Sciences, 9(12): (2014).

[10] Janani and P. Meena, "photograph segmentation for tumor detection using fuzzy inference system", International Journal of Computer Science and Mobile Computing, 2(5): 244 – 248 (2013).

[11] Sergio Pereira et al, "Brain Tumor Segmentation the use of Convolutional Neural Networks in MRI Images", IEEE Transactions on Medical Imaging, (2016).
[12] Jiachi Zhang et al, "Brain Tumor Segmentation Based on Refined Fully Convolutional Neural Networks with A Hierarchical Dice Loss", Cornell university library, pc imaginative and prescient and pattern popularity, (2018).
[13] [Radiopaedia] http:// radiopedia.Org.
[14] [BRATS 2015] https://www.Smir.Ch/BRATS/

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details