Advances in Artificial Intelligence and Machine Learning; Research 4 (3) 2648-2664 Received 16-07-2024; Accepted 21-09-2024; Published 28-09-2024

EHR Innovations: Shedding Light on Anemia in the Healthcare Paradigm

Souhardya Das souhardyadas11@gmail.com
Department of Electronics & Tele-communication Engineering, Jadavpur University, Kolkata, India

Proma Mondal promamondal32@gmail.com
Department of Electronics & Tele-communication Engineering, Jadavpur University, Kolkata, India

Shambhab Chaki shambhabc@gmail.com
Department of Electronics & Tele-communication Engineering, Jadavpur University, Kolkata, India

Pratyusha Rakshit pratyushar1@gmail.com
Department of Electronics & Tele-communication Engineering, Jadavpur University, Kolkata, India

Archana Chowdhury chowdhuryarchana@gmail.com
Department of Computer Science and Engineering, Christian College of Engineering and Technology, Bhilai, India

Corresponding Author: Shambhab Chaki

Abstract

This study introduces a novel approach to Electronic Health Record (EHR) analysis, extending the use of phenotyping with machine learning (ML) models to enhance the recognition and treatment of anemia. It first examines the healthcare scenario in India and suggests potential improvements through data-driven personalized care. Using the MIMIC-III dataset, the research involves extensive data preprocessing and analysis to uncover key insights into anemia's prevalence, gender distribution, comorbidities, and Intensive Care Unit (ICU) stays. Partitioning clustering algorithms like K-Means, K-medoids, Fuzzy C-means, and hierarchical clustering algorithms such as Agglomerative Clustering, DIANA, and HDBSCAN are used to identify groups of patients with similar medical profiles. The distance metrics employed are Levenshtein and Euclidean distances combined with TF-IDF Vectorization. The effectiveness of these algorithms is evaluated based on Length of Stay (LoS) estimation, a critical parameter in EHR studies. To predict a new patient's LoS, the patient is at first classified into an existing cluster, which shows the highest support to the patient's clinical activities. A decision tree regressor is then trained using data from the selected cluster to predict the new patient's LoS, significantly improving predictive accuracy and reliability. Notably, the HDBSCAN algorithm, applied to the Tf-Idf Vectorizer object, achieves a 50.82% reduction in Root Mean Squared Error (RMSE) compared to baseline model. The

Citation: Souhardya Das, et al. EHR Innovations: Shedding Light on Anemia in the Healthcare Paradigm. Advances in Artificial Intelligence and Machine Learning. 2024;4(3):154.

novelty of this study lies in proposing an efficient approach for EHR analysis, specifically for predicting ICU patients' LoS, and identifying the most effective clustering algorithm to improve healthcare delivery for anemic patients in healthcare scenario of India.

Keywords: Anemia, Electronic health records, Length of stay, HDBSCAN, Levenshtein distance, Tf-Idf vectorizer.

1. INTRODUCTION

In contemporary healthcare, the significance of comprehensive medical records is paramount. Electronic Health Records (EHR) [1], epitomize the pinnacle of record-keeping advancements, providing exceptional benefits through digitization. EHR systems streamline data management, enhance accuracy, and enable seamless communication among healthcare providers by digitizing records, thereby improving patient care outcomes.

Traditional methods of EHR analysis using machine learning techniques and clustering (FIGURE 1), involve leveraging algorithms to analyze and process patient data. Machine Learning (ML) models are trained on extensive datasets that include medical records, diagnostic tests, treatment plans and patient demographics to identify patterns and relationships within the data. Through this learning process, ML algorithms can predict treatment processes for new patients. By grouping similar patient records together (phenotyping), healthcare providers can streamline EHR analysis, improve data organization, and enhance patient care.

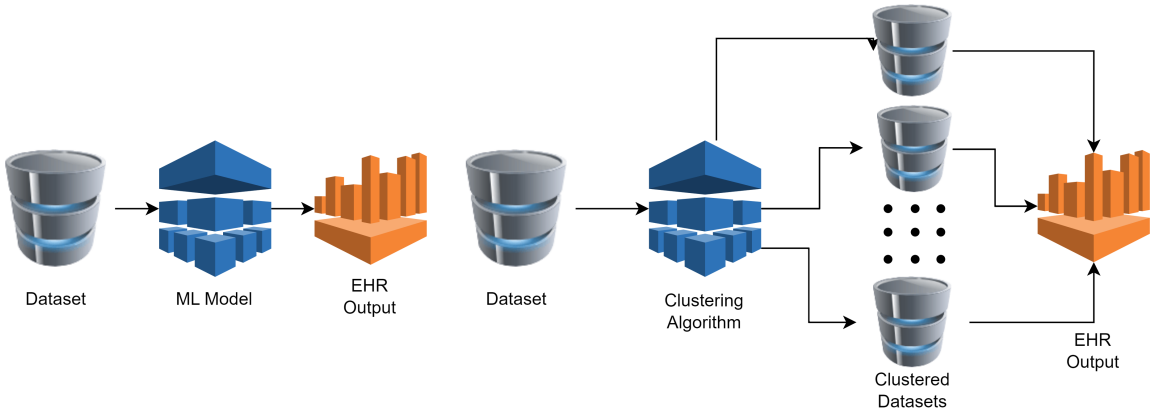


Figure 1: Traditional Methods of EHR analysis

Focusing on anemia and utilizing ML models for EHR improvement highlights a proactive strategy to address underrepresented health concerns. Anemia, a condition characterized by deficiency of red blood cells or hemoglobin, significantly impacts an individual's quality of life. Despite its prevalence, anemia remains undertreated, especially in diverse cultural contexts such as India. A significant reason for the under-treatment of anemia may be the shortage of doctors in India. The doctor-patient ratio in India falls below the WHO-prescribed limit of 1:1000. On average, a government doctor attends to 11,082 people, which is over 10 times the WHO recommendation.

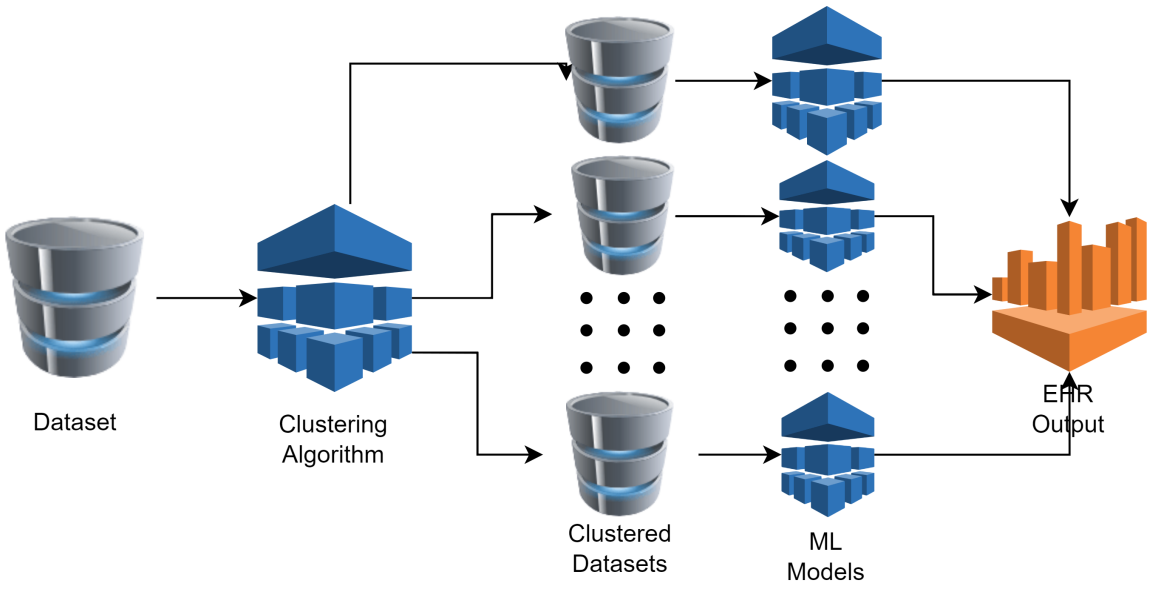


Figure 2: Proposed Method

The shortage of doctors in India contributes significantly to the high rate of misdiagnosis. This scarcity often results in long waiting periods for appointments and rushed consultations, increasing the risk of diagnostic errors. Limited access to healthcare facilities and overburdened healthcare providers lead to symptom oversight and incomplete assessments, compromising diagnostic accuracy. Our study aims to mitigate diagnostic errors in India by leveraging advanced data analytics and machine learning techniques to enhance the accuracy and efficiency of EHR systems. By facilitating more informed diagnostic decisions, we seek to improve patient outcomes and reduce the burden of misdiagnosis in the Indian healthcare landscape.

It is noteworthy that the significant parameter influencing in treatment of patients admitted in hospital with anemia, is to predict the corresponding *length of stay* (LoS). The early prediction of LoS enhance the time-effective utilization of hospital resources and also portrays the possible future cost to incur. Early prediction of LoS and hence the future cost play immense significant role in developing country like India. The paper proposes a novel method to predict the LoS of a newly admitted anemic patient with limited medical records. First, the existing anemic patients' data in the EHR are clustered into different groups. The new patient is assigned to the cluster with patients exhibiting the maximum similarity in clinical activities with new patient. Next, the data of the patients of the selected cluster are used to predict the LoS of the of the new patient (FIGURE 2).

The merit of the paper lies in the following counts:

a. The paper explores different partitioning and hierarchical clustering algorithms to categorize the medical records of the anemic patients in the EHR. The categorization of the patients helps on minutely scrutinizing the patients' conditions to correctly predict the LoS of a new patient.

b. Though there are literatures on clustering patients based on their EHR data, there is no trace of literature providing the guideline on selecting the clustering method. To the author's best

knowledge, this is internationally the first work on exploring different clustering algorithms in the present context including partitioning clustering (K-means [2], K-medoids [3], Fuzzy C-means [4]) and hierarchical clustering (agglomerative [5], DIANA [6], HDBSCAN [7]).

Diverse experiments are undertaken with EHR data of patients suffering from five types of anemia. MIMIC-III database [8], is used for this purpose. Comparative analysis with the state-of-the-art techniques revealed that the HDBSCAN method yields the best result with respect to the root mean squared error during LoS prediction. It substantiates the merit of the present study.

This paper is organized into five sections. Section 2 reviews the existing literature. Section 3 depicts the proposed methodology for clustering the EHR data. Experiments undertaken and the corresponding results are reported in section 4. Section 5 concludes the paper and highlights the potential for personalized healthcare and improved outcomes through data-driven strategies.

2. LITERATURE REVIEW

Anemia is a prevalent medical condition that typically requires invasive blood tests for diagnoses and monitoring. It has the potential to hinder or diminish the blood's ability to effectively transport oxygen throughout the body. The feasibility of using EHR data for the surveillance of anemia, iron deficiency (ID) and ID anemia (IDA) among pregnant women in the first trimester is studied in [9].

Zhi *et al.* in [10], modelled a machine learning algorithm to escalate clinical decision-making processes for hemoglobin level/ anemia degree prediction. The scope of different classification algorithms to identify iron deficiency anemia is studied in [11]. Similarly, in [12], the efficacy of eight machine learning algorithms is studied in the context of prediction of anemia among young girls in Ethiopia.

One of the most significant parameters for timely treatment of anemic patients is length of stay (LoS) at hospital. Predicting LoS for anemic patients in healthcare setting is important from the perspective of: (i) resource allocation, (ii) cost management, (iii) plan for personalized patient care, (iv) improved planning strategy for timely discharge and (v) early identification of high risk. ICU length-of-stay scores, such as APACHE, SAPS, and SOFA are widely used to estimate ICU resource use and predict patient mortality. However, these methods are not designed to address specific diseases [13]. Consequently, there is a need for research into more accurate and reliable prediction systems for ICU resource consumption. Machine learning models based on EHRs can offer more precise and disease-specific assessments compared to traditional methods. Therefore, developing alternative ICU LoS assessment systems is stressed on in [14], as a key method for optimizing resource allocation and better patient care.

Clinical information systems (CISs), including clinical decision support systems (CDSS), leverage electronic health record (EHR) data to enhance healthcare. Recently, there has been a growing interest in developing non-knowledge-based CDSS systems that employ machine learning (ML) and statistical learning techniques. However, the "black-box" nature of deep learning models makes it difficult to interpret the predictions. In this context, it is worth mentioning that the personalized medicine (LDAPPM) approaches are the new trend.

Ma *et al.* in [15], combined the just-in-time learning and extreme machine learning methods to predict LoS for 10 days or more. Su *et al.* [16], compared XGBoost, Random Forest (RF) and Logistic Regression (LR) models for LoS prediction of patients suffering from septicemia. SMOTE was used to address imbalance in dataset. RF came out on top.

Staziaki *et al.* in [17], studied the efficacy of SVM and ANN in case of LoS prediction of ICU patients. They include CT imaging data along with other clinical activities such as vital signs, laboratory test results and patient demographics to classify the patients with respect to long and short LoS in ICU. This study revealed that incorporating more features leads to better prediction of LoS. Alghatani *et al.* [18], used six classifiers to predict shorter or longer LoS using MIMIC-III database. The limitation of the study is, it only considered vital signs as features. Gentmis *et al.* in [19], used 25 features from MIMIC-III database to predict long (> 5 days) or short (< 5 days) LoS including diagnoses and other procedural events. Although they did not consider medications and did not use important model performance metrics like AUC, sensitivity etc. Steele & Thompson in [20], studied seven predictive models for similar task. Bayesian Network outperformed others. Although, they lack the description of data processing and did not provide enough information on which features were selected in the study and how they were selected. In [21], Alsinglawi *et al.* proposed a predictive model combining XGBoost, Random Forest, Logistic Regression and Multilayer Perceptron models for classifying patients in short LoS (< 7 days) and long LoS (> 7 days) groups. They focused on explainability and integration into CIS pipeline. However, the limitation of their method is that they treated the Length of Stay (LoS) prediction as a binary classification task, similar to the studies mentioned above [15–20].

In this paper, authors have explored various clustering methods to cluster patients' records in EHR. The data contained within a cluster, which is the most similar to a new patient's limited medical records, is utilized to train a decision tree regressor. That regressor predicts the LoS of that new patient. The predicted value in our approach is the actual Length of Stay (LoS), unlike existing literature that considers LoS prediction as a binary or multiclass classification task. Inclusion of medication as one of the features, also provides valuable insight into the patient clusters and improve the predictive accuracy, paving the way for more efficient personalized patient care mechanisms.

3. METHODOLOGY

3.1 Clustering Basics: Partitioning Models Used

- **K-means:** The K-means clustering algorithm [2], partitions data into K clusters by iteratively assigning data points to the nearest centroid and updating the cluster centers. Its goal is to minimize the within-cluster sum of squared distances, resulting in compact, homogeneous clusters.
- **K-medoids:** K-medoids [3], measures dissimilarity among data points by taking absolute distance between them. Medoid points are defined as the points that show the least dissimilarity with all other points in a cluster. At every iteration, every other data points other than medoid points are considered as a medoid point and costs are calculated. The cost function is the summation of dissimilarity measure. If cost decreases, the corresponding selection kept intact and the algorithm goes into the next iteration until convergence.

- **Fuzzy C-means:** Fuzzy C-means [4], employs fuzzy membership functions for each cluster. It begins by randomly assigning a fuzzy membership value to each data point, indicating its degree of belonging to each cluster. The cluster centroids are then iteratively updated using an expression that takes the membership values of each data point as input. The fuzzy membership values are adjusted accordingly. This process continues until convergence.

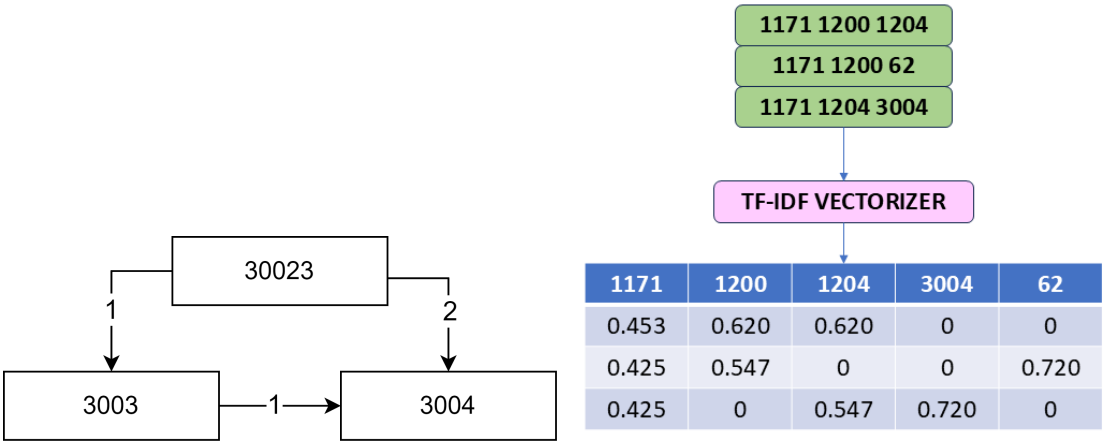
3.2 Clustering Basics: Hierarchical Models Used

- **Agglomerative Clustering:** This algorithm [5], employs bottom-up approach. It first starts with considering each data point as one cluster. Then it proceeds by merging them together. The merger process makes use of specific linkage criterion. Those are as follows: ward, average, single linkage and complete. They try to minimize distances within all clusters, minimize distances between all data points within pairs of clusters, minimize distances between the closest data point within pairs of clusters, maximize average distance between all data points within pairs of clusters respectively.
- **DIANA:** This algorithm [6], is the reverse of agglomerative clustering. It employs top-down approach. It starts with assuming a single cluster containing all data points and then split them up. The splitting process maximizes the dissimilarity within pairs of clusters. It makes use of similar kinds of linkage criterion in reverse manner.
- **HDBSCAN:** HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [7], is a density-based hierarchical clustering. It first computes core distances between pairs of data points utilizing the minimum number of neighborhood points hyperparameter. Then it makes use of mutual reachability as a distance metric, built upon core distances of data points. Then minimum spanning trees are built. Next, trees are pruned according to the minimum cluster size hyperparameter. Finally, the clusters are chosen using 'excess of mass' concept. Its tree-based hierarchical approach offers a flexible framework for clustering analysis, making it particularly useful for exploring high-dimensional datasets and uncovering hidden data structures.

3.3 Distance Metrics Used

The length of patient data in EHR datasets varies, as patients visit healthcare facilities for different numbers of days. Each day, they undergo various clinical activities. Therefore, distance metrics are carefully chosen to capture the dissimilarities in these varying lengths and patterns.

- **Levenshtein Distance:** The Levenshtein distance metric [22], measures string similarity by quantifying the minimum number of single-character edits needed to transform one string into another. By using this edit distance, it effectively groups text data into clusters with similar patterns or characteristics, as illustrated in FIGURE 3a.
- **TF-IDF Vectorizer clubbed with Euclidean Distance:** The TF-IDF vectorizer metric [23], uses the Term Frequency-Inverse Document Frequency (TF-IDF) method to convert text data into numerical representations, highlighting important terms while minimizing the impact of common ones (FIGURE 3b). After getting numerical representations of same dimensionality



(a) Levenshtein Distance (b) TF-IDF Vectorizer Object

Figure 3: (a) Levenshtein Distance, (b) TF-IDF Vectorizer Object.

using TF-IDF vectorizer, which is in this paper referred to as TF-IDF vectorizer objects in further discussions, Euclidean distance can be applied to measure the distance between two such objects.

3.4 Proposed Models

The model architectures employing various transformation processes and clustering models are described in this section. For each patient, the diagnosis sequence string (in ICD9 format) and medication sequence string are concatenated. The following models are deployed on these concatenated strings to assign cluster labels to patients: a) K-means with Vectorizer object, b) K-means with Levenshtein distance, c) K-medoids with Vectorizer object, d) K-medoids with Levenshtein distance, e) Fuzzy C-means with Vectorizer object, f) Fuzzy C-means with Levenshtein distance, g) Agglomerative clustering with Vectorizer object, h) Agglomerative clustering with Levenshtein distance, i) DIANA with Vectorizer object, j) DIANA with Levenshtein distance, k) HDBSCAN with Vectorizer object, l) HDBSCAN with Levenshtein distance.

3.5 Classification of A New Patient with Limited Entries in EHR

Let us assume the new patient has an incomplete record of length N days. The Levenshtein distance between the new patient's record and those of all patients within the current cluster is calculated. The support value of a cluster is determined by computing the mean of these distances. This process is repeated for all other clusters. The new patient is then assigned to the cluster with the highest support value, ensuring optimal classification accuracy within the dataset. Following this step, the length of stay (LoS) is predicted using the regression model described next.

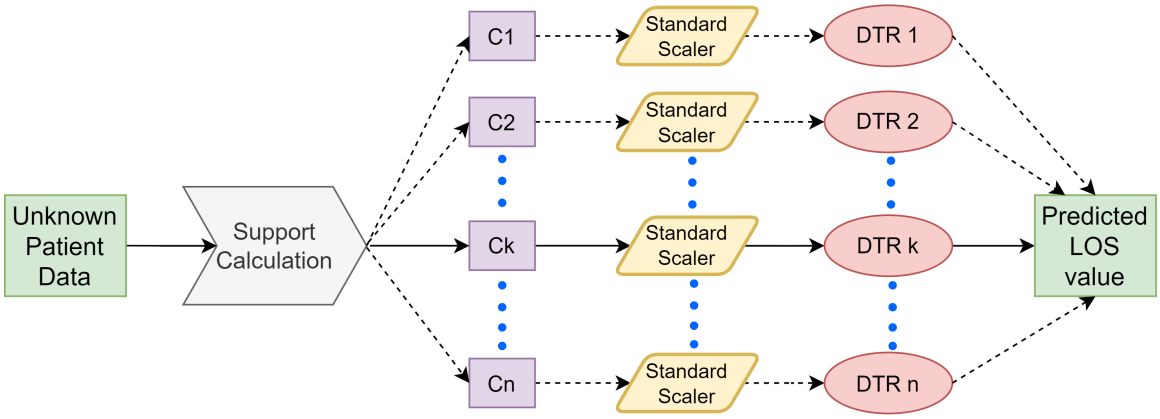


Figure 4: LoS Estimation

3.6 LoS Estimation

This evaluation process, illustrated in FIGURE 4, involves deploying a regression model once the patient is classified into cluster k. For each sub-dataset in a clustering model, a different instance of the same regressor model is trained, and the length of stay (LoS) for the test patients within chosen sub-dataset is predicted.

Pre-processing: During the pre-processing stage, the gender attribute undergoes one-hot encoding. Additionally, the concatenated diagnoses lists and medication lists are subjected to multilabel binarization. This process transforms the lists into binary arrays. Furthermore, a train-test split is executed with a ratio of 0.2. Finally, standard scaling is applied, ensuring that each feature contributes equally to the analysis.

Regression Model: A Decision Tree Regressor [24], is used in the regression process for all datasets. In a Decision Tree Regressor, the goal is to create a model that predicts the target value by learning simple decision rules inferred from the data features. The model splits the data into subsets based on the value of input features, forming a tree structure with nodes and branches.

At each node, the data is divided into two or more sub-nodes based on a feature that results in the best split, which is typically determined by minimizing the mean squared error (MSE). The splitting process continues recursively until a stopping criterion is met, such as a maximum depth or minimum number of samples per leaf node. The final prediction is made by averaging the target values of the data points in each leaf node.

Decision Tree Regressors are intuitive and easy to visualize, making them a popular choice for regression tasks in various domains. In a Decision Tree Regressor, the regression equation is not a single formula but rather a set of decision rules that define how the input space is partitioned. However, the basic idea can be summarized by following piecewise constant function:

$$f(x) = \sum_{\mathrm{m}=1}^{\mathrm{M}} c_m \cdot I(x \in R_m) \tag{1}$$

where M is the number of leaf nodes, c_m is the constant prediction value of the m-th leaf, R_m is the region (set of input feature conditions) corresponding to the m-th leaf, and $I(\cdot)$ is an indicator function that returns 1 if x belongs to R_m and 0 otherwise.

Performance Metric: Root Mean Squared Error (RMSE) is used as the performance metric for evaluating our algorithms. RMSE calculates the square root of the average squared differences between the predicted values and the actual values. The formula for RMSE is:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{2}$$

Here, n is the number of patients in test data, y_i is the actual value of LoS of patient i, and $\widehat{y_i}$ is the predicted value of LoS of patient i.

4. RESULTS AND DISCUSSION

4.1 Database

The MIMIC-III (Medical Information Mart for Intensive Care III) dataset [8] is a crucial resource in the field of medical research. It contains a vast collection of de-identified health records from over 40,000 critically ill ICU patients admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012.

Key Files in the Database Used for Evaluation:

Patients.csv: This excel sheet contains key demographic information and characteristics of patients, providing a foundational understanding of the study population.

Admissions.csv: This excel sheet includes detailed information about patient admissions, such as admission type, source, and discharge disposition, which are crucial for tracking patient journeys through the healthcare system.

ICUstays.csv: Focusing on critical care, this excel sheet provides insights into ICU stays, including admission and discharge times, ICU type, and duration of stay, allowing for analysis of intensive care utilization patterns.

Prescriptions.csv: This excel sheet examines prescription data to reveal medication usage trends, drug classes, and dosage frequencies, offering insights into treatment protocols and pharmaceutical interventions.

Diagnoses_ICD.csv: This excel sheet contains diagnoses records encoded with ICD-9 codes, enabling a detailed understanding of prevalent medical conditions, comorbidities, and disease trajectories among patients.

D_ICD_Diagnoses.csv: This excel sheet is the dictionary of ICD-9 codes present in the database. It is used to correlate ICD-9 codes present in the *Diagnoses_ICD* TABLE with their corresponding medical descriptions.

ICD-9 is the abbreviated form of International Classification for Diseases, 9^{th} Revision. It usually consists of five characters. The first three characters of these codes stand for disease types. The fourth and fifth ones specify the disease further.

By detailing our use of these datasets, we emphasize the methodological rigor of our study, ensuring both transparency and reproducibility in our research approach.

4.2 Data Pre-processing

The dataset for clustering patients diagnosed with anemia is sourced from the MIMIC III database tables mentioned above. Total patient cohort for anemia comprises records of 13573 patients, obtained after cleaning the dataset, includes the following types of anemia:

(i) 1962 patients are diagnosed with 'iron deficiency anemias', ICD-9 indicator: '280'.

(ii) 279 patients are diagnosed with 'other deficiency anemias', ICD-9 indicator: '281'.

(iii) 16 patients are diagnosed with 'hereditary hemolytic anemias', ICD-9 indicator: '282'.

(iv) 132 patients are diagnosed with 'acquired hemolytic anemias', ICD-9 indicator: '283'.

(v) 68 patients are diagnosed with 'aplastic anemias', ICD-9 indicator: '284'.

(vi) 10631 patients are diagnosed with 'other and unspecified anemias', ICD-9 indicator: '285'.

(vii) 70 patients are diagnosed with 'anemia complicating pregnancy, childbirth or the puerperium', ICD-9 indicator: '6482'.

(viii) 19 patients are diagnosed with 'congenital anemia', ICD-9 indicator: '7765'.

(ix) 396 patients are diagnosed with 'anemia of prematurity' ICD-9 indicator: '7766'.

We have selected only those types of anemia for which the patient cohort consists of at least 100 patients. This criterion leaves us with five different types of anemia, represented by the following ICD-9 codes:

(a) '280' for 'iron deficiency anemias,' with 1,962 patient records

(b) '281' for 'other deficiency anemias,' with 279 patient records

(c) '283' for 'acquired hemolytic anemias,' with 132 patient records

(d) '285' for 'other and unspecified anemias,' with 10,631 patient records

(e) '7766' for 'anemia of prematurity,' with 396 patient records.

The subject ID and gender attributes are taken from the *patients* tables, with the subject ID serving as the primary key of the database. The age attribute is calculated from the subject's birth year and the year of first admission, as recorded in the *admissions* table. Similarly, LoS, the list of prescribed

medications, and the list of diagnoses in ICD-9 code format, are combined from the *ICUstays*, *Prescriptions*, and *Diagnoses_ICD* tables in MIMIC III. The medication and diagnoses lists are maintained sequentially during the patient's treatment.

4.3 Data Analysis of the Whole Patient Cohort

The age distribution graph in FIGURE 5, illustrates the distribution of ages among patients. The histogram shows a noticeable right-skewness, indicating a higher prevalence of anemia among middle-aged individuals. Additionally, the smooth blue curve overlaid on the histogram represents the scaled Kernel Density Estimation (KDE) distribution pattern, estimating the probability density function (PDF) of the age demographics. These observations underscore the importance of age as a key demographic variable in clinical research and practice.

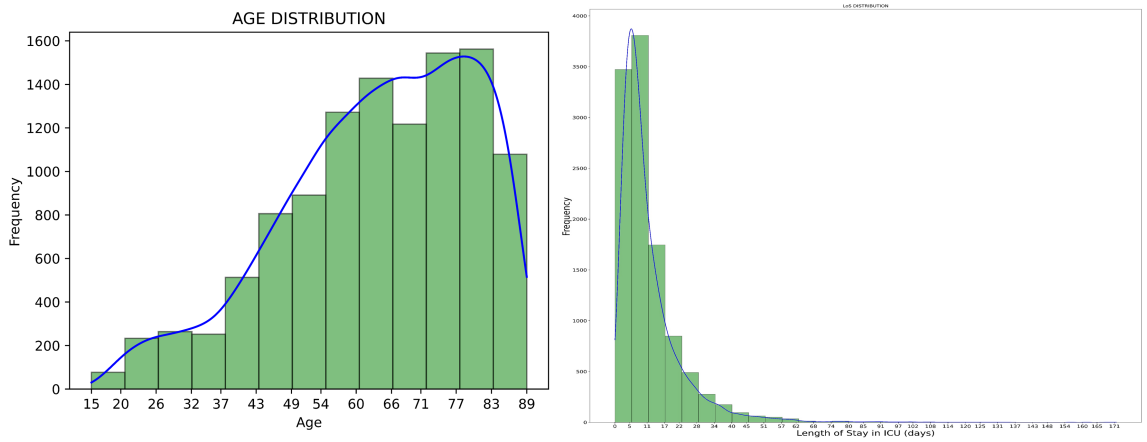


Figure 5: Age distribution and LoS of Patients in Training Dataset

In the dataset, dysthymia is found to be more prevalent in females, with a rate of 56.66%, compared to males, who had a prevalence of approximately 43.34%. This finding is supported by previous international studies. The underlying reasons for this gender difference in dysthymia are multifaceted and may include hormonal, social, and genetic factors. The LoS distribution in FIGURE 5, shows the frequency of ICU stays across different durations. The right-skewness indicates that most ICU stays are concentrated within shorter durations, with a notable peak at approximately 1 to 4 days.

FIGURE 6 highlights the most prevalent comorbidities among the patients, with hypertension as the primary comorbidity, affecting 45% of patients, followed by congestive heart failure at 43% and acute kidney failure at 36%. These findings are consistent with prior research, which has identified hypertension as a common comorbidity with anemic ICU patients. Research on anemia has increasingly explored its association with other comorbid conditions, including hyperlipidemia and respiratory problems.

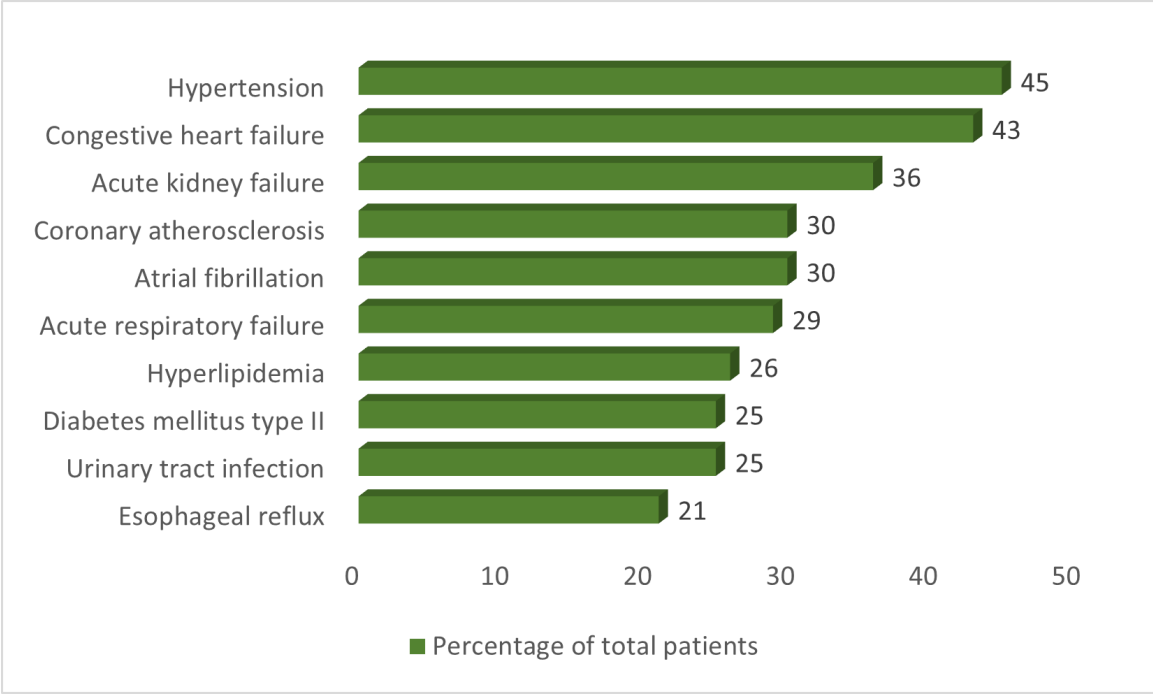


Figure 6: Most Prevalent Comorbidities

4.4 Cluster Analysis

This subsection discusses about the clusters of patients obtained by employing different clustering algorithms and transformers for different sub-classes of anemic patients. CID refers to cluster-ID. Each cluster is characterized by distinct patient feature patterns, offering insights into subgroups within the anemic population. The analysis underscores the heterogeneity of anemia and the potential implications of the proposed method in personalized treatment approaches. We here provide insights into the analysis of one type of anemia, the 'iron deficiency anemia'. The patients diagnosed with this particular type are identified with '280' as the first three digit of ICD9-code.

4.4.1 Patients diagnosed with 'iron deficiency anemias'

This group of patients are identified with ICD-9 codes starting with '280'. The pi-charts in FIGURE 7, show how many patients are in different groups identified by the corresponding clustering algorithm and transformer model in terms of the percentage of total training population. TABLES 1-4 illustrate the most common comorbidity (diagnoses other than 'iron deficiency anemia' and other types of anemia) and the most common medication used for a particular cluster of patients identified by HDBSCAN-Vectorizer, HDBSCAN-Levenshtein, Fuzzy C Means-Vectorizer and Fuzzy C Means-Levenshtein models. It is worth mentioning that Fuzzy C-Means has consistently outperformed other partitioning algorithms and HDBSCAN has been the most successful in LoS prediction task. These findings correspond to the fact that different clusters correspond to different clinical

activities and training ML models on top of these data allows in improving efficiency of personalized care.

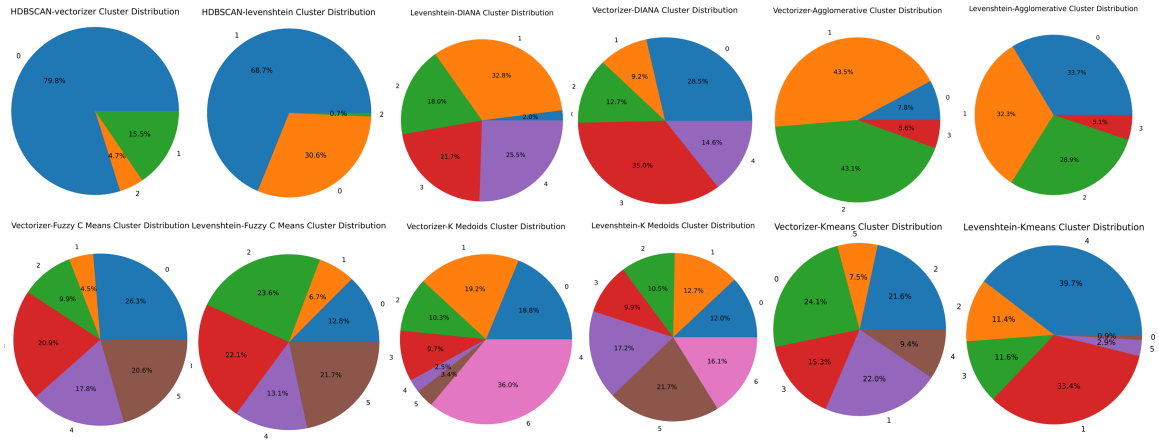


Figure 7: Patients, diagnosed with 'iron deficiency anemias' and associated with the corresponding CID in terms of percentage of training population identified by different clustering algorithms and transformer models

Table 1: HDBSCAN-Vectorizer clusters analysis

CID	Comorbidity	Medication
0	Hypertension	Sodium Chloride 0.9% Flush
1	Coronary atherosclerosis	Acetaminophen
2	Diabetes with ketoacidosis, type I	Sodium Chloride injection 0.9% 1000 ml

Table 2: HDBSCAN-Levenshtein clusters analysis

CID	Comorbidity	Medication
0	Hypertension	Sodium Chloride 0.9% Flush
1	Acute kidney failure	Acetaminophen
2	Diabetic retinopathy	Insulin

Table 3: Fuzzy C Means-Vectorizer clusters analysis

CID	Comorbidity	Medication
0	Acute respiratory failure	Heparin
1	Hypertension	Docusate Sodium
2	Congestive heart failure	Sodium Chloride 0.9% Flush
3	Esophageal reflux	Sodium Chloride injection 0.9% 1000 ml
4	Coronary atherosclerosis	Metoclopramide
5	Acute kidney failure	Acetaminophen

Table 4: Fuzzy C Means-Levenshtein clusters analysis

CID	Comorbidity	Medication
0	Hypertension	Sodium Chloride 0.9% Flush
1	Acute respiratory failure	Iso-Osmotic Dextrose
2	Acute kidney failure	Acetaminophen
3	Atrial fibrillation	Senna
4	Congestive heart failure	Heparin

4.5 Comparative Analysis with State-of-the-Art Techniques

To demonstrate the efficacy of the proposed model, it is compared with four state-of-the-art algorithms in the LoS prediction setup. They are: (i) the SVM model proposed by Staziaki *et al.* [17], (ii) the XGBoost model, as proposed by Alghatani *et al.* [18], (iii) the Bayesian Network model proposed by Steele & Thompson [20], (iv) the G+C+L+M model proposed by Alsinglawi *et al.* [21]. For all contender algorithms, the evaluated RMSE values, averaged over the five different LoS prediction tasks for five groups of anemic patients as discussed in the previous sub-section, are reported in TABLE 5. The 'Reduction' column in these tables stands for the reduction in RMSE with respect to baseline Decision Tree Regressor model.

Table 5: Comparative Analysis

Algorithms			RMSE	Reduction
Decision Tree Regressor (Baseline Model)			**3.361**	
	Transformer	**Clustering Algorithm**		
Proposed Method	Vectorizer	HDBSCAN	**1.653**	**50.82%**
	Levenshtein	HDBSCAN	1.758	47.69%
	Vectorizer	DIANA	2.111	37.19%
	Levenshtein	DIANA	2.078	38.17%
	Vectorizer	Agglomerative	2.157	35.82%
	Levenshtein	Agglomerative	2.143	36.24%
	Vectorizer	Fuzzy C-means	2.139	36.36%
	Levenshtein	Fuzzy C-means	2.172	35.38%
	Vectorizer	K-medoids	2.345	30.23%
	Levenshtein	K-medoids	2.491	25.89%
	Vectorizer	K-means	2.666	20.68%
	Levenshtein	K-means	2.622	22.05%
SVM by Staziaki *et al.* [17]			2.457	26.90%
XGBoost by Alghatani *et al.* [18]			2.365	29.63%
Bayesian Network by Steele & Thompson [20]			2.031	39.57%
G+C+L+M by Asinglawi *et al.* [21]			1.980	41.09%

TABLE 5 reveals that the realization of the proposed algorithm with Vectorizer & HDBSCAN achieves the best rank by providing the minimum RMSE. Substitution of Vectorizer with Levenshtein distance, keeping the clustering algorithm same as HDBSCAN yields the second-best result.

However, it is worth mentioning that the remaining models proposed in this paper is outperformed by the G+C+L+M model. The results reported in these tables also revealed that HDBSCAN is outperforming other algorithms consistently for all five groups of patients and Fuzzy C-means consistently outperformed other partitioning methods.

5. SUMMARY AND CONCLUSION

This paper proposed a novel approach towards analyzing EHRs. The novelty lies in the following counts: (a) The paper explores different partitioning and hierarchical clustering algorithms to categorize the medical records of the anemic patients in the EHR. The categorization of the patients helps on minutely scrutinizing the patients' conditions to correctly predict the LoS of a new patient. (b) Though there are literatures on clustering patients based on their EHR data, there is no trace of literature providing the guideline on selecting the clustering method. To the author's best knowledge, this is internationally the first work on exploring different clustering algorithms in the present context including six clustering methods. (c) The EHR data is effectively handled using Levenshtein distance and Tf-Idf Vectorizer object. (d) Deep insights into clusters are provided to demonstrate the need of personalized care of anemic patients.

In future research, we plan to explore a variety of other features including radiology image reports, clinical procedures, clinical texts etc. Additionally, we aim to address the undertreatment of anemia in India by leveraging advanced data analytics and machine learning techniques on both international and Indian datasets.

References

[1] Birkhead GS, Klompas M, Shah NR. Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health. Annu Rev Public Health. 2015;36:345-359.

[2] Hamerly G, Elkan C. Learning The K in K-Means. Adv Neural Inf Process Syst. 2003;16281-16288.

[3] Arora P, Deepali S, Varshney S. Analysis of K-Means and K-Medoids Algorithm for Big Data. Procedia Comput Sci. 2016;78:507-512.

[4] Suganya R, Shanthi R. Fuzzy C-Means Algorithm-a Review. Int J Sci Res Publ. 2012;2:1.

[5] Murtagh F, Contreras P. Algorithms for Hierarchical Clustering: An Overview. Wiley Interdiscip Rev Data Min Knowl Discov. 2012;2:86-97.

[6] Giordani P, Ferraro MB, Martella F, Giordani P, Ferraro MB, et. al. Hierarchical Clustering. In: An Introduction to Clustering With R. 2020:9-73.

[7] Campello RJ, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pacific-Asia conference on knowledge discovery and data mining. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013:160-172.

[8] https://physionet.org/content/mimiciii/1.4/

[9] Sharma AJ, Ford ND, Bulkley JE, Jenkins LM, Vesco KK, et. al. Use of the Electronic Health Record to Assess Prevalence of Anemia and Iron Deficiency in Pregnancy. J Nutr. 2021;151:3588-3595.

[10] Zhi Z, Elbadawi M, Daneshmend A, Orlu M, Basit A, et al. 2024. Hgbnet: Predicting Hemoglobin Level/Anemia Degree From EHR Data. arXiv preprint arXiv:2401.12002.

[11] Hafiz A, Rai S. Anemia Prediction Using Machine Learning Algorithms. In: Rajagopal S, Popat K, Meva D, Bajeja S, editors Advancements in Smart Computing and Information Security. ASCIS 2023. Commun Comput Inf Sci. Springer; Cham . 2024:2037 .

[12] Zemariam AB, Yimer A, Abebe GK, Wondie WT, Abate BB, et al. Employing Supervised Machine Learning Algorithms for Classification and Prediction of Anemia Among Youth Girls in Ethiopia. Sci Rep. 2024;14:9080.

[13] Li C, Chen L, Feng J, Wu D, Wang Z, et al. Prediction of Length of Stay on the Intensive Care Unit Based on Least Absolute Shrinkage and Selection Operator. IEEE Access. 2019;7:110710-110721.

[14] Levesque E, Hoti E, Azoulay D, Ichai P, Samuel D, et. al. The Implementation of an Intensive Care Information System Allows Shortening the ICU Length of Stay. J Clin Monit Comput. 2015;29:263-269.

[15] Ma X, Si Y, Wang Z, Wang Y. Length of Stay Prediction for ICU Patients Using Individualized Single Classification Algorithm. Comput Methods Programs Biomed. 2020;186:105224.

[16] Su L, Xu Z, Chang F, Ma Y, Liu S, 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. Front Med (Lausanne). 2021;8:664966.

[17] Staziaki PV, Wu D, Rayan JC, Santo ID, Nan F, et. al. Machine Learning Combining CT Findings and Clinical Parameters Improves Prediction of Length of Stay and ICU Admission in Torso Trauma. Eur Radiol. 2021;31:5434-5441.

[18] Alghatani K, Ammar N, Rezgui A, Shaban-Nejad A. Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation. JMIR Med Inform. 2021;9:21347.

[19] Gentimis T. Ala'J, A., Durante, A., Cook K, Steele R. Predicting Hospital Length of Stay Using Neural Networks on Mimic III Data. In: 2017 IEEE 15th international conference on dependable, autonomic and secure computing, 15th international conference on pervasive intelligence and computing, 3rd international conference on big data intelligence and computing and cyber science and technology congress (DASC/PiCom/DataCom/CyberSciTech). IEEE PUBLICATIONS; 2017:1194-1201.

[20] Steele RJ, Thompson B. Data Mining for Generalizable Pre-admission Prediction of Elective Length of Stay. In: 9th Annual Computing and Communication Workshop and Conference (CCWC). IEEE PUBLICATIONS. 2019;2019:127-133.

[21] Alsinglawi BS, Alnajjar F, Alorjani MS, Al-Shari OM, Munoz MN, et. al. Predicting Hospital Stay Length Using Explainable Machine Learning. IEEE Access. 2024;12:90571-90585.

[22] Ristad ES, Yianilos PN. Learning String-Edit Distance. IEEE Trans Pattern Anal Mach Intell. 1998;20:522-532.

[23] Robertson S. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. J Doc. 2004;60:503-520.

[24] Belson WA. Matching and Prediction on the Principle of Biological Classification. J R Stat Soc C. 1959;8:65-75.