An Interactive Web Solution for Electronic Health Records Segmentation and Prediction

Sudeep Mathew[1], Mithun Dolthody Jayaprakash[2] and Rashmi Agarwal[3]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com

Abstract. A vast variety of patient data has been collected and monitored through Electronic Health Records (EHR) using various tools in the clinical research industry and it is a concern for healthcare providers to ensure the safety of the patients who are participating in the clinical trials. It is evident that need for a centralized analytics solutions for EHR datasets that deliver insights and predictability.

The paper focuses on the healthcare industry, which can benefit immensely by allowing medical practitioners to gain insights into the EHR data. The paper aims to provide a platform to explore and gain descriptive statistics and to provide patient segmentation and recommendation.

The objective of the paper is to start data acquisition and data understanding and then create a web interface for data exploration and segmentation and classification. In the data modeling phase, the objective is to create machine learning models for segmentation and classification.

The first step is data acquisition from the *MIMIC-III v1.4* (Clinical database) data mart. In the data understanding phase, the relationship of multiple tables is evaluated. In the data wrangling phase, SQL and Python are used to combine different tables to create a single dataset for analyzing the data and modeling the data. The combined dataset is then used for k-means clustering techniques for obtaining chest heart failure patients clusters. In the following phase, the diagnosis text data is extracted from the diagnosis dataset and performed text cleaning by removing punctuation, numbers, and stopwords. The cleaned text data is used for data modeling and for that TFIDF (Term Frequency Inverse Document Frequency) vectors and count vectors are created and then multiple classification techniques are applied for predicting the occurrences of death and the best model is considered for the model deployment.

In the model evaluation phase, it is observed that six clusters were optimal while training the model and it is incorporated into the application for predicting the segments of the patients based on the risk levels. Few machine learning models were trained on patient's historic diagnosis text data and the logistic regression model indicated 89 % of AUC score in test data and is deployed into the application for the prediction.

Finally, a web interface is created using the python *streamlit* framework which allows the users to bring raw EHR datasets to explore the data. The created models for segmentation and classification are deployed with the web application and thus will provide a recommendation to the business.

Keywords: Natural Language Processing, EHR, Segmentation, Serious Adverse Event Prediction

Keywords: Natural Language Processing, EHR, Segmentation, Serious Adverse Event Prediction

1 Introduction

Electronic health records (EHRs) contain patient diagnostic records, physician records, and records of hospital departments. For heart diseases, we can receive huge unstructured data from EHR time series. By analyzing and mining, we can identify the links between diagnostic events and ultimately predict the probability of the occurrence of a serious adverse event. The adoption of EHR datasets and the increase of digitized information about patient data revolutionize the emergence of clinical research in oncology research [1]. One of the applications of EHR data is an improvising learning system for clinical research and which helps in various applications of patient selection, dosing, drug target, etc. as discussed [2]. Standardizing electronic health records in the Indian health record system is implemented by [3]. The comprehensive techniques for modeling EHR data are provided by [4].

The web app is developed by aiming to help the medical or clinical team to monitor the safety of the patients during the clinical trials by allowing the users to explore the data through visualization and statistics, clustering, and the probability of the occurrence of SAE.

Clustering or segmentation techniques are helpful to find out the underlying hidden association between each data point. This helps the business to decide on each cluster. Detection anomalies or outliers in EHR data was implemented by [5]. The primary objective is to apply unsupervised clustering techniques to EHR data the result indicated that clustering techniques produced high sensitivity and specificity.

The occurrence of serious adverse events or SAE is one of the primary concerns that pharmaceutical companies face during the clinical trial. This application intended to solve this problem by providing the probability of the occurrence of SAE by analyzing patient's diagnosis data. In one of the works, a prediction model was implemented to detect the occurrence of an adverse event such as cardiac arrest by utilizing patient data [6].

The paper primarily focuses on the clinical research industries team which helps to improve overall the patient's safety concerns and address key issues during the clinical trial.

2 Background

2.1 Application of ML on HER Data

One of the works by Ziyi was referring to the challenges and the perspectives of machine learning multimodal in electronic health records. And this work suggests that including structured data is not enough to achieve a good result instead this study seeks to use machine learning and deep learning models on structured and unstructured EHR datasets [7]. The machine learning models in electronic health records could outperform conventional survival models for predicting mortality in coronary disease. These works include multiple machine learning models such as the cox model, and random forest in the 80000 patients EHR dataset and the output indicated that it outperforms conventional models [8]. Adeler proposed a risk prediction model for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. This work was to combine the time series model and cox proportional model to get a range of risk prediction models and the evaluation of the model is based on discriminatory statistics [9]. Predicting the risk of heart failure with EHR sequential data modeling is developed and used patient's diagnosis data with the LSTM sequential model used and the evaluation is based on the utility and efficacy of the proposed solution [10].

2.2 Unsupervised Techniques on HER Data

Lutz Has performed an unsupervised machine learning model to detect patient subgroups in electronic health records. This project used agglomerative hierarchical clustering and *k-means* clustering on patient's lab and coded datasets. The results indicated that natural grouping is present in the dataset and hierarchical clustering provides higher quality clusters than *k-means* clustering [11]. One of the works implemented by Gabriele is for private hospital workflow optimization through *k-means* clustering. This work is to optimize work allocation-based staff members, patients, hospitals, and locations to cluster the staff members utilizing the frequency of the facing time [12]. *k-means* clustering is used for healthcare knowledge discovery is one of the works and the objective is used to discover hidden patterns by applying *k-means* clustering and a self-organizing map (MAP)[13]. One of the works is an unsupervised machine learning model used for the discovery of latent disease clusters using electronic health records [14]. Prediction of health outcomes for pediatric patients is one of the works implemented and the approach of the project was to implement a Bayesian model and clustering for predicting the risk of type 2 diabetics for children between the ages of 10 and 14 [15]. Another work implemented using an unsupervised LDA approach to cluster patient subgroups into multiple clusters using patient's health records [16].

2.3 Text Analytics on HER Data

Natural language processing is used in the field of unstructured text data and this work explores the possibilities of applying NLP models in EHR datasets. The author discussed various applications of NLP such as classification models, question answering, phenotyping, knowledge graphs, medical dialogue, etc. [17]. A comparative analysis of text classification approaches in electronic health records indicated that text classification in traditional approaches could exceed the performance of contextual embedding models such as BERT [18]. A work mentioning the application of deep learning models in an electronic health record by developing various deep neural network models. The proposed solution is a deep learning model for predicting the health risk of the patient [19]. On other hand, a work implemented by Mascio suggests different text classification approaches to patient text data [20]. This work focused on various traditional machine learning models and compared them with contextual embedding BERT and identified that traditional models performed well on the text data. Bittar tried to implement suicide risk assessment using text data and the work implemented to predict the tendency to commit suicide by extracting text features from the clinical notes and the data trained SVM model [21].

Various machine learning techniques on electronic health records were discussed above on segmentation as well as text classification. This paper seeks to predictability of death of the patient based on the historic text data as well segmentation of HER data.

3 Methodology

In contrast to the above-mentioned methods, we develop a *k-means* clustering technique to group Congestive Heart Failure (CHF) patients based on risk levels. We also developed a Serious Adverse Event (SAE) prediction, model for predicting the probability of death using the text classification technique. We present the details of our approach in the following session.

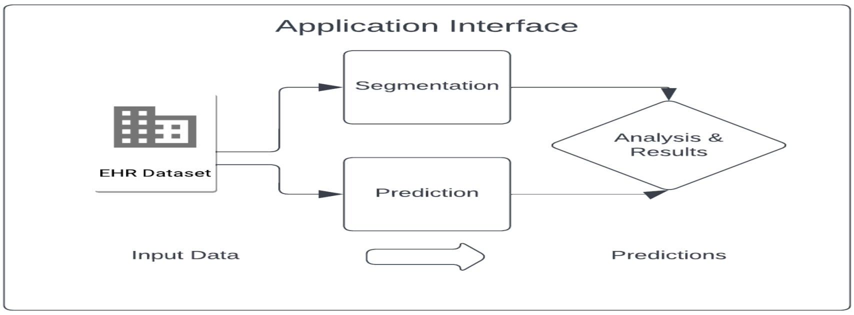

Fig. 1. Application Interface

In Fig. 1 the details of the application are represented. In the following sessions, the details of the application describing.

3.1 K-means Clustering on CHF Data

The input data for the clustering or segmentation technique is combined data of patient demographic, admission, diagnosis, and drug administration. The following flowchart in Fig. 1. depicts the proposed methods for the *k-means* clustering technique.

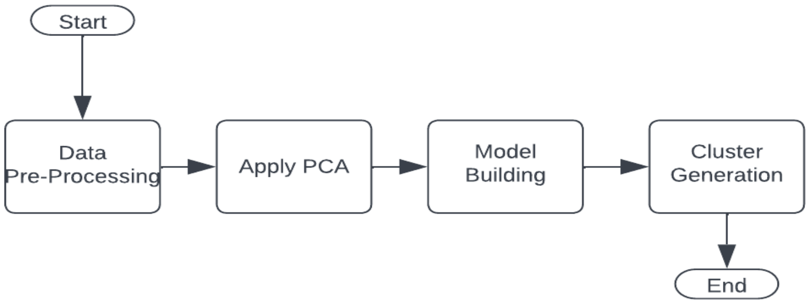

Fig. 2. Methodology for K-means Clustering

The input features containing patient's demographic, admission, prescription, and diagnosis datasets were joined and transferred to Principal Component Analysis (PCA) model for dimensionality reduction, and afterward, the principal components are transferred to the *k-means* model for the segmentation of the patients. This model helps the business to identify the risk levels of the patients to take a valid decision for each cluster.

Data Pre-Processing and Feature Engineering. The output of the data wrangling created a single data frame for the data analysis and data modeling. The primary diagnosis of the patients with this CHF disease formed a data frame in the previous data wrangling steps. To perform segmentation, data scaling and missing records removal were the two methods performed before the data modeling phase. Data scaling is the approach to normalizing the range of independent features of the dataset. And finally, all the missing values are removed from the data frame by the missing value removal method in python. Table No.1 lists all the features that were generated.

Table 1. Features Of Segmentation

No	Name	Description
1	Count of diagnosis	Aggregated diagnosis count for each patient
2	Drug administrated days	Aggregated total days of drugs given for each patient
3	No of drugs	Total number of drugs given to each patient
4	Age Group	Grouped patients based on age such as >90 as very old, 60 -80 as a senior citizen, >18 as an adult, <18 as young
5	Ethnic Group	All the non-white people grouped as other as others as white
6	Is hyper	Whether hypertension present for patient 0 as not present and 1 as present
7	Is kidney	Whether kidney diseases present for patient 0 as not present and 1 as present
8	Is diabetic	Whether diabetic present for patient 0 as not present and 1 as present
9	Is resp	Whether respiratory diseases present for patient 0 as not present and 1 as present

Applying PCA Technique. Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a set of in a dataset into a smaller number of features called *principal components* while at the same time trying to retain as much information in the original dataset as possible. Preprocessed data was transferred to the PCA model and the output was four components.

Model Building. The *k-means* algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known as the clustering or segmentation technique. Below are the stages of clustering.

1. First provided the number of the cluster that we need to generate from the algorithm
2. Next, choose K data points at random and assign to each cluster
3. The cluster centroid is computed
4. Iterate the below steps until the ideal centroid is met, which the assigning of data points similar into the same clusters and heterogenous into other clusters

The sum of the squared distance between data points and cluster centroid is calculated first and allocated to the data points similar in the same clusters. Clustering is an unsupervised approach where the hidden association in the data can be extracted. The input PCA components are transferred to the *k-means* algorithm and as an outcome, the ideal clusters were obtained.

3.2 SAE Text Data Classification

The input data for SAE classification is extracted from the patient diagnosis dataset and it consists of text data features. The text data classification process flow chart is in Fig. 3. provided below.

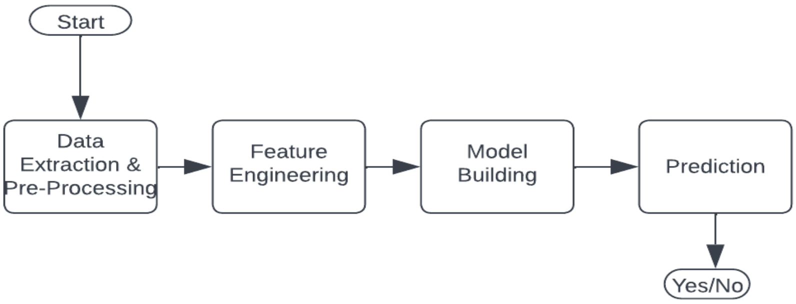

Fig. 3. Methodology for SAE Classification

Text Data Pre-Processing and Feature Engineering. The data frame after the wrangling step contains one feature containing patient's diagnosis text data and the label for the serious adverse event is a binary field called "expire flag" which tells whether the patient had observed any serious adverse event or not. The input text data had gone through various text pre-processing techniques before data modeling. Text pre-processing steps include text normalization, removing special characters and numbers, and stop ward removal. These steps are applied to each patient's diagnosis text data for better results in modeling. Text normalization is the process of converting all the text data into lower cases. Removal of unwanted special characters and integers is an important step in text pre-processing hence those should not need for text prediction. Finally, stop words are non-important words in the text and are the most occurred words and which may not provide the overall semantic meaning of the text. The cleaned text diagnosis data is converted into input features for the SAE prediction model. Count vectorizer and Term Frequency Inverse Document Frequency (TFIDF) vectorizer are used for creating features for the text data. Count vectorizer is a method to convert text to numerical data by considering the count of the word in each sentence. On the other hand, TFIDF is better than the count vectorizer because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. Fig. 4. displays the features created for building the text classification model.

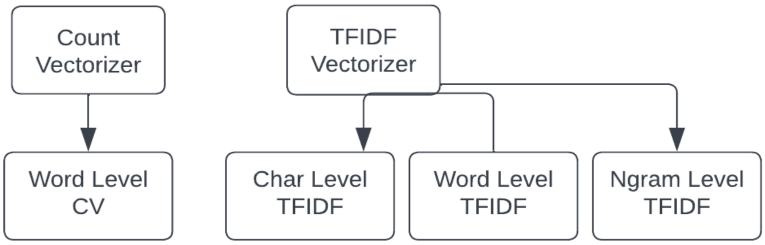

Fig. 4. Text Feature Engineering

Model Building. The classification model was developed using text data classification. Patient's historic diagnosis text data was collected and joined together for all the patients in the database and multiple classification techniques were trained on the dataset to obtain the best-fitted model. The output of the model is a binary classification that tells whether the probability of a serious adverse can occur in the future.

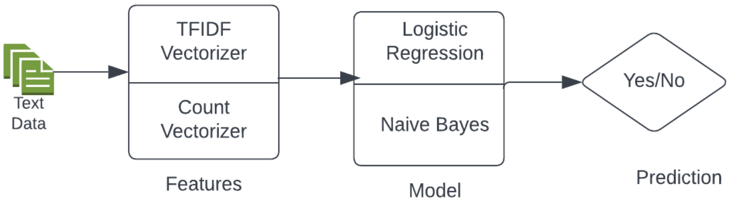

Fig. 5. Design of Text Classification Model

Fig. 5. depicts the workflow of the patient SAE prediction application. Patient's diagnosis text data converted to TFIDF and count vectorizer as features for training the model. Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF). A count vectorizer is used to transform a given text into a vector based on the frequency (count) of each word that occurs in the entire text. Converted features are then passed as input for the model such as logistic regression and a naïve Bayes which in this case is supervised binary classification techniques. The term binary classification is referred to because the outcome of the model is binary or yes/no classification. In statistics, the (binary) logistic model (or logit model) is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying the Bayes theorem with strong (naive) independence assumptions between the features. Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector representing some *n* features (independent variables), it assigns to this instance probabilities. Features are trained on both models and the best-resulting model is considered for deployment.

4 **Data Analysis**

The data set is collected from MIMIC III data mart v.1.4 and it consists of 46000 patient's data [22]. The data contains patient's admission to the hospital. The data used in this work consists of patient's demographics, admission, diagnosis, and prescription datasets. Table No. 2 lists the datasets and the description.

Table 2. Datasets And Description

No	Dataset	Description
1	Patients	Contains the demographic data for each patient's
2	Admissions	Consists of unique records of patient's admission to the hospital
3	D_icd_diagnosis	Standard 1cd9_code and label for different diagnoses which is a standard dataset
4	Diagnoses_Icd	Diagnosis contains icd9_code for each patient's visits
5	Prescriptions	Data is related to the drugs administrated to each patient during the admission to the hospital

The Patients. Admission, D_icd_diagnosis, Diagnosis_ICD, and prescription dataset is combined using SQL for data wrangling and extracted chest heart failure dataset for segmentation. Diagnosis dataset contained the patient's diagnosis data which were used for SAE classification.

5 Deployment

The usability of the solution is something that businesses looking for as there are a lot of solution builds but all of them cannot be used. Hence it is important to consider that all solutions that are provided must be in user-friendly format that it must easily interact with the users of the app. A web solution is build using *streamlit* framework by integrating the model and all other widgets for the usability of the app. The developed models were promoted for deployment by using python *streamlit* framework which is lightweight framework used to construct a python-based web dashboard and analytical widget for the data science solution. Application consists of three distinctive features which are exploratory data analysis app, a chest heart failure clustering model build on the *k-means* algorithm and a logistic regression model for SAE classification are exported as a model package for the deployment.

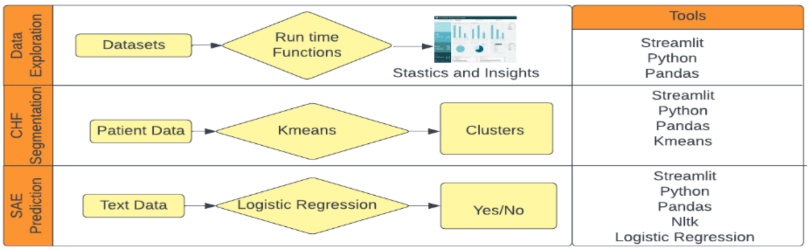

Fig. 6. Deployment Flow and Tools Used

5.1 Software Used

Python. Python is used for the building of the application. Several functions build in python for handling data wrangling and data transformation primarily for dataset exploration. Various dataset was joined together using pandas merging and aggregating functionality.

Streamlit. *Streamlit* framework provides widgets and graphs to display effectively and integrate all the building blocks of the apps into a web solution and the streamlit server is used for deploying the App.

Nltk. *Nltk* is a set of packages used for creating a text classification model which is mainly used for data cleaning steps such as the removal of characters, and stop words.

Keras. *Keras* library provides python classes and objects for various machine learning model building and Keras provided *k-means* and Logistic regression classes for building the model and later exported as a package for the deployment of the application.

6 Evaluation and Results

The evaluation of the performance of the models is essential to pick the best performance model for deploying the solution. Each modeling techniques have different types of the evaluation model. In contrast to supervised learning where we have the ground truth to evaluate the model's performance, clustering analysis does not have a solid evaluation metric that we can use for evaluating the outcome of the model. Elbow method and silhouette analysis where the two types of evaluation metrics used in the *k-means* clustering algorithm. Unlike clustering for SAE prediction, we have the outcome of the patients in the test dataset to evaluate the performance of the model. There are various performance metrics are widely used in the supervised classification model in which the most common ones are the F1 ratio, AUC score, AUC Curve plotting, Log ratio, etc. In this paper, Elbow method is used for clustering and the AUC curve is used for classification.

The Elbow method gives us an idea on what a good *k* number of clusters would be based on the Sum of Squared distance (SSE) between data points and their assigned clusters' centroids. AUC - ROC curve is a performance measurement for classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

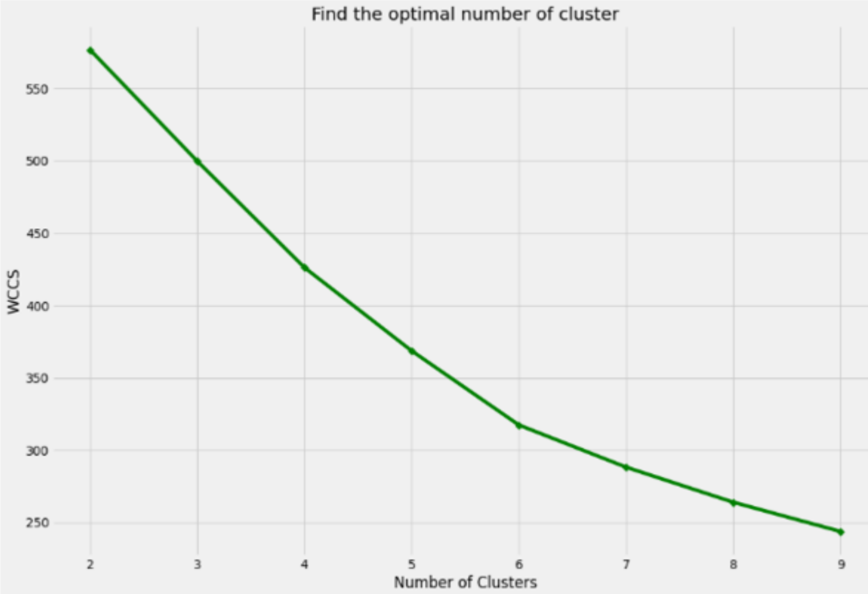

Fig. 7. WCCS Plot for Clustering

From Fig. 7. the y-axis shows the WCCS (Within-Cluster Sun of Square) and the x-axis plot the number of clusters the ideal number of clusters that are used in this work is six as the WCCS point is steady releases till the cluster number six and afterward the decentness of the cluster of reduced slightly which implies the optimal cluster might be six.

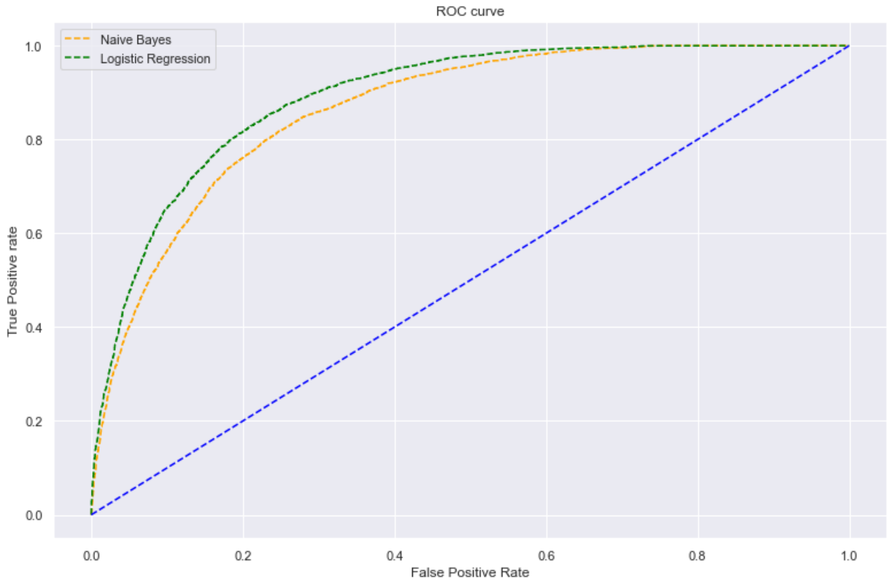

Fig. 8. SAE Classifier AUC Curve

Fig. 8. shows the ROC curve plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. An excellent model has AUC near the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In this work Logistic Regression and Naïve Bayes model's performance was good and per the figure, the logistic regression shows the higher the curve and so it is considered the best model for the classification problem.

7 Conclusion

In this paper, we have presented an efficient way of *k-means* clustering technique that can be used to identify groups and associations for chest heart failure. The outcome of the clustering technique indicated that it succeeded in grouping six clusters for grouping the data effectively. In addition, SAE classification techniques are implemented by which patient diagnosis text data is used for modeling and the outcome flags the serious threat to the life of the patient. The logistic Regression algorithm showed 89 % of the AUC score in successfully classifying the data.

References

[1] M. L. Berger, M. D. Curtis, G. Smith, J. Harnett, and A. P. Abernethy, “Opportunities and challenges in leveraging electronic health record data in oncology,” *Future Oncol*, vol. 12, no. 10, pp. 1261–1274, May 2016, doi: 10.2217/FON-2015-0043.

[2] H. G. Eichler *et al.*, “Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals?,” *Clin Pharmacol Ther*, vol. 105, no. 4, p. 912, Apr. 2019, doi: 10.1002/CPT.1226.

[3] M. M. M. Pai, R. Ganiga, R. M. Pai, and R. K. Sinha, “Standard electronic health record (EHR) framework for Indian healthcare system,” *Health Serv Outcomes Res Methodol*, vol. 21, no. 3, pp. 339–362, Sep. 2021, doi: 10.1007/S10742-020-00238-0/FIGURES/9.

[4] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, “Mining electronic health records (EHRs): A survey,” *ACM Comput Surv*, vol. 50, no. 6, Jan. 2018, doi: 10.1145/3127881.

[5] H. Estiri, J. G. Klann, and S. N. Murphy, “A clustering approach for detecting implausible observation values in electronic health records data,” *BMC Med Inform Decis Mak*, vol. 19, no. 1, Jul. 2019, doi: 10.1186/S12911-019-0852-6.

[6] M. M. Churpek, T. C. Yuen, S. Y. Park, R. Gibbons, and D. P. Edelson, “Using Electronic Health Record Data to Develop and Validate a Prediction Model for Adverse Outcomes on the Wards,” *Crit Care Med*, vol. 42, no. 4, p. 841, 2014, doi: 10.1097/CCM.0000000000000038.

[7] Z. Liu, J. Zhang, Y. Hou, X. Zhang, G. Li, and Y. Xiang, “Machine Learning for Multimodal Electronic Health Records-based Research: Challenges and Perspectives,” Nov. 2021, doi: 10.48550/arxiv.2111.04898.

[8] A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, and N. M. Luscombe, “Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease,” *PLoS One*, vol. 13, no. 8, Aug. 2018, doi: 10.1371/JOURNAL.PONE.0202344.

[9] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, “Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis,” *J Am Med Inform Assoc*, vol. 22, no. 4, pp. 872–880, Jul. 2015, doi: 10.1093/JAMIA/OCV024.

[10] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, “Predicting the Risk of Heart Failure with EHR Sequential Data Modeling,” *IEEE Access*, vol. 6, pp. 9256–9261, Jan. 2018, doi: 10.1109/ACCESS.2017.2789324.

[11] E. Lütz, “Unsupervised learning to detect patient subgroups in electronic health records,” *DEGREE PROJECT COMPUTER SCIENCE AND ENGINEERING*, 2019.

[12] G. Spini, M. van Heesch, T. Veugen, and S. Chatterjea, “Private Hospital Workflow Optimization via Secure k-Means Clustering,” *J Med Syst*, vol. 44, no. 1, pp. 1–12, Jan. 2020, doi: 10.1007/S10916-019-1473-4/TABLES/5.

[13] M. Zubair, M. Asif Iqbal, A. Shil, E. Haque, M. Moshiul Hoque, and I. H. Sarker, “An Efficient K-means Clustering Algorithm for Analysing COVID-19,” *Advances in Intelligent Systems and Computing*, vol. 1375 AIST, pp. 422–432, Dec. 2020, doi: 10.48550/arxiv.2101.03140.

[14] Y. Wang *et al.*, “Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records,” *J Biomed Inform*, vol. 102, Feb. 2020, doi: 10.1016/J.JBI.2019.103364.

[15] R. A. Hubbard, J. Xu, R. Siegel, Y. Chen, and I. Eneli, “Studying pediatric health outcomes with electronic health records using Bayesian clustering and trajectory analysis,” *J Biomed Inform*, vol. 113, Jan. 2021, doi: 10.1016/J.JBI.2020.103654.

[16] W. Cui, D. Robins, and J. Finkelstein, “Unsupervised Machine Learning for the Discovery of Latent Clusters in COVID-19 Patients Using Electronic Health Records,” *Stud Health Technol Inform*, vol. 272, pp. 1–4, 2020, doi: 10.3233/SHTI200478.

[17] I. Li *et al.*, “Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review,” Jul. 2021, doi: 10.48550/arxiv.2107.02975.

[18] A. Mascio *et al.*, “Comparative Analysis of Text Classification Approaches in Electronic Health Records,” pp. 86–94, Jul. 2020, doi: 10.18653/V1/2020.BIONLP-1.9.

[19] J. R. Ayala Solares *et al.*, “Deep learning for electronic health records: A comparative review of multiple deep neural architectures,” *J Biomed Inform*, vol. 101, p. 103337, Jan. 2020, doi: 10.1016/J.JBI.2019.103337.

[20] A. Mascio *et al.*, “Comparative Analysis of Text Classification Approaches in Electronic Health Records,” May 2020, Accessed: Sep. 22, 2022. [Online]. Available: http://arxiv.org/abs/2005.06624

[21] A. Bittar, S. Velupillai, A. Roberts, and R. Dutta, “Text classification to inform suicide risk assessment in electronic health records,” *Stud Health Technol Inform*, vol. 264, pp. 40–44, Aug. 2019, doi: 10.3233/SHTI190179.

[22] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci Data*, vol. 3, May 2016, doi: 10.1038/SDATA.2016.35.