ISSN 1533-9211

DOI: 10.5281/zenodo.10029158

ELECTRONIC HEALTH RECORDS ANALYSIS OF LEPROSY PATIENTS USING MACHINE LEARNING TECHNIQUES

JALPA DARSHIT MEHTA*

Research Scholar, Sir Padampat Singhania University, Udaipur, India.
*Corresponding Author Email: jalpadarshit.mehta@spsu.ac.in

Dr. MUKESH KALLA

Assistant Professor, Sir Padampat Singhania University, Udaipur, India. Email:mukesh.kalla@spsu.ac.in

Abstract

Electronic Health Records (EHRs) are rapidly being implemented by health care providers in the recent years. This has given rise to increase in the availability and quality of EHR data. Leprosy is one of the main public health problems and listed among the neglected tropical diseases in India. It is also called Hansen's Diseases (HD), which could be a long-term contamination by microorganisms, mycobacterium leprae. The delay in the diagnosis of leprosy can lead to increase disability rate among various patients. This paper intends to identify type of leprosy by applying Machine Learning based classification techniques on various leprosy cases from the first sign of symptoms recorded in clinical text included in Electronic Health Records (EHRs). Electronic Health Records (EHRs) of Leprosy patients from verified sources have been generated. The clinical notes included in EHRs have been processed through various Natural Language Processing techniques. In order to predict type of leprosy, Rule based classification method has been applied in this paper. Further the classification results of various Machine Learning (ML) algorithms like Support Vector Machine (SVM), Logistic regression (LR), K-nearest neighbor (KNN) and Random Forest (RF) are compared and their performance parameters are analyzed.

Keywords: Electronic Health Records, Natural Language Processing, Clinical notes, Machine Learning, Support Vector Machine, Logistic Regression, K-nearest neighbor, Rule based classification, SNOMED-CT, Random Forest

I. INTRODUCTION

According to IBM big data analytics volume of data in the world approximately doubles every two years. 90% of the world's data generated within the last two years. 2.5+ Exabyte of information has been produced day after day. With the massive use of the Internet, the mobile internet, digital medicine, social media, finance, and Internet of Things the quantity of data has expanded drastically. Big data not only depicts the massive size of data but also implies rapid processing ability and novel methodologies for handling the data.

The health care data are one of the major driving forces for big data. The availability of Electronic Health Records (EHRs) worldwide with massive extent of details is enhancing health care chain at each stage. Currently the problem has been moved from gathering data to understanding of these data. To discover hidden knowledge, it should be extracted and integrated in a structured way as the EHR data are complex heterogeneous and mostly unstructured. In order to do this various machine learning techniques and statistical techniques have been used by researchers in past few years. In the Data Driven Health care (DDH) area, the secondary use of EHR data enables possibility of forecasting imminent development of a disease by predicting the risk factors from patients' records [1]. The Global Leprosy Strategy 2016–2020: "Accelerating towards a leprosy-free world" was settled in April 2016 by World Health Organization [2]. Every national program has agreed upon 3 key targets in supporting the worldwide system: (I) zero disabilities (G2D) among children determined to have leprosy; (II) decrease of new instances of leprosy with Grade 2 Disability (G2D) to less than one case per million people and (III) zero nations with enactment permitting discrimination based on leprosy. Leprosy is one of the recorded ignored tropical illnesses which proceed as a significant medical issue in India. As per worldwide leprosy update by World Health Organization (WHO), there are 210,670 new leprosy cases detailed from 150 nations worldwide in 2017.There are approx. 1,35,485 new leprosy cases were distinguished in India. About half (67,160) cases have been analyzed as critical stages of new recognized cases. [3].

One of the main causes for the increase in disability rate is the delay in the diagnosis of leprosy and lepra reactions which lead to determined neuritis and eventually to disability. There is a requirement for more extensive alertness about the signs and symptoms of leprosy and reactions among general health care staff as well as in the community to promote self-reporting. Early diagnosis and proper management of the disease and its complications in an integrated setting will be helpful to reduce disability caused by leprosy.

The proposed study is aiming to recognize classes of leprosy from the different cases of leprosy patients at a referral center in India. Leprosy patients dependent on their symptoms of different elements of leprosy have been broken down from electronic health records of the leprosy patients. The health records include diagnoses, the first sign of symptoms and clinical notes. There are several factors by which the type of leprosy can be determined. Such factors are analyzed and a rule-based classification algorithm is developed. Before applying rule-based algorithm data is preprocessed. To provide trustworthy EHRs, concept identification and annotation is done. The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) has been applied for this purpose. Further rule-based algorithms are compared with other machine learning algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), K-nearest neighbor (KNN) and Random Forest (RF). The classification algorithms are applied to the EHRs of 250 patients to classify leprosy cases based on guidelines suggested by WHO [4] and Ridley-Jopling [5].

II. BACKGROUND

A. Electronic Health Records (EHRs)

The health care providers are implementing Electronic Health Records (EHRs) at a very fast rate in recent times. This has caused in an exceptionally significant growth within the amount and availability of EHRs data. While the key use of EHRs data is enhancing clinical competence, there is an improved and developing interest in the secondary use of EHRs data has been created recently. This includes several clinical tasks, such as disease diagnosis and risk predictions [6], prediction of patients' readmission [7], predicting medical concept names and applying standardized coding schema [8] and predicting future of the patients [9].

ISSN 1533-9211

DOI: 10.5281/zenodo.10029158

EHR systems were mainly considered for core hospital managerial tasks. There are numerous classification schemas and terminologies are available for recording medical information and procedures. There are some examples like International Statistical Classification of Diseases and Related Health Problems (ICD) for diagnosis codes, Current Procedural Terminology (CPT), United Medical Language System (UMLS) and the Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT) for procedure codes. EHR systems handle numerous forms of patient information including personal information, diagnoses, examination records, sensor measurements, laboratory test results, patient history, Doctors' prescriptions, and clinical notes.

B. Types of Leprosy

Basic side effects present in the various kinds of leprosy incorporate a runny nose, dry scalp, eye issues, skin injuries, muscle shortcoming, rosy skin, smooth, sparkling, diffuse thickening of facial skin, ear and hand, loss of sensation in fingers and toes, thickening of peripheral nerves, a flat nose because of the decimation of nasal ligament, phonation and resounding of sound during discourse. Individuals may start to see indications inside the primary year or as long as 20 years after infection. Leprosy is broadly classified as WHO classification and Jopling classification.

Leprosy can be assembled dependent on clinical appearances of skin lesions and skin smear results. In the request reliant on skin smears, patients demonstrating -ve smears at all sites are collected as paucibacillary leprosy (PB), while those showing +ve smears at any site are assembled as having multibacillary leprosy (MB). The clinical plan of request with the ultimate objective of treatment joins the usage of number of nerves and skin lesions required as the explanation behind common case leprosy patients into paucibacillary (PB) and multibacillary (MB) leprosy. Jopling classification of leprosy is the arrangement that separates 5 structures dependent on the bacteriological list. These structures connect with the immunological reaction to M. leprae. Patients with Tuberculoid leprosy (TT) are impervious to the bacillus and contamination is limited. Patients with Lepromatous leprosy (LL) are very touchy to the bacillus and the disease is spread. Borderline structures (BT, BB, and BL) are between the two parts of the types (TT and LL). In this study SNOMED CT coding applied to provide annotation to various types of leprosy as shown in the Table I.

Table I: Snomed-Ct Codes for Types of Leprosy

Type of Leprosy	SCTID
Paucibacillary Leprosy (PB)	416483009
Multibacillary Leprosy (MB)	416257001
Tuberculoid Leprosy (TT)	70143003
Borderline Tuberculoid (BT)	240402003
Mid Borderline (BB)	400154003
Borderline Lepromatous (BL)	240403008
Lepromatous Leprosy (LL)	21560005

DOI: 10.5281/zenodo.10029158

III. LITRATURE REVIEW

The EHR is used to automate and update the workflow of clinical systems. The EHR has the capability to produce a comprehensive record of a clinical patients' meeting, along with supporting other care-related events mainly or secondarily using interface which includes Decision Support System, quality management, laboratory results and reports. EHR can be characterized based on the functionalities: (i) basic EHRs without clinical notes (ii) basic EHRs with clinical notes and (iii) comprehensive systems [24]. While doing the literature survey, we found many terms that have developed together with EHR. One of these is the Electronic Medical Record (EMR). EMR is often used in equivalent with EHR [23].

Pre-processing of clinical unstructured data is converted into a structured format using NLP approach this helps to understand clinical notes or clinical information. Data cleansing processes like stop word removal helps to reduce computational time of the algorithm as well [10]. Importance and methods of pre-processing in text mining by removal of stop words on an unstructured data to get better output has been introduced [11]. How SVM algorithm is used in identifying the medical terminologies in the field of medical science and by labelling important terms [12]. EHR data pre-processing is done using regular expressions and it shows various types of vectorization and clustering techniques applied on medical texts [25].

Creating bags of words, considering combinations of cases, tokenizing and putting it in SVM trained model and measuring using classification parameters such as F1 score [13]. To study the comparative analysis of various machine learning algorithms some tenable parameters are considered such as F1 score, accuracy, recall and precision to get the reliability level of algorithm [20]. Numerous disease risk prediction is performed based on some Machine Learning technology [14]. To improve classification of large collection of biomedical documents, K-nearest Neighbours (KNN) algorithm and Explicit Semantic Analysis (ESA) technique are implemented [26].There is a different dynamic method using Random Forest (RF) algorithm to predict risk from clinical text for survival, longitudinal, and multivariate (SLAM) results has been proposed named as RF-SLAM [27].

Text mining is done on Electronic Patient Records (EPR) to get the proper pre-processed data and how labelling, classifiers and feature extraction is performed on the EPR.[15]. The problem of knowledge retrieval from EHR has been discussed and interpretation of medical records has been generated by applying various text mining algorithms [16]. A rule based feature classification along a deep learning technique was studied for effective disease classification [17]. Using rule classification and logistic regression are hybrid approach models of both the algorithms in applied and compared through classification parameters [18].

Comparison of Ridley-Jopling classification and WHO classification with other clinical classification for leprosy and their operational methods are described [19]. A rule-based model was compared with other machine learning models like SVM, logistic regression, and decision tree [21]. Text mining and data processing of EMR patients has been applied using named entity recognition, data cleansing, data transformation, reduction and integration [22].

IV. METHDOLOGY

A. Data Collection

A web based EHR system has been developed, where doctors who are treating leprosy patients insert details which included personal details, allergies, addiction (for example tobacco, alcohol, etc.), known leprosy contacts, signs and symptoms, type of leprosy, grade of disability, nerves affected, skin smears test results, drugs prescribed and many more. Data was collected from the web-based patient history form and transformed into a .csv file.

The list of fields included in input dataset are patient_id, record_id, gender, age, symptoms, assessment of skin lesion and diagnosis code of respective leprosy cases. After gaining some domain knowledge several factors are analysed, by which leprosy can be classified into different types. Several factors like first sign and symptom, number of smears, nerves damaged and doctors' comment on assessing skin lesions were taken into consideration in order to predict the type of leprosy.

B. Data Pre-processing

Removing stop words: Stop words are considered as good for nothing, which are selected through to decrease the processing time. This rundown comprises the relational word, articles, conjunctions, punctuation, phrase removal, etc.

Lemmatization: Lemmatization typically refers to doing things appropriately with the utilization of jargon and morphological investigation of words, ordinarily, meaning to evacuate inflectional endings just and to restore the base or lexicon type of a word, which is known as lemma.

Tokenization: Tokenization is a typical job in NLP, it is fundamentally an undertaking of slashing a character into pieces, called as a token, and discarding the specific character simultaneously. Figure 1 shows the word cloud representing sign and symptoms of leprosy included in the clinical notes of patients in EHR system.

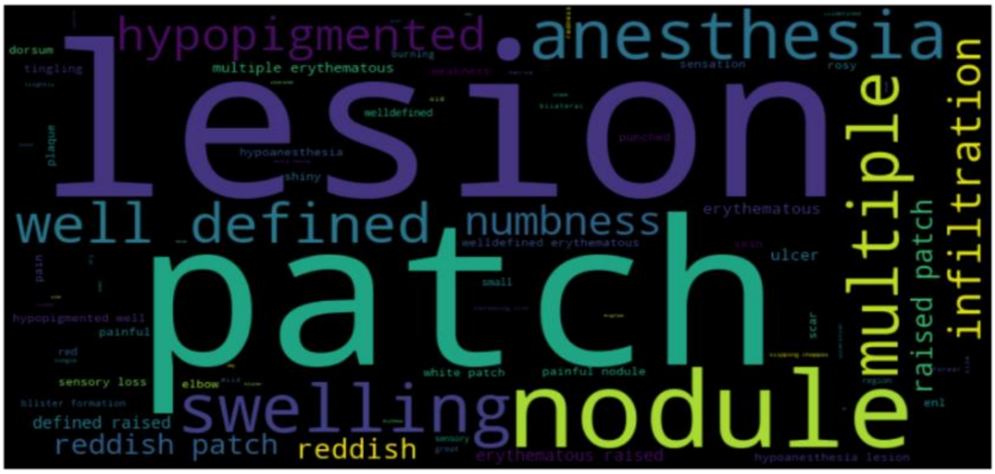

Figure 1: Word Cloud of Clinical Note

DOI: 10.5281/zenodo.10029158

C. Classification using Rule based Algorithm Equations

Rule-based classification models can be effectively upgraded and supplemented by including new guidelines from area specialists dependent on their space information. Rules can be easily expressed as logic in IF-THEN format, for instance in our study, IF more than four smear sites are present on the patient's body THEN it is a MB type of leprosy. It follows a general pattern of IF case/condition THEN answer/conclusion. IF the following condition or case is satisfied THEN only the conclusion can be predicted. Some of the rules generated in our study described below.

R1: IF (no of. thickened nerves > 2) ^ (no.of skin smears > 4) ^ (corpus.find ('bilateral')>=0) ^ (corpus.find ('lagophthalmos') >=0) THEN predicted _MB = 1.

R2: IF (corpus.find ('blister)>=0) ^ corpus.find ('nodule') >=0 & corpus.find('ulcer') >=0 THEN predicted = 1 predicted_BL = 1.

The key aspects of the algorithms are

1. Generate word corpus of symptoms and skin lesions assessment by pre-processing clinical text.
2. Build set of rules by using IF-THEN for classifying type of Leprosy such as PB/MB or TT/BT/BB/BL/LL.
3. Extraction of rules by applying Sequential Covering on training dataset.
4. Evaluation of the algorithm by calculation accuracy.

Main characteristic or advantage of rule-based algorithms is that we can generate our own rules which should be logical according to the data and it should be able to predict and give proper results. Another advantage of rules-based algorithm is that rules can be modified according to the changes and requirements.

Several bags or words, conditions and combinations are examined in order to generate rule-based algorithms to give a better logical conclusion. Moreover, it provides a good data model which is human understandable

D. Applying Machine Learning based Classification Algorithms

The Electronic Health Records (EHR) data is taken from verified sources. Further, analysis of the clinical texts is performed on the EHR. Clinical features like swelling, erythematous, numbness, number of smears, nerve thickened are analysed and data is pre-processed by removal of stop words and further lemmatizing data in order to get clinical root words from the clinical notes and further tokenization is performed. That data is further split into training data (80%) and testing data (20%). The 10-fold cross validation is performed on the training set so that every observation from dataset gets a chance to perform in the training set and ensures that the input data is limited to few observations. Data is fed to machine learning models like Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbour (KNN), Random Forest (RF) and accuracy of these algorithms are compared with the rule based model.

DOI: 10.5281/zenodo.10029158

Other classification parameters are also used such as f1 score, precision and recall for all algorithms for better comparison. The classification applied on 250 patients EHRs. Overall working of the system is shown in Figure 2.

V. RESULT AND DISCUSSION

After applying all the algorithms, it can be clearly seen that rule-based algorithm stays to be highest in both cases which is for MB, PB classification and for other types. On the other hand, Support Vector Machine shows accuracy of 89.47% in (PB/MB) type of classification and 89.58% in (TT/BT/BB/BL/LL) cases. Logistic Regression gives 96.6% and 83.4% accuracy in WHO and Jopling classification respectively. KNN classification algorithm achieves 96.26% accuracy in (PB/MB) type classification and 90.40% in (TT/BT/BB/BL/LL) type classification. Table II and Table III show overall comparison results of classification algorithm applied in the study.

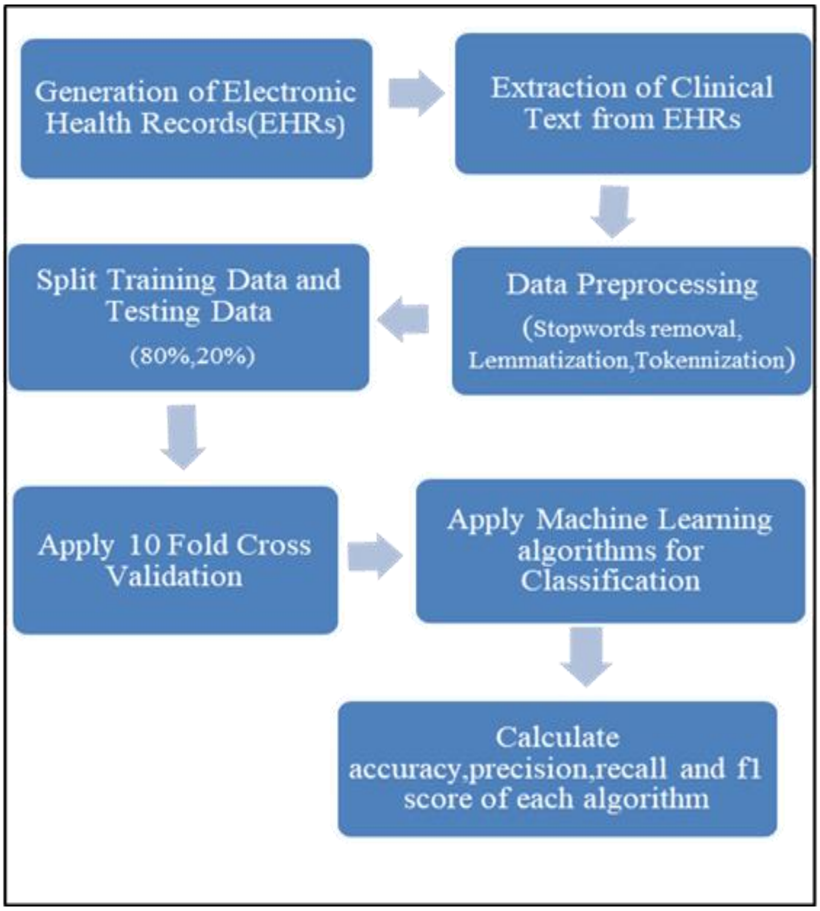
Generation of Electronic Health Records(EHRs)
Extraction of Clinical Text from EHRs
Data Preprocessing (Stopwords removal, Lemmatization,Tokennization)
Split Training Data and Testing Data (80%,20%)
Apply 10 Fold Cross Validation
Apply Machine Learning algorithms for Classification
Calculate accuracy,precision,recall and f1 score of each algorithm

Figure 2: Overall Workflow of the System

Table II: Comparisons of Classification Results of (PB/MB) Cases

Classification algorithm	Accuracy	Precision	Recall	F1-score
Rule-based Algorithm	99.15%	99.5%	99.5%	99.5%
Support Vector Machine	89.47%	89.47%	100%	94.07%
Logistic Regression	96.6%	95.8%	95.8%	95.8%
K-Nearest Neighbor	96.26%	95.83%	95.83%	95.83%
Random Forest	98.59%	99.5%	99.10%	99.29%

Table III: Comparison of Classification Results of (TT/BT/BB/BL/LL) cases

Classification algorithm	Accuracy	Precision	Recall	F1-score
Rule-based Algorithm	94.5%	97.8%	92%	94.6%
Support Vector Machine	89.58%	87.29%	87.29%	87.29%
Logistic Regression	83.4%	79.1%	79.1%	79.1%
K-Nearest Neighbour	90.4%	89.5%	89.5%	89.5%
Random Forest	91.66%	91.66%	91.66%	91.66%

VI. CONCLUSION AND FUTURE SCOPE

In the field of medical science Leprosy is known as a contagious disease and there lacks any algorithm specifically for predicting the type of Leprosy. Thus, this set of conditions in a rule-based system is helping to get a better output of the type of leprosy with minimum factors.

Considering the unstructured data rule-based algorithm is applied, on the other side the same dataset is applied on the other machine learning algorithms and the results are obtained.

Hence, it can be proved that rule-based algorithms are giving better performance on small corpus, but as the number of clinical notes will increase, machine learning algorithms can have better accuracy than rule-based algorithms.

Future work can be advanced by extending this work on large number of clinical notes using deep learning models like Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN).

Acknowledgment

The authors are thankful for positive support received from the Bombay Leprosy Project for collecting patients' history, clinical notes, smear test results, and providing Corpus of leprosy patients. We sincerely like to express our gratitude to Dr. V.V. Pai and Dr. Ashish Khodke for providing basic domain knowledge and validating our research work.

References

1) L. B. Madsen, Data Driven healthcare: How analytics and BI are transforming the industry. John Wiley & Sons, 2014.

2) WHO / Department of Control of Neglected Tropical Diseases.: Global leprosy update, 2015: time for action, accountability and inclusion. 2016.

3) https://www.indiaspend.com/leprosy-is-making-a-comeback-in-india-but-the-govt-wants-to-deny-it/.

4) https://www.who.int/lep/classification/en/.

DOI: 10.5281/zenodo.10029158

5) D. S. Ridley and W. H. Jopling, “A classification of leprosy for research purposes,” Lepr. Rev., vol. 33, no. 2, pp. 119–128, 1962.

6) C. Zhengping, Y. Cheng, S. Zhai, S. Zhaonam, and Y. Liu, “Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records,” in IEEE International Conference on Data Mining, 2017.

7) B. K. Reddy and D. Delen, “Predicting hospital readmission for Lupus patients: An RNN LSTM-based deep-learning Methodology, Computers in Biology and Medi-cine,” Computers in Biology and Medi-cine”, 2018.

8) F. Zheng and L. Cui, “Exploring deep learning-based approaches for predicting concept names in SNOMED CT,” in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018.

9) R. Miotto, L. Li, and J. T. Dudley, “Deep Patient: An unsupervised representation to Predict the future of Patients from the Electronic Health Records, Advances in In-formation Retrieval,” pp. 768–774, 2016.

10) K. Kreimeyer et al., “Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review,” J. Biomed. Inform., vol. 73, pp. 14–29, 2017.

11) K. V. Ghag and K. Shah, “Comparative analysis of effect of stop words removal on sentiment classification,” in 2015 International Conference on Computer, Communication and Control (IC4), 2015.

12) K. Takeuchi and N. Collier, “Bio-medical entity extraction using support vector ma-chines,” Artificial Intelligence in Medicine, vol. 33, no. 2, pp. 125–137, 2004.

13) W.-H. Weng, K. B. Wagholikar, A. T. Mccray, P. Szolovits, and H. C. Chueh, Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach BMC Medical Informatics and decision making. 2017.

14) M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Ma-chine Learning Over Big Data from Healthcare Communities,” IEEE Access, vol. 5, pp. 8869–8879, 2017.

15) S. Tsumoto, T. Kimura, H. Iwata, and S. Hirano, “Mining Text for Disease Diagnosis,” Procedia Computer Science, vol. 122, 2017.

16) M. Oleg, B. Ekaterina, Y. Alexey, B. Anastasia, and F. Sergey, Pattern-based Min-ing in Electronic Health Records for Complex Clinical Process Analysis. .

17) Y. Liang, M. Chengsheng, and L. Yuan, “Clinical Text Classification with Rule-based Features and Knowledge-guided,” Convolutional Neural Networks, pp. 1–10, 2018.

18) S. C. Virgeniya and E. Ramaraj, “Predictive Analytics Using Rule Based Classification and Hybrid Logistic Regression (HLR),” Algorithm For Decision Making, pp. 1–5, 2019.

19) J. I. A. Rodrigues et al., “Leprosy classification methods: a comparative study in a referral centre in Brazil,” International Journal of Infectious Diseases, vol. 45, pp. 118–122, 2016.

20) S. Mohammadian, A. Karsaz, and Y. M. Roshan, “A comparative analysis of classification algorithms in diabetic retinopathy screening”,” in 7th International Conference on Computer and Knowledge Engineering, 2017.

21) T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, “A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL),” International Journal of Computer Applications, vol. 68, pp. 975–8887, 2013.

22) W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, “Data processing and text mining technologies on electronic medical records: A review,” J. Healthc. Eng., vol. 2018, pp. 1–9, 2018.

23) A. Diween and M. Garg, Electronic health record - Improving quality of care, reducing costs and empowering patients” Calance. 2006.

24) L. Caroprese, P. Veltri, E. Vocaturo, and E. Zumpano, “Deep learning techniques for electronic health record analysis,” in 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), 2018.

25) S. Sergey, M. Oleg, Y. Alexey, and K. Sergey, “Machine Learning Based Text Mining i Electronic Health Records: Cardiovascular Patient n Cases,” International Conference on Computational Science ICCS, pp. 818–824, 2018.

26) K. Dramé, F. Mougin, and G. Diallo, “Large scale biomedical texts classification: a KNN and an ESA-based approaches,” J. Biomed. Semantics, vol. 7, no. 1, p. 40, 2016.

27) S. Wongvibulsin, K. C. Wu, and S. L. Zeger, “Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis,” BMC Med. Res. Methodol., vol. 20, no. 1, p. 1, 2019.