Clinical Text Mining of Electronic Health Records to Classify Leprosy Patients Cases

Jalpa Mehta, Jaydeep Dharamsey, Pravalika Domal, V. V. Pai

Abstract*: Leprosy is one of the major public health problems and listed among the neglected tropical diseases in India. It is also called Hansen's Diseases (HD), which is a long haul contamination by microorganisms Mycobacterium leprae or Mycobacterium lepromatosis.Untreated, leprosy can dynamic and changeless harm to the skin, nerves, appendages, and eyes. This paper intends to depict classification of leprosy cases from the main indication of side effects. Electronic Health Records (EHRs) of Leprosy Patients from verified sources have been generated. The clinical notes included in EHRs have been processed through Natural Language Processing Tools. In order to predict type of leprosy, Rule based classification method has been proposed in this paper. Further our approach is compared with various Machine Learning (ML) algorithms like Support Vector Machine (SVM), Logistic regression (LR) and performance parameters are compared.

Keywords: Clinical Text Mining, Natural Language Processing, Leprosy, Support Vector Machine, Logistic regression, Rule based, Electronic record, Clinical Notes.

I. INTRODUCTION

Leprosy is one of the recorded ignored tropical illnesses which proceeds as a significant medical issue in India. As per worldwide leprosy update by World Health Organization(WHO) In year 2015, India revealed more than 127,000 new cases, representing more than half of the worldwide new leprosy cases; Brazil, announced 26,396 new cases, which is 13% of the worldwide new cases; and Indonesia detailed 17,200 new cases, 8% of the worldwide case load. No different nations announced more than 10,000 new cases [13] (Report of WHO 2016). There are 210,670 new leprosy cases detailed from 150 nations worldwide in 2017, according to the World Health Organization (WHO). In 2017, there are approx. 135,485 new leprosy cases were distinguished in India. Of the new cases recognized, about half (67,160) have been analyzed at a propelled arrange [11] (Menon Ramesh 2019).

The Global Leprosy Strategy 2016–2020: "Quickening towards a without leprosy world" was discharged in April 2016 by WHO.

Revised Manuscript Received on March 02, 2020.

Ms. Jalpa Mehta, Assistant Professor, Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, India.

Mr. Jaydeep Dharamsey, Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, India.

Ms. Pravalika Domal, Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, India.

Dr. Vivek Vasudev Pai, Director, Department of Bombay Leprosy Project, Mumbai, India.

The methodology depends on the standards of starting activity, guaranteeing responsibility and advancing consideration. It is planned around three columns: expanding government possession, coordinated effort and organization; finishing leprosy and its inconveniences; and forestalling separation and encouraging consideration. Every national program have acknowledged 3 key targets in supporting the worldwide technique: (I) zero incapacities (G2D) among youngsters determined to have leprosy; (II) disposal of new instances of leprosy with G2D to < 1 case per million populace; and (III) zero nations with enactment permitting separation based on leprosy.

One of the key purposes behind the ascent in inability is a deferral in determination of leprosy and lepra responses which lead to tireless neuritis and at last to incapacity. There is a requirement for more extensive mindfulness about the signs and manifestations of leprosy and responses among general human services staff just as in the network to advance self-detailing, just as early determination and appropriate administration of the ailment and its intricacies in a coordinated setting.

The proposed examine is intending to recognize sorts of leprosy from the different instances of leprosy patients at a referral community 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, 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. Further rule based algorithms are tested with other machine learning algorithms such as SVM and Logistic Regression. Rule based algorithm is applied on the EHRs of 236 patients to classify leprosy based on guidelines of WHO and W.H.Jopling.

II. BACKGROUND

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 fringe nerves; a level nose because of the decimation of nasal ligament; phonation and resounding of sound during discourse. Individuals may start to see indications inside the principal year or as long as 20 years after infection.

The primary recognizable indication of leprosy is frequently the improvement of pale or pink shaded patches of skin that might be obtuse toward temperature or pain. Patches of stained skin are now and again joined or went before by nerve issues remembering deadness or delicacy for the hands or feet. Secondary diseases (extra bacterial or viral contaminations) can bring about tissue misfortune, making fingers and toes become abbreviated and distorted. The nerve harm continued is reversible when treated early, however becomes changeless when suitable treatment is begun following a postponement of a while. Harm to nerves may cause loss of muscle work, prompting loss of motion. It might likewise prompt sensation variations from the norm or deadness, which may prompt extra contaminations, ulcerations, and joint distortions. Leprosy is broadly classified as WHO classification and Jopling classification.

A. WHO Classification:

Leprosy can be assembled dependent on clinical appearances and skin smear results. In the request reliant on skin smears, patients demonstrating -ve smears at all goals 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 social occasion leprosy patients into paucibacillary (PB) and multibacillary (MB) leprosy.

B. Jopling classification:

This arrangement 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, BL) are between the two parts of the bargains (TT and LL).

Table 1. Classification of Leprosy

Paucibacillary forms (WHO Classification)		Multibacillary forms (Jopling Classification)		
Tuberculoid	**Borderline Tuberculoid**	**Borderline**	**Borderline Lepromatous**	**Lepromatous**
TT	BT	BB	BL	LL

III. LITERATURE REVIEW

Preprocessing 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 helps to reduce computational time of the algorithm as well [23]. To study the comparative analysis of various machine learning algorithms some tunable parameters are considered such as F1, accuracy, recall and precision to get the reliability level of algorithm [22]. Modern detailed information on the type of leprosy and its causes or effects [21]. Classification of leprosy using Ridley-Jopling and WHO classification with its detailed information.[20]

Comparison of Ridley-Jopling and WHO classification which is better in comparison with other clinical classification and their operational methods [19].Importance and methods of preprocessing in text mining by removal of stop words on an unstructured data to get better output [17]. 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 [16]. How SVM algorithm is used in identifying the medical terminologies in the field of Medical science and by labeling important terms [15].

Several disease risk prediction is performed based on several Machine Learning technology [12]. Comparison of various leprosy cases all over the world and its awareness [11]. The problem of knowledge retrieval from EHR and interpreting medical records and applying various algorithms using text mining [14]. Text mining is done on electronic patient records to get the proper preprocessed data and how labeling ,classifiers and feature extraction is performed on the EPR. [9]. The importance of text and data mining in health and medical information systems [8]. Information by WHO about the neglected tropical disease [13]. Text mining of cancer pathology report A rule based system is developed for comparing manual encoding pathology report to validate the performance of rule based system [2]. Information of WHO classification on leprosy [6]. A rule-based model was compared with other machine learning models like fph, logistic regression, and decision tree [5]. In this, a rule based feature classification along a deep learning technique was studied for effective disease classification [3].Text mining and data processing of EMR patients using named entity recognition, data cleansing, data transformation,reduction and integration [1]. Using rule classification and logistic regression are hybrid approach models of both the algorithms in applied and compared through classification parameters [4]. Medical terms and spell check along with corpus creation, used to get a proper clinical text [10]. Using NLP technique medical features are extracted from stroke patients and applied on data mining algorithms such as SVM to find accurate stroke patients within a limited timespan to prevent if from severe consequences[18]. Disease prediction which are cases related to obesity by lexical analysis using hybrid machine learning approach and text mining [7].

IV. METHODOLOGY

A. Data Collection

A web system was developed where doctors who treated leprosy filled patient's details which included personal details, allergies, addiction (eg: tobacco, alcohol, etc), known leprosy contacts, signs and symptoms, type of leprosy, grade of disability, nerves affected, skin smears, drugs prescribed and many more. Data was collected from the developed web system and transformed into a csv file.

B. Data review

After gaining some domain knowledge we analyzed several factors 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.

C. Data Preprocessing

- **Removing stop words**

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

- **Lemmatization**

Lemmatization typically alludes 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 errand in NLP, it is fundamentally an undertaking of slashing a character into pieces, called as a token, and discarding the specific character simultaneously, similar to accentuation.

D. Generating Rule based Algorithm

Rule-based characterization models can be effectively upgraded and supplemented by including new guidelines from area specialists dependent on their space information. This has been effectively executed in numerous master frameworks. In many cases, rule based algorithms are competitive and better when compared to other machine learning algorithms. Undoubtedly, rules can speak to data or information in a basic and viable manner. They give a generally excellent information model that individuals can see well indeed.

Rules can be easily expressed as logic in IF-THEN format, for instance IF more than five 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.

E. Prediction based on Rule Based Algorithm

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 arrested in order to generate rule-based algorithms to give a better logical conclusion. Moreover, it provides a good data model which is human understandable.

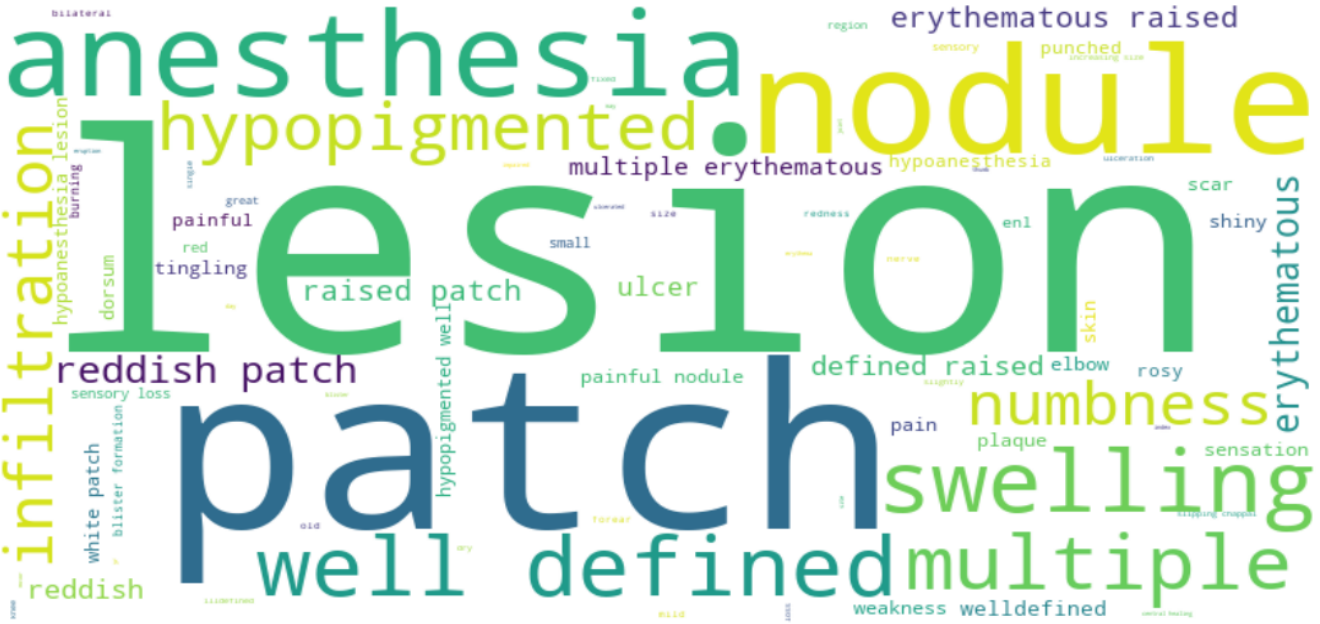

Fig 1. Word Cloud of Clinical text

IV. APPLYING MACHINE LEARNING ALGORITHM

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 ,etc. are analyzed and data is preprocessed by removal of stopwords 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 and Logistic Regression and accuracy of these algorithms are compared with the rule based model.Other classification parameters are also used such as f1 score, precision and recall for all algorithm for better comparison. A simple flowchart is displayed in Figure 2.

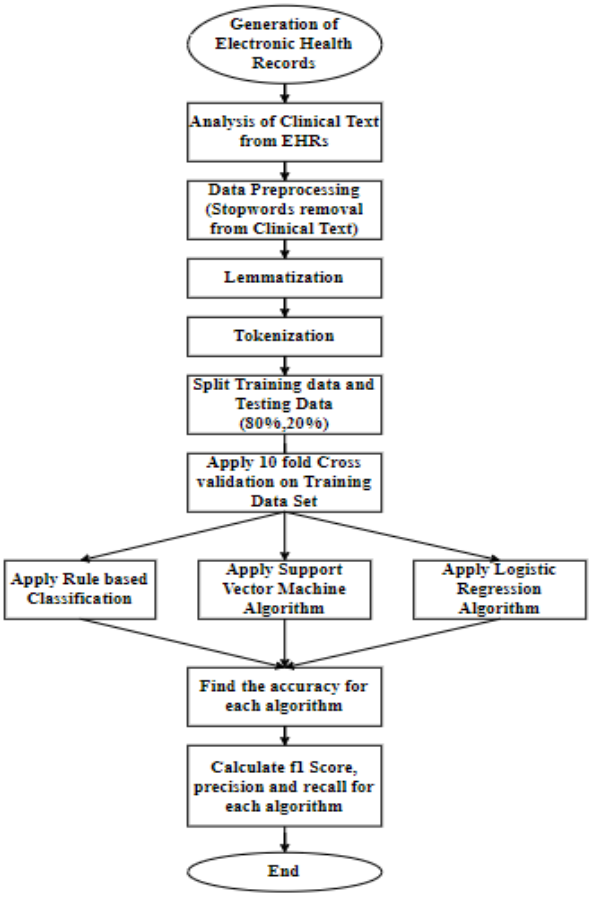

Fig 2. Process Model

V. RESULTS 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 classification with 99.2% and 94.9% respectively. On the other hand Support Vector Machine shows accuracy of 98.7% in WHO type of classification and 87.7% in Jopling. Whereas, Logistic Regression gives a 96.6% and 83.4% accuracy in WHO and Jopling classification.

Table 3. MBPB (WHO Classification)

Classification Algorithm	Accuracy	Precision	Recall	F1 Score
Rule-based Algorithm	99.2%	99.5%	99.5 %	99.5 %
Support Vector Machine	98.7%	97.8%	100%	98.9 %
Logistic Regression	96.6%	95.8%	95.8 %	95.8 %

Table 4. Others (Ridley-Jopling classification)

Classification Algorithm	Accuracy	Precision	Recall	F1 Score
Rule-based Algorithm	94.9%	97.8%	92%	94.6 %
Support Vector Machine	87.7%	83.4%	83.4%	83.4 %
Logistic Regression	83.4%	77.1%	77.1%	77.1 %

The given figure 3 and 4 are two confusion matrix based on the prediction results using rule based algorithm where figure represents confusion matrix of PB and MB type of leprosy and the figure represents confusion matrix of other types of leprosy

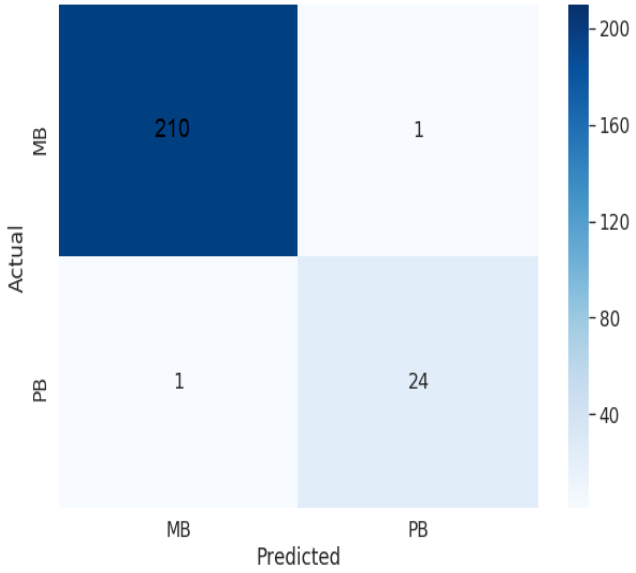

Fig 3. MBPB Confusion Matrix (WHO Classification)

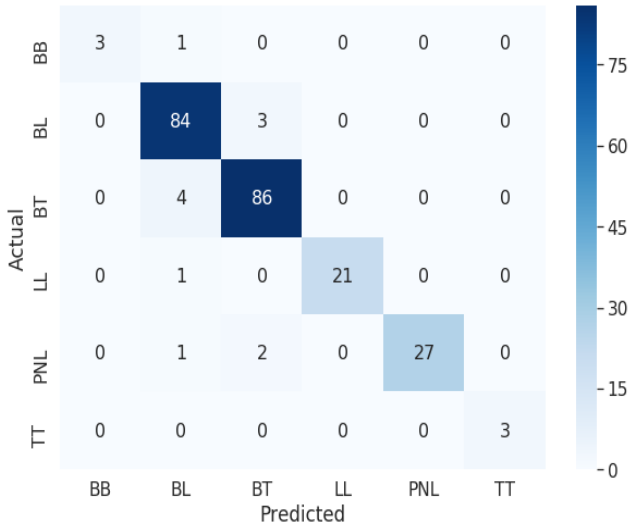

Fig 4. Others Confusion Matrix (Jopling Classification)

VII. CONCLUSION AND FUTURE WORK

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 developed 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

This work is partially funded by Mumbai University under Minor Research Grant 2018-19. We would like to thank the Bombay Leprosy Project for collecting patients' history, clinical notes and providing Corpus of leprosy patients. We sincerely like to express our gratitude to Dr.Ashish Khodke for providing basic domain knowledge and validating our research work.

REFERENCES

1. Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). *Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. Journal of Healthcare Engineering, 2018, 1–9.* doi:10.1155/2018/4302425
2. Rebecka Weegar, Hercules Dalianis.Creating a rule based system for text mining of Norwegian breast cancer pathology reports .Computer Science, Published in Louhi @ EMNLP 2015. DOI:10.18653/v1/W15-260.
3. Liang Yao , Chengsheng Mao,Yuan Luo.(2018) Clinical Text Classification with Rule-based Features and Knowledge-guided Convolutional Neural Networks ,2018 , 1-10. arXiv:1807.07425
4. S.Clement Virgeniya, E. Ramaraj. Predictive Analytics Using Rule Based Classification And Hybrid Logistic Regression (HLR) Algorithm For Decision Making , 2019, 1-5. ISSN 2277-8616
5. Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu.A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL), 2013, 1-5.
6. https://www.who.int/lep/classification/en/
7. Yang, H., Spasic, I., Keane, J. A., & Nenadic, G. (2009). *A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries. Journal of the American Medical Informatics Association, 16(4), 596–600.* doi:10.1197/jamia.m309
8. Kocbek, S., Cavedon, L., Martinez, D., Bain, C., Manus, C. M., Haffari, G., … Verspoor, K. (2016). *Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. Journal of Biomedical Informatics,64, 158–167.* doi:10.1016/j.jbi.2016.10.008
9. Tsumoto, S., Kimura, T., Iwata, H., & Hirano, S. (2017). *Mining Text for Disease Diagnosis. Procedia ComputerScience,122,1133–1140.*doi:10.1016/j.procs.2017.11.483
10. Dalianis, H. (2018). *Clinical Text Mining.* doi:10.1007/978-3-319-78503-5
11. Ramesh Menon., "Do More Detected Cases Mean Leprosy Is Making A Comeback In India?Experts,GovernmentDiffer",4thJanuary 2019 https://www.indiaspend.com/leprosy-is-making-a-comeback-in-india-but-the-govt-wants-to-deny-it/
12. Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). *Disease Prediction by Machine Learning Over Big Data From HealthcareCommunities.IEEEAccess,5,8869–8879.* doi:10.1109/access.2017.2694446
13. WHO/ Department of Control of Neglected Tropical Diseases., "Global leprosy update, 2015: time for action, accountability and inclusion", 2nd September 2016
14. Oleg Metskera , Ekaterina Bolgovaa , Alexey Yakovleva,b, Anastasia Funknera , Sergey Kovalchuka. Pattern-based Mining in Electronic Health Records for Complex Clinical Process Analysis
15. Takeuchi, K., & Collier, N. (2005). *Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine, 33(2), 125–137.* doi:10.1016/j.artmed.2004.07.019
16. Weng, W.-H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). *Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Medical Informatics and Decision Making, 17(1).* doi:10.1186/s12911-017-0556-8
17. Ghag, K. V., & Shah, K. (2015). *Comparative analysis of the effect of stopwords removal on sentiment classification. 2015 International Conference on Computer, Communication and Control (IC4).* doi:10.1109/ic4.2015.7375527
18. Mining clinical text for stroke prediction. Journal: Network Modeling Analysis in Health Informatics and Bioinformatics > Issue 1/2015. Elham Sedghi, Jens H. Weber, Alex Thomo, Maximilian Bibok, Andrew M. W. Penn
19. Rodrigues Júnior, I. A., Gresta, L. T., Noviello, M. de L. M., Cartelle, C. T., Lyon, S., & Arantes, R. M. E. (2016). *Leprosy classification methods: a comparative study in a referral center in Brazil. International Journal of Infectious Diseases, 45, 118–122.* doi:10.1016/j.ijid.2016.02.018
20. Tasleem Arif, Konchok Dorjay, Mohammad Adil, Marwa Sami. Classification of leprosy-From past to present.Journal of Pakistan Association of Dermatologists, 28(1):95-99 · July 2018
21. Nunzi, E., & Massone, C. (Eds.). (2012). *Leprosy.* doi:10.1007/978-88-470-2376-5
22. Mohammadian, S., Karsaz, A., & Roshan, Y. M. (2017). *A comparative analysis of classification algorithms in diabetic retinopathy screening. 2017 7th International Conference on Computer and Knowledge Engineering(ICCKE).* doi:10.1109/iccke.2017.8167934
23. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Botsis, T. (2017). *Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics, 73, 14–29.* doi:10.1016/j.jbi.2017.07.012

AUTHORS PROFILE

Ms. Jalpa Mehta, Assistant Professor, Information Technology Department, Shah and Anchor Kutchhi Engineering College, Mumbai, India**.** She has received M.Tech. degree in Computer Engineering from VJTI, University of Mumbai. Her research interests include Data science, Image processing, natural language processing, machine learning, recommender systems, and healthcare informatics.

Mr. Jaydeep Dharamsey, Student, Information Technology Department, Shah and Anchor Kutchhi Engineering College, Mumbai, India. He is in the final year of completing his B.E. degree. His research interest includes natural language processing, big data analytics, data science, text mining etc.

Ms. Pravalika Domal, Student, Information Technology Department, Shah and Anchor Kutchhi Engineering College, Mumbai, India. She is studying in the last year of her B.E. degree. Her research interests include machine learning, natural language processing, data mining , deep learning etc.

Dr. Vivek Vasudev Pai, Indian physician and the Director of Bombay Leprosy Project, Mumbai, India. He has been named as Best Leprosy Worker, Bombay Municipal Corporation Indian Leprosy Foundation**.** He has been listed as a noteworthy physician by Marquis Who's Who.