Text Mining In Healthcare

Pranita Mahajan, Dipti P. Rana

Abstract: ***In healthcare, data mining intensively and extensively becoming essential. Data mining applications can benefit all patients and the healthcare professionals. This paper starts with introducing data mining and the healthcare paradigm. This study confers various techniques of data mining in healthcare application domain. As the scope of the study is limited to text mining classification, state of art in particular to healthcare text mining classification is studied in detail with suggested improvements. Various issues and challenges owing to the type of data in healthcare are also discussed in detail with possible solutions. Finally, the paper highlights the need for personalized prescriptive systems for patients and healthcare professionals.***

Keywords: NLP, Text Mining, Healthcare, Image Processing, Mobile App, Ontology, Prescriptive analysis, Descriptive Analysis, Predictive Analysis.

I. INTRODUCTION

Data mining is the process of extracting useful, novel, understandable and valid patterns from large databases or data warehouses [58]. This paper concerns Data Mining research work to date in the Healthcare domain. Healthcare mining distinguishes itself from traditional market-driven data mining applications. This paper focuses on the study of Data Mining methods in detail, Healthcare applications and the current state of art, open research challenges and issues in Healthcare Mining.

A. Data Mining in Healthcare

The healthcare industry has emerged with advanced data analytic techniques to find hidden patterns from clinical data. The health industry generates significant sized medical data, leading to giving rise to proper analysis of resources, which can give useful insight using data mining and knowledge discovery (KDD) [1].

B. Data Mining Models

The healthcare industry's expectations are moving from quantitative business to qualitative business by providing enhanced healthcare quality to patients and healthcare professionals, in this view traditional Data Mining models can be visualized with different perspective [2], [3], [4].

- **Descriptive Analytics**

The models analyze data with primary statistical descriptive methods on charts and reports to get meaningful information out of it.

- **Predictive Analytics**

Predictive analytics find relationships in health data and with advanced techniques in data mining it can also detect hidden patterns predict behavior and detect trends.

- **Prescriptive Analytics**

Prescriptive analytics includes both descriptive and predictive analytics. It makes use of domain knowledge along with data or information.

Figure 1 is the evolution of data sources to analysis results. Medical data need important cleaning process and transformations, to be useful [57]. Raw data is assessed, cleaned and transformed for effective results of data analytics techniques. Missing values are treated during pre-processing phase.

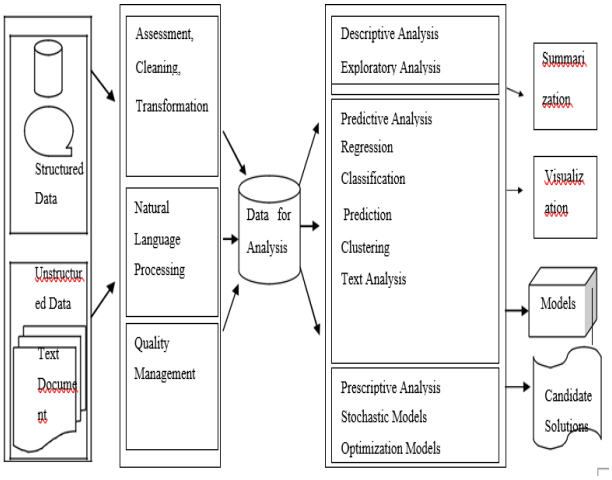

Figure 1Data Analytics Techniques and Data Mining Models

Figure 1 illustrates categorization of Data Analytics techniques as per Data Mining Models and it also highlights representation of analysis results with respect to Data Mining Models. Data Mining techniques will be explained in detail in following Sections.

C. Motivation

Data collected by healthcare information systems can be available in text, graph, images and numeric formats. With intelligent processing many more insights can be revealed which intern may help patients and doctors to receive and provide quality of results respectively. Plethora of work revolves around categorizing the Biomedical research knowledge base by using clustering or classification techniques which can help domain researchers in finding and reading relevant documents [5], [6], [7], [8], [9]. Research in the area of identifying user query intent is flourishing [10].

Revised Manuscript Received on December 14, 2019.

Pranita Mahajan, Assistant Professor, SIES Graduate School of Technology, Mumbai, Maharashtra, India.

Dr. Dipti P. Rana, Assistant Professor Sardar Vallabhbhai National Institute of Technology Surat. Gujarat, India

Recent advancement in Data Mining technologies shows wide scope for systems based on text analytics for patient-centric prevention, awareness and diagnosis. Which gives rise to need of Prescriptive System, which can help decision making process to achieve best possible results.
Major goal is to deliver Person-Centered Care, to find patterns for applications such as, patient/drug similarities and differences. There is high need of technical support which can help not only clinical people but also patients [11].

D. Objectives

The objectives of this paper are to study Data Mining techniques in detail and find its applications in Healthcare domain. Reviewing Health Mining state of art and future challenges of it is the major involvement. This study concentrates on various application domains of Health Mining such as text mining, wearable technologies and image processing.
The rest of the paper is organized as follows. Section 2 is review of data mining in healthcare. Section 3 is detailed literature review of text mining techniques in healthcare. Section 4 explains issues in text mining concerning healthcare domain and the study is summarized.

II. LITERATURE REVIEW OF DATA MINING TECHNIQUES IN HEALTHCARE

This section is a study and review state of art related to Data Mining Techniques in Healthcare. Existing literature is categorized as text based IoT (Internet of Things) based and image-based work. Patient care can be extensively increased while reducing cost by applying various data mining algorithms and techniques which are discussed in this section.

Nisha et al. reviewed literature from 2005-2015 for predictive and descriptive data models in Healthcare domain. It is a detailed survey of data mining techniques in predicting, Liver disorder, probability of readmission, pressure ulcer, degree of abnormality, Parkinson disease. Techniques described in detail are, Mahalanobis Distance (MD), Linear Discriminant Analysis (LDA), Decision Tree classification, Particle Swarm Optimization (PSO), K-Nearest Neighbor, Logistic Regression, Bayesian Classifier, Support Vector Method (SVM). The author has concluded that Healthcare Datasets are highly imbalanced datasets and hybrid approach should be used to enhance the prediction accuracy [12]. Jangwon et al. studied Prescriptive Analysis Framework for Human Care Services Based on CKAN Cloud. Data collected form sensors and cloud is integrated to identify common concept. The system is developed for fire-fighters with the wearable to identify the emergency situation by recording parameters such as heart rate and temperature [4].

Durairaj et al. work is detailed survey of different Data Mining tools and techniques used in Healthcare domain and their impact. It is comparative research with respect to various tools, diseases and type of analysis performed. The author emphasizes on combining more than one technique to achieve better performance [13]. Kashfia et al. proposed cloud based solution in case of emergency. Patient is given a portable device in first visit, all the data related to patient is stored in DB and with the device patient is monitored wirelessly through WiFi and sensors to give them special suggestions, reminders of next visits and sending messages to relatives and local health professionals in case of alarming and emergency situations [14].

Harkiran et al. focused on comparative analysis of descriptive analytical algorithms and their applications in different domains [15]. Javazmaa et al. work is comparative evaluation of K-mean and K-Medoid clustering algorithms to predict probability of occurrence of disease in consideration with the month of the season. Experiments are performed on four season data of Mongolia country. The author has predicted based on facts recorded from 2009 to 2015. After studying correlation between weather and disease month Index, The author has concluded that cardiovascular disease has more effect on people above age 50 and in middle of autumn and Spring. This study is about measuring pathological index by considering weather parameters from 2009 to 2015 [16].

Frantisek et al. developed a system for Heart Disease diagnosis on three freely available UCI Machine Learning dataset. The author developed and compared SVM, Naïve Baye's, Decision Trees and Neural Network algorithms with existing system and claimed that their results are reasonably improved [17]. Kaynath et al. surveyed preventive Healthcare mobile apps on Cardiovascular Diseases during 2011-16. This paper is survey of preventive Healthcare mobile apps on Cardiovascular Diseases during 2011-16. The author reviewed 100 Android and iOS Apps related to diabetes, cancer and cardiovascular diseases. The author developed single mobile app after surveying existing apps and people of different age and gender. The author mentioned they received positive feedback for their app after taking feedback from 100 users. App monitors blood pressure, heart rate, cholesterol, Glucose level and exercise level and generates charts and gives tips for food/nutrition and exercises to be done [18].

Nimna et al. summarize and evaluate the current state of knowledge in data mining techniques which helps to prevent and diagnosis of Non-Communicable Diseases (NCD). The author categorized existing data mining algorithms based applications in diagnosis and prevention of Diabetes, Heart diseases, Breast Cancer and Hypertension. For Diabetes, clustering, association rule, decision tree, regression and ANN based algorithms were compared and The author found that decision tree algorithms gave better accuracy with stress level has most significant feature. For Heart diseases ANN gave better performance highlighting the relationship between disease and features such as drinking, smoking eating vegetables and physical activity. The best algorithm for Breast cancer prediction was decision tree. For Hypertension The author compared CART, CHAID, Exhaustive CHAID and discriminant analysis [19].

Javzmaa et al. reviewed Cluster based algorithms for prevention of NCD by studying relation between atmosphere and disease. The author evaluated performance of K-mean and K-Medoids by using Hopkins statistics, silhouette coefficient and rand index measures[16].

This study categorizes data mining healthcare literature based on their applications into three broad domains as Text Mining in Healthcare, which can further be classified into Text Summarization and Question – Answer System, IoT in Healthcare and Medical Image Processing.

A. Text Mining in Healthcare

Critical clinical information, extracted from EMR and physician's notes can be used for knowledge discovery using text mining techniques. Aaron and William's surveyed various papers in biomedical text mining and categorized existing work as Synonym and abbreviation extraction, Named Entity Recognition, Relationship extraction, hypothesis generation and Text Classification. The author has identified lack of interdisciplinary coordination and cooperation in the field of Biomedical Text Mining [5]. Existing literature on text mining can be further categorized based on application into Text Summarization and Question Answer based systems.

- **Text summarization for Clinical Research**

NLP have been used for various application in healthcare system. For example [21] identifies Biomedical concept by using ontological concept and Named Entity Recognition techniques. They have automated the process of disease diagnosis by applying K-Nearest Neighbor (KNN) algorithm for Cluster Analysis for document clustering, KNN algorithm to find best suitable document for discharge sheet and Vector space model to find word to corpus. Recognizing biomedical ontology concepts in full text journal articles using deep learning techniques originally developed for machine translation is a two-stage concept recognition system, which is a conditional random field model for span detection followed by a deep neural sequence model for normalization, improves the state-of-the-art performance for biomedical concept recognition. Treating the biomedical concept normalization task as a sequence-to-sequence mapping task similar to neural machine translation improves performance [21]. Parser based system produced better performance with less ambiguity [6].

Review of application of various clustering algorithms in Biomedical domain is done by [7]. The Interaction Network Ontology (INO) is one such method to identify interaction between keywords using SPARQL queries [20].

Biomedical expert finding, BMExpert is language based model which finds expert considering relevance, importance and associations. Authors have claimed it to be better than JANE, GoPubMed and eTBLAST [22]. A CNN-LSTM attention predict user intents using an unsupervised clustering method. The author claims better results over baseline models [23]. CRF based model have shown satisfactory performance in extracting important clinical parameters [9]. Colorado Richly Annotated Full-Text (CRAFT) Corpus is a collection of biomedical journal articles selected bioinformatics resource. CRAFT identifies concepts listed as: the NCBI Taxonomy, the Cell Type Ontology, the Protein Ontology, the Sequence Ontology, the entries of the Entrez Gene database, the Gene Ontology and the Chemical Entities of Biological Interest ontology [24]. Convolutional Neural Network (CNN) along with feature engineering and hyperparameters tuning gave better results [25]. Table -I is work done in text summarization for clinical research.

Table- I: Text Summarization State of Art

Sr. No	Paper Title	Technique /Algorithm	Dataset	Advantages	Improvements /Drawbacks
1	A Text Mining Approach To Automated Healthcare For The Masses	KNN, Vector space model	http://www.ncbi.nlm.nih.gov/pmc, Patient's discharge sheets	User can place written or oral query. Identified disease and treatment based on word similarity between summary and corpus	Need of Semantic similarity calculation.
2	Biomedical Concept Recognition Using Deep Neural Sequence Models	LSTM Model with CRF, OpenNMT	(CRAFT) corpus	Pre-processing of medical text data	Recognized concepts can be further used to identify relations between symptoms and diseases
3	Text Mining From Biomedical domain Using A Full Parser	NL processing to identify actor, actee andcondition	A collection of MEDLINE journal abstracts	Forms cluster of documents based on semantic similarities between the documents	Use of grammar mining for improved postprocess
4	Text Mining In Biomedical Domain With Emphasis On Document Clustering	Wordcloud, Hierarchical clustering	5,644 Title, abstract and keywords of SCOPUS, PubMed IEEE Xplore, ABI/Inform, ACM-DL journals.	Keyword based capturing Biomolecules interactions	Combined clustering and classification can make results more useful to researchers
5	The Interaction Network Ontology Supported Modeling And Mining Of Complex Interactions Represented With Multiple Keywords In Biomedical Literature	Support Vector Machine	INO, VO	Improved results by finding the similarity between query and document. System finds expert and importance of document	Machine learning algorithm can be used to analyze multiple sentences and their interactions

6	BMExpert: Mining MEDLINE For Finding Experts In Biomedical Domains Based On Language Model	Weighted Language Model	ISMB dataset	Intent in the query detection	Reformulation and expansion of input queries
7	An CNN-LSTM Attention Approach To Understanding User Query Intent From Online Health Communities	Sentences clustering with DBScan. CNN-LSTM	HCQI	Feature engineering resulted in improved dataset quality	Does not generalize well.
8	Extracting Formulaic And Free Text Clinical Research Articles Metadata Using Conditional Random Fields	CRF ++	Collected dataset from PubMed Central articles	The two relations between the documents are identified, COREF (coreferentiality) and APPOS (appositive)	Pre-processing with NL techniques will improve the accuracy and efficiency
9	Concept Annotation In The CRAFT Corpus	Ontology based semantic similarity calculation	CRAFT Corpus,	Hallmarks are used as class labels, Comparative experiments on CNN base and CNN tuned	Can be extended to find semantic relations between Entities
10	Cancer Hallmark Text Classification Using Convolutional Neural Networks	SVM with Bag of Words Features and Rich Features, CNN	Corpus of 1852 Biomedical publication, Abstracts Annotated		Can be further extended for other disease text classification.

Table II: Literature Review of Question – Answer System

Sr. No	Paper Title	Technique/ Algorithm	Dataset	Advantages	Improvements/ Drawbacks
1	Training IBM Watson Using Automatically Generated Question - Answer Pairs	DeepQA Engine provides appropriate actions to be performed. It is not Domain specific	Free Base Knowledge	Parallel decomposition of question increases accuracy	Does not generalize for medical domain
2	A Survey Of Medical Question Answering Systems	NL pre-processing, CRF for keyword identification, SVM for question Classification	AskHermes MedQA HONQA	Comparative result analysis of QA system on three different data sets	Survey paper
3	AskHermes: An Online Question Answering System For Complex Clinical Questions	Longest Common Subsequence	PubMed repository	Natural language queries can be answered	Manual identification of evidence component
4	BIOAMA: Towards An End To End Biomedical Question Answering System	Indri: A language-model based search engine for complex queries	BioASQ and SquaD dataset	Ranking with unsupervised candidate identification to identify best answer for given question	Only yes/no type questions
5	A Novel Approach For Medical Assistance Using Trained ChatBot	Rule based system	--	Given the name and manufacturing company , the system can describe the medicine	Static rule based system
6	MEANS: A Medical Question-Answering System Combining NLP Techniques And Semantic Web Technologies	CRF and SVM SPARQL to query RDF	MESA Ontology : Medical question Answering	Hybrid method to find global result	Finding syntactic In question and decomposing it to answer complex questions
7	Natural Questions: A Benchmark For Question Answering Research	Entity identification and word similarity	Created question-answer dataset	Not domain specific, generates long and short answers for given question	Can be used with Corpus for medical domain
8	A Medical ChatBot		UCI dataset	Android App for patients to ask query	

Table III: Survey of IoT in Healthcare

Sr. No	Paper Title	Technique/ Algorithm	Dataset/Sensor	Advantages	Improvements/ Drawbacks
1	Automated Medical Diagnosis from Clinical Data	Bag-of-words and TF/IDF. Indri Toolkit	Set of discharge sheets, PubMed Central (PMC).	Simple system to find association between discharge sheets and disease	Use of advanced Analytical techniques can be used to extend it to predictions
2	Profiling of Artificial Breathalyzer To early Diagnosis of Non- communicable Diseases	C4.5, SVM and Back Propagation NN	Samples collected from healthy and non-healthy people	Accuracy is high	Not portable
3	MediNet: A Mobile Healthcare Management System for the Caribbean Region	Text data transmission using mobile phone app	Patient's HER	Patient interface, web interface and electronic diary with healthcare provider	More of information transfer based than analytical
4	An Evaluation of Smartphone Apps for Preventive Healthcare Focusing on Cardiovascular Disease	Feature selection based on importance	Appmetadata collected from android and iOS	Categorical review of mobile apps android and iOS platform	More analytical Techniques will give better insight
5	Blood Triglyceride Monitoring With Smartphone as Electrochemical Analyzer for Cardiovascular Disease Prevention	If-then-else rules	Triglyceride (TG) sensory Module	Patient checking results side and	More analytical techniques will give better insight
6	IoT Based Health Monitoring Systems	If –then-else rules	Pulse, temperature and sweat sensor	Use of portable and inbuilt phone sensors	More analytical techniques will give better insight
7	Review of Wearable Device Technology and Its Applications to the Mining Industry	Rule based Systems	Smart helmet, clothes, eyewear and smart watch	Detail survey of existing devices and sensors	--

1) *Question - Answer System*

Question - Answer (QA) system retrieves accurate answers for natural language queries. In clinical domain QA systems can assist patients and physicians in action taking. Researchers have utilized IBM Watson, system capable of question answering in natural languages, to implement general purpose QA system. The author has trained IBM Watson using automatically generated question-answer pairs [26]. QA system developed by Yong Gang et al. is a comparison of four biomedical systems MiPACQ, AskHermes, HONQA and MedQA. System is five phase system, Question classification, Query Generalization, Document Retrieval, Answer Extraction and Text Summarization [27].

AskHERMES uses structural clustering to identify multiple answer topics, LCS-based ranking to identify best answer [28]. BioAMA: "Biomedical Ask Me Anything" focus on summary type questions. The author proposed a novel Natural Language Inference (NLI) based framework to answer the yes/no questions. In first phase system make use of NER taggers and ranking and generating a candidate set using using both supervised and unsupervised techniques [29].

Chatbots are current trends in providing medical assistance, [30] is a system with artificial intelligence predicting disease given symptoms. It can also give the composition of medicine. It is simple rule based system. Advanced version of this system in implemented by using SVM and word order similarity to provide personalized results considering features such as age of a patient [31]. To achieve more precise results medical ontologies are used by researchers, developed system pre-process question with NL pre-processing and extracts relation between the symptoms and diseases, medicine with SPARQL query on RDF ontology [32], [33]. Table - II is literature review of existing QA system in medical domain.

The right diagnosis will also lessen the need of hospitalization. IoT lead to provide remote medical assistance, tracking and alerts, data assortment and analysis. There is tremendous growth in the form of mobile apps, wearable devices, smart clothes etc.

B. IoT in Healthcare

Table III is in detail survey of IoT based systems for automated medical data analysis. Literature shows high need of data analytical techniques to be used in IoT based systems. Table - III contains survey of wearable based, sensor based and mobile based systems. Most of the work is based on sensors or wearable devices using which data can be collected, these samples are then compared with existing annotated corpus to predict the results [38], [39], [40]. On the other hand there are systems which make use of inbuilt sensors of mobile device and provide user interface to doctors and patients [34], [35], [36], [37]. User can post a text based query from mobile phone, these queries are processed with techniques such as bag-of-words, TF-IDF etc to identify and extract meaningful features, these features are compared with the existing dataset and using classification algorithms such as C4.5, SVM and back propagation, diseases can be predicted and treatments are suggested to the patients.

C. Medical Image Processing

Image processing has a wide application scope in medical domain, such as identifying Birth, Craniofacial Fractures, Defects Brain Tumour Detection, Breast Cancer Detection, Diagnosis Heart Valve Diseases, Lung infections, Congenital Heart Defects and many more.

Popularly used algorithms are SVM, Neural Networks, K-mean, Fuzzy clustering, PCA [41]. General flow of the analysis is Image Pre-processing and Segmentation, Feature Reduction and Classification. AS images are used as input much of efforts are needed for feature extraction. Researchers are working on various methods such as filtered and reconstructed Features, Autoencoder Features, Regularization and PCA [42]. The next paradigm in image processing is combining the results of image analytics with textual reports to increase the accuracy [43]. In this Section data mining techniques in healthcare are explained. Existing literature in healthcare topic modeling, Question-Answer system, Mobile Apps, wearable devices and IoT based research are studied in detail.

III. RESULTS AND DISCUSSIONS

Study shows dramatic increase in textual data generated by health organizations and users giving high opportunities to get the insight of this text data by applying text mining techniques. This Section describes several of the most fundamental text mining tasks. As discussed in earlier Sections various data mining techniques such as regression, clustering, anomaly detection, association mining and classification can be applied on clinical data to identify hidden knowledge and provide better service to patients and healthcare organizations. Scope of this study is limited to classification; this Section gives detail process of classification of medical text to predict diseases from extracted symptoms.

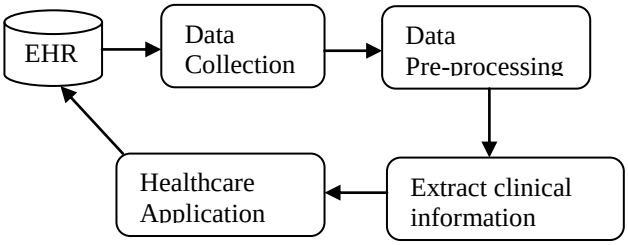

Figure II EMR data processing flow

Figure II shows general process of EMR data processing. Data stored in hospital information system is collected, pre-processed, analyzed and evaluated for specific operations. Every block of this process is a complex task, which will be explained in following sections of this section, EMR database is composed of inconsistence and incomplete data and hence needs pre-processing to ensure accuracy, consistency and most important privacy. Text classification using Machine-learning algorithms automate feature selection techniques to identify meaningful features from EMR. Features are ranked statistically for example correlation and similarity with respect to corpus. A plethora of work is available on feature ranking and selection using bag-of-words, feature identification with semantic similarity, ontology-guided feature ranking. Study shows better results for ontology-guided feature ranking. In medical domain, use of ontologies is prominent when it comes to finding relation between concepts and their ranking [45]. Authors developed Unified Medical Language System (UMLS) based feature ranking and regulation achieve better classification accuracy. Natural language processing (NLP) technologies such as part-of-speech tagger, chunking, NER and context detection are used to convert complex unstructured clinical information in to machine readable structured form [46]. Text data is rich with multiple features; important features are selected with dimensionality reduction techniques such as MaxEnt, Naive Bayes and SVM classifiers and PCA. Identified features/symptoms from EMR has high potential of predicting the disease, the next phase of process is to analyze these extracted symptoms to predict disease. Most of the related work has used classification techniques to achieve this goal [44], [45], [46]. Literature shows good results with classification techniques such as KNN, SVM, HMM, decision tree, Naive Bayes and CRF [47].

A. Issues and Challenges in Healthcare Text Mining

Previous section is detailed study of current state of art in healthcare domain. As discussed in earlier sections healthcare data has its own features and challenges, this section gives detail overview of issues and challenges related to healthcare data, Electronic Medical Record (EMR) [44].

- **Information Extraction from EMR**

The process of extracting knowledge from text constitutes, Information retrieval, information extraction, knowledge discovery and knowledge application. Methods such as Named-Entity Recognition (NER) and Relation Extraction (RE) need to be performed on medical data to prepare it for further analysis. Unstructured data need to be preprocessed to find valuable insights and patterns.

1) Data Cleaning – Missing value imputations, Normalization are few of the techniques to clean data while pre-processing.
2) Noise Processing – Irrelevant data need to be identified and corrected or removed.
3) Inconsistent Data Processing – Technique such as normalization can be used to achieve consistency while merging various data sources.
4) Data Integration – As medical data can be stored at various places and in various formats, it is important to integrate cautiously.
5) Heterogeneous Data Processing – Data can be in the form of image, pathological reports, clinical summaries leading to heterogeneity.
6) Redundant Data Processing – AS data is merged information may get redundant.
7) Data Reduction – When dealing with unstructured data, one of the important task is to find important and meaningful features.
8) Data Transformation – There may be need to transform data from one form to another, for example X-ray images can be transformed to summaries of values.

9) Privacy Protection – As medical data is sensitive data lot of care need to be taken to maintain anonymity of patients.

IV. CONCLUSION

Existing literature scope can be extended for disease prediction and early stage detection which can greatly help in reducing death penalties caused by diseases which can be cured when detected in early stage such as Non Communicable diseases (NCD). Here we highlight NCD as major concern disease domain where text mining techniques can show significant impact. State of art revolves around a particular disease, there are many parameters which can be collectively utilized to predict and guide for set of diseases. Extensive use of EMR systems has given rise to abandoned available information related to medical stakeholders. This data in combination with patient's history can be utilized to design a system which can work as single interface for patient to make them aware and prevent from life threatening diseases which can be cured if detected in early stages such as Non Communicable Diseases.

Relevant research gaps and future challenges as discussed here briefly.

Scarcity of open access corpus for training and testing. Discussed literature shows high need of healthcare text corpus which can be utilized to perform analytics to get better insight.

Need of predictive system for personalized medical system. Personalized medical system is the new open domain in healthcare system. With the technology at their fingertip people prefer services to be available at their doorstep which gave rise to online transactional system.

These systems can be extended to provide précised healthcare solution to the people, which greatly help elderly people and bedridden patients. As medical predictions are highly influenced by medical history of a patient, giving rise to need of storing personalized medical data to achieve high precision.

REFERENCES

1. Parvez Ahmad, Saqib Qamar, Qasim Afser Rizvi, "Techniques of Data Mining In Healthcare: A Review", International Journal of Computer Applications (0975 – 8887) Volume 120(15), June 2015.
2. Harkiran Kaur, Aanchal Phutela, "Commentary Upon Descriptive Data Analytics", Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018) IEEE Xplore, 2018.
3. Babic, Frantisek, Olejár, Jaroslav, Pella, Zuzana, Paralic, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", 2017 Federated Conference on Computer Science and Information Systems, Jan. 2017.
4. Jangwon Gim, Sukhoon Lee, Wonkyun Joo, "A Study of Prescriptive Analysis Framework for Human Care Services Based On CKAN Cloud", Hindawi Journal of Sensors, 2018.
5. Aaron M. Cohen, William R. Hersh, "A Survey Of Current Work In Biomedical Text Mining", Briefings In Bioinformatics. Volume 6(1) 57–71. March 2005.
6. P. Govindarajan, K. S. Ravichandran, "Text mining from biomedical domain using a full parser," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, IEEE Xplorer 2016.
7. Renganathan V., "Text Mining in Biomedical Domain with Emphasis on Document Clustering", Healthcare Information Research, Volume 23(3), 2017.
8. Z. Jiang, L. Li, D. Huang, Liuke Jin, "Training word embeddings for deep learning in biomedical text mining tasks", 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, 2015.
9. Sein Lin, "Extracting formulaic and free text clinical research articles metadata using conditional random fields", Louhi '10 Proceedings of the NAACL HLT Second Louhi Workshop on Text and Data Mining of Health Documents, June 2005.
10. R. Cai, B. Zhu, L. Ji, T. Hao, J. Yan and W. Liu, "An CNN-LSTM Attention Approach to Understanding User Query Intent from Online Health Communities", 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017.
11. Roghayeh Azimzadeh, Leila Valizadeh, Vahid Zamanzadeh, Azad Rahmani, "What are important for patient centered care? A quantitative study based on perception of patients' with cancer", Journal of Caring Sciences, Volume 2(4), 2013.
12. Neesha Jothia, Nur'Aini Abdul Rashidb, Wahidah Husain, "Data Mining in Healthcare – A Review", The Third Information System International Conference, Published by Elsevier, 2015.
13. M. Durairaj, V. Ranjani, "Data Mining Applications In Healthcare Sector: A Study", International Journal Of Scientific & Technology Research Volume 2(10), October 2013. Khondaker A. Mamun, "Cloud Based Framework for Parkinson's Disease Diagnosis and Monitoring System for Remote Healthcare Applications", Future Generation Computer Systems, Volume (6), November 2015.
14. H. Kaur, A. Phutela, "Commentary upon descriptive data analytics," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018.
15. Tsend, Javzmaa, Oyunbileg, Bat Enkh, Luvsan, Ajnai, Tsagaan, Baatarkhuu, "Comparative study based on famous clustering algorithms of non-communicable disease prevalence in Mongolian urban area", Proceedings - 2017 International Conference on Green Informatics, ICGI 2017.
16. Babič, František, Olejár, Jaroslav, Vantová, Zuzana Paralič, Ján, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", Proceedings of the Federated Conference on Computer Science and Information Systems, 2017.
17. K. Zaman, K. A. A. Mamun, "An evaluation of smartphone apps for preventive healthcare focusing on cardiovascular disease," 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, 2017
18. Nimna Jeewandara, Ppg Dinesh Asanka, "Data Mining Techniques In Prevention And Diagnosis Of Non Communicable Diseases", International Journal Of Research In Computer Applications And Robotics, Volume 5, Nov 2017.
19. Arzucan, "The Interaction Network Ontology supported modeling and mining of complex interactions represented with multiple keywords in biomedical literature", BioData Mining, 2016.
20. Negacy D. Hailu, Michael Bada, Asmelash Teka Hadgu, and Lawrence E., "Biomedical Concept Recognition Using Deep Neural Sequence Models", bioRxiv 530337; doi: https://doi.org/10.1101/530337, Jan 2019.
21. Beichen Wang, Xiaodong Chen, Hiroshi Mamitsuka, Shanfeng Zhu, "BMExpert: Mining Medline For Finding Experts In Biomedical Domains Based On Language Model", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Volume 12(6), November/December 2015.
22. R. Cai, B. Zhu, L. Ji, T. Hao, J. Yan and W. Liu, "An CNN-LSTM Attention Approach to Understanding User Query Intent from Online Health Communities," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017.
23. Michael Bada, "Concept annotation in the CRAFT corpus", BMC Bioinformatics, 2012.
24. Baker S, Korhonen A, Pyysalo S. Cancer hallmark text classification using convolutional neural networks. Proceedings of the fifth workshop on building and evaluating resources for biomedical text mining (BioTxtM2016), 2016.
25. J. Lee, G. Kim, J. Yoo, C. Jung, M. Kim and S. Yoon. Training ibm watson using automatically generated question-answer pairs. arXiv preprint arXiv:1611.03932, 2016.
26. Samrudhi Sharma, Huda Patanwala, Manthan Shah, Khushali Deulka, "A survey of medical question answering systems", International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869, Volume 3(2), February 2015.
27. YongGang Cao, "AskHERMES: An online question answering system for complex clinical questions", Journal of Biomedical Informatics, Volume 44, 2011.
28. Vasu Sharma, Nitish Kulkarni, "BioAMA: Towards an End to End BioMedical Question Answering System", Published in BioNLP, 2018.
29. Rashmi Dharwadkar, Neeta A. Deshpande, "A Medical ChatBot", International Journal of Computer Trends and Technology (IJCTT) – Volume 60(1), June 2018.

30. Divya Madhu, “A Novel Approach for Medical Assistance Using Trained Chatbot”, International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017.
31. Asma Ben Abacha, “MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies”, Published in Inf. Process. Manage, 2015.
32. Tom Kwiatkowski, “Natural Questions: a Benchmark for Question Answering Research”, Transactions of the Association of Computational Linguistics, 2019.
33. V. S. Pendyala, S. Figueira, "Automated Medical Diagnosis from Clinical Data", 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), San Francisco, CA, 2017.
34. D. A. P. Daniel, K. Thangavel, "Profiling of artificial Breathalyzer to early diagnosis of non-communicable diseases", 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2015.
35. P. Mohan, S. Sultan, "MediNet: A mobile healthcare management system for the Caribbean region", 2009 6th Annual International Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous, Toronto, ON, 2009. K. Zaman, K. A. A. Mamun, "An evaluation of smartphone apps for preventive healthcare focusing on cardiovascular disease," 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, 2017.
36. J. Wang, X. Huang, S. Tang, G. M. Shi, X. Ma and J. Guo, "Blood Triglyceride Monitoring With Smartphone as Electrochemical Analyzer for Cardiovascular Disease Prevention", in IEEE Journal of Biomedical and Health Informatics, volume 23(1), Jan. 2019.
37. D. S. R. Krishnan, S. C. Gupta and T. Choudhury, "An IoT based Patient Health Monitoring System", 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, 2018.
38. Mardonova, M, Choi, Y, “Review of Wearable Device Technology and Its Applications to the Mining Industry”, Energies, Volume 11, 2018.
39. N. Goel, A. Yadav, B. M. Singh, "Medical image processing: A review," 2016 Second International Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH), Ghaziabad, 2016.
40. C. Liu, Y. Huang, J. A. Ozolek, M. G. Hanna, R. Singh and G. K. Rohde, "SetSVM: An Approach to Set Classification in Nuclei-Based Cancer Detection," in IEEE Journal of Biomedical and Health Informatics, volume 23(1), Jan. 2019.
41. Anjali Sahu, “Integrating Text Mining with Image Processing”, IOSR Journal of Computer Engineering (IOSR-JCE), 2018.
42. W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, G. Wang, “Data processing and text mining technologies on electronic medical records: a review,” Journal of Healthcare Engineering, volume 2018, Article ID 4302425, 2018.
43. Vijay N. Garla, Cynthia Brandt, “Ontology-guided feature engineering for clinical text classification”, J. of Biomedical Informatics, Volume 45(5), October 2012.
44. Kevin Buchan, Michele Filannino, zlem Uzuner, “Automatic prediction of coronary artery disease from clinical narratives”, J. of Biomedical Informatics, Volume 72©, August 2017.
45. D. Lakshmipadmaja, B. Vishnuvardhan, “Classification performance improvement using random subset feature selection algorithm for data mining”, Big Data Res., 2018.
46. Vateekul, P, Kubat, M., “Fast induction of multiple decision trees in text categorization from large scale, imbalanced, and multi-label data”, In Data Mining Workshops, ICDMW'09. IEEE International Conference, 2009.
47. Xu, Z., Li, P., Wang, Y, “Text classifier based on an improved SVM decision tree”, Physics Procedia, Volume 33, 2012.
48. Ageev, M. S., Dobrov, B. V, “Support Vector Machine Parameter Optimization for Text Categorization Problems”, In ISTA, 2003.
49. Jiang, S., Pang, G., Wu, M., Kuang, L, “An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications”, Volume 39(1), 2012.
50. Al-Shalabi, R., Obeidat, R., “Improving KNN Arabic text classification with ngrams based document indexing”, In Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, 2008.
51. Ting, S. L., Ip, W. H., Tsang, A. H, “Is Naive Bayes a good classifier for document classification”, International Journal of Software Engineering and Its Applications, 2011.
52. Maneesh Singhal, Ramashankar Sharma, “Optimization of Naïve Bayes Data Mining Classification Algorithm”, International Journal for research in applied Science and Engineering Technology (I JRAS ET). Volume 2(8), August 2014.
53. Frasconi, P., Soda, G., Vullo, A, “Hidden Markov models for text categorization in multi-page documents”, Journal of Intelligent Information Systems, Volume 18(2), 2002.
54. Murugesan, Suguna, “Optimization of Hidden Markov Model using Minimum Message Length Estimator”, International Journal of Emerging Technology and Advanced Engineering, August 2013.
55. Gasthaus, Jan, Teh, Yee Why, "Improvements to the Sequence Memoizer" (PDF). Proc. NIPS, 2010.
56. Guduru, N, “Text mining with support vector machines and non-negative matrix factorization algorithms”, Doctoral dissertation, University of Rhode Island, 2006.
57. National Academies of Sciences, Engineering, and Medicine. 2017. Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions. Washington, DC: The Nationa Academies Press.
58. Jiawei Han, Micheline Kamber, Jian Pei Data Mining: Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791.

AUTHORS PROFILE

Pranita Mahajan, PhD Scholar, SVNIT Surat and Assistant Professor at SIESGST. She has completed her M.E (Computer Engineering) from Mumbai University. She has published 10 papers in various conferences and journals. Her area of interest are Machine Learning, Text Mining, NLP etc.

Dr. Dipti P. Rana is Assistant Professor at SVNIT, Surat. She is guiding many PhD and Master’s students at SVNIT, Surat. Her area of interest are, association mining, parallel computing, text mining etc.