Machine Learning Applications in Medical Field: A Review

Rutuja A. Gulhane[1], Sunil R. Gupta[2]
[1]*Research Scholar, PRMIT&R, Badnera, Amravati, India*
[2]*Assistant Professor, PRMIT&R, Badnera, Amravati, India*
[1]gulhanerutuja@gmail.com
[2]sunilguptacse@gmail.com

Abstract— Machine Learning and predictive analytics techniques could transform the entire healthcare industry by providing accurate analysis and predictions related to symptoms, diagnosis, procedures, and medications. Machine learning techniques are broadly used in several fields and predominantly the medical industry has been profited a lot through its usage. With a growing population and large datasets of patient's data being collected, it is more critical that Machine Learning technologies be deployed to provide quick, automated and deeper understanding of the healthcare data to improve patient's wellbeing. With the necessity to lower the healthcare costs and the need for personalized healthcare, the healthcare sector faces challenges in areas like Electronic Medical Record organization, data analysis, software-based diagnosis and disease predictions. This paper presents the perceptions involved, the current related work, and the important open topics for research within the analysis of electronic medical records, using data and text mining techniques. This article also demonstrates how medical will benefit from machine learning and deep learning technology.

***Keywords*— Machine Learning, Deep Learning, Data Mining, Text Mining, Electronic Medical Records (EMR).**

I. Introduction

Analysis of medical records is a challenging field demanding substantial effort and time [1]. Many sources of data such as symptoms, patient history, various tests, treatments, and medications need to be considered for an accurate analysis. In addition to this, the analysis requires knowledge in various other fields, namely, the medical specific areas, hospital and clinical procedures, text mining, data mining, medical records. Also, mining medical records and developing a process to do it is complex. Mostly these records are in the form of free texts, can be in the much-unstructured format and can be from a complex and specific domain. In fact, each physician has typically his/her own style of defining symptoms, based on his/her previous medical practices and learning experiences [2]. There is a need for additional tools to understand the designated terms and symptoms and to extract information from medical records as each physician uses a specific language [3].

In the need for meaningful use of patient data, the popularity of electronic health records (EHR) and electronic medical records (EMR) has been increasing lately. The scrutiny of EMR is a major task since they are usually unstructured and generally presented in the form of plain text and usually have a very specific medical vocabulary. Presently, Electronic Medical Records (EMR) are generally used by medical institutes and hospitals to record patient's health conditions, comprising diagnostic information, medical history, operations performed, and treatment results [4]. EMR contains various medical data which can be classified into three types: structured data, semi-structured data, and unstructured data [5]. EMR has characteristics of diversity, redundancy, incompleteness, and privacy, which make it hard to straightaway do information analysis and mining. Different processing techniques are required for different types of information. Complex and challenging processing procedures are required for semi-structured or every unstructured data such as medical texts that contain more health information, Machine learning (ML) is a kind of algorithm that enables software to forecast the output without being explicitly programmed for it using artificial intelligence to learn automatically to become more and more accurate in predicting outputs and without the involvement of human. The procedures involved in machine learning identify patterns through incoming data and take actions accordingly similar to the procedures of data mining and predictive modeling [6]. At present machine learning is the most emerging sub-domain in computer science and Healthcare is it's the greatest challenging application [7]. The ability of Machine learning to process huge datasets that are beyond the scope of human capability and in real-time adds great value in the healthcare sector. Machine learning can also aid to reliably convert the analyzed data into clinical perception which help physicians to take care of the patient's health resulting in increased patient satisfaction and decreased cost of care.

Data mining (DM), also known as Knowledge Discovery and Data Mining (KDD), is the process of identifying important and implicit patterns from large data sets. Text mining is a sub-domain of data mining [8]. Data mining is the extensive process of extracting valid, potentially useful, understandable patterns in the data. It is the process of identifying and analyzing patterns in huge data sets to extract useful information [9]. In healthcare sector, this information can be useful to improve the quality of treatments, lower costs or increasing revenues for healthcare organization such as a hospital. Text mining is crucial to scientific research as the very large volume of scientific literature being produced every year. More and more new articles are being added on a daily basis resulting in these large archives of online scientific articles. Even though researchers can now easily assess more scientific information, it is more difficult for them to identify articles more relevant to their interests. Thus, mining and processing this enormous amount of text is of great necessity for patients [10]. The more complex structure and larger scale of EMR will make it tougher to process data in EMR, but the economic and social benefits it carries will be more outstanding and the EMR plays a bigger role in the healthcare sector.

The review paper of machine learning applications in medical field is organized as follows: Section I introduces to the development of medical field in text mining, data mining, machine learning, electronic medical records with its need and objective. In Section II, the literature survey on text mining, data mining, machine learning has been discussed. There is a need to understand how the methodology of data mining, text mining and machine learning with its purpose and application and how it can be incorporated in EMR. The summary and discussion of research is stated in Section III. Section IV concludes the topic with the future scope mentioned in Section V.

II. Literature Survey

In the recent years, there is substantial increase in the usage of machine learning for solving biomedical mining and EMRs or EHRs. But still there are some major areas such as text mining and data mining which are used to recognize the disease. Various Novelists defined data collection from several data sources and also discuss uncompromising data explanation with several applications in real time. Studying the various machine learning techniques helps to select the right method for a specific application. The authors in [11] discussed about text mining and its association with data mining.

The healthcare data has grown exponentially from terabytes to Exabyte. The medical data from numerous domains has increased by leaps concerning the volume of the big data. With the advancement of computing power, information technology and HIS (hospital information system), EMR has also been popularized [12]. EMR (electronic medical record) or EHR (electronic health record), refers to medical records which medical staff practices to document texts, graphs, charts, information and other digital information generated by HIS. This record can be stored, analysed, managed, transmitted, and reproduced efficiently. EMR has been established as a treasure box for large-scale data analysis in healthcare sector as numerous sources of medical data including diagnostic history, medications, laboratory test results, are getting available [13].

This paper presents a comprehensive collection of works in the field of text mining, data mining, machine learning applied to Electronic Medical Records (EMR).

TABLE I
TEXT MINING APPROACH IN THE BIOMEDICAL DOMAIN

Sr. No.	Methodology	Purpose	Application	Future Scope
1	CRAB System - Fully integrated text mining pipeline tool [14].	To support chemical health risk assessment in biomedicine.	Cancer Risk Assessment.	The taxonomy can be extended to cover other types of health risks (e.g. allergy, endocrine disruption, among many others) with a minimum of effort to improve the efficiency and quality of chemical risk assessment.
2	PKDE4J - Text mining system that integrates dictionary-based entity extraction and rule-based relation extraction [15].	To extract entities and relations from unstructured text based on a flexible and extensible framework for public knowledge discovery.	PKDE4J can serve as the middleware for many applications. It can be used to create a knowledge graph for knowledge discovery in biomedical field and previously unknown relationships can be discovered with the help of this system.	To improve the speed of PKDE4J for large-scale extraction tasks by incorporating a parallel distributed architecture such as Hadoop.
3	BeFree – Text mining system that helps in extraction of relations between genes and diseases from text and large-scale data analysis using data priority strategy [16].	To identify relationships between biomedical entities with a special focus on genes and their associated diseases. Joint analysis of text mined data with data curated by experts, approaches to both assess data quality and highlight novel and interesting information.	Identification and Extraction of gene-disease, drug-disease and drug-target associations with state-of-the-art performance like depression.	A system with a higher Recall that provides better definition of relationships from the semantic point of view and to explore the identification of contextual information of the relationship.
4	Pathway-based evaluation by text mining using statistical model that subsets from focal adhesion, pathways in cancer, natural killer (NK) cell-mediated cytotoxicity and immunosuppression along with epithelial-mesenchymal transition (EMT) [17].	To enhance the molecular understanding of the predictor gene set using text-mining and significant subpathways related to the early-onset colorectal cancer (CRC) cases.	Rectal cancer susceptibility.	To be validated experimentally on early-onset CRC in terms of dedifferentiation or differentiation, which is underscored in EMT and immunosuppression.
5	PWTEES system - A molecular interaction network by using large-scale text mining of gene and pathway events [18].	To enhance the molecular context of diseases by extracting gene and pathway events (TEES). To identify interactions involving both genes/proteins and pathways by applying pathway NER.	Molecular Pathogenesis of complex diseases like thyroid cancer.	To run PWTEES on whole MEDLINE and PMC Open Access set to generate a large-scale dataset that can provide a searchable disease-sensitive interface for interaction events involving pathways.
6	Molecular profiling of thyroid cancer subtypes by classification evaluation using large-scale text	To identify key genes and biological pathways and to find molecular information associated with different	Molecular biology of thyroid cancer.	Based on the molecular differences between subtypes, biologists can look for better diagnostic biomarkers by using

	mining system [19].	subtypes of thyroid cancer and therapeutic targets.		molecular information to build molecular networks of thyroid cancer subtypes, which would enable further systems biology analyses and stimulate development of targeted therapeutics.
7	MeTAE (Medical Texts Annotation and Exploration) - A knowledge and linguistic-based approach for the extraction of semantic relations between medical entities [20].	To have a high precision in relation extraction by extracting and annotate medical entities and relationships from medical texts and by exploring semantically produced Resource Description Framework (RDF) annotations.	Efficient question-answering systems.	Design a method which automatically extracts contextual information such as the status of the relation (e.g. hypothetical, established known) and information about patients (e.g. gender, age).
8	Turku Event Extraction System (TEES 2.1): A SVM based text mining system for extraction of events and relations from natural language texts [21].	To speed up development and enable application of the system to novel event corpora.	The generalization of event extraction techniques by fully automating task-specific adaptation via automated learning of event annotation rules.	To improve the automated annotation learning system to overcome its current limitations with the ontology concept application.
9	A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. A co-occurrence-based text-mining approach through Naive Bayesian & SVM [22].	To filter false-positive interactions by integrating heterogeneous kinds of evidence from genomic and proteomic datasets.	To explore the biological functions of mouse protein Interactions.	The Naive Bayesian model can perform the genome-scale prediction in the future. It can be helpful for the molecular mechanisms of human disease at a system level.
10	HiPub: a seamless Chrome browser plug-in that automatically recognizes, annotates and translates biomedical entities from texts into networks for knowledge discovery [23].	To recognize genes, proteins, diseases, drugs, mutations and cell lines in texts, and achieve high precision and recall using a combination of two different named-entity recognition resources.	To extract biomedical entity-relationships from texts to construct context-specific networks, and integrates existing network data from external databases for knowledge discovery.	Extracting all relevant information of biomedical publications for the research and use them in this knowledge discovery cycle.
11	Data Processing and Text Mining Technologies on Electronic Medical Records [24].	Information extraction of EMR based on text mining.	Medical Decision Support and Disease Risk Prediction, Mobile Health, Network Medical Treatment, and Personalized Healthcare, Disease Evolution Prediction and Drug Reaction Detection.	Public Annotated Corpus, Professional Dictionary and Knowledge Base, Privacy Protection, Reasonable Selection of Processing Tools have to be addressed for the larger scale and more complex structure of EMR.
12	Text-mining approaches in molecular biology and biomedicine [25].	To extract relevant information using automatic techniques, text-mining and information-extraction approaches.	The identification of biologically relevant entities in free text and the construction of literature-based networks of protein–protein interactions.	Biomedical text mining might provide new approaches for drug discovery that exploit efficiently indirect relationships derived from bibliographic analysis of entities contained in biological databases (e.g. genes, proteins and chemical compounds).
13	Text mining techniques like text summarization, classification, Clustering [26].	To enhance speed and decreases time and effort required to extract valuable information from large amount of unstructured data.	Mining the text and discover valuable information for future prediction and decision-making process.	Design algorithms which will help to resolve issues like domain knowledge integration, varying concepts granularity, multilingual text refinement,

				and natural language processing ambiguity.
14	A series of text mining techniques that conforms to the analytical process used by patent analysts which includes text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification, and information mapping [27].	To automate the whole process, to create final patent maps for topic analyses, and facilitates or improves other patent analysis tasks such as patent classification, organization, knowledge sharing, and prior art searches.	verify the usefulness of segment extracts as the document surrogates and an automatic procedure to produce generic cluster titles.	Patent segments, predict business trends.
15	Text Mining Technique Using Association Rules Extraction (EART) [28].	To automatically extract association rules from collection of documents based on the keyword features.	Find relations between the features such as the disease, its spreading location, victim and extract useful information about the outbreak of disease.	medical domain to focus on the disease treatments (pharmaceuticals), their effectiveness and side effects. Moreover, to visualize the extracted association rules in graphical representation in two or three-dimension association networks.
16	Applies ontologies and named entities Recognition for text mining process using real electronic health records, and apply K-Nearest Neighbors as a white-box lazy method to classify each instance [29].	To support paediatric physicians' decisions in a real context.	Support medical decisions relating to epilepsy diagnosis and classification in children.	Expand the real dataset and deal with dynamic issues for adult epilepsy.
17	The ProMiner method uses a pre-processed synonym dictionary in the domain of protein and gene name detection [30].	To identify potential name occurrences in the biomedical text and associate protein and gene database identifiers with the detected matches.	Rule-based protein and gene entity recognition of fly, mouse and yeast.	Recognition of phenotypic descriptions or functional categories in the development of an application which efficiently detects a multitude of biologically relevant named entities in biomedical text.
18	Dictionary-based approach, which expands the bio-entity name dictionary via Abbreviation Definitions identifying algorithm and improves the recall rate through the improved edit distance algorithm [31].	To exploit the performance of the dictionary-based bio-entity name recognition.	Dictionary-based bio-entity name recognition in biomedical literature.	Improve recall by approximate string match and improve performance by the expansion of the dictionary and post-processing.
19	A named entity recognition system to deal with biomedical domain using Hidden Markov Model (HMM) [32].	To resolve the data sparseness problem using Support Vector Machine (SVM) plus sigmoid.	Deep Knowledge Resources in Biomedical Name Recognition.	Improve performance by investigating more on conjunction & disjunction construction and combination of coreference resolution.

TABLE II
DATA MINING APPROACH IN THE BIOMEDICAL DOMAIN

Sr. No.	Methodology	Purpose	Application	Future Scope
1	Data Mining Technology on Electronic Medical Records (EMR) [33].	To increase Data Quality, Data Sharing and Privacy, EMR Management.	Quality Health Services.	EMR will be useful for privacy protection.
2	Semantics-driven frequent data pattern mining algorithm on semantics-enhanced Electronic Health Records data for effective adverse drug event (ADE) monitoring [34].	To enhance the identification of the drug ADE causality, embedded in noisy data, out of large amounts of heterogeneous data sets to improve human health.	Effective Adverse Drug Events (ADE) monitoring and prediction in a population.	Construct a comprehensive pharmacovigilance Knowledge base, which includes a list of drugs, their cellular and molecular activity profiles, their associated ADEs, and related analytic software for efficient data mining EHRs.
3	A clustering based data mining approach to temporal data in hospital information system [35].	To reuse the stored data for hospital management system.	Hospital management to improve in quality of hospital services.	Detailed analysis of clustering-based data mining system in hospital information system.

TABLE III
MACHINE LEARNING APPROACH IN THE BIOMEDICAL DOMAIN

Sr. No.	Methodology	Purpose	Application	Future Scope
1	A method of machine learning-based named entity recognition (NER) to filter out false recognitions of disease/ gene names [36].	To improve the precision of recognizing gene and disease names using domain dictionaries and machine learning.	Relation extraction between diseases and genes.	Encompass increasing the size of the annotated corpus and enriching annotation.
2	Cox Regression with Correlation based Regularization for Electronic Health Records using cyclic coordinate descent and Alternate Direction Method of Multipliers algorithms [37].	To handle correlation and structured sparsity in high dimensional EHR data effectively.	To provide the healthcare physician with a more thorough understanding on the expected probability of readmission risk for heart failure diagnosed patients from a hospital.	Building cox regression models which can deal with multiple outputs at the same time and combine cox with multi-task learning and multi-output regression models.
3	A Cost-Benefit Analysis of Electronic Medical Records in Primary Care [38].	To estimate the net financial benefit or cost of implementing electronic medical record systems in primary care.	An ambulatory electronic medical record system can yield a positive return on investment to health care organizations.	Risk-sharing arrangements in health care expenditures such as partial capitation, risk pools, and pharmacy withholds.
4	Electronic Health Records & Clinical Decision Support Systems using multivariable logistic regression [39].	To improve clinical care via clinical decision support and electronic guideline–based reminders and alerts.	Better outpatient quality of care.	Translate CDS benefits in randomized controlled trials into national quality improvement.
5	Natural Language Processing using Regenstrief Extraction tool (REX) [40].	To Improve Accuracy and completeness of automated electronic laboratory reporting (ELR) of notifiable conditions.	Improve accuracy of Automated Notifiable Disease in reporting methicillin-resistant Staphylococcus Aureus (MRSA) infection.	Further improvements in completeness and timeliness of ELR by increasing amounts of patient medical data.
6	Predictive modeling with a dedicated medical pre-	To develop a dedicated pre-processing pipeline that is able	Prediction of colorectal cancer (CRC).	Further improve prediction quality by applying

	processing pipeline on routine electronic medical records using machine learning techniques [41].	to address all the aforementioned issues independent of the EMR used. To enhance disease prediction, and early detection and intervention in medical practice.		temporal pattern mining on EMR datasets.
7	Electronic Health Record (HER) identity-management architecture which links the patient's records indirectly via pseudonym identifiers [42].	To link only those records belonging to the same patient, to allow patients to keep certain linkages private, to override privacy rules in special circumstances.	Privacy-Preserving Electronic Health Record (EHR) Linkage to satisfy security needs of patients.	Integrate this approach with general access control mechanisms, and define an identity management approach for linking entire family trees to allow diagnosis of, and research into, genetic diseases.
8	Comprehend Medical: Health Insurance Portability and Accountability Act (HIPAA) eligible Named Entity Recognition (NER) and Relationship Extraction (RE) service launched under Amazon Web Services (AWS) [43].	To reduces the cost, time and effort of processing large amounts of unstructured medical text with high accuracy.	To identify patterns of occurrence and knowledge discovery in order to develop polices and actions to improve health of a group or population. Medical language entity recognition and relationship extraction web service.	To leveraged unstructured clinical data for prediction of high-risk patients and epidemiologic studies on outbreaks of diseases.
9	Machine Learning and the Electronic Health Record (EHR) predictive model [44].	To estimate patient risk for a complicated course of CDI by considering all structured data available in the HER.	Prediction of Clostridium (Clostridioides) difficile infection (CDI).	Investigate the appropriateness of complicated CDI derived solely from the EHR. Integrating both biological and administrative data for improving models of disease progression.

III. Summary And Discussion

In this research, an introduction to the field of data mining, text mining, machine learning has been given. An overview of some of the most fundamental techniques and applications which are extensively used in the text domain has been discussed. This research also overviewed some of important text mining, data mining, machine learning approaches in the biomedical domain. The research subject is covered by a wide range of different mining concepts and techniques in addition to description of the background concepts and paramount techniques related in medical field. The main areas include categorizing of the disease diagnosis, decision support systems, treatments and wellness management systems. Even though, it is difficult to describe all different methods thoroughly, the paper gives an overview of the research in the field of Medicine. Table I provides brief literature about text mining approaches, Table II states work done under data mining approaches in the biomedical domain and Table III provides machine learning approaches using various algorithms in the medical field. With the help of this literature survey, new researchers get the knowledge about which algorithm or methodology is used by previous researchers and it also provides future scope to the particular problem.

Text mining can recognize adverse events in case of high-risk patients and can develop algorithms to propose proceedings. However, developing a process capable of proposing proceedings involves a complex understanding of several knowledge fields, such as medical symptoms, clinical texts, medical proceedings, data mining, text mining, and machine learning and so on. The identification and extraction of data may offer a huge complexity in this field because of the unstructured form of information. However, according to many researchers, text mining techniques to analyse medical information can largely reduce the time and effort to diagnose a patient or propose treatments to various healthcare problems. Many researchers observed that the text mining in Electronic Medical Records (EMR) can alert the physicians of the patients on adverse signals on the patients' health and also can suggest on proceedings for high-risk patients; though there are several other techniques and theories that can be applied for the diagnosis of the illnesses.

IV. Conclusion

The aim of this paper is to motivate more researchers to analyse and experiment with machine learning and apply it for solving various medical problems. This paper performed a review of text mining, data mining, and machine learning in the medical field. Based on this paper, the researchers can be informed about which types of medical applications and work has been done on Text Mining, Data Mining, and Machine Learning. Table I, II and III summarizes various methodologies used, their purposes, applications and future scope for specific challenges in the area of text mining, data mining and machine learning under biomedical field. In this, a study of different methodologies used from past years for different purposes in biomedical domain have been discussed wherein, their applications and their future scope are described. The study will help visualize the technology used for emerging health problems that would be useful for public health practice and researchers to solve medical problems.

V. Future Scope

There are lots of future challenges and open problems in dealing with diverse, increasingly large amounts of non-standardized and unstructured information and immense amounts of scattered, heterogeneous, highly dynamic data sets. There is a need to develop and apply innovative tools for the effective analysis, integration and interpretation of multifaceted biomedical data with the intention to build the accurate models for diagnosis and prediction of various types of diseases and their reappearances, which is one of the most important challenges in Healthcare industry. Algorithms must also handle noisy, incomplete, even ambiguous or conflicting information. Hence, effective machine learning approaches becomes crucial in the health care industry to address these challenges.

References

[1] Brown, R. J., & Trimble, M. R., *"Dissociative psychopathology, non-epileptic seizures, and neurology"*, Journal of Neurology, Neurosurgery, and Psychiatry, Vol. 69, No. 3, pp. 285–289, Sept. 2000.

[2] Weiner, E. L., Swain, G. R., Wolf, B., & Gottlieb, M., *"A qualitative study of physicians' own wellness-promotion practices"*, The Western Journal of Medicine, Vol. 174, No. 1, pp. 19–23, Jan. 2001.

[3] Fernandez, A., Schillinger, D., Grumbach, K., Rosenthal, A., Stewart, A. L., Wang, F., & Stable, E. J. P., *"Physician Language Ability and Cultural Competence"*, General Internal Medicine, Vol. 19, No. 2, pp. 167–174, Feb. 2004.

[4] Ludwick, D. A., & Doucette, J., *"Primary Care Physicians' Experience with Electronic Medical Records: Barriers to Implementation in a Fee-for-Service Environment"*, International Journal of Telemedicine and Applications, Vol. 78, No. 1, pp. 22–31, Jan. 2009.

[5] Feldman, K., Hazekamp, N., & Chawla, N. V., *"Mining the Clinical Narrative: All Text are Not Equal"*, IEEE International Conference on Healthcare Informatics, pp. 271-280, Oct. 2016.

[6] Mohssen, M. Z. E., Muhammad, B. K., Eihab, B. M. B., *"Machine Learning Algorithms and Applications"*, published by CRC Press, Taylor & Francis Group, New York, July 2016.

[7] Jordan, M.I., Mitchell, T.M., *"Machine learning: trends, perspectives, and prospects"*, Science Vol. 349, No. 6245, pp. 255–260, July 2015.

[8] Navathe, B. S., & Ramez, E., *"Fundamentals of Database Systems"*, Vol. 6, Pearson Education, 2000.

[9] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., *"From Data Mining to Knowledge Discovery in Databases"*, American Association for Artificial Intelligence, Vol. 17, No. 3, pp. 37-54, 1996.

[10] Cohen, A. M., & Hersh, W. R., *"A survey of current work in biomedical text mining"*, Briefings in Informatics, Vol. 6, No. 1, pp. 57–71, Oct. 2004.

[11] Hearst, M. A., *"Untangling text data mining"*, in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Stroudsburg, USA, pp. 3-10, June 1999.

[12] Luo, J. S., *"Electronic Medical Records"*, Primary Psychiatry, MBL Communications, Vol. 2, No. 13, pp. 20–23, Feb. 2006.

[13] Mcdonald, C. J., *"The Barriers to Electronic Medical Record Systems and How to Overcome Them"*, Journal of the American Medical Informatics Association, Vol. 4, No. 3, pp. 213–221, June1997.

[14] Korhonen A, Séaghdha D O, Silins I, et al. *"Text mining for literature review and knowledge discovery in cancer risk assessment and research"*, Plos One, Vol. 7, No. 4, April 2012.

[15] Song M, Kim W C, Lee D, et al., *"PKDE4J: Entity and relation extraction for public knowledge discovery"*, Journal of Biomedical Informatics, Vol. 57, pp. 320-332, Oct. 2015.

[16] Bravo A., Pinero J, Queraltrosinach N, et al., *"Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research"*, BMC Bioinformatics, Vol. 16, No. 55, Feb. 2015.

[17] Nam S, Park T., *"Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with epithelial-mesenchymal transition"*, Plos One, Vol. 7, No. 4 2012.

[18] Wu C, Schwartz J M, Brabant G, et al., *"Constructing a molecular interaction network for thyroid cancer via large-scale text mining of gene and pathway events"*, BMC Systems Biology, Vol. 9, No. 6, pp. 1-10, Dec. 2015.

[19] Wu C, Schwartz J M, Brabant G, et al., *"Molecular profiling of thyroid cancer subtypes using large-scale text mining"*, BMC Medical Genomics, Vol. 7, No. 3, pp. 1-11, Dec. 2014.

[20] Abacha A B., Zweigenbaum P. *"Automatic extraction of semantic relations between medical entities: a rule-based approach"*, Journal of Biomedical Semantics, Vol. 2, No. 5, Oct. 2011.
[21] Bjorne J, Salakoski T., *"TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task"*, Proceedings of Bionlp Shared Task Workshop, pp.16-25, Aug. 2013.
[22] Li X, Cai H, Xu J, et al., *"A mouse protein interactome through combined literature mining with multiple sources of interaction evidence"*, Amino Acids, Springer-Verlag, Vol. 38, No. 4, pp. 1237-1252, Sept. 2009.
[23] Lee K, Shin W, Kim B, et al., *"HiPub: translating PubMed and PMC texts to networks for knowledge discovery"*, Bioinformatics, Oxford, Enland, Vol. 32, No.18, pp. 2886-2888, Aug. 2016.
[24] Sun W., Cai Z., Li Y., Liu F., Fang S., Wang G., *"Data Processing and Text Mining Technologies on Electronic Medical Records: A Review"*, Hindawi, Journal of Healthcare Engineering Volume, April 2018.
[25] Krallinger, M., Erhardt, R. A., & Valencia, A., *"Text-mining approaches in molecular biology and biomedicine"*, Drug Discovery Today, Vol. 10, No. 6, pp. 439–445, March 2005.
[26] Ramzan T., Muhammad K. H., Shaeela A., and Fakeeha F., *"Text Mining: Techniques, Applications and Issues"*, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 11, pp. 414-418, 2016.
[27] Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I., *"Text mining techniques for patent analysis"*, International Journal of Information Processing and Management, Vol. 43, No. 5, pp. 1216-1247, Sept. 2007.
[28] Mahgoub, H., Rösner, D., Ismail, N., & Torkey, F., *"A Text Mining Technique Using Association Rules Extraction"*, International Journal of Computational Intelligence, Vol. 4, No. 1, pp. 21-28, Jan 2007.
[29] Pereira, L., Rijo, R., Silva, C., & Agostinho, M., "*Using Text Mining to Diagnose and Classify Epilepsy in Children*", IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 345-349, Oct. 2013.
[30] Hanisch D., Fundel K., Mevissen H. T., Zimmer R., Fluck J., "*ProMiner: rule-based protein and gene entity recognition"*, Bmc Bioinformatics, Feb. 2005.
[31] Yang Z, Lin H, Li Y., *"Brief Communication: Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature"*, Computational Biology & Chemistry, Vol. 32, No. 4, pp. 287-291, Aug. 2008.
[32] Guodong Z, Jian S., *"Exploring deep knowledge resources in biomedical name recognition"*, Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and ITS Applications, Association for Computational Linguistics, pp. 96-99, Jan 2004.
[33] Sun, W., Cai, Z., Liu, F., Fang, S., and Wang, G., *"A survey of data mining technology on electronic medical records"*, IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 1–6, Dalian, China, Dec. 2017.
[34] Huang, J., Huan, J., Tropsha, A., & Dang, J., *"Semantics-driven frequent data pattern mining on electronic health records for effective adverse drug event monitoring"*, IEEE International Conference on Bioinformatics and Biomedicine, pp. 608-611, 2013.
[35] Tsumoto, S., & Hirano, S., *"Clustering-based Analysis in Hospital Information Systems"*, Proceedings of the International Conference on Granular Computing, pp. 669-674, 2011.
[36] Chun H. W., Tsuruoka Y, Kim J D, et al., *"Extraction of gene-disease relations from Medline using domain dictionaries and machine learning"*, Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, Vol. 11, pp. 4-5, Oct. 2005.
[37] Vinzamuri, B., & Reddy, C. K., *"Cox Regression with Correlation Based Regularization for Electronic Health Records"*, IEEE, International Conference on Data Mining, pp.757-766, Dec. 2013.
[38] Wang, S. J., Middleton, B., Prosser, L. A., Bardon, C. G., Spurr, C. D., Carchidi, P. J., & Bates, D. W. et al., *"A cost-benefit analysis of electronic medical records in primary care"*, The American Journal of Medicine, Vol. 114, No. 5, pp. 397–403, April 2003.
[39] Romano, M. J., & Stafford, R. S., *"Electronic health records and clinical decision support systems: Impact on national ambulatory care quality"*, Archives of Internal Medicine, Vol. 171, No. 10, pp. 897–903, May 2011.
[40] Friedlin, J., Grannis, S., & Overhage, J. M., *"Using natural language processing to improve accuracy of automated notifiable disease reporting"*, AMIA Annual Symposium Proceedings, pp. 207-211, Feb. 2008.
[41] Kop R., Hoogendoorn M., Teije A. T., et al., *"Predictive modeling of colorectal cancer using a dedicated preprocessing pipeline on routine electronic medical records"*, Computers in Biology & Medicine, Vol. 76, pp. 30- 38, June 2016.
[42] Alhaqbani B, Fidge C., *"Privacy-preserving electronic health record linkage using pseudonym identifiers"*, International Conference on E-Health Networking, Applications and Services, Healthcom, IEEE Xplore, pp. 108-117, July 2008.
[43] Parminder B., Busra C., Mohammed K., Selvan S., *"Comprehend Medical: a Named Entity Recognition and Relationship Extraction Web Service"*, Oct. 2019.
[44] Benjamin Y. Li, Jeeheh Oh, Vincent B. Young, Krishna Rao, Jenna Wiens, "*Using Machine Learning and the Electronic Health Record to Predict Complicated Clostridium Difficile Infection*", Open Forum Infectious Diseases, Vol. 6, No. 5, May 2019.