Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 3188-3196

Research Article

A Machine Learning Artery to Credentials of Syndrome Management Relation Approach

Raja Suganya PV[a], Nalayini CM[b], Sathyabama AR[c]

[a,b,c] Assistant Professor,
[a] Artificial Intelligence and Data Science,
[b,c] Information Technology, Velammal Engineering College.

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: Machine learning has been used in every research and it is an effective tool in the medical field. This kind of automatic learning is used in the extraction of medical knowledge, patient management care and medical decision support. In machine learning the computerized algorithm can be integrated in the health care field to get an efficient and effective medical care. In the proposed system a machine learning based methodology is described for building an application for the identification and propagation of medical information. It extracts the sentences from database containing medical journals that contains various details about disease and treatment information. This task involves the scanning of the database where the unnecessary sentences are eliminated and only the required information is given to the user. It also derives a semantic relation that exists between disease and treatments and the non-informative sentences is not taken into consideration. Hence accurate results are obtained from the database providing effective retrieval. Finally, the result of the task produces an efficient outcome that can be integrated in an application that can be used in medical field.

Keywords: Health care, Machine learning, Natural language processing

1. Introduction

Artificial intelligence has a branch called Machine Learning, whereby the term refers to the ability of IT systems to independently find solutions to problems by recognizing patterns in databases. Empirical denotes the information gained by means of observation and experiments. Hence empirical evidence is information that verifies the truth or falsity of a claim.

People rely on internet for every aspect. All individual care deeply about their health and want to be, now more than ever, in charge of their health and health care. Life is more hectic than has ever been, the medicine that is practiced today is an Evidence-Based Medicine (EBM) in which medical

Proofs is not only based on the number of years of practice but on the latest discoveries carried out. Tools that can help all to manage and better keep follow up of our health such as Google Health and Microsoft Health Vault are reasons and source that make people more powerful when it comes to knowledge in healthcare and management.

The existing health care system is also becoming the trend that rules the Internet and the electronic world. Records related to health are maintained electronically are becoming the standard in the healthcare domain. Researches and studies show that the effective benefits of having an EHR system are:

Health information recording and health data repositories-Diagnosing the patients in emergency conditions, allergies, and lab test results that enable better and time-efficient medical decisions.

Maintaining medical records-Quick access to information related to potential adverse drug reactions, immunizations, supply of medicines.

Support in decision making-The way of capturing and application of quality medical data for decisions in the healthcare domain.

Effective treatments that are meant to specific health needs for patients-Access to information that is focused on certain topics in shorter time span.

In order to achieve the standard of EHR system we need a better, faster, and more reliable access to the information. When the query is typed it hits the Medline database. Medline term which is coined for Medical Literature analysis and retrieval system online. Medline is a collective database of wide variety of articles published on Life science. It consists of journal citations and abstracts for biomedical literature around the world.

There are generally two tasks involved in the work. The first one is identification of sentences published in literature related to Biomedicine to find out whether it is containing or not information about diseases and treatments, and automatically

of their approach is on entity recognition for diseases and treatments. Their representation techniques are based on words used in context, part of speech information, phrases, and a medical lexical ontology. The task involved in the system are extraction of information and relation extraction. Instead of classification of the entire datasets identifying meaningful relations that exist between various diseases and treatments. The second task is mainly based on three semantic relations Cures, Prevent and Side Effect.

The main objective involves showing what Natural Language Processing (NLP) and Machine Learning (ML) methodologies guides us what representation of information and what Classification algorithms are suitable in identifying and classifying relevant medical information in short texts. We acknowledge the fact that tool capable of identifying reliable information in the medical domain stand as the basic building block for the healthcare system to update with latest discoveries. Medical professionals keep need to be up to date with all new discoveries about a certain treatment, in order to identify if it might have side effects for certain types of patients. It consists of around 21 million citations which keeps on updating continuously.

The work carried out in the system presents an extensive study of various ML algorithms and textual representations for classifying short medical texts and identifying semantic relations existing with two medical entities: diseases and treatments.ML shows that in short texts when identifying semantic relations between diseases and treatments a substantial improvement in results is obtained when using a hierarchical way of approaching the task. The output of one task becomes the input for the other. The best part is to identify and eliminate first the sentences that do not contain relevant information, and then classify the rest of the sentences by the relations of interest and not by doing everything in one step by classifying sentences into one of the relations of interest and the other class as uninformative sentences.

2. Related Work

Like the proposed system a related work carried out by Rosario and Hearst approach is used. They used less data sets which helped in the classification of prescribed tasks. The dataset always has sentences from Medline literature abstracts annotated with disease and treatment entities and with eight semantic relations between diseases and treatments. The main target only the informative sentences are taken for classified.

While ascending the process of information extraction and relation the three approaches used are co-occurrences analysis, rule-based approaches, and statistical methods. The methods of co-occurrence analysis are mostly based only on lexical knowledge and context of words, and even though we try to obtain good levels of recall, the amount of precision is low. Rule-based approaches have been widely used for solving extraction of relation in task. Syntactic source of information like part-of-speech (POS) and syntactic structures. Syntactic rule-based relation systems for extraction are complex systems based on additional tools used to assign POS tags or to extract syntactic parsing trees. It has been derived in literature of Biomedicine and such tools are not yet at the state-of-the-art level as they are for general English texts, and therefore their performance on sentences is not always considered as best. Statistical methods tend to be used to solve various NLP tasks when annotated bulk of data are available. Rules are extracted in automation by the learning algorithm when using statistical approaches to solve various tasks. In general, statistical techniques can perform well even with less training data. Identification of informative sentences from medical abstracts involves summarization of tasks and extraction of meaningful information.

3. The Proposed Approach

The proposed system provides an efficient technique for identifying disease and treatment relations. The main enhancement is the usage of Stemming algorithm. When the Query is typed into the search space, however lengthy maybe it does not retrieve all the results relating the query, but the unwanted words gets filtered by means of Stemming algorithm. Stemming process consolidates a keyword by making it into a meaningful query. The verbal words get eliminated and the main keyword is alone considered as query.

Then the query is sent to the Medline database consisting of journal citations. Once the Query hits the Medline then two tasks namely sentence selection and relation identification are performed. In the First task the system automatically identifies sentences published in medical journal's abstract as containing or not information about diseases and treatments, and automatically identifying semantic relations that are found between diseases and treatments, as expressed in these texts. Second task is focused on three relations Cures, Prevent and Side effect. The proposed system uses an enhancement of Rosario and Hearst approach. In that approach all the datasets were classified. But in the proposed system first Informative sentences Vs. Non-informative is obtained. Only the informative sentences are used in the classification. The non-informative sentences are the one which does not contain the semantic relation but contain some information regarding the disease or cure or side effect. Three kinds of classification are formed between cure, prevent and side-effect.

Finally, exact results are given to the user. These processes can be carried out in six modules.

(a) Admin Registration

(b) Affix Removal

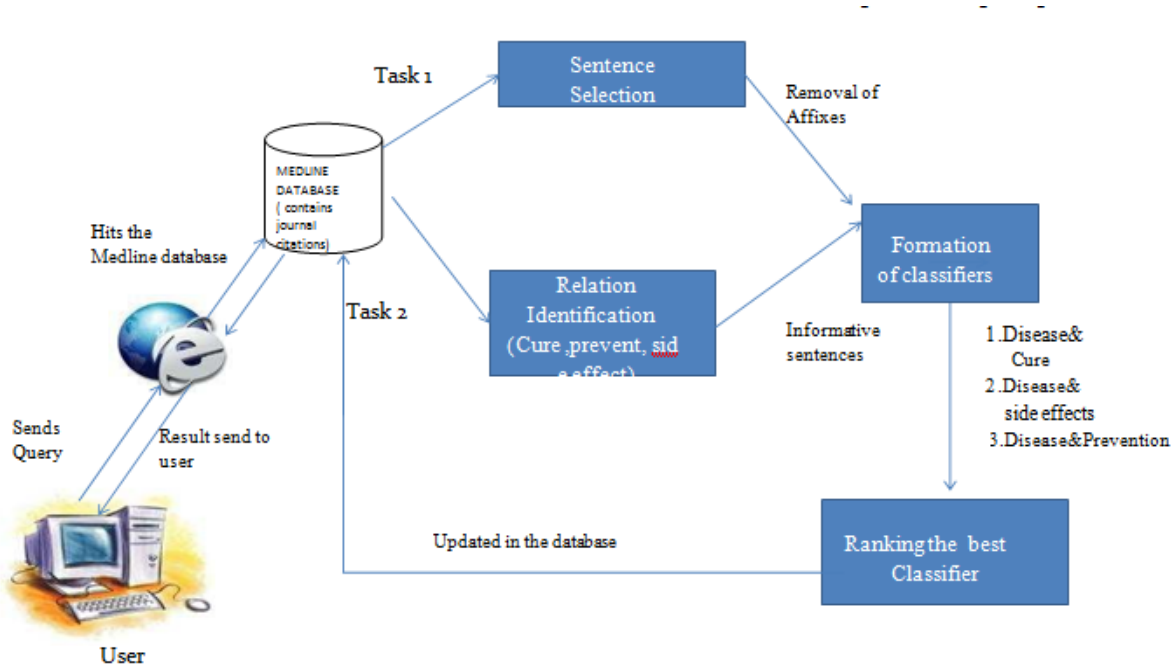
(c) Formation of classifiers

(d) Ranking the best Classifiers.

(e) Comparison and Extraction.

(f) Data Retrieval

The following architecture diagram explains the process of proposed system



Admin Registration

The admin is provided with the user name and password. Only the Admin has full access to login to the system. Only the doctors and Medical professional updates the server. Whenever a citation is found the Medline is updated as the database is not stable and keeps on changing with new citations and journals. An ordinary person with his views about the disease treatment relation is not allowed to update the database server.

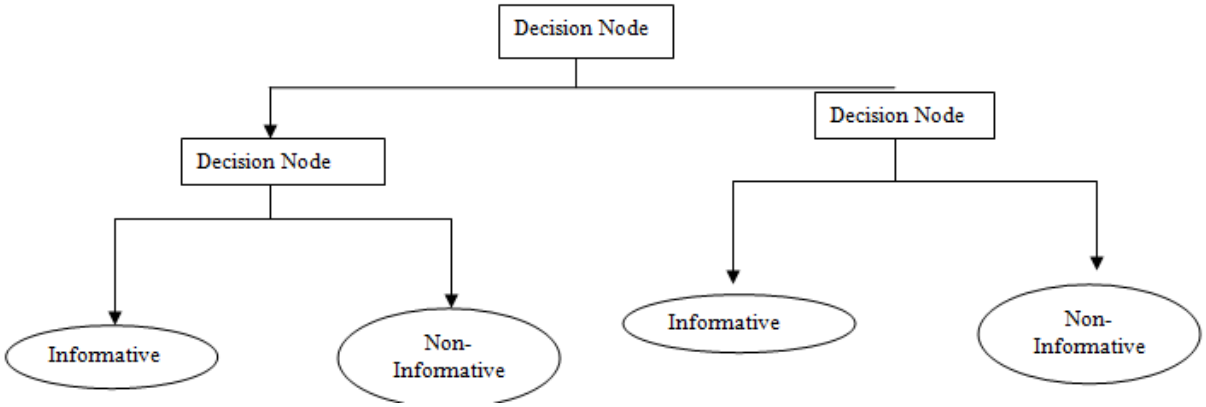
Affix Removal

This is module where the stemming algorithm is applied. People's search involves huge number of queries which may convey the same meaning. Hence by the usage of Stemming algorithm unwanted leaf words are eliminated and only the main keywords are taken into consideration. Leaf words are the words like suffixes or prefixes which are attached to the main keyword. Affix removal conflation technique are referred to stemming process. By using this process, the suffices and prefixes are removed, and the keyword is reduced to a consolidated word. This process suggests the distinction between existing and the proposed system, wherein the existing system no stemming process is involved. The stemming process will remove the suffices one at a time starting at the end of the word and working towards beginning.

We also use Bloom filter in order to find whether an element is a member of set. The process of stemming helps in providing correctness by two actions. Over stemming in which too much of a term is removed and in Under stemming a term which occurs too little of times is eliminated.This helps in finding accurate relation.

Formation of classifiers

The classifiers segment a relation into distinct classes. This involves searching the journals and extracting the information from the journals and storing the result in database server.



In these three kinds of classifiers are obtained. They are:

(i)Disease and Treatment

(ii)Disease and Symptoms

(iii)Disease and Side effect

Whenever a new classifier is formed it gets updated in the database and the best classifier gets the highest rank. Hence three classifiers are formed between treatment, symptoms and side effects we can easily analyze the accurate relations.

In the decision-making system, two types of classification are made based on the informative and non-informative sentences. The informative sentences are the one which consists of semantic relationship with the contents in the med-line data base and the non-informative sentences which consists of some details about the query but not the exact semantic relation with the query.

The informative sentences again classified further to check any relationship exists with the classified sentences. Based on the resulting sentences we again try to extract any meaningful relation existing in association with the query. The further classified sentences are suggested in bags of words (BOW) representation in zero's and one's.

The sentences with zero value are considered as non-information sentences and not further included in the classification again. But the sentences containing one's are used to form three types of classifiers like prevent, cure and side effects. So, the users will get the accurate results based on the effective solutions proposed. These ranking helps in finding accurate solutions.

4. The Proposed Algorithm

1: After detecting number and pause accumulate distinctive words from records.
2: Segregate the glossary into a number of clusters (which is comparable groups) Z1, Z2,Z3,Z4....Zn such that each cluster surrounded by words which is having a specific length of an universal append.
3: For each cluster Zi= $\{r_1,r_2,r_3...r_n\}$ do
4: For every duplet(r_a,r_b) Zi(a<b)do
5: Figure out unwritten-sim(r_a,r_b), Synchronize-sim(r_a,r_b), Prospective-postfix-duplet-freq(r_a,r_b) Figureout equivalent outcome ie., X(r_a,r_b)
6: end For
7: Result equivalent matrix X: for each cluster
8: Build weighted undirected graph vi for class Ci whose s= $\{r_1,r_2,r_3...r_n\}$ and note the edge between the nodesr_a and r_b if potential survival of postfix duplet that convince r_a,r_b or if(r_a,r_b)>0; set weight r(r_a,r_b)=x(r_a,r_b)
9: while Ui! =NULL do
10: Observe the median node(m) with the elevated grade and assign C={V}
11: For all nodes,S Neighbour(V) with elevated grade of weights r(s,v)do
12:C=CV{V}
13: else
14: Detach the fringe between S and V
15: end if
16: end for
17: Mark class C

18: Detach the nodes in C from matrix Si and let the sub matrix be S_i

19: Set $S_i=S_i$

20: end while

21: end for

Ranking the best Classifiers

Ranking is mostly done when a query consists of more answers. The answers are ranked based on the order of relevance and importance. It suggests the relationship between a set of relation such that for any two relation the first is either ranked higher than or ranked lower than or ranked equal to the second. In the system the relation which is used by majority of professionals is given the highest rank and its priority is higher than the others. In this we differentiate between the existing and proposed system. In the existing system if webpage containing a new product is visited again and again by the same person the rating of that page gets increased whether the person may or may not have knowledge about the product. Hence it does not provide the best ranking. But in the case of proposed approach the ranking can be done only by professionals' experts in that field. If a relation is used by many professionals then the relation is given higher priority than the other. Similarly, if the same relation occurs twice or thrice in the same journal then the rating is considered as one only. Hence the new users using the system can easily justify which the best relation exists between the disease and treatment. Lay people can also easily identify the best relation. Existing between the disease and treatment and concluded which to be used.

Comparison and extraction

This module helps to compare the values depending upon the classifiers which is already stored in the database. Whenever a person uses the system he can also view the users who all used the particular relation previously. User can compare the relation which is newly classified with the already used one and obtain the best relation for a particular disease and treatment.

Data retrieval

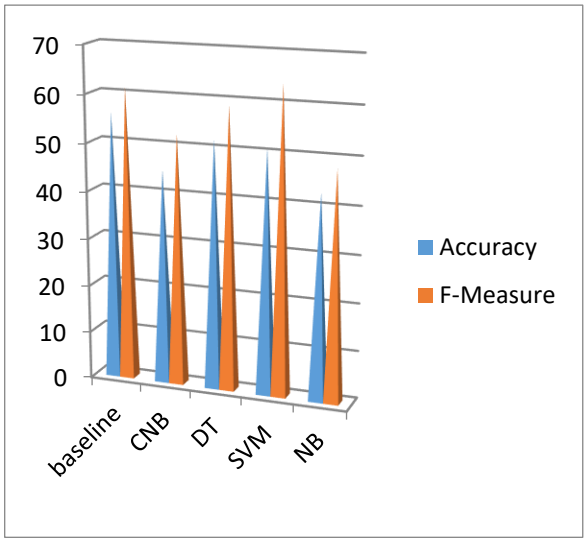
The system uses BOW representation for classification purposes. In the system we use frequency feature representation which suggests the no of times a relation appears in the instance or 0 if it does not appear. By this the user can identify whether a relation existing or not. If present, they get retrieved efficiently.

5. Evaluation And Results

Evaluation measures most commonly used in the ML settings are accuracy, precision, recall, and F-measure. In the system two kinds of results are obtained one for identifying informative sentences and other for the identification of semantic relationship. The evaluation measures Accuracy is the total number of correctly classified instances. Recall is the ratio of correctly classified positive instances to the total number of positives. Precision means correctly classified positive instances to the total number of classified as positive. F-measure is the technique to detect the harmonic mean between precision and recall.

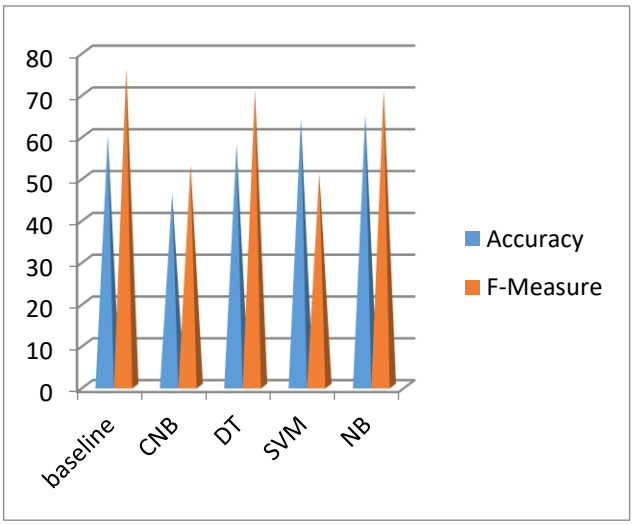
Two results can be obtained for the task of identification of informative sentences and for the task of identification of semantic relationship. It can be used as a pipeline of tasks where the output of task one is used as the input to task two. Two setting can be obtained for easier classification of tasks. The first setting deals with usage of all the sentences, including those that do not contain information about the three relations of interests. The second setting uses the Sentences that contain one of the three Relations. Thus, observation of the second setting also validates the choice of proposing the first task to identify which sentences are informative and which not. For good performance level in the relation classification task, we need to weed out non-informative sentences.

6. Results

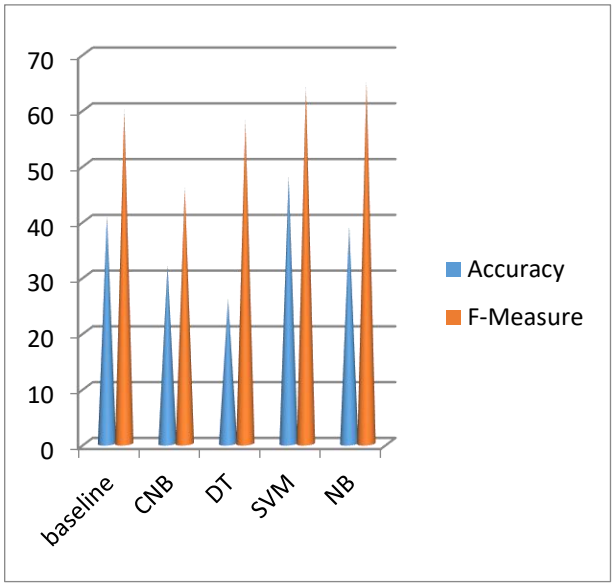


Graph: 1 comparison of accuracy and f-measure when verb phrases are used.

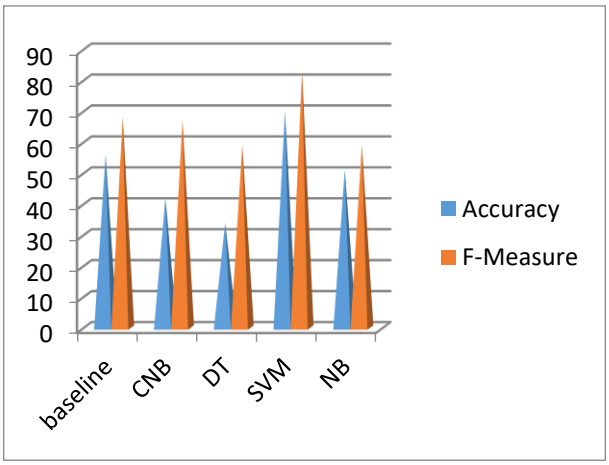
Note: All the values of graphs are rounded off to two decimal points and mentioned as percentage.



Graph : 2 comparison of accuracy and f-measure when biomedical are used.



Graph : 3 comparison of accuracy and f-measure when noun phrases are used.



Graph : 4 comparison of accuracy and f-measure when npl and biomedical are used.

Classifiers	Accuracy	F-Measure
Baseline	53.61	58.71
CNB	42.4	49.9
DT	50.6	57.7
SVM	48.7	60.3
NB	40.6	45.6

Table 1 comparison of accuracy and f-measure when verb phrases are used in tables.

Classifiers	Accuracy	F-Measure
Baseline	58.3	72.6
CNB	44.1	51.6
DT	56.9	69.9
SVM	61.3	47.4
NB	62.1	68.7

Table 2 comparison of accuracy and f-measure when biomedical are used.

Classifiers	Accuracy	F-Measure
Baseline	39.6	57.3
CNB	29.9	43.7
DT	25.7	56.1
SVM	47.6	62.6
NB	37.8	63.7

Table 3 comparison of accuracy and f-measure when noun phrases are used.

Classifiers	Accuracy	F-Measure
Baseline	52.1	63.4
CNB	39.9	63.9
DT	31.6	57.1
SVM	66.1	78.9
NB	48.7	57.6

Table 4 comparison of accuracy and f-measure when npl and biomedical are used.

7. Conclusion

The new type of relation extraction and information retrieval system for the classification of relation is obtained. It is identified that potential improvements in results is obtained when more information is brought in the representation technique for the task of classifying short medical texts. By usage of BOW approach, reliable results on text classification tasks can be obtained. The second task is obtained by solving the first task. Best results are obtained by focusing on three semantic relations between diseases and treatments. Thus an efficient and effective system for retrieval of best relation for disease and treatment is obtained.

References

1. Bunescu.R and Mooney.R ĐTC" Ujqtvguv" Rcvj"Fgrgpfgpe{"Mgtpgn"hqt"Tgncvkqp"GzvtcevkqpÑ."Rtqe0"Conf. Human Language Technology and Empirical Methods in Natural Language Processing vol. 14, no. 4 pp. 724-731,2005.
2. Bunescu.R,Mooney.R,Weiss.SÑSubsequence Kernels for Relation Extraction Advances in Neural Information Processing SystemsÑ." xqn0"3:." pp. 171-178,2006.
3. T.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter, "EDGAR: Extraction of Drugs, Genes, and Relations from the Biomedical Literature," Proc. Pacific Symp. Biocomputing, vol. 5, pp. 514-525, 2000.
4. Cohen A.M. and Hersh W.R. and Bhupatiraju R.T(2004) ĐFeature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage,Ñ Proc. 13th Text Retrieval Conf. (TREC),2004.
5. J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S. Mark, and L. Brilliant, "Detecting Influenza Epidemics Using Search Eng
6. Frunza.O and Kpmrgp0F." ĐVgzvwcn" Kphqtocvkqp"kp"Rtgfkevkpi"Hwpevkqpcn"Rtqrgtvkgu"qh"vjg"Igpgu.Ñ"Proc. Workshop Current Trends in Biomedical Natural Language Processing (BioNLP) in conjunction with Assoc. for Computational Nkpiwkuvkeu"*CEN"Ó2:+,2008.
7. Ginsberg .J, Mohebbi Matthew, Rajan S.P, Lynnette B,Mark U0U.ĐFgvgevkpi" Kphnwgp¦c" EpidemicsUsingUgctej"Gpikpg"Swgt{"Fcvc.Ñ""xqn0"457, pp. 1012-1014, 2009.
8. A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, "Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage," Proc. 13th Text Retrieval Conf. (TREC), 2004
9. J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-Based Learning for Biomedical Relation Extraction," J. Am. Soc. Information Science and Technology, vol. 59, no. 5, pp. 756-769, 2008.
10. L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner Jr., H.L. Johnson, P.V. Ogren, and K.B. Cohen, "OpenDMAP: An Open Source, Ontology-Driven Concept Analysis Engine, with Applications to Capturing Knowledge Regarding Protein Transport, Protein Interactions and Cell-Type-Specific Gene
11. B.J. Stapley and G. Benoit, "Bibliometrics: Information Retrieval Visualization from Co-Occurrences of Gene Names in MEDLINE Abstracts," Proc. Pacific Symp. Biocomputing, vol. 5, pp. 526-537, 2000.
    A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event Extraction from Biomedical Papers Using a Full Parser," Proc. Pacific Symp. Biocomputing, vol. 6, pp. 408-419, 2001.
12. Hunter and Cohen K.B ĐBiomedical Language Rtqeguukpi<" YjcvÓu" dg{qpf" RwdOgfAÑ Molecular

13. Cell, vol. 21-5, pp. 589-594,1998.
14. Kohavi .R and Provost0H"ÐMachine Learning, Editorial for theSpecial Issue onApplications of Machine Learning and the knowledge Discovery ProcessÑ, vol. 30, pp. 271-274,2005.
15. P. Kaur, N. Sharma, A. Singh, and B. Gill, ―CI-DPF: A Cloud IoT based Framework for Diabetes Prediction‖, In IEEE, 2018, pp. 654- 660. [Dig. 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)]
16. K. Chaudhary, O.B. Poirion, L. Lu and L.X. Garmire, ―Deep learning–based multi-omics integration robustly predicts survival in liver cancer‖ in Clinical Cancer Research, 2018, 24(6), pp.1248-1259.
17. V. Krishnaiah, G. Narsimha, N. Subhash, ―Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques‖, (IJCSIT) International Journal of Computer Science and Information Technologies, 2013, 4 (1), 39 – 45.
18. J.H. Oh, R. Al-Lozi and I. El Naqa, ―Application of machine learning techniques for prediction of radiation pneumonitis in lung cancer patients‖, In 2009 International Conference on Machine Learning and Applications (pp. 478-483). IEEE (2019).
19. G.Niranjana, Dr.M.Ponnavaikko, "A Review on Image Processing Methods in Detecting Lung Cancer using CT Images" ,International Conference on Technical Advancements in Computers and
20. Communications,IEEE 2017.
21. Nooshin Hadavi, Md Jan Nordin, Ali Shojaeipour, "Lung Cancer Diagnosis using CT-scan Images based on Cellular Learning Automata",
22. International conference on Computer and Information Sciences (ICCOINS), IEEE, 2014.
23. S. Kalaivani, Pramit Chatterjee, Shikhar Juyal ,Rishi Gupta,"Lung Cancer Detection Using Digital Image Processing and Artificial Neural Network", International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.
24. O. Gevaert, J. Xu, C.D. Hoang, A.N. Leung, Y. Xu, A. Quon, D.L. Rubin, S. Napel and S.K. Plevritis, ―Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results‖, in Radiology, 2012, 264(2), pp.387-396.
25. X. Luo, X. Zang, L. Yang, J. Huang, F. Liang, Rodriguez, ―Comprehensive computational pathological image analysis predicts lung cancer prognosis‖, in Journal of Thoracic Oncology, 2017, 12(3), pp.501-509.
26. T. Hoang, R. Xu, J.H. Schiller, P. Bonomi and Johnson D.H., ―Clinical model to predict survival in chemonaive patients with advanced non–small-cell lung cancer treated with third-generation chemotherapy regimens based on Eastern Cooperative Oncology Group data‖, in Journal of Clinical Oncology, 2005, 23(1), pp.175- 183
27. T. Win, K.A. Miles, S.M. Janes, B. Ganeshan, M. Shastry, R. Endozo, M. Meagher, R.I. Shortman, S. Wan, I. Kayani and P. Ell, ―Tumor heterogeneity as measured on the CT component of PET/CT predicts survival in patients with potentially curable non-small cell lung cancer‖, in Clinical Cancer Research, 2013
28. B. Zhu, N. Song, R. Shen, A. Arora, M.J. Machiela, L. Song, M.T. Landi, D. Ghosh, N. Chatterjee, V. Baladandayuthapani and H. Zhao, ―Integrating clinical and multiple omics data for prognostic assessment across human cancers‖, in Scientific reports, 2007, 7(1), p.16954.