Automating data mining of medical reports

Geetha Mahadevaiah[1], Dinesh M.S [1], Amogh Hiremath[1], Vani Agarwal[2] ,Ponnurangam Kumaraguru[2] and Andre Dekker[3]

[1]Philips Research India, Philips Innovation Campus, Manyata Tech Park, Nagavara Bangalore - 560045, India.

[2]Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi – 110020

[3]Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Dr Tanslaan 12, 6229ET, Maastricht, The Netherlands

Abstract

Medical reports contain large amounts of clinicalinformation which is noteasily mined due to its unstructured and free flowing format. In addition, the medical terminologyis context sensitive and varies between entities.

In this research, a method to convert data from unstructured medical reports into structured format and techniquesto identify cancer cases including classificationon the basis of its type, stage, occurrence in anorgan, are devised.

Different NLP techniqueswereevaluatedfor feature extraction. A machine learning based algorithm for automatic information extraction was developed. The system performs well for malignant/benign cancer classification with 0.89 F1-score.

Semantic structure of reports in the form of ontology was developed, enabling machine comprehensive data storage and retrieval of semantic information. For illustration purpose, Semantic representation of Lung cancer with mapping from clinical report is shown.

Key words

Clinical reports, report classification, cancer classification, RDF files, ontologies, natural language process

1. Introduction

Medical reports contain detailed clinical information about patient's medical condition. They include patient information like findings, impressions, demographics, past medical history of patient, brief hospital course, diagnosis etc. Large portion of these reports are in unstructured free-text format.During the medical journey of the patient, based on the need, medical practitionersaccess information in the medical reportsat various points. During this process, medical practitionersor clinician has to read many reports to gain insights about the patient's condition. There are chances that the doctors might miss critical information which is aggravated by time pressure and resource constraints.

The information in the unstructured text data may provide additional medical insights and therefore are of interest to clinical research community. An example of information one would like to extract from free-text clinical reports is the disease of the patient, which would allow automatic grouping of clinical reports into various diseases for further study. The discharge summary notes of patients are labelled according to ICD (International Classification of Disease) codes[2] . Since 1967, the World Health Organization (WHO) has developed ICD codes to monitor the incidence and prevalence of diseases, observe reimbursements and resource allocation trends, and to keep track of safety and quality guidelines. In the hospital setting, to perform annotation, techniciansshave to first classify reports according to their disease types and ICD codes. In medical field, ontology is used to represent knowledge about symptoms, diseases and treatments. Mapping the disease information to predefined ontologies is done manually which requires manpower and is prone to errors.

There is a need for a system that automatically groupsmedical reports into various diseases to help in early identification of symptoms and treatment of disease. Semantic web and Natural Language Processing provides methods to convert unstructured data into structured format and store data with semantic context, a format that machines can process.Automated process of identification of information from medical reports can help medical practitioners to derive clinical insights and provide treatment in a more precise manner.

In this work, the authors have devised techniques to process and extract information from medical reports, convert them to structured format and map disease information in semantic form. Cancer is among the leading causes of death worldwide[4].Cancer care is multidisciplinary and is suited for wider use of electronic health records to manage oncology data and workflows [3]. There is an increase in the occurrence of cancer with related morbidity, thus the need of the hour is to build an automated system for its early detection. To develop and validate proposed techniques, cancer related clinical reports were used from Mimic database and ICD codes.

2. Methods

This section provides a brief description of related work in the area of automated medical report annotation. It provides details of the methods and techniques of different approaches and their benefits.

2.1. Related Work

To automate the process of tagging medical reports,previous works have used many techniques for automatically tagging ICD9 codes to medical reports.Rule based techniques are used to tag reports by using pattern matching and supervised machine learning algorithms [3]. Dublin S[4], et.al., have used Natural Language Processing techniques to validate pneumonia cases from chest radiographic reports.

Due to unstructured text, there is ambiguity in medical reports and errors such as misspelled words, use of phrases, abbreviations, lexical variations,thus pattern matching methods fail to provide comprehensive results. Rule based systems are similar to the manual annotation of reports. TYanshan Wang et al [5] showed that NLP techniques have gained power and competence compared to rule-based techniques.

Machine Learning methods have also been used in Imon Banerjeeet. Al [6] mainly focuses on word embeddings using an unsupervised hybrid method. Word embeddingsare formed by training word2vec on text data. The method proposed by author combines word embeddings with a semantic dictionary mapping technique for creating a dense vector representation of unstructured radiology reports. Further, they have applied intelligent Word embeddings to generate embedding of chest CT radiology reports from two health care organizations and utilized the vector representations to semi-automate report categorization based on clinically relevant categorization related to the diagnosis of pulmonary embolism (PE).

HeinerOberkampfet. al.[7] implemented a prototype to demonstrate structured representation of findings from unstructured reports which allows successful review and more efficient comparison among various EHRs. The role of the information model proposed is to define the schema according to which the terminology is used. In previous work, Sonja Zillner[8] created a Model for Clinical Information (MCI) that is based on ontologies from the Open Biological and Biomedical Ontologies (OBO) library[9,10]. RadLex [11], and other ontologies are employed as reference terminologies. Further, Oberkampf Het. al. [12] demonstrated how structured representations of measurement findings can be extracted from free-text radiology reports. ThusithaMabotuwana[13], proposeda measure to determine similarity between two individual concepts extracted from a free text document is studied using ontological parent-child (is-a) relationship as matching techniques, as lexicon based comparisons is typically not sufficient to determine an accurate measure of similarity. The addition of semantic context into the document vector space model improves the ability to differentiate between radiology reports of different

anatomical and image procedure-based classes. This effect when leveraged for document classification tasks, can be used for efficient biomedical information retrieval.

Mahadevaiah G[14] demonstrated the use of semantic technology constructs to store clinical relevant features from DICOM (The Digital Imaging and Communications in Medicine (DICOM) standard is widely used in medicine for storing and transmitting medical information.**)** files. Natural Language processing is used for mining and retrieval of information. The proposed technique stores the clinical relevant information,from a DICOM RT dataset, as triples in a Resource Description Framework (RDF) repository.

Literature discussed in above paragraphs describeNLP techniques, to demonstrate the automatic text processing.Butnone of the discussed techniques provide a disease specific classification with the representation of sub-categories of disease as ontology. Proposed work as shown in Figure1, builds an automated text classifier which can identify malignant or benign cancer types and its sub classification by processing unstructured text data. This information is represented as an ontology to enable easy retrieval of insights on a patient'sdisease progression to a clinician.

Medical records from MIMIC-III (Medical Information Mart for Intensive Care III) dataset [1] were used to design and validate proposed techniques. It contains de-identified medical records of patients suffering from various diseases within the intensive care units at Beth Israel Deaconess Medical Center from 2001 to 2012 in free-text format. The MIMIC-III dataset contains ICD9 codes. ICD9 labelling of clinical reports was done manually under the expert supervision of doctors. Manually tagging of medical reports is a labor intensive task.

MIMIC-III dataset contains diverse information. In this work, discharge summaries of patients that are present in free-text format wereused (The NOTEEVENTS table in MIMIC-III contains the discharge summary). Rule-based regular expressions techniques were used to convert these reports into structured format.

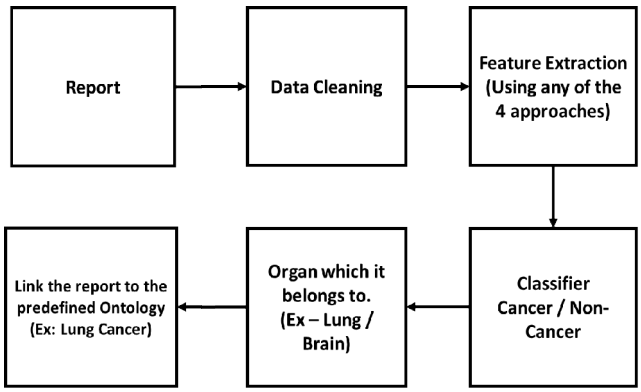

Figure 1. Overview of the approach used for classification

2.2. Data and Ground Truth

The sample reports from MIMIC database, contains discharge diagnoses label from which the findings can derived. Thecorresponding ICD9 codes for the findingsare also stored in DIAGNOSES_ICD table of MIMIC-III database and this is used as a ground truth for cancer identification.

In ICD9 code ranges specified by National Cancer Institute,there are 389 ICD9 codes related to cancer [15,16].The reports were scanned for all the ICD9 codes and valid text and 6228 reports with a cancer related ICD code were shortlisted for further processing. Among the shortlisted 6228 reports, 2500 reports had information on the cancerstage. The distribution of reports based on organ and type of cancer are listed in Table 1 and 2 respectively. Reports were sanitized by removing references to ICD9 codes. Figure 2 shows the overview of the steps followed for ground truth extraction.

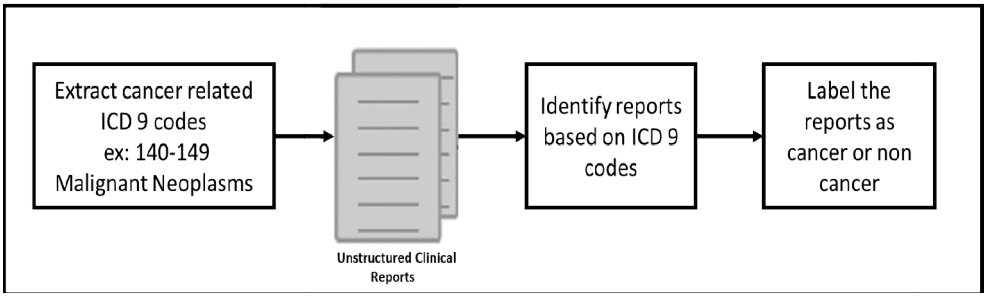

Figure 2.Ground truth extraction overview

CancerType	Count
Lymphoma	537
Sarcoma	327
Mastocytoma	70
Carcinoma	57

Table 1. Cancer Type based count

Organ	Count
Head and Neck	1317
Lung	1177
Lymph Node	709
UrinaryTract	376
Colon	305
Prostate	302
Brain	294

Table 2. Organ based count

2.3. Data Preprocessing

Unstructured reports were converted into structured format by carefully analyzing discharge summaries and processingvarious sections like findings, impressions, medical history etc..The processing was doneusing NLTK - Natural Language Processing Toolkit. This is a Python library software for text processing, such as, case conversion, removing special characters, canonizing numerals and tokenizing. The processingsteps aredescribedin Figure 3.

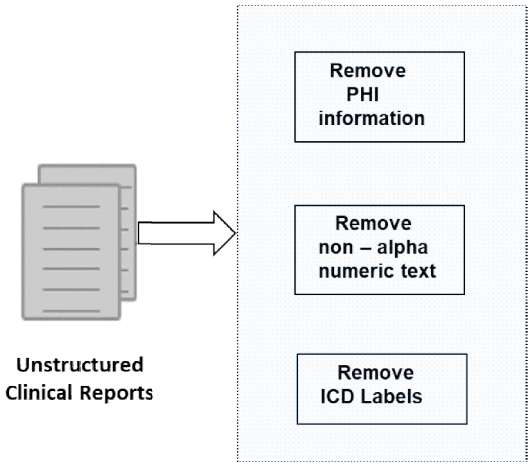

Figure 3. Data Preprocessing Steps

Organ classification was based on the ICD9 codes.The discharge summaries were processed and the corresponding ICD9 code of an organ was used to identify the organ mentioned in the report. The dataset thus obtained was sparse as seen in Table 1,hence, up-sampling and down-sampling techniques were used to balance the dataset.

2.4. Feature Extraction

In this work, four approaches for feature extractionnamely, term frequency-inverse document frequency (TF-IDF), word2vec model, a combination of TF-IDF and Word2vec model. Word2vec is a set of pre-trained models to generate word embedding [21]. TF-IDF is used to weigh the terms based on frequency of occurrence.

TF-IDF is a measure used to evaluate the significance of a word in a document within a collection or corpus. TF is the frequency of word in that document and IDF measures the commonness of word among various documents. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

$$IDF(w) = \log(Nd / n(d,w))$$

whereNd is the total number of documents and n(d,w) is the number of documents that contain word w. To calculate TF-IDF, as a first step tokenize all the notes, then at later stages document-word matrix which stores count of each word (TF) multiplied by the IDF weight were created.

Word2vec models[17] are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words to produce word embeddings. Word2vec takes as input a large corpus of text and produces a vector space, typically of several hundred dimensions and similar words are assigned vectors in close proximity. Continuous Bagof Words (CBOW), a feed-forward neural network model that consists of inputs, projection and output layerswere used in this feature extraction technique. The inputs are words and the CBOW model predicts the target word based on the context, that is, words that precede and follow the target word. As corpus to train Word2vec model, text notes from MIMIC-III andpre-trained word vectors from PubMed [18,19] were used. PubMed contains more than 27 million records of articles in the biomedical literature and items from the NCBI books database.

2.5. Approach 1 - TF-IDF

To generate feature vector for a report, TfIdfVectorizer[20] is trained on clinical terms present in the report, representing clinical terms as n-grams (unigrams, bigrams, trigrams). To extract clinical terms, Named Entity Recognizer was used. The clinical terms extracted are combination of words or single word like x-ray, lung cancer etc. Fragmentation of these wordsinto unigrams would resultis loss ofmeaning of the term lung cancer, therefore bigram was chosen. Similarly, trigrams for medical terms containing three termswas selected. TF-IDF fit on clinical termswas performed to obtain thefeature vector corresponding to training reports and this was transformed

to test reports using TF-IDF fit model. Figure 4 shows the data set used to extract TF-IDF feature extraction and its output format.

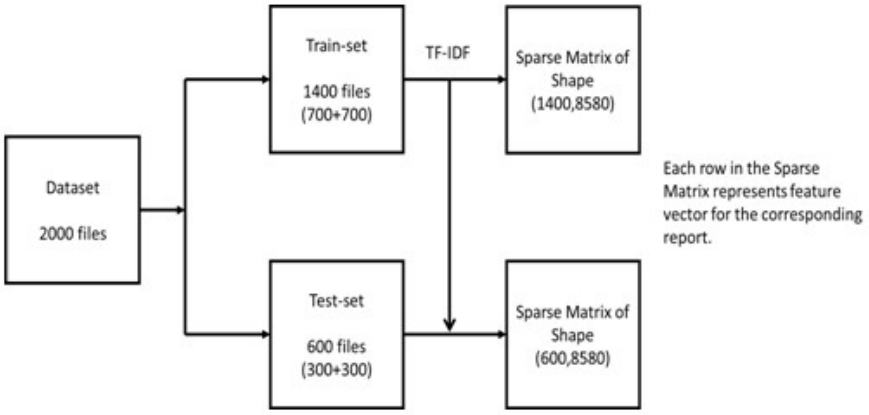

Figure 4. TF-IDF Medical Terms Extracted Using NER

Approach 2 - Word2Vec

Feature vector corresponding to a report is formed using the following steps:

1. For computing word vectors,text from all the reportswere combined, Word2vec was trained to generate word vectors for all the words in corpus.
2. To obtain vector representation for each sentence in a report, mean average pooling was performed.
3. K-means clustering algorithm was trained to cluster all the sentences.
4. For each report:

   i) Sentence vectors were computed.

   ii) K-means clustering of these sentences using trained k-means model.

   iii) Sentences count in each cluster used to create a histogram and generate a Probability Distribution Function (PDF).

Figure 5 provides an overview of the proposed method and a sample word vector is givenin Figure 6.

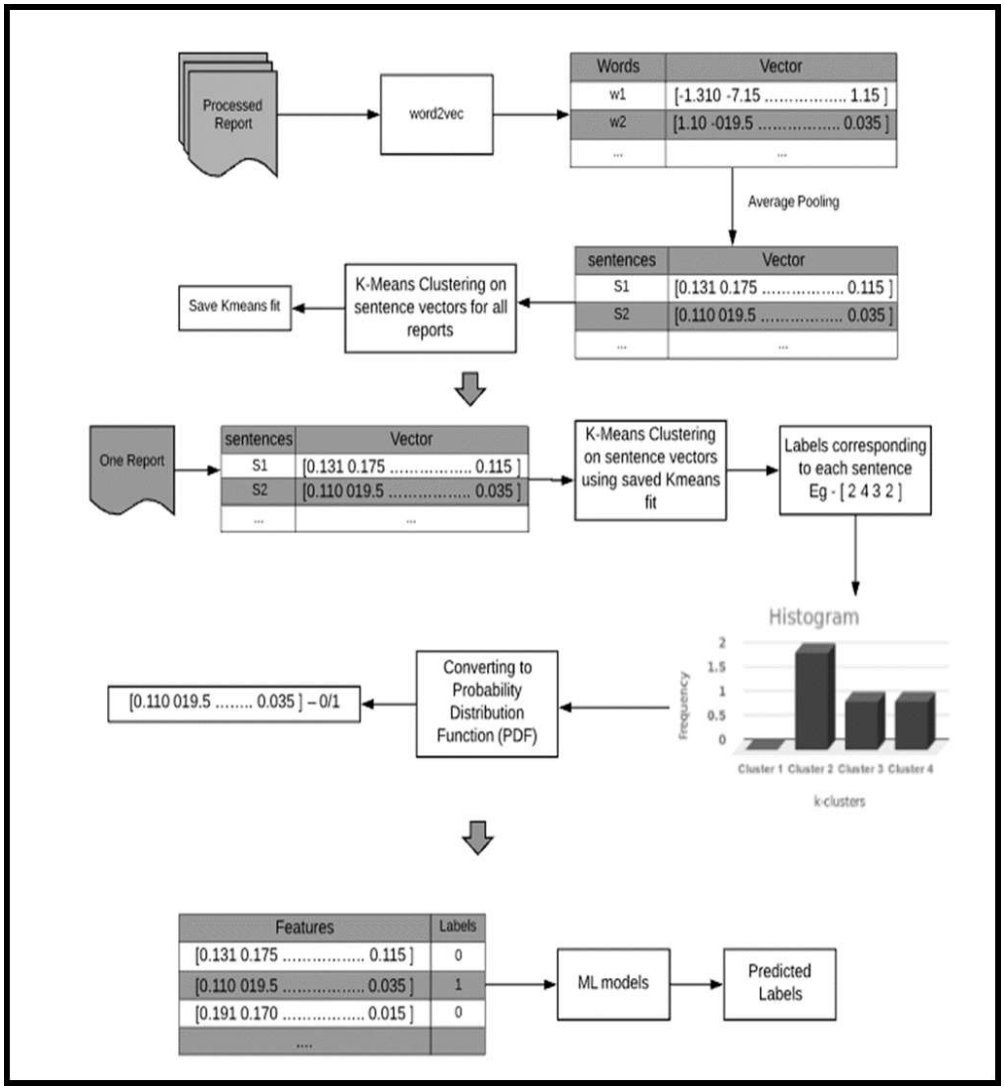

Figure 5. Word2Vec & K-means feature extraction flow

```
Bone: 100 features using Word2Vec

array([-0.45568326, -0.42008284,  0.5879811 , -0.5954721 ,  0.242069  ,
        0.04314294, -0.11443035, -0.42404965, -0.48740718,  0.7659249 ,
       -0.5828809 ,  0.7553276 , -1.1470004 , -0.1757206 , -0.18863104,
        0.21828651, -0.5719981 ,  0.16566484,  0.20187922, -0.41965342,
        0.37324256, -0.10080674,  0.01068049, -0.2813971 , -0.2829634 ,
        0.42982036, -0.18310162,  0.13920984,  0.5431863 , -0.65572494,
        0.20396811,  0.00256491,  0.03591614, -0.5485845 , -0.5153154 ,
        0.46385816,  0.22171667, -0.11218242, -0.1582615 , -0.08939845,
        0.06131884, -0.3892101 ,  0.23531616,  0.27197105, -0.5130427 ,
       -0.24943025,  0.10215823, -0.89288765, -0.42185077,  0.14471734,
        1.0374857 , -0.008071  ,  0.02999489, -0.29102188, -0.50396645,
        0.08793559,  0.3439966 , -0.34348166,  0.26452625,  0.38907006,
        0.4787299 ,  0.3284534 , -0.04414903, -0.04706165,  0.03913996,
        0.04705186,  0.09013759, -0.08766278,  0.09889007,  0.27696326,
        0.24847467,  0.21624993, -0.4111728 , -0.08203211, -0.21875288,
        0.01415944, -0.00842295,  0.02118845,  0.04208755, -0.23537743,
        0.23402141, -0.01921754,  0.30300575, -0.05118107, -0.03348159,
       -0.24223523, -0.12935163,  0.47220245, -0.0772469 ,  0.23612452,
       -0.3045119 , -0.03456276,  0.18997438, -0.37388992, -0.18503377,
        0.32238147,  0.4318316 , -0.2188602 , -0.46486193,  0.11918008],
      dtype=float32)
```

Figure 6. Word Vector – Example (Len = 100)

2.6. Feature Analysis

Based on the fisher score analysis of cancer and non-cancer discrimination, final set of features were selected for classification.

1. Average feature set reduced from [1000x300] to [1x300] is shown in Figure 7a.
2. Using fisher discrimination analysis, out of 300 features 10 features which gives good discrimination between cancer and non-cancer were selected as final set of features for classification. Figure 7b shows the discrimination details of few selected features.

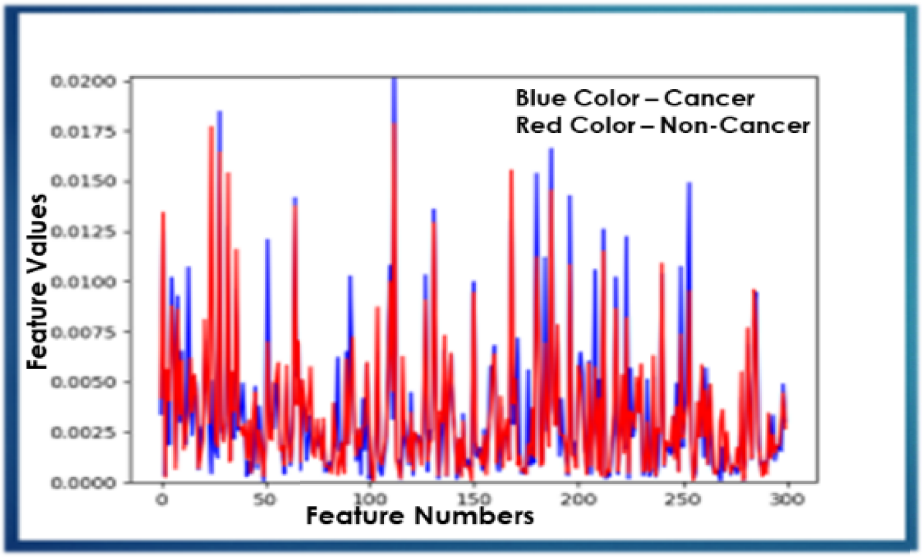

Figure 7a. Feature Analysis Plot

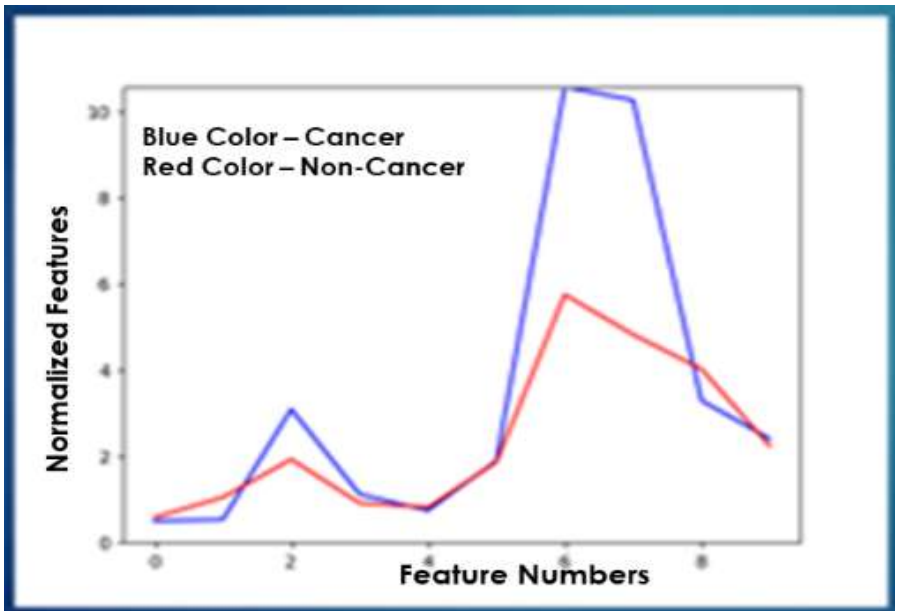

Figure 7b. Down selected features

2.7. Approach 3: Hybrid Model consisting of both TF-IDF score and Word Vectors

In this work, Word2vec and TF-IDF features were combined to get stronger features. Following process explains the hybrid model approach.

The processing steps for the hybrid model is describedbelow:

1. Train TF-IDF on clinical terms to obtain TF-IDF score. TF-IDF score was obtained from Named Entity Recognizer.
2. Train Word2vec on sentences to obtain word vectors for each word. (As described in Approach2).
3. Multiply TF-IDF weight of clinical terms to their corresponding word vector and perform average pooling of these new word vectors to make a sentence vector.
4. Follow K-Means clustering process as in Approach 2.

2.8. Approach 4: Using Pre-trained Word Embedding (PubMed & PMC)

In this approach, pre-trained PubMed and PMC word embedding [18,19] of the Word2vec modelwere used and the remaining steps are similar to the Word2Vec and K-means Approach.By this method, features with corresponding labels were obtained.

Different classification models were trained using Grid search, Word2vec model has various hyper-parameters [23] such as, num_features, num_workers, context_size, min_word_count etc. Among these num_features and context_sizeplays a crucial role as num_features is the size of a word vector and context_size is the size of context window (8). Experimentswere performed with different values of num_features (100, 200, 300, 400, 500, 600, 800) and context_size (5, 10). Also for k-means clustering different k values were tested, ranging from 100 to 1000.

To build model for classification into Malignant/Benign cancer, training was performed on different machine learning models, analyzed performance of each model with different hyper parameters and selected the model with best performance. Models used for the grid test were based onLogistic Regression, Support Vector Machines, Random Forest classifier and Gradient Boosting Classifiers.

Different oversampling techniques [24] have been used to classifyon the basis ofcancer occurrence in a particular location/organ. In the proposed approach, experiments were conducted

with Random Over Sampling, Synthetic Minority Oversampling Techniques (SMOTE), and ADASYN at the Word2vec processing step to increase the sample size.

2.9. Semantic Information Extraction

Semantic Web was used to find self-describing interrelations of data in a form that machine can process. The structured formatfacilitates storage and information retrieval based on the meaning and logical relationships. Instead of retrieving matching text for a query, the technology permits us to find related text.

The following steps describe the method to build ontology for Lung cancer reports:

1. Extracted semantic information from MIMIC reports and represented them as Ontology structure. Corresponding to each term selected, the relevant predefined ontologies were extractedusing Bio OntologyAPI [25].
2. Existing ontologies from SNOMED[26], RADLEX [11], LOINC [27]were used to link medical findings to ontology structure.
3. Resource Description Framework (RDF)[28] was used to create triples (subject, predicate, object) where subject is the URI corresponding to ontology, predicate describes relationship between subject and Object is either a URI or a string or literal.
4. For building ontology [29,30], the information in lung cancer reports were converted into the graph structure as shown in Figure 8 (image source [31]). APython scripts was developed to convert reports to RDF structure. Various namespaces were added for existing ontologies.For each sentence from a report, triples are formed as shown in example below. These sentences are linked as an RDF graph as shown in Figure 9.

The following are the triples for the example sentence: “Patient1 with small bilateral pleural effusions”.

(Patient1, has, findings1)

(finding1, consist, pleural effusions)

(pleural effusion, is, bilateral)

(pleural effusion, effect/size, small)

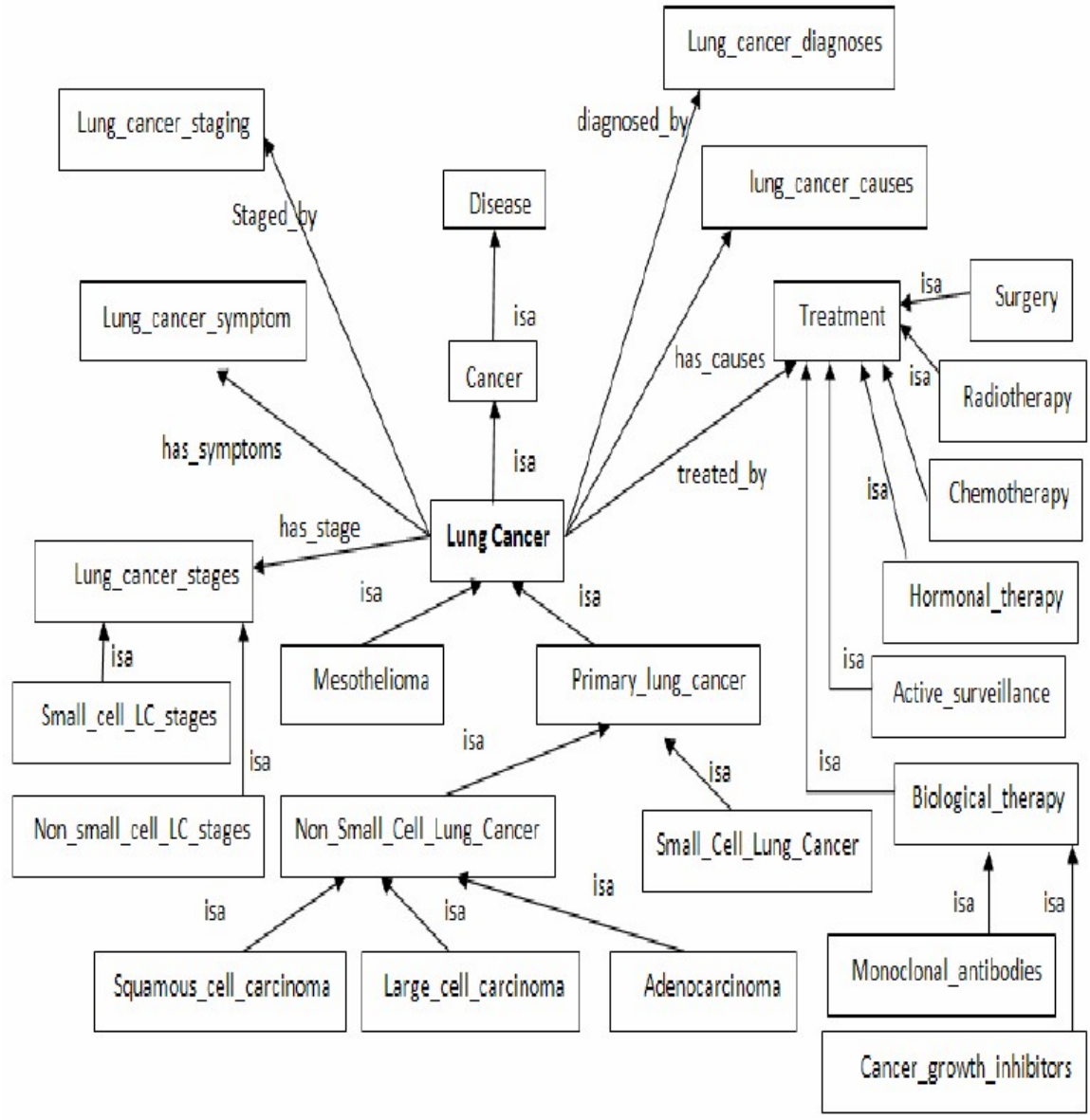

Figure8.Lung cancer graph structure

2.10. Ethical considerations

This research did not involve human subjects. Anonymized patient data from publicly available database, such as MIMIC and ICD codeswere usedto develop and validate the proposed methods.

3. Results

For malignant/benign cancer classification the dataset was balanced for 12456 samples and the results obtained are as shown in Table 3. For organ based classification of cancer, the dataset was imbalanced, based on the comparative analysis of results captured in Table 4, to overcome this imbalance problem oversampling techniques were used to balance the data. By analyzing values on different hyper-parameters, the best parameter value for num_featuresare from 300 to 600. Gradient boosting classifier performed well over other classifiers for both TF-IDF and word2vec models.

A comparative analysis of different approaches shows word2vec produces improved result than combined word2vec and TF-IDF. TF-IDF and (word2vec + TF-IDF) produces almost similar results for gradient boosting classifiers. The results using word embeddings trained on reports are improvedthan pre-trained word embeddings.Wikipedia PubMed + PMC articles produces comparable accuracy to word2vec but word2vec produces improved F1-score as the embeddingsare trained on the MIMIC-III reports. This is the general trend among all the classifiers on which the experiments were done.

Figure 9 shows the ontology based semantic graph built for lung cancer report.This figure contains Lung cancer (Adenocarcinoma) graph structure identified in one of the report on the left and its corresponding RDF structure on the right. The nodes in the RDF structure are id values of predefined ontology taken into consideration while building the graph. The ids are obtainedfrom NCBO API [25]

Approach	Classifier							
	Logistic Regression		SVM		Random Forest		Gradient Boosting	
	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy
TF-IDF	0.70	0.839	0.85	0.876	0.80	0.817	0.87	0.889
Word2vec	0.75	0.791	0.87	0.887	0.87	0.852	0.89	0.899
Word2vec+ TF-IDF	0.69	0.75	0.84	0.87	0.79	0.86	0.84	0.887
PMC pre-trained	0.80	0.823	0.81	0.875	0.87	0.867	0.83	0.877
PubMed pre-trained	0.75	0.83	0.80	0.866	0.75	0.87	0.81	0.878
Wikipedia PubMed + PMC articles	0.78	0.811	0.83	0.87	0.81	0.872	0.80	0.891

Table 3. Classifier performance for malignant/benign cancer

Approach	Classifier			
	Logistic Regression		Multilayer Perceptron (4 layers)	
	F1 Score	Accuracy	F1 Score	Accuracy
Original Dataset with no sampling (imbalanced)	0.12	0.27	0.32	0.37
Dataset built by Random Over Sampling	0.37	0.39	0.75	0.76
Dataset built using SMOTE	0.39	0.42	0.74	0.765
Dataset built using ADASYN	0.37	0.416	0.73	0.72

Table 4. Classifier performance for cancer specific to organ

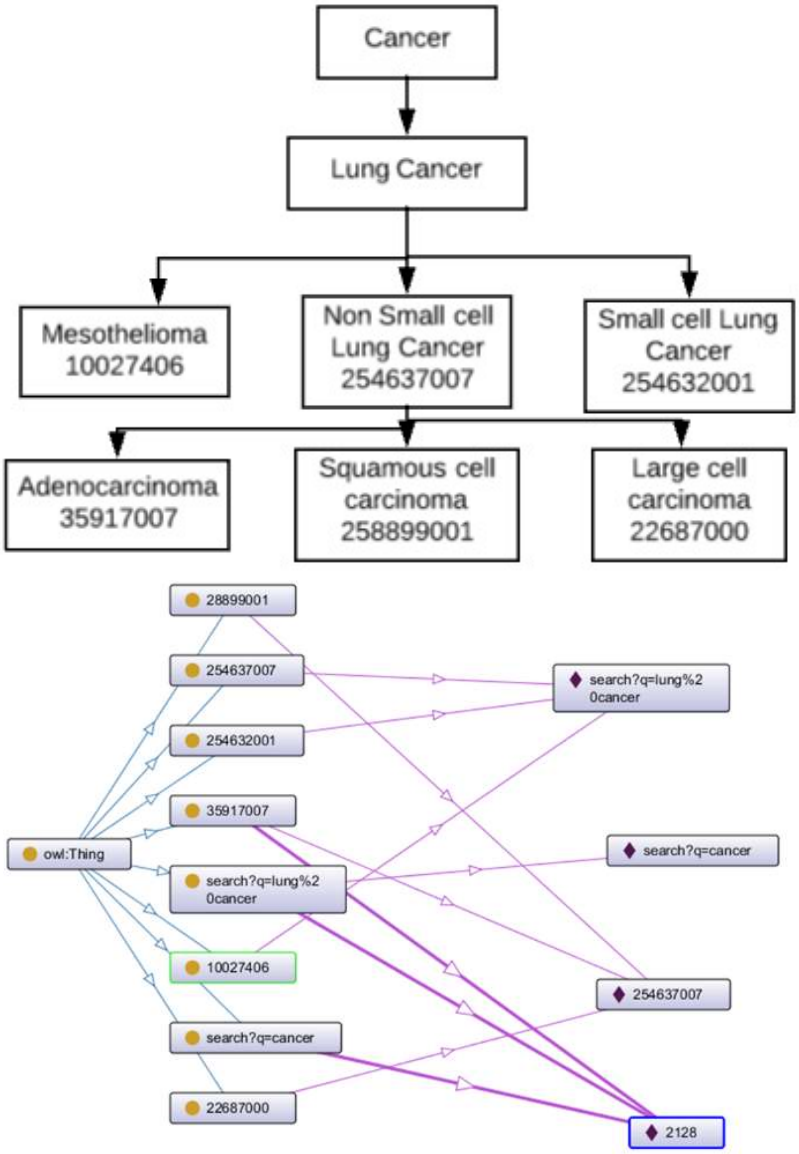

Figure 9. Semantic representation of report

4. Conclusion

In this research work, experimentation was performed on different NLP techniques for feature extraction and developed various machine learning models for classification.Anend to end solution from classifying the report to a disease type and mapping disease information to disease ontologies is proposed.

The system performs well for malignant/benign cancer classification with 0.89 F1-score. Random oversampling method produced 0.75 F1 score. Figure 9 shows the semantic representation of reports for lung cancer and Adenocarcinoma ontology. Proposed techniques will help doctors to diagnose cancer early and brings in efficiency in the workflow.

In the current scenario, a typical medical practitioner has to manually studythe historical reports to arrive at conclusions and there are chances that medical practitioners might miss critical information, aggravated by time pressure and resource constraints. Proposed approach reduces the burden of medical practitioners by improving their throughput by presenting the information in a semantic graph. Visualization of graph data improves readability of information extracted and makes the analysis easy.

The novelty of the proposed work lies in customizing and extending existing NLP techniques to approaches similar to the one used for extracting features from images (histogram). A new methodology is proposedto represent reports into ontology. This method captures the relationships and links similar concepts. In this work, experiments were conducted by combining TF-IDF and word2vec by weighted averaging of word2vec vector by TF-IDF weights. Theproposed work has significant potential for new clinical applications to targeted cancer treatments.

As a future work, one can further improve the performance of the system by experimenting with deep learning models like RNN and LSTM. Current ontology structure is defined for lung cancer, in a similar way the ontology structure can be extended to different cancer types and further to different disease types as well.

References

[1] Johnson AEW, Pollard TJ, Shen L, et al., MIMIC-III, a freely accessible critical care database. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. https://mimic.physionet.org/

[2] "The International Classification of Diseases, 9th Revision, Clinical Modification" (ICD-9-CM), Sixth Edition, issued for use beginning October 1, 2008 for federal fiscal year 2009. http://icd9.chrisendres.com/

[3] Alan R. Aronson, Olivier Bodenreider, Dina Demner-Fushman, Kin Wah Fung, Vivian K. Lee1, James G. Mork, AurelieNeveol, Lee Peters, Willie J. Rogers From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. BioNLP 2007: Biological, translational, and clinical language processing, pages 105–112

[4] S. Dublin, E. Baldwin, R.L. Walker, L.M. Christensen, P.J. Haug, M.L. Jackson, J.C. Nelson, J. Ferraro, D. Carrell, W.W. Chapman, Natural language processing to identify pneumonia from radiology reports, Pharmacoepidemiol. Drug Safety 22 (8) (2013) 834–841.

[5] T Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng 1 , Saeed Mehrabi 2 , SunghwanSohn, Hongfang Liu. Clinical information extraction applications: A literature review.

[6] Imon Banerjee, Matthew C. Chen, Matthew P. Lungren, Daniel L. Rubin, Department of Biomedical Data Science, Stanford University, Stanford, CA, United States Department of Radiology, Stanford University, Stanford, CA, United States. Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort.

[7] HeinerOberkampf, Sonja Zillner, James A. Overton , Bernhard Bauer, Alexander Cavallaro , Michael Uder and Matthias Hammon. Semantic representation of reported measurements in radiology.

[8] HeinerOberkampf , Sonja Zillner, Bernhard Bauer and Matthias Hammon. An OGMS-based Model for Clinical Information (MCI). In: Proceedings of International Conference on Biomedical Ontology. 2013. p. 97–100.

[9] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.

[10] The Open Biological and Biomedical Ontologies Foundry - http://www.obofoundry.org Accessed 31 July 2014.

[11] RadLexontology entity, Version 3.11. http://radlex.org Accessed 31 July 2014.

[12] Oberkampf H, Bretschneider C, Zillner S, Bauer B, Hammon M. Knowledge- based Extraction of Measurement-Entity Relations from German Radiology Reports. IEEE International Conference on Healthcare Informatics; 2014. p.149–154.

[13] ThusithaMabotuwana, Michael C. Lee, Eric V. Cohen-Solal Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY 10510, USA. An ontology-based similarity measure for biomedical data- Application to radiology reports.

[14] Mahadevaiah G, Soest JV, Dekker A, Udupa N, Rao SV, Kumar YK, et al. Semantic Representation of Radiotherapy data for effective data mining. Proc. Fifth Int. Conf. Adv. Appl. Sci. Environ. Eng. - ASEE 2016 [Internet]. Kuala Lumpur, Malaysia: Institute of Research Engineers and Doctors, USA; [cited 2017 Jun 8]. p. 12–5. Available from: http://www.seekdl.org/nm.php?id=7421

[15] International Classification of Diseases, 9th Revision, Clinical Modification, Sixth Edition,2014. https://seer.cancer.gov/tools/casefinding/case2014.html

[16] Mark G. Weiner, M.D., Alice Livshits, Carol Carozzoni, Pharm.D., Erin McMenamin, Gene Gibson, Pharm.D., Alison W. Loren, M.D., Sean Hennessy, Pharm.D., M.S.C.E. Division of General Internal Medicine, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104 3Department of Pharmacy Services, Hospital of the University of Pennsylvania. Derivation of Malignancy Status from ICD-9 Codes.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space

[18] Billy Chiu, Gamal Crichton, Anna Korhonen, "How to Train GoodWordEmbeddings for Biomedical NLP" Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 166–174, Berlin, Germany, August 12, 2016.

[19] Kevin Patel, Divya Patel, Mansi Golakiya, Pushpak Bhattacharyya, NileshBirari 1 Indian Institute of Technology Bombay, India ,Dharmsinh Desai University, India, 3ezDI Inc, India. Adapting Pre-trained Word EmbeddingsFor Use In Medical Coding.

[20] Fabian Pedregosa, GaëlVaroquaux, Alexandre Gramfort, et al., "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research 12 (2011) 2825-2830 12(Oct):2825−2830, 2011.

[21] Tomas Mikolov, Karl Chen, Greg Corrado, et al. "Efficient Estimation of Word Representations in Vector Space", arXiv:1301.3781v3 [cs.CL] 7 Sep 2013

[22] Joseph Lilleberg,Computer Science Department Southwest Minnesota State University Marshall, MN 56258 USA joseph.lilleberg@smsu.edu Yun Zhu, Yanqing Zhang Computer Science Department Georgia State University Atlanta, Georgia 30302-5060 USA yzhu7@student.gsu.edu, yzhang@gsu.edu. Support Vector Machines and Word2vec for Text Classification with Semantic Feature

[23] Gamal Crichton Anna KorhonenSampoPyysalo Language Technology Lab DTAL, University of Cambridge {hwc25|gkoc2|alk23}@cam.ac.uk, sampo@pyysalo.net. How to Train Good Word Embeddings for Biomedical NLP

[24] Imbalanced-learn documentation https://github.com/scikit-learn-contrib/imbalanced-learn

[25] Bio ontology term search API, http://data.bioontology.org/documentation#nav_search

[26] U.S. National Library of Medicine,https://www.nlm.nih.gov/healthit/snomedct/

[27] Unified Medical Language System (UMLS) https://www.nlm.nih.gov/research/umls/

[28] Decker S, Mitra P, Melnik S. IEEE Internet Computing 2000;4:68–73. Framework for the semantic Web: an RDF tutorial.

[29] International Classification of Diseases, Version 9 - Clinical Modificationhttps://bioportal.bioontology.org/ontologies/ICD9CM?p=classes\&conceptid=http\%3A\%2F\%2Fpurl.bioontology.org\%2Fontology\%2FICD9CM\%2F782.3

[30] LungCancer Ontology, https://bioportal.bioontology.org/ontologies

[31] ABDEL-BADEEH M. SALEM, MARCO ALFONSE Computer Science Department Ain Shams University Faculty of Computers and Information Systems CAIRO, EGYPT. Ontology versus Semantic Networks for Medical Knowledge Representation.

Authors Biography:

1. **Geetha Mahadevaiah, (Corresponding Author)**
Senior Director at Philips, Research Department, PIC, Bangalore.
30+ years of experience in software engineering and management.
Areas of interest : Clinical decision support systems, Semantic Web, healthcare applications
Bachelor of Engineering in Computer Science and Technology
Bangalore University
Master of Business Administration, Bangalore University

2. Dinesh M.S.
Senior Principal Scientist at Philips, Research Department, PIC, Bangalore.
19+ years of post-doctoral experience in applied research (healthcare domain).
Areas of interest: Machine Learning, Pattern recognition and Image Processing
Education: Bachelor of Engineering, Master of Technology and Doctor of Philosophy from University of Mysore

3. André Dekker
Professor of clinical data science at Maastricht University and has been leading the development of prediction models in radiation therapy for many years. He is also coordinator of the Personal Health Train project, aiming to facilitate citizen science.
*Areas of interest :*Radiomics, Semantic Web, Radiotherapy, machine learn*ing*