A Multi-Label Machine Learning Approach to Support Pathologist's Histological Analysis

Antonia Azzini
Consortium for the Technology Transfer – C2T, Italy
Nicola Cortesi
Consortium for the Technology Transfer – C2T, Italy
Stefania Marrara
Consortium for the Technology Transfer – C2T, Italy
Amir Topalović
Consortium for the Technology Transfer – C2T, Italy

Abstract

This paper proposes a new tool in the field of telemedicine, defined as a specific branch where IT supports medicine, in case distance impairs the proper care to be delivered to a patient. All the information contained into medical texts, if properly extracted, may be suitable for searching, classification, or statistical analysis. For this reason, in order to reduce errors and improve quality control, a proper information extraction tool may be useful. In this direction, this work presents a Machine Learning Multi-Label approach for the classification of the information extracted from the pathology reports into relevant categories. The aim is to integrate automatic classifiers to improve the current workflow of medical experts, by defining a Multi-Label approach, able to consider all the features of a model, together with their relationships.

Keywords: machine learning, health problems, knowledge extraction, data mining, classification
JEL classification: I10, I12

Introduction

In the medical field, the use of information technologies plays an important role since more than 50 years. Thanks to this interaction, nowadays it is possible to refer to several medical applications, which improve doctors and patients life. This paper focuses on telemedicine, defined as a specific branch where IT supports medicine, in case distance impairs the proper care to be delivered to a patient. Especially in case of remote diagnosing and teleconsulting systems, data (including medical texts and images) are acquired locally and shared via web to physicians, which can be anywhere else and still be able to analyse the data and send the diagnosis back (Combi et al., 2016).

The information contained into medical texts (such as patient records or discharge summaries) is relevant to several different retrieval, coding and inference purposes. It should, for instance, provide support for medical decision making, for mapping data into medical coding systems, or for quality assurance of medical treatment. According to the growing availability of medical documents in machine readable form, procedures for automatically analysing and formatting textual data gain more and more importance, since hand-coding and manual indexing are time-consuming and usually error-prone.

In fact, pathology reports, i.e. the output document of a tissue sample analysis, play an important role in cancer diagnosis and staging (describing the extent of cancer within the body, especially whether it has spread). These reports are usually written by the pathologist in natural language, and then the relevant information has to be extracted and organized in a form suitable for statistical analysis to be stored in a proper data structure. Anyway, even if the use of structured information may help the data sharing among institutions, integrating structured and unstructured data information remains a challenge (Garcia-Remesal et al., 2009). Moreover, clinicians need time to learn the different standards available, hence they prefer the flexibility of free text to record their analyses and conclusions. Ideally, natural language texts would be then used as input to automatically extract the data required by different protocols.

This project aims to introduce an innovation in the field of telemedicine, with a particular focus on the diagnosis of samples in the oncology field. During oncological surgery, in fact, it often happens that the surgeon has to remove tissue samples for histological examination. While waiting for the result of this examination, the surgery is suspended, clearly lengthening the action time and precluding the possibility to complete the surgery in a single step. If the hospital does not have a pathological anatomy laboratory inside, equipped for these analyses, two scenarios may occur: the pathologist is moved to the facility where the operating block is located for the entire duration of the operations that may require this service; the samples are sent to an external analysis laboratory, which will have its own costs and time.

These issues may create great inefficiencies and significant costs. Moreover, pathology records contain sensitive information, and often it is not easy to make them widely available. In order to reduce errors and to improve quality control, a dedicated tool may be useful. Natural Language Processing (NLP) and Machine Learning (ML) approaches represent in this scenario a promising solution to handle, respectively, unstructured data and to extract useful information.

This work is then based on a Machine Learning approach for the classification of relevant information contained in pathology reports into categories related to cancer diagnosis, by applying Information Extraction and Text Mining techniques to extract the features of the classifiers.

The remaining of the paper is organized as follow. After a brief summary of the related works into "Related Work" Section, the approach is presented into "The Multi-Label Machine Learning Approach" Section, by detailing the data pre-processing and the knowledge extraction phases, and by describing the implemented classifiers. Preliminary experiments are then presented and discussed, together with the obtained results, into "Preliminary Experiments and Prototype" Section, and final remarks conclude into "Conclusion" Section.

Related Work

As reported by the literature, all the reports are usually written by the pathologist in natural language. Such a scenario is then considered promising for text mining research. The state of the art in text mining from pathology reports has mainly relied on domain-specific lexicons and rules (Cohen & Hersh, 2005). Anyway, different solutions have been proposed for several specific problems.

An interesting review presented by Meystre and colleagues in (Meystre et al., 2007) underlines the importance of information encoding in order to reduce errors and improve the quality control of medical records. In particular, the authors emphasize that NLP techniques, and more precisely Information Extraction (IE), are essential in this

domain. Moreover, pre-processing such as spell checking, document structure analysis, sentence splitting and contextual features are crucial for the accurate interpretation of the extracted information. Li and colleague, for example, define in (Li & Martinez, 2010) a comparative work that considers different supervised text classification systems to predict a set of defined categories, encoded as string values, in the domain of pathology records, while in the work carried out by Sariuglu and colleagues (Sarioglu et al., 2013), an approach, based on a recommender system with unsupervised techniques, has been implemented in order to support the clinical decision-making activity. A similar work has been presented in the literature (Jouhet et al., 2012), aimed at constructing and evaluating functions (classifiers), produced by supervised ML techniques. This approach is also based on an automatic categorization of pathology reports by using only their content, divided between two levels of granularity obtained from the data pre-processing. Even though the work a single label methodology has been considered, the authors report that the text pathology reports could be useful as a data source for automated systems in order to identify, classify and notify new cancer cases.

Coden and colleagues (Coden et al., 2009) present in their contribution a different approach that automatically instantiates a knowledge representation model starting again from textual pathology reports. Their work is based on an open-source framework by using NLP, ML and rules to discover and populate elements of a defined model.

On the other hand, Pestian and colleagues define in (Pestian et al., 2007) an approach aimed at collecting and pre-processing the textual data, through a corpus definition and a coding process, which refers to critical aspects like ambiguity and anonymization, manual inspection and majority annotation. Zhou and colleagues present in (Zhou et al., 2004) an exploratory work on adapting an existing HMM-based named entity recognizer to the biomedical domain. Various lexical, morphological, syntactic, semantic and discourse features have been incorporated to cope with the so-called entity recognition problem. A K-NN algorithm is proposed by the authors to effectively resolve the data sparseness problem.

Several different techniques may be used to extract information, from simple pattern matching to complete processing methods based on symbolic information and rules or based on statistical methods and ML. In this direction, the approach implemented in this work implements an automatic learning system of the analyses carried out by the pathologist on several samples in the oncological field.

Machine Learning based Text Classifiers

The approach adopted into this paper lies in the domain of "Text Classification". In the recent literature, Text Classification (TC) has proven to give good results in extracting knowledge from many real-life Web-based data such as, for instance, those gathered by institutional scientific information platforms or microblogs and other social media platforms (Ceci & Malerba, 2007) and also in many different research areas such as opinion spam detection (Viviani & Pasi, 2017) and sentiment analysis (Bifet & Frank, 2010).

For decades, constructing a machine learning system required considerable expertise to design the feature extraction phase to transform the raw data into input features for a (machine learning) classifier (LeCun et al., 2015).

The Word Representation before feeding a classifier can be obtained by performing word selection or by replacing words with continuous value representations (e.g., word embedding (Turian et al., 2010) like word2vec (Mikolov et

al., 2013)), or by using a classifier able to discover word representations, or a combination thereof.

Automatic text classification can be performed by using a supervised, unsupervised, or semi-supervised machine learning approach. Supervised learning is based on training a classifier over a set of texts previously labelled by domain experts.

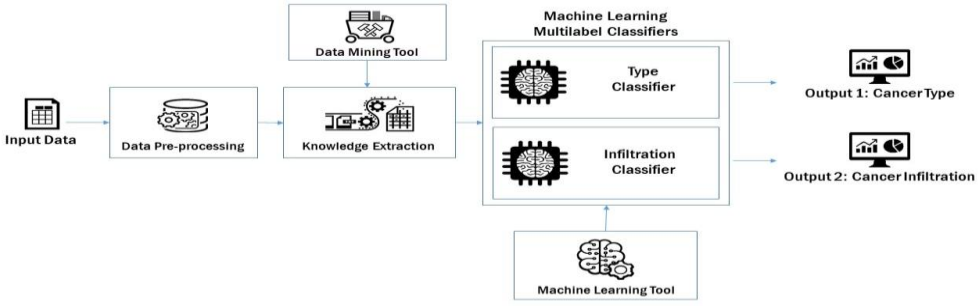
Unsupervised learning does not require an already labelled dataset, while semi-supervised approaches rely on both labelled and unlabelled data.
Supervised learning approaches performs better in text classification w.r.t unsupervised ones, but the task of labelling a dataset requires a huge effort to domain experts. Unsupervised learning is used to identify categories or discover hidden structure in texts (e.g., clustering). The prototype presented into this approach implements an example of a supervised learning application.

Generally, a supervised approach is mainly divided into single, multi and/or multi-label classification (Tsoumakas & Katakis, 2007). In the first case each instance is associated only to one category; the second one considers multiple categories, but each instance is assigned only to one; while in the latter case each instance can be assigned to more than one category. Multi-label classification can be mainly divided into: "Binary Relevance", where a multi-label problem is divided into n single label problems; "Classifier Chains", the problem is transformed into "n" different correlated problems, i.e. in which the output of the problem "n-1" is the input of the problem "n"; "Label Powerset", that turns a multi-label problem into a single multi-class problem, and finally "Adaptive algorithms", where multi-label algorithms have been applied to solve the problem.

The Multi-Label Machine Learning Approach

As previously introduced, the idea behind this work is to support the pathologist in the process of classifying the information acquired from the analysis of stains (markers) carried out on the tissue samples analyzed. The approach has been built on a set of 2186 cases of breast cancer described in pathologist reports that are the output of tissue samples analyses. Each record contains, in textual form, observations on the markers used during the analysis, the values of these markers, and notes regarding the cancer diagnosis. The overall architecture is reported in Figure 1.

Figure 1
Overall Architecture of the Approach



Source: Authors' work

After the acquisition phase, the data are preprocessed and the information useful for the classification processes are obtained by applying text mining techniques (as n-gram encoding (generally speaking, an n-gram is a set of n consecutive words) and by defining rules applied by the knowledge extraction module. Such rules are then applied in order to extract the text knowledge and to define a structure for it. From an

empirical analysis carried out by the expert, it is then possible to notice that both the text and their defined keywords have a fixed structure. This information is used to select the most informative tokens among the extracted n-grams and to create regular expressions able to capture interesting interdependencies among the markers.

The resulting dataset is composed by couples (input-expected output), used to train two multi-label supervised classifiers. The input values correspond, as detailed in Figure 2(b), to the seven most useful markers, their positivity value and the tumour staging value. The expected output, (i.e. those defined by the pathologist) instead, differ for each of the supervised classifier, and correspond, respectively to the ductal/lobular analyse values for the Type classifier, and to the infiltrating/in situ analyse values for the Infiltration classifier. These classifiers, are able to acquire in parallel the entire dataset, and provide, respectively, the type of cancer and the type of infiltration for each clinical record stored into the repository.

Data Preprocessing and Knowledge Extraction

As introduced, the clinical data from the pathological analysis are in text form. The final dataset is structured as an Excel table whose columns describe the results of the analysis. The "Clinical Code" is represented by an alphanumeric identification value of the medical record. The "Marker Code and Description" are textual objects used for analysis and the marker. The "Textual Description" corresponds to the description of the free text of the pathologist's analysis result, while "Positive" is the boolean value of the analysis result. Each medical record is divided into several files, one for each marker considered by the pathologist to analyze the tissue sample. Consequently, the first operation to be performed is to aggregate the different markers and identify the corresponding records in order to create the vocabulary associated with them.

As introduced, the clinical data resulting from the pathological analyses are in textual form. Since the texts are in the form of short quick notes (more similar to keywords lists than proper text), it is not necessary to perform a traditional preprocessing, by including activities such as tokenization (i.e., sentences are split into separate words and punctuation is removed), lower case reduction, stop-words removal (e.g., elimination of common and low informative words as "the", "of", "as", by using a predefined list), or stemming (words are reduced to their stem) (McCallum, 2005). Instead, the operations performed in this approach are focused on eliminating redundant and unuseful information as duplicates. The final dataset is structured as an Excel table whose columns detail the items reported in Table 1. Each clinical record is divided into several rows, one for each marker considered by the pathologist to analyse the tissue sample. Consequently, the first operation to be performed is to aggregate the different markers and recognize the corresponding records in order to create the vocabulary associated with them.

Table 1
Formatting of the Data Provided as Inputs to the Implemented Prototype

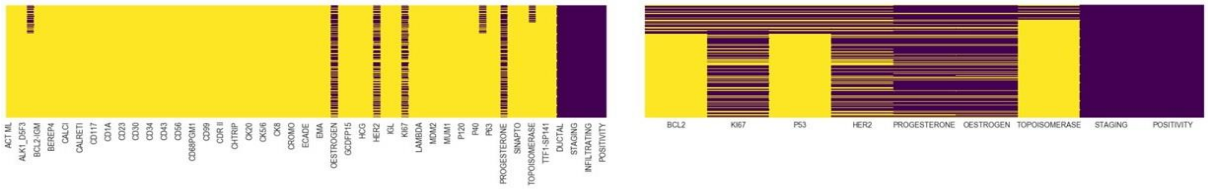
Item	Type	Description
Clinical Code	**Text**	**Alfanumeric identification value of the medical record.**
Marker Code	**Text**	**Marker code used for the analysis.**
Marker Description	**Text**	**Marker description used for the analysis.**
Textual Description	**Free Text**	**Free text with the description of the results of the pathologist's analysis.**
Positive	**Boolean**	**Boolean value of the analysis result.**

Source: Authors' work

The Clinical code is represented by an alphanumeric identification value of the medical record. The Marker Code is a textual item used for the analysis, and the Marker defines the marker description used for such analysis. The textual description corresponds to the free text description of the result of the pathologist's analysis, while Positive is the Boolean value of the analysis result.

The Figure 2(a) shows the distribution of the presence assumed by the tumour markers that are detailed into the Excel sheet. Missing or null values are reported by yellow areas, while the purple ones indicate the presence of values. However, as Figure 2(b) shows, it is possible to observe how, among all the markers reported by Figure 2(a), only seven (BCL2, KI67, P53, HER2, PROGESTERONE, OESTROGEN, TOPOISOMERASE), together with the tumour staging and positivity value, represent the most significant information. These data are used in the prototype as input of the Machine Learning system.

Figure 2:
Markers distribution Representation (a) and Most Useful Markers Extraction (b)



(a) (b)

Source: Authors' work

This observation may, even though partially, reduce the dataset dimension, since only a few cases can be considered complete (i.e. those for which all fields are valued).

The words extracted from the text are finally processed by using a "bag-of-words" representation. The n-grams are extracted in the form of unigrams, bigrams, and 3-grams frequencies, which have been stored in ad-hoc data structures (McCallum, 2005).

The data pre-processing also shows that the notes reported by the pathologist highlight information like the cancer type. Note that it is possible that the tokens position may be switched within the considered n-gram (eg: "lobular carcinoma" vs "ductal carcinoma", or "infiltrating carcinoma" vs "in situ carcinoma").

From an empirical analysis performed by the experts, it comes out that there exist several correlations among data that co-occur in a given diagnosis. Different regular expressions are implemented in order to define those that better fulfil the representation and correlations observed among the input data of the ML classifier. In particular, the main information acquired by each instance of the dataset corresponds to the extraction of the following information:

- Marker value: this information is a couple (marker, value), where significant markers, by a manual analysis of pathological findings, such as BCL2, KI67 (that indicates tumour growth), P53 (it works as tumour suppressor; also called the "guardian of the genome"), HER2, OESTROGEN, PROGESTERONE, TOPOISOMERASE.
- Tumour score: represents the score of the tumour. The resultant of the corresponding regular expression represents a grade value, associated with a score value as shown below, indicating a different growing of the tumour cells.

This system grades breast tumours based on the following features: Tubule formation: how much of the tumour tissue has normal breast (milk) duct structures; Nuclear grade: an evaluation of the size and shape of the nucleus in the tumour cells; Mitotic rate: how many dividing cells are present, which is a measure of how fast the tumour cells are growing and dividing.
Each of the categories gets a score between 1 and 3; a score of "1" means that the cells and the tumour tissue look mostly like normal cells and tissue, while a score of "3" means that the cells and tissue look mostly abnormal. The scores for the three categories are then added, yielding a total score ranged between 3 and 9. Three grades are possible:

- Grade 1 o well differentiated (score (3-5): slow growing, similar to a normal breast tissue.
- Grade 2 o Moderately differentiated (score 6-7): the growing is faster than those identified with a lower score.
- Grade 3 o poorly differentiated (score 8-9). Tumour cells appear to be very different from normal cells and have a higher and faster probability of growing and spreading w.r.t. the two previous grades.

- Tumour Staging: it is based on the TNM classification, which expresses information related to certain anatomical characteristics of the tumour itself, such as: T (Tumour): indicates the size and extent of the tumour. N (Nodes) indicates the evaluation of the involvement of regional lymph nodes. M (Metastasis): indicates the presence/absence of metastases far from the primary tumour.
- Type of tumour: correspond to the types classified by the multi-label Machine Learning approach; they are: lobular, ductal, in situ and infiltrating.

The correlations among data observed by the experts have been expressed by defining regular expressions. An example of the regular expression for the extraction of the tumour staging value is reported as follows, where "t_n" represents the tumour dimension, "tx" an unidentifiable tumour location, and "ptis" an "in situ" tumour location.

```
staging_regex = r'' + staging_type
staging_types = ['t1','t2','t3','t4','tx', 'ptis']
```
(1)

Another example of regular expression is reported as follows, in order to extract the marker's value: "marker_regex'' is a 171oolean value representing the presence/absence of a marker. "value_regex'' corresponds to the value of the marker reported in the pathology report.

```
marker_regex = r'' + marker + ' .*\n(.*)'
value_regex = r''([.,]?[0-9][.,]?[0-9]*).*%''
```
(2)

Regular expressions may be nested. In this case the output of the evaluation of the first expression is used as input of the second one. The first expression searches for the marker label in the text, and returns the text line following the line containing the marker to the second expression. The second expression searches in the input text line for a token like "number %" and returns the number (value of the marker).

The info extracted in this step are used as features of the classifiers. The aim of this approach is to classify the markers w.r.t. the cancer diagnosis in order to understand which are the most important for each class. This classification is important to give useful suggestions to the pathologist during the tissue sample analysis. Based on these

suggestions the pathologist chooses on which markers to focus after a first set of preliminary results, avoiding to observe un-useful markers and optimizing the analysis process.

Description of the classifiers adopted in the approach

The dataset has been divided into two subsets, with, respectively, 80% of the data used for the training (randomly chosen), and the remaining 20% for the test. In this project we adopted a step by step strategy, in which several solutions were studied, developed and compared to identify the most promising. In the first solution four single-label binary classificators were developed, one for each category involved into the classification problem, i.e., lobular, ductal, in situ and infiltrating. Anyway, it was soon clear that with this solution none of the different dependencies of the outputs have been taken into account during the classification. The result is an approach too simple, that leads to an unrealistic view of the complexity of the problem being analyzed.

For this reason a multi-label approach has been then defined, by considering the overlapping of the features provided as input and of the corresponding output classes, respectively divided into tumour type (ductal/ lobular) and tumour infiltration (infiltrating/in situ), since a tumour could be both ductal and lobular and simultaneously in situ and infiltrating. As reported in Table 2, several multi-Label algorithms have been applied and tested during the classification process (Madjarov et al., 2012).

Among them, the best results are obtained, for both the classifiers, by using the adaptive Multi-Label-KNN algorithm. The main idea of such an approach, indeed, is that an instance's labels depend on the number of neighbours that possess identical labels (Liu & Cao, 2015). In particular, given an instance x with an unknown label set defined as "L(x) <= L", the algorithm first identifies the K nearest neighbours in the training data and counts the number of neighbours belonging to each class. Then the maximum a posteriori principle is used to determine the label set for the test instance. In other words, KNN, which searches within the example space for the nearest neighbour to the element being classified, i.e. the most similar k elements, and associates it with the most common class among them. If "K=1", the element under examination is assigned the same class as its closest neighbour.

As reported in the literature (Liu & Cao, 2015), two main advantages arise from the ML-KNN and regard, respectively, the decision boundary that can be adaptively adjusted due to the varying neighbours identified for each new instance, and the usage of probabilities, estimated for each class label, in order to solve the class-imbalance problem.

Preliminary Experiments and Prototype

The prototype is implemented with the Python v3 programming language, by using Scikit-Learn as Machine Learning Library.

The prototype is then tested on a dataset of 2186 medical records provided by national institutes for pathological analysis (omitted for privacy reasons), with a number of positive cases equal to 1952 and a number of negative cases equal to 234. The dataset is unbalanced towards positive cases because these analyses are carried out on people suspected of the presence of a percentage (%) disease, in our case breast cancer.

The results of a set of preliminary experiments are detailed in Table 2. The accuracy values of the implemented classifiers confirm that, as previously reported, the better

accuracy is reached by considering an adaptive algorithm, as K-NN, while lower accuracy is obtained for all the other implemented algorithms, that are led to develop a "transformation" approach, from multi to single label.

Table 2
Preliminary Comparative Results

Multi Label	Lobular/Ductal		Infiltrant/in Situ	
Machine Learning	Accuracy	MSE	Accuracy	MSE
KNN	0.810502	0.189498	0.625570	0.374430
OneVsRest	0.745313	0.254687	0.566529	0.433471
Gradient Boosting	0.273519	0.726481	0.541107	0.458893
Binary Relevance	0.381278	0.618722	0.545662	0.454338
Label Powerset	0.504566	0.495434	0.461187	0.538813

Source: Authors' work

Figure 3 shows the interface of the prototype, histograms highlight for each type of classification, the data divided into positive and negative (true/false) cases, predicted and observed for each dataset analyzed.

The pre-processing phase drastically reduces the size of the dataset, by eliminating all the duplicates, and by grouping the markers used by the pathologist, for each medical record. Therefore, the markers used in the final dataset are 109.

The data are then processed by the multi-label ML Classifier by using the "Process Data" command, as shown in Figure 3 (a). The result produced by the application is graphically represented by means of four histograms, corresponding to the four types of cancer classification adopted (two for tumour type and two types of tumour infiltration).

Both Type and Infiltration classifiers show with different colours, blue and orange respectively, the expected results and the results predicted by the ML classifier.
The predicted results obtained from the implemented classifiers are very close to the observed results of the pathologist. The differences between the histograms are mainly due to the differences in the vocabulary used in the textual reports by different pathologist. A semantic approach would be useful to overcome this issue.

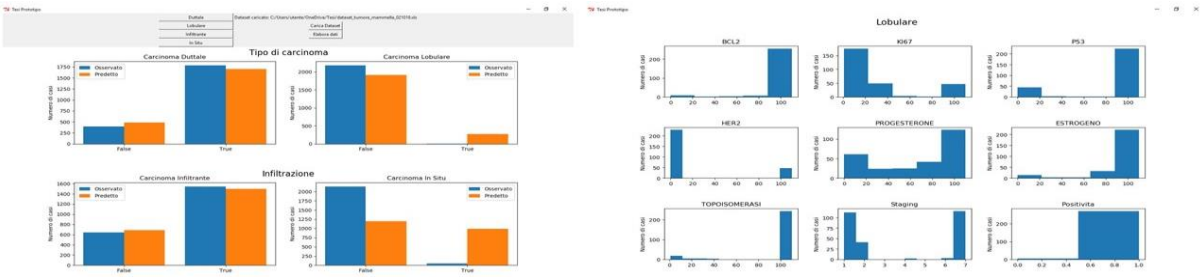
Finally, the four classification types (i.e. ductal, lobular, infiltrating and in situ) reported on the left side of the prototype window, allow the user to visualize further information regarding the marker values that have been selected for the classification process. In particular, as shown in Figure 3(b), for each marker for the lobular type, statistics regarding the percentage of positivity are plotted into histogram representations, together with the number of cases registered in the entire analyzed dataset.

For instance, the histogram of positivity, shown in Figure 3(b), indicates that a lobular type of tumour has been detected. Moreover, for most of the patients affected by such a tumour, the histograms highlight how some markers reach high values of positivity percentage, as for example by BCL2, P53, Oestrogen, and Topoisomerase, while, on the other hand, for others markers, a low percentage positivity is reached, as for KI67 and HER2.

A further analysis has also emphasized the potential impact of the coexistence of different information (e.g. tumour and non-tumour pathologist) over the report content. It is therefore essential to identify reports containing several information.

These elements could have a strong impact over the final decision carried out by a human expert, since several target classes could be allocated into single clinical report, by simplifying the overall evaluation process. Such an advantage suggests the overall benefits of a multi-label classification.

Figure 3
Example of Prototype Screenshots: Predictive Classification (left side (a)) and Example of Extracted Marker values for the Lobular Classification (right side (b))



(a) (b)

Source: Authors' work

Conclusion

The work proposes the design, implementation and start-up of services aimed at supporting cancer telepathology, with the aim of improving the quality of health services in the territories of reference, supporting the construction of solutions for cancer care and encouraging the creation of excellence in the regional and national oncology areas.

The results obtained show that even in the presence of a poor quality dataset, little data available to train the machine learning model, unbalanced dataset, and difficulty in interpreting the free text written by the pathologist, it was possible to create a tool that would support the pathologist in writing the report and optimizing the analysis steps, thus reducing times and costs.

One of the future developments will be the definition and adoption by pathologists of a standard synoptic report for reports generation. All these reports will be automatically validated through the use of the multi-Label Machine Learning prototype, currently under development and the subject of this contribution.

References

1. Bifet, A., Frank, E. (2010), "Sentiment knowledge discovery in twitter streaming data", in the Proceedings of the International Conference on Discovery Science, Canberra, ACT, Australia, Springer, pp. 1-15.
2. Ceci, M., Malerba, D. (2007), "Classifying web documents in a hierarchy of categories: a comprehensive study", Journal of Intelligent Information Systems, Vol. 28, No. 1, pp. 37-78.
3. Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W., De Groen, P.C. (2009), "Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model", Journal of biomedical informatics, Vol. 42, No. 5, pp.937-949.
4. Cohen, A. M., Hersh, W. R. (2005), "A survey of current work in biomedical text mining", Briefings in Bioinformatics, Vol. 6, No. 1, pp. 57-71.
5. Combi, C., Pozzani, G., Pozzi, G. (2016), "Telemedicine for developing countries", Applied clinical informatics, Vol. 7, No. 4, pp. 1025-1050.

6. Garcia-Remesal, M., Maojo, V., Billhardt, H., Crespo, J. (2009), "Integration of Relational and Textual Biomedical Sources A Pilot Experiment Using a Semi-automated Method for Logical Schema Acquisition", Methods of Information in Medicine, Vol. 49, No. 5, pp. 337-348.
7. Jouhet, V., Defossez, G., Burgun, A., Le Beux, P., Levillain, P., Ingrand, P., Claveau, V. (2012), "Automated classification of free-text pathology reports for registration of incident cases of cancer", Methods of information in medicine, Vol. 51, No. 3, pp. 242-251.
8. LeCun, Y., Bengio, Y., Hinton, G. (2015), "Deep Learning", Nature, Vol. 521, No. 7553, pp. 436-444.
9. Liu, C., Cao, L. (2015), "A coupled k-nearest neighbor algorithm for multi-label classification", in Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Ho Chi Minh City, Vietnam, Springer, pp. 176-187.
10. Li, Y., Martinez, D. (2010), "Information extraction of multiple categories from pathology reports", in the Proceedings of the Australasian Language Technology Association Workshop, Melbourne, Australia, University of Melbourne, pp. 41-48.
11. Madjarov, G., Kocev, D., Gjorgjevikj, D., Deroski, S. (2012), "An extensive experimental comparison of methods for multi-label learning", Pattern Recognition, Vol. 45, No. 9, pp. 3084-3104.
12. McCallum, A. (2005), "Information extraction: Distilling structured data from unstructured tex", Queue, Vol. 3, No. 9, pp. 48-57.
13. Meystre, S., Savova, G., Kipper-Schuler, K. (2007), "Extracting information from textual documents in the electronic health record: A review of recent research", Yearbook of Medical Informatics, Vol. 17, No. 1, pp. 128-144.
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013), "Distributed representations of words and phrases and their compositionality", in the Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, USA, Curran Associates, Inc, pp. 3136-3145.
15. Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Cohen, K. B., Duch, W. (2007), "A shared task involving multi-label classification of clinical free text", in the Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Prague, Czech Republic, Association for Computational Linguistics, pp. 97-104.
16. Sarioglu, E., Yadav, K. Choi, H. (2013), "Topic modeling based classification of clinical reports", in the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, ACL, pp. 67-73.
17. Tsoumakas, G., Katakis, I. (2007), "Multi-label classification: An overview", International Journal of Data Warehousing and Mining, Vol. 3, No. 3, pp. 1-13.
18. Turian, J., Ratinov, L., Bengio, Y. (2010), "Word representations: a simple and general method for semi-supervised learning", in the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, Association for Computational Linguistics, pp. 384-394.
19. Viviani, M., Pasi, G. (2017), "Credibility in social media: opinions, news, and health information - a survey", Data Mining and Knowledge Discovery, Vol. 7, No. 5.
20. Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C. (2004), "Recognizing names in biomedical texts: a machine learning approach", Bioinformatics, Vol. 20, No. 7, pp. 1178-1190.

About the authors

Antonia Azzini received the Master Degree and the PhD in Computer Science at the University of Milan. She has a long experience as PostDoc at the Computer Science Department of the University of Milan. Actually, she is an ICT Researcher at Consortium for the Technology Transfer, and She is responsible of the cooperation with universities and other research centres. She is member of the Editorial Board of IJKL Journal, and she is a member of the IFIP 2.6 Working Group on Databases. In the last years She was also Co-Organizer of several events. The author can be contacted at **antonia.azzini@consorzioc2t.it.**

Nicola Cortesi, is a Junior researcher at Consortium for Technology Transfer C2T and PhD student at University of Bergamo in Computer Science. The theme of his thesis regards frequent itemset mining for Big Data. He has a degree in Computer Science Engineering at University of Bergamo (Italy). His main research interests are data analysis, data mining, machine learning, distributed algorithms, geoinformatics, fintech. The author can be contacted at **nicola.cortesi@consorzioc2t.it.**

Stefania Marrara is an ICT researcher at the Consortium of Technology Transfer. With a PhD in Information Engineering at Politecnico di Milano, and with a long experience as a PostDoc in both the University of Milan and the University of Milano Bicocca, she is responsible of the cooperation with Universities and other research centres, and she collaborates within several projects of the Consortium and of the Start-Up Find Your Doctor. She is member of the Editorial Board of IJWET. In the last years Stefania Marrara was also Co-Organizer of several events. The author can be contacted at **stefania.marrara@consorzioc2t.it.**

Amir Topalovic received his Master Degree in Finance and Risk Management at the University of Parma and he is actually a PhD student in Business and Administration at the University of Sarajevo under the supervision of Prof. Nijaz Bajgoric. His Thesis title is "A TOE based Study on Data Mining employment in Italian Small and Medium Enterprises". Actually, he is CEO at C2T, Consortium for the Technology Transfer, and his research interests are related to data mining in corporate and banking data. The author can be contacted at **amir.topalovic@consorzioc2t.it**.