IJBPAS, September, 2015, 4(9), Special Issue: 327-342

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

DESIGN AND IMPLEMENTATION OF A STRUCTURED ELECTRONIC FORM FOR CELIAC DISEASE PATHOLOGY REPORT: AUTOMATIC MARSH CLASSIFICATIONDIAGNOSIS

FARZANEH KHADEM SAMENI[*1], AZADEH KAMEL GHALIBAF[2], KOBRA ETMINANI[3]

[*1] Assistant Professor, Pathologist, Faculty of medicine, Islamic Azad University, Zahedan brunch, Iran.

[2] PhD Candidate in Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran

[3]Faculty Member of Mashhad University of Medical Sciences, Mashhad, Iran

[*]**Corresponding Author: E Mail: fsameni@iauzah.ac.ir**

ABSTRACT

Pathology is generally reported in unstructured text format and contains a complex web of relations among medical concepts. In order to enable computers to understand and analyse report's free text, attempts were made to convert concepts and their relations in a structured form. To this sense, the training, validation and evaluation of this implementation study was based on a corpus of 258 pathology reports, randomly selected from two pathology labs, with a diagnosis of celiac disease on the report. The proposed system consists of three phases: standardization of celiac disease pathology report (Delphi technique with three experts), information extraction from free text reports (Text mining techniques using Stanford parsers), and automatic classification of celiac disease stages in marsh system (Decision tree classifier j48 algorithm). Information were extracted from free text pathology report and each piece of information was assigned to the associated pre-defined field in standardize template form with 76% accuracy. After determining marsh stage for each report in the third phase, our system overall accuracy showed 62% on average. Evaluation of the third phase as an independent system with manually

corrected, gold-standard input reached an amount greater than 84% accuracy. Findings indicated that benefits of standardized synoptic pathology reporting include enhanced completeness and improved consistency, avoidance of confusion and error, and facilitation of faster and safer transmission of critical pathology data in comparison with narrative reports.

Keywords: Text Mining, Celiac disease, Synoptic Reporting, Delphi Techniques, Decision Tree

INTRODUCTION

Celiac disease is an autoimmune disease one of its symptoms is Mucous Tissue Damage of small intestine occurring as a result of consumption of foods containing gluten [1]. 'Gluten' is an umbrella term referring to one type of protein in wheat, rye and barley. When this protein is taken by people with Celiac, it can cause degradation and damage to villous of small intestine and consequently dyspepsia and lack of absorption food and vitamins and, if not cured pertinently, other diseases such as anemia, lack of bone compaction, weight reduction and many other problems [2]. Different studies have reported that about 1% of population has this disease [3]. Furthermore, probability of catching Celiac is more for those whose close relatives have the same disease or for those who have autoimmune disorders, Diabetes 1, and turner and down syndromes [4]. Most people with Celiac have whether no symptom or non-specific digestive symptoms such as indigestion, abdominal aches, bloat and disorder in intestines processes and this makes diagnosis of Celiac disease more difficult [5]. Diagnosis of the disease is not possible only through clinical symptoms; results of Serologic and biopsy tests of small intestine help doctor in coming at a correct diagnosis.

After Celiac has been rectified by the laboratory tests and the present results, a small part of small intestine might be removed through biopsy after final diagnosis and identification of the amount of damage to intestine villous. Biopsy is considered as critical criterion in diagnosis of Celiac and the pathologists report their observations of biopsy characteristics in form of a written report [1]. In 1992, an English doctor named Marsh introduced a four-level classification system for standardization of the amount of damage to the intestine texture [6]. Few years later, in 1999, Oberhuber made some revisions to Marsh classification and categorized these damages into three levels, namely Marsh I, Marsh II and Marsh III; Marsh III includes three sub-classes of a, b

and c [7]. Nowadays, in field of pathology, use is usually made of this classification as a criterion for diagnosis in referenced pathology texts [8, 9 and 10]. Reports of pathology are in form of 'free text' describing results observed in cells. Content of the report include a network of relations among medical concepts and the doctor uses it for diagnosis and reasoning. However, if one would like to seek help from a computer for diagnosis and analysis of these reports, they need to present relations and concepts in form of a pre-established and standard structure. Data processing and extraction is an important step in content analysis and exploration of the texts [11].

So far, many systems have utilized text processing for the automatic processing of pathology texts, one of which was McCowan et al. [12] who designed a system for automatic determination of the step of lung cancer in 2006. Using methods of natural tongue processing, this system replaced report text with standard terms in UMLS (Unified Medical Language System) and then, using weighting method called LTC, transforms data into a numerical vector. This vector which the compacted form of the report is considered as the input to classification algorithm SVM (Support Vector Machine) so as to determine the level involved in cancer. Overall accuracy of the system has been reported to be 74%. Similarly, 13 and 14 have made use of text processing methods to analyze pathology reports of lung cancer.

Another study carried out on 1038 instances of pathology reports of lymphoma in Massachusetts Public Hospital in 2014 [15] showed that automatic classification of lymphoma from pathology reports had 85% accuracy in a model in which the sentences were analyzed in form of a correlation graph; this was the best performance in comparison with other characteristics sets.

In 16, a study which was carried out in 2006 in France, researchers classified 5121 pathology reports of free texts which were dedicated to 35 pathologists using SVM and Naive Bayes based on cancer textures. The criterion of report classification in this study was the method suggested by IARC and results indicated 96% for F1. In this study, to represent the report, use was made of weight frequency TF-IDF which transforms text into a numerical vector based on the frequency of words.

Cancer Biomedical Information Network (caBIG), which was developed by National Cancer Institute of America and aimed at providing easy access and safe transaction of cancer findings, used a system named

caTIES [17] to extract coded information from free text clinical documents such as pathology reports of surgery. This system has been designed and developed in Subscribe to Pathology Information Network (SPIN) in Pittsburg University and allows researchers to ask and make research works such as clinical trials on extracted data from texts besides offering facilities for graphic demonstration of obtained results of concept-based content analysis. In addition, many processing systems of medical language processing systems have come together in a same group aiming at identifying patients with specific clinical characteristics [18, 19, 20 and 21]. As it was previously noted, many symptoms, usually irrelevant, has made diagnosis of Celiac difficult to the doctor and so they need to request different tests, collect complementary data and finally come to correct diagnosis by putting these next to each other. Structuring celiac disease pathology reports is significant since reasoning based on these volumes of data is very time-consuming and increases possibility of error. One standard proposed by Cancer Commission of college of American Surgeons for more validation of laboratories is making use of scientific validated data elements (SVDE) which necessitates pathologists prepare synoptic reporting using structured data [22]. However, no study has so far made use of text mining to analyze and standardize celiac pathology reports. Therefore, two main objectives have been aimed by designing a standard form for pathology reports: 1) making sure of registering necessary information by the pathologist and 2) faster and more accurate restoration of data information from reports by the physician and making an easier and more accurate decision.

Following sections are as follows: in section 2 methodology and method of implementation of these steps are explained. In section 3, findings and results are presented and finally section 4 discusses and investigates findings and influential factors.

MATERIALS AND METHODS

In this section, firstly data are introduced and described. Then, total structure of the system are explained and each step is comprehensively stated.

Data sets

Data set in the present study includes 258 Duodenum biopsy reports collected by two pathologists whose final diagnosis has been celiac disease. These reports were written from 2009 to 2014 in English in form of free texts, each report was in one page and includes three parts: Macroscopy,

microscopy and final diagnosis. The first part includes outward characteristics of the sample like size, weight, color and outward changes which can be seen by eye. Second part, i.e. microscopic part, takes account of characteristics of cells and textures under the microscope which serves as the criteria for diagnosis and determination of damage. As for observation of moral principles, demographic information of patients has been omitted.

General structure of system

Suggested system in this study includes three main phases which are shown in figure 1. In the first phase, a standard structured form has been prepared for pathology report of Celiac using Delphi method.

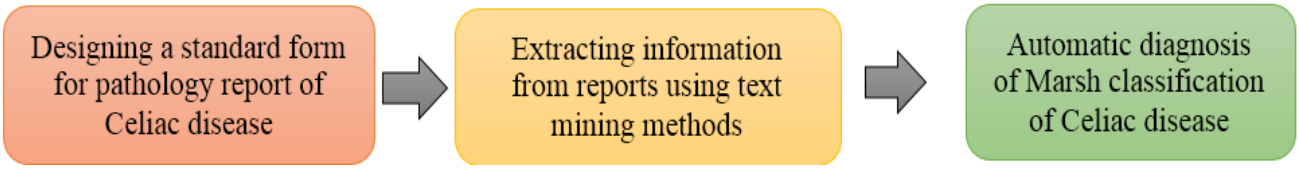


Figure 1: proposed model for automatization of diagnosis process of pathology reports of Celiac disease

Delphi method is a structured process for collection and classification of experts' and specialists' knowledge by distributing questionnaires among them and then welcoming their comments and responses [23]. Its validation is not based on number of participants but it is based on their scientific knowledge. Participants can range from 5 to 20 people.

In the second phase, information are extracted from reports and saved in standard forms designed in the previous section; to this end, use has been made of many language processing methods and text mining procedures. Natural Language Processing has many potential applications in healthcare and research fields. Although many patients' data are restorable from their electronic files, some parts which are in form of text such as hospitalization reports, file summaries, radiology and pathology reports are not directly accessible [24]. To solve this problem some techniques need to be designed so as to be able to organize and extract data in the texts. NLP methods help management of immense volume of texts such as patients' reports by extracting relevant data at the appropriate time [25].

Finally for each report, related Marsh class would be defined automatically using Machine learning methods. Machine learning is a very useful branch of artificial intelligence which offers methods for education and instruction of computer. One of these methods is learning by a supervisor where sets of 'input-output' pairs are offered

to the system for learning and the system attempts to acquire a function of input to the output. In what follows, implementation of each phase will be explained in detail.

Phase 1: designing a standard form

Early steps involved in designing a structured form for celiac reports are shown in figure 2. In the first step, list of necessary fields are extracted and prepared in form of a checklist before investigating the existing reports and with coordination of a pathologist. Then, the list would be given to three expert pathologists so as to have their ideas about necessity or no necessity of each item. Also, the experts can give opinion about addition of a new item or format of data input (checkbox, drawer lists, etc.). After collection and analysis of responses items with the most agreement, i.e. more than two positive ideas, would be incorporated into a standard form. Attached can be seen a standard form and a sample text of celiac pathology report.

Creating a standard format allows the doctor or the computer search specific type of data or figure out what information element belongs to what information group and consequently make restoration, transference and interpretation of data easier.

Phase 2: extraction of data from free text reports

In this phase, attempts have been made to extract information from reports using a set of methods and tools of text mining and organize them in a standard form prepared in the first phase. Overall illustration of the model can be seen in figure 3 which will be explained more in what follows.

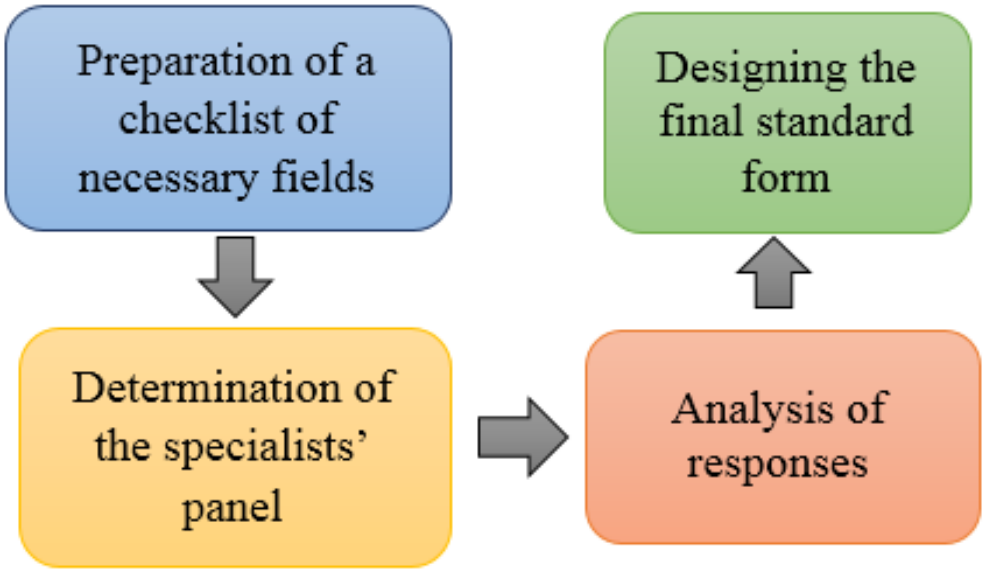


Figure 2: designing a standard form for celiac pathology report

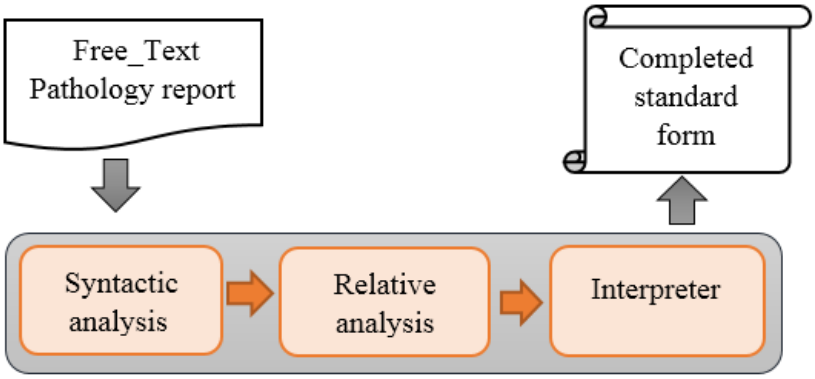


Figure 3: steps involved in extraction of data from reports

Main purpose of text processing is to transform it into a form readable for the computer. To this end, use is made of calculation theories, algorithms and data structures in computer sciences. Often, the first step involved in text mining is pre-processing through which information are saved in an appropriate data structure for next processes. One task performed in this step is identification of word limitations, setting sentence boundaries, determining parts of speech of words, etc.

In syntactic analysis, sentences are synthesized into their components such as noun phrase (NP), verb phrase (VP), adjective phrase (ADJP), adverb phrase (ADVP), etc. Moreover, Syntactic Parser allocated part of speech (POS) of each word in the sentence. Some tags referred to in the figure include determiners (DT), nouns (NN), adjectives (JJ). A comprehensive list of POS tags and their related grammatical roles can be observed in [27].

Relative syntax is one field of linguistics and has been formulated based on Theory of Vocabulary Capacity. This theory states that each word has relations according to its capacity and its place in the sentence. Relative syntax identifies these syntax/semantic relations in the sentence.

With help of POS tags syntactic groups and their relative functions one can come to meaningful information about semantic structures in the sentence. After synthesizing the sentence into syntactic groups and identification of their relations, parts appropriate for each place must be inserted. To this aim, a mediating program was designed with Visual Basic in Visual Studio 2013. This program analyzed text content using information of POS tags, syntactic groups and their relative function and then completes the standard form with appropriate data and present them in the output. Since values for each field are usually in sentence structure as adjective or adverb, the program

firstly searches keyword in each field in relative analysis of the text and extracts all adverb-adjective relations with the keyword as the core. If findings are more than one, to find the most related pair, we would search for a noun phrase, adjective phrase or adverb phrase with both core and relative words in it. In this case, relative would be placed across the related field in the form as the identifier. Typing errors in the report and structural diversity of the sentences can be counted as reasons of error in this system. In addition, it is expected that we could reach more accuracy in text mining by employing more developed techniques.

Phase 3: determination of Marsh class

After text information are transferred to a standard form, the appropriate Marsh class would be automatically set for the mentioned features in this phase and this makes pathologist's job easier. According to Oberhuber revised system, Marsh class has been set based on three characteristics of number of Intra-epithelial, Crypthyperplasia and amount of analysis of intestine villus; these are present in figure 1.

Learning algorithm used in the present study is decision tree J48. Input values are a set of features defined in table 1 and output class is appropriate according to Marsh class. Procedure is in a way that firstly sets of data which have been correctly determined by the specialists are entered into data algorithm as the instructional data sets and then decision tree decides about new data with no pre-set Marsh class by creating a set of 'if-then' rule. Procedure has been shown in figure 4.

Table 1 – Marsh Revised Classification

Marsh Type	ILE*	Crypts	Villi
0	<40	Normal	Normal
1	>40	Normal	Normal
2	>40	Increased	Normal
3a	>40	Increased	Mild atrophy
3b	>40	Increased	Moderate
3c	>40	Increased	Severe

*Intraepithelial Lymphocytes (per 100 Enterocytes)

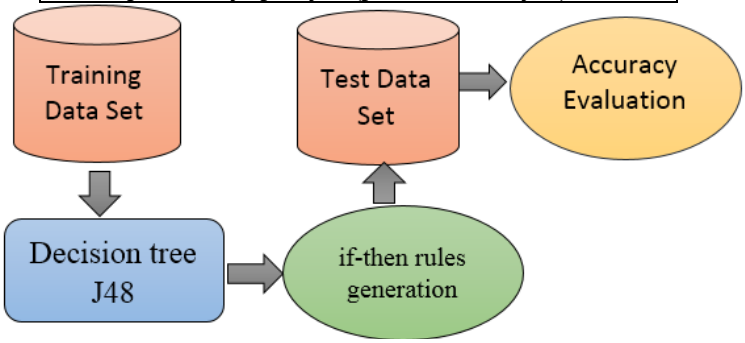


Figure 4: automatic allocation of Marsh class to each report

RESULTS

Out of 258 reports in text, 186 belonged to Marsh I class, 14 belonged to Marsh II class and 58 belonged to Marsh III class. Moreover, regarding sex, 159 reports belonged to female patients and 99 belonged to male patients. Here, to analyze content of reports, use has been made of tools offered by Linguistics Center of Oxford University [28]. Figure 5(a) illustrates syntactic synthesis for a sample sentence and part b shows synthesis of grammatical relations for the sentences. As it is evident in the figure, output of this parser is as 'core-relative' pairs and type of relativeness is stated as a function of this pair (numbers next to words indicate ordinal numbers of that word in the sentence). For example, relationship between words 'hyperplastic' and 'mildly' in the sentence has been determined by a relative function of adverb modifier (ADVMOD) which shows that hyperplasia or crypts distribution have been balanced.

Figure 5: output of syntactic analysis and synthesis of relativeness for a sample sentence

"The lamina propria is expanded with numerous lymphocytes and plasma cells and the crypts are mildly hyperplastic."

a- analysis of syntax	B- analysis of relativeness
(ROOT (S (S (NP (DT the) (NN lamina) (NN propria)) (VP (VBZ is) (VP (VBN expanded) (PP (IN with) (NP (JJ numerous) (NNS lymphocytes) (CC and) (NN plasma) (NNS cells)))))) (CC and) (S (NP (DT the) (NNS crypts)) (VP (VBP are) (ADJP (RB mildly) (JJ hyperplastic)))) (. .)))	det(propria-3, the-1) nn(propria-3, lamina-2) nsubjpass(expanded-5, propria-3) auxpass(expanded-5, is-4) root(ROOT-0, expanded-5) amod(lymphocytes-8, numerous-7) prep_with(expanded-5, lymphocytes-8) nn(cells-11, plasma-10) prep_with(expanded-5, cells-11) conj_and(lymphocytes-8, cells-11) det(crypts-14, the-13) nsubj(hyperplastic-17, crypts-14) cop(hyperplastic-17, are-15) advmod(hyperplastic-17, mildly-16) conj_and(expanded-5, hyperplastic-17)

As for evaluation of the accuracy of function of interpreter's program, 54 sample reports were randomly given to the program and, out of 432 input fields, 327 were properly valued; in general, accuracy rate was 76%. Regarding size of sample according to [29], results can be over generalized to all data with 90% of confidence interval.

In figure 6, graphic illustration of decision tree and part b in the same figure indicate a sample form of produced rules in the tree; numbers in parenthesis shown number of all instances reached this nod in their order and the second number shows number of instances have been misclassified. To implement algorithm of decision tree J48, use

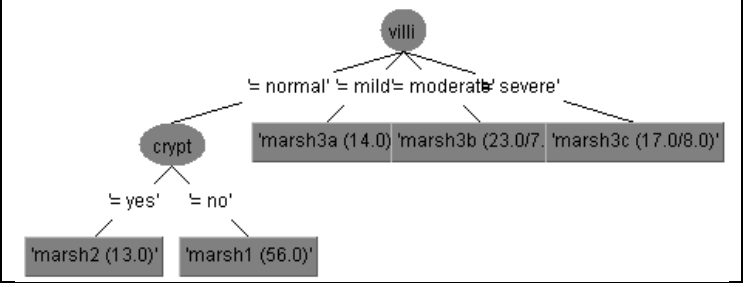
was made of WEKA software version 11, 6, 3. Weka is a free text mining software and supports many learning algorithms [30].

Figure 6: learning algorithm of decision tree J48

b- some rules of decision tree

villi = normal
| crypt = yes: marsh2 (13.0)
| crypt = no: marsh1 (56.0)
villi = mild: marsh3a (14.0)
villi = moderate: marsh3b (23.0/7.0)
villi = severe: marsh3c (17.0/8.0)

a- graphic illustration of decision tree



Findings related to automatic classification of Marsh class have been reported in two states. One concerns a situation in which obtained findings from previous phases enter the system as input and error in previous steps affects findings in subsequent steps. In other state, system input has no error and has been completed by an expert and in this case efficiency of the phase is evaluated independently. Accuracy of the system function is 62% for the first state and 84% for the second state. Table 3 shows details of implementation of the system in the second state with values of other evaluation factors such as 'precision' and 'recall' and F factor. Standard accuracy rate (TP) in the table indicates proportion of correct diagnosis to total number of diagnoses; this value is indicated by FP for incorrect diagnoses. Comparison of values of these

two factors have been observed for different Marsh classes it was observed that regarding Marsh classes 11 and 2, a total of 77% of sample, nearly all instances were diagnosed correctly by the system. So one can expect that if we increase number of instances for Marsh class III, system would have a better function in diagnosis of these cases.

Classification method of instructional and test packages for learning algorithm was through Fold Cross Validation method where primary data sets are divided into 10 equal parts. Then, in each classification, one part is chosen for test step and other parts are used in the instruction step. This process is reiterated 10 times in a way that each data recession occurs exactly one time in the test.

Table 3- results of automatic classification of Marsh class for input with no error

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	class
1	0	1	1	1	1	**Mars1**
1	0	1	1	1	1	**Marsh2**

0/7	0	1	0/7	0/82	0/95	**Marsh3a**
0/66	0/11	0/59	0/66	0/62	0/91	**Marsh3b**
0/5	0/071	0/38	0/5	0/43	0/88	**Marsh3c**
0/84	0/027	0/87	0/84	0/85	0/96	**.Weighted Avg**

DISCUSSION

Studies so far carried out on application of computer-based methods in pathology can be categorized into two categories, namely pathology photo processing [31] and pathology text processing. Studies regarding processing of pathology texts are less than the other category and have often made use of statistics-based methods; these methods are not able to show deep relations and concepts in the text. For example, Jouhet et al [16] utilized 'frequency vector of terms in the text (TF-IDF) to restore texts. In another study carried out by Li and Martinez [32] use was made of regular expression to extract information from pathology report texts via BOW (bag of words) model. In this way, making use of linguistics methods and language processing techniques can be a strong point of the present research.

The present model here aimed at introducing a framework for structuring free text reports in medicine, particularly pathology reports of celiac disease. Findings indicated acceptable performance of system in automatic diagnosis of outcome. Leo et al. [15] aimed at categorizing pathology reports of lymphoma cancer with 87% accuracy and [14] proved an accuracy of 95% concerning lung cancer. Among all English and Persian texts done up to now, no study has taken account of investigation and analysis of pathology reports of celiac disease and therefore comparison is not possible.

One advantage of structuring free text report is possibility of supervision and evaluation of the pathologist's performance. In the second phase of study it became that at least 30% of instances in one or two fields have no value during the process of transforming reports from text to form. This indicates information errors in report texts. This can be due to doctor's ignorance in reporting them.

Concentration on a specific disease with a limited and fixed set of specific vocabulary makes text processing better and more efficient. Also, pathologists usually use similar syntactic and semantic structures and this can lead to the most accurate processing of the text. On the contrary, relatively low prevalence of this disease makes data accessibility more difficult which is in itself one limitation of the present system.

CONCLUSION

Since celiac disease is hard to diagnose and there are this might lead to dangers (20% of

celiac patients might get intestine cancer if not cured properly) necessity of designing facilitating and supportive systems in really felt here. In the present study, a model was designed and implemented for structuring pathology reports of the celiac disease so as to facilitate entry and restoration of information, improvement of quality of data, increase of readability of reports, possibility of computer processing of data and finding relations and patterns with research-based goals. As it was also expressed in third phase of the study, after data transference to structured electronic forms use was made of leaning machine algorithm of decision tree for data processing and automatic classification of Marsh class of reach report and finally great function of the system was seen. Possibility of report data processing helped research and instructional activities greatly.

Suggested model in this study can be applied for structuring reports related to pathology reports, other diseases and cancers or any other free text report in medicine such as radiology, patients' charts, etc. As free text reports improve efficiency, accuracy and speed of doctors' decision-making, report forms understandable to the patients can be designed in the future.

Moreover, it is recommended that the present model here can be replicated with more diverse and greater number of reports in future researches. In other studies, performance of other text mining and processing methods can be investigated in analysis of texts. In addition, future researches can take account of doctors' satisfaction and efficiency of the structured form when incorporating free texts. Since only few researches have accounted for this, the present authors hope that the present findings would serve as basics of more comprehensive future studies.

ACKNOWLEDGEMENT

Gratitude of the present authors goes to pathobiology personnel in Zahedan University, particularly Mr Tabatabaei, chief of laboratory, for their sincere cooperation.

MACROSCOPY:Received specimen in formalin consists of 4 tiny creamy colored soft tissue fragments T.M 0.6 cm. SOS=4/1 E=100%

MICROSCOPY; Sections show duodenal mucosa with partial villous shortening and atrophy. The lamina propria is expanded with numerous lymphocytes and plasma cells and the crypts are mildly hyperplastic. The intraepithelial lymphocytes are increased (50 per 100 enterocytes).

DIAGNOSIS: Duodenum mucosa ,D2, biopsy;
-Partial villous shortening and increased intraepithelial lymphocytes,
compatible with celiac disease Marsh IIIA.

Attachment 2: Assuming that demographic data are registered at the time reception; in this form, necessary agreed fields mostly agreed by panel of specialists of pathology report of celiac disease have been considered.

Date __/__/__

No. of biopsies ..... Oriented ..... Non-Oriented.....

Villi: ☐ normal atrophy ☐ Mild ☐ Severe

Villus/Crypt Ratio: normal [1:3] ..... altered .....

Intraepithelial Lymphocytes: normal..... increased .....

Evaluation with CD3 .....

Glands: normal ..... hyperplastic .....

Lamina Propria .....

Diagnosis (Oberhuber-marsh):

☐ Type 1 ☐ Type 3a ☐ Type 3c
☐ Type 2 ☐ Type 3b

Note: .....

REFERENCES

[1] Catassi C, and Alessio Fasano. Celiac disease diagnosis: simple rules are better than complicated algorithms. The American journal of medicine 123, no 8. 2010:691-3.

[2] Julio C. Bai MF, Gino Roberto,Detlef Schuppan. Celiac disease. World Gastroenterology Organisation Global Guidelines. April 2012.

[3] Catassi C, and Alessio Fasano. Celiac disease. Current opinion in gastroenterology 24, no 6. 2008:687-91.

[4] Lionetti E, Stefania Castellaneta, Alfredo Pulvirenti, Elio Tonutti, Ruggiero Francavilla, Alessio Fasano, Carlo Catassi, Italian Working Group of Weaning, and Celiac Disease Risk. Prevalence and natural history of potential celiac disease in at-

family-risk infants prospectively investigated from birth. The Journal of pediatrics 161, no 5. 2012:908.14-

[5] Ensari A. Gluten-sensitive enteropathy (celiac disease): controversies in diagnosis and classification. Archives of pathology & laboratory medicine 134, no 6 2010:826-36.

[6] Marsh MN. Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). Gastroenterology. 1992.

[7] Oberhuber G, Gerhard Granditsch, and Harald Vogelsang. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. European journal of gastroenterology & hepatology 1110 1999.

[8] Odze RD, and John R. Goldblum, eds. Surgical pathology of the GI tract, liver, biliary tract, and pancreas. Elsevier Health Sciences. 2009.

[9] Corazza GR, and V. Villanacci. Coeliac disease. Journal of clinical pathology 58, no 6. 2005:573-4.

[10] Tytgat GN, and Stefaan HAJ Tytgat. Grading and staging in gastroenterology. Thieme,. 2011.

[11] Zhang R, Yan Wang, and Genevieve B. Melton. Natural Language Processing in Medicine .Medical Applications of Artificial Intelligence 2013:375.

[12] McCowan I, Darren Moore, and M-J. Fry. Classification of cancer stage from free-text histology reports. In Engineering in Medicine and Biology Society, 2006 EMBS'06 28th Annual InternationalConference of the IEEE, pp 5153-5156 IEEE, . 2006.

[13] McCowan IA, Darren C. Moore, Anthony N. Nguyen, Rayleen V. Bowman, Belinda E. Clarke, Edwina E. Duhig, and Mary-Jane Fry. Collection of cancer stage data by classifying free-text medical reports. Journal of the American Medical Informatics Association 14, no 6. 2007:736-45.

[14] Nguyen AN, Michael J. Lawley, David P. Hansen, Rayleen V. Bowman, Belinda E. Clarke, Edwina E. Duhig, and Shoni Colquist. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. Journal of the American Medical Informatics Association 17, no 4. 2010:440-5.

[15] Luo Y, Aliyah R. Sohani, Ephraim P. Hochberg, and Peter Szolovits. Automatic lymphoma classification with sentence subgraph mining frompathology reports. Journal of the American Medical Informatics Association 2014.

[16] Jouhet V, Georges Defossez, Anita Burgun, Pierre Le Beux, P. Levillain, Pierre Ingrand, and Vincent Claveau. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods of information in medicine 51, no 3 2012.

[17] Crowley RS, Melissa Castine, Kevin Mitchell, Girish Chavan, Tara McSherry, and Michael Feldman. caTIES: a grid based system for coding and retrieval of surgicalpathology reports and tissue specimens in support of translational research. Journal of the American Medical Informatics Association 17, no 3. 2010:253-64.

[18] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, p 17 American Medical Informatics Association. 2001.

[19] Savova GK, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical Text Analysis andKnowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17, no 5 2010:507-13.

[20] Liao KP, Tianxi Cai, Vivian Gainer, Sergey Goryachev, Qing Zeng-treitler, SoumyaRaychaudhuri, Peter Szolovits et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research 62, no 8. 2010:1120-7.

[21] Uzuner Ö, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association 15, no 1. 2008:14-24.

[22] Srigley JR, Tom McGowan, Andrea MacLean, Marilyn Raby, Jillian Ross, Sarah Kramer, and Carol Sawka. Standardized synoptic cancer pathology reporting: A population-based approach. Journal of surgical oncology 99, no 8. 2009:517-24.

[23] Stitt-Gohdes WL, and Tena B. Crews. The Delphi technique: A research strategy for career and technical education. 2005.

[24] Shortliffe EH, and James J. Cimino. Biomedical informatics. Springer Science+ Business Media, LLC,. 2014.

[25] Shortliffe EH, and James J. Cimino. Biomedical informatics. Springer Science+ Business Media, LLC. 2006.

[26] Liron Pantanowitz JMT, Ulysses Balis. Pathology Informatics: Theory & Practice: American Society of Clinical Pathologists; 2012.

[27] Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). 1990.

[28] De Marneffe M-C, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In Proceedings of LREC, vol 6, no. 2006:449-54.

[29] calculator ss. raosoft.

[30] Hall M, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11, no 1. 2009:10-8.

[31] Hegenbart S, Andreas Uhl, and Andreas Vécsei. Survey on computer aided decision support for diagnosis of celiac disease. Computers in Biology and Medicine. 2015.

[32] Yue Li DM. Information Extractionof Multiple Categories from Pathology Reports. In Australasian Language Technology Association Workshop 2010:41.