journal homepage: www.ijmijournal.com

Artificial intelligence techniques applied to the development of a decision–support system for diagnosing celiac disease

Josceli Maria Tenório[a,*], *Anderson Diniz Hummel*[a], *Frederico Molina Cohrs*[b], *Vera Lucia Sdepanian*[c], *Ivan Torres Pisa*[d], *Heimar de Fátima Marin*[e]

[a] *Graduate Program on Health Informatics, Universidade Federal de São Paulo, São Paulo, SP, Brazil*
[b] *Graduate Program on Public Health, Universidade Federal de São Paulo, São Paulo, SP, Brazil*
[c] *Division of Pediatric Gastroenterology, Universidade Federal de São Paulo, São Paulo, SP, Brazil*
[d] *Department of Health Informatics, Universidade Federal de São Paulo, São Paulo, SP, Brazil*
[e] *Graduate Program of Health Informatics, Nursing School, Universidade Federal de São Paulo, São Paulo, SP, Brazil*

ARTICLE INFO

Article history:
Received 13 March 2011
Received in revised form
13 July 2011
Accepted 2 August 2011

Keywords:
Decision support systems, clinical
Celiac disease
Artificial intelligence

ABSTRACT

Background: Celiac disease (CD) is a difficult-to-diagnose condition because of its multiple clinical presentations and symptoms shared with other diseases. Gold-standard diagnostic confirmation of suspected CD is achieved by biopsying the small intestine.

Objective: To develop a clinical decision–support system (CDSS) integrated with an automated classifier to recognize CD cases, by selecting from experimental models developed using intelligence artificial techniques.

Methods: A web-based system was designed for constructing a retrospective database that included 178 clinical cases for training. Tests were run on 270 automated classifiers available in Weka 3.6.1 using five artificial intelligence techniques, namely decision trees, Bayesian inference, *k*-nearest neighbor algorithm, support vector machines and artificial neural networks. The parameters evaluated were accuracy, sensitivity, specificity and area under the ROC curve (AUC). AUC was used as a criterion for selecting the CDSS algorithm. A testing database was constructed including 38 clinical CD cases for CDSS evaluation. The diagnoses suggested by CDSS were compared with those made by physicians during patient consultations.

Results: The most accurate method during the training phase was the averaged one-dependence estimator (AODE) algorithm (a Bayesian classifier), which showed accuracy 80.0%, sensitivity 0.78, specificity 0.80 and AUC 0.84. This classifier was integrated into the web-based decision–support system. The gold-standard validation of CDSS achieved accuracy of 84.2% and $k = 0.68$ ($p < 0.0001$) with good agreement. The same accuracy was achieved in the comparison between the physician's diagnostic impression and the gold standard $k = 0.\ 64$ ($p < 0.0001$). There was moderate agreement between the physician's diagnostic impression and CDSS $k = 0.46$ ($p = 0.0008$).

Conclusions: The study results suggest that CDSS could be used to help in diagnosing CD, since the algorithm tested achieved excellent accuracy in differentiating possible positive from negative CD diagnoses. This study may contribute towards developing of a computer-assisted environment to support CD diagnosis.

[*] *Corresponding author at*: Universidade Federal de São Paulo, Department of Health Informatics, Rua Botucatu, 862, São Paulo, SP 04023-062, Brazil. Tel.: +55 11 5576 4347.
E-mail address: josceli.tenorio@unifesp.br (J.M. Tenório).

1386-5056/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2011.08.001

1. Introduction

Celiac disease (CD) is defined as permanent intolerance to gluten, a protein found in wheat, rye and barley. It is characterized by development of T cell-mediated enteropathy that affects genetically susceptible individuals [1,2]. It is not a gastrointestinal condition but, rather, a systemic autoimmune disorder involving a combination of genetic and environmental factors [3]. In addition, the disorder develops when an individual exposed to gluten has abnormally high small-intestine permeability, which allows large quantities of gluten fragments to penetrate into the mucosa and induce immune system cells [4].

CD is a difficult-to-diagnose condition because of its multiple clinical presentations and symptoms that are also shared with other diseases. It is typically characterized by chronic diarrhea, often accompanied by abdominal distension, weight loss, fatigue and malnutrition [5]. A significant number of patients with CD have an atypical form with no gastrointestinal symptoms. Such individuals only present isolated symptoms or symptoms in different systems. CD is also more likely to develop in high-risk groups such as first-degree relatives and individuals with diabetes mellitus, Down syndrome, Turner syndrome, Sjögren syndrome and Williams syndrome [1,6]. According to Fasano and Catassi [6], the European Society of Pediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) recommends that a definitive diagnosis of CD should be made based on a consistent history with clinical manifestations of CD, positive serological tests and Marsh III and IV mucosal villous atrophy of the small intestine [7], which is considered to be the gold standard. Biopsying the small intestine is thus a mandatory diagnostic test.

The only treatment for CD is a gluten-free diet. Compliance with the diet is essential for the therapy. Hill et al. [2] stressed the importance of monitoring such patients through serological tests after confirmation of the diagnosis confirmation, given the potential for new symptoms to emerge.

Today, development of a clinical decision–support system (CDSS) to assist in diagnosing and clinically monitoring CD is a feasible strategy. A CDSS is a computer program designed to help healthcare providers make clinical decisions [8,9]. Denekamp [10] claimed that CDSSs can close the gap between evidence and clinical practice by providing relevant data and knowledge at the point of care. Artificial intelligence (AI) techniques have proved effective in medicine for the purpose of analyzing medical data and making diagnostic predictions [11,12]. CDSSs can be a useful tool at the point of care.

The present study aimed to describe the development of a CDSS in a web environment integrated with an automated classifier, to identify cases of CD and assist medical providers in making positive diagnoses.

2. Materials and methods

This study was approved by the Research Ethics Committee of the Universidade Federal de São Paulo (protocol number 0927/08), and was conducted at the outpatient clinic of the Department of Pediatrics, in Hospital São Paulo, the university's teaching hospital.

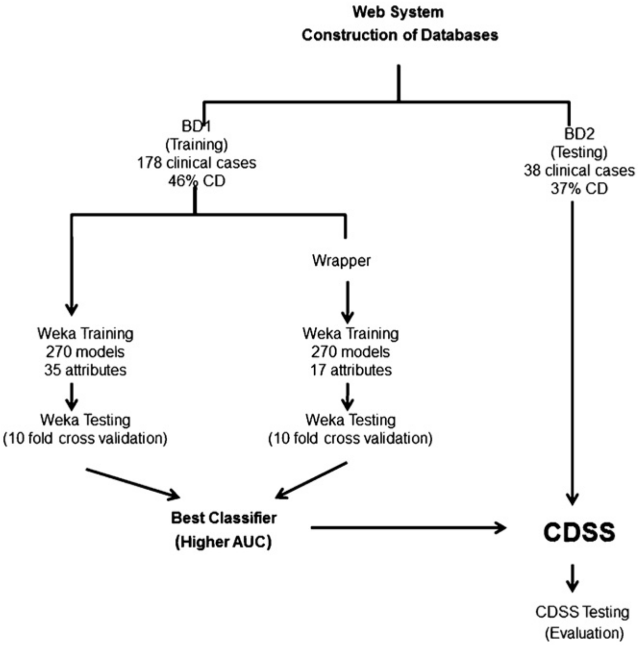

Fig. 1 – Phases of CDSS development.

The development of a CDSS for diagnosing CD consisted of three phases: I, II, and III. In Phase I, a web-based system for acquisition and retrieval of clinical data was developed, assessed and implemented in the outpatient clinic. Web system usability was evaluated by ten attending physicians using the System Usability Scale [13] questionnaire. Training and testing databases were constructed. In Phase II, the data were coded. A training database was tested with a set of automated classifiers. A wrapper approach for attribute selection was then applied and the same set of classifiers was retrained. The model with the most accurate parameters was integrated into the web-based system. In Phase III, the developed CDSS was evaluated. Fig. 1 shows all the phases of development of this CDSS.

2.1. *Phase I – database construction*

A web-based system was developed and implemented for recording and retrieving clinical data relating to the first patient appointment, and for clinical monitoring. The system was designed to reproduce the clinical diagnostic workflow, in which attending physicians who were trainees in Pediatric Gastroenterology discussed patient diagnosis and clinical management with their tutors.

A retrospective database (DB1) comprising the first-appointment records of 178 clinical cases was created for algorithm training. Confirmed CD cases (CD group) accounted for 46% of all records, while the remaining other diagnoses had signs, symptoms and complaints that were shared with clinical cases of CD (non-CD group; NCD). Patients in the CD group were classified according to the small intestine biopsy results, based on the gold standard [7]. Similarly, a testing database (DB2) was created for classifier evaluation. This included 38 clinical cases, among which the proportion of CD cases was 37%.The diagnosis provided by the physician and tutor during the first appointment for each clinical case in DB2 was collected. The patients included in the NCD group were attended by 13 different physicians, assisted by five tutors. The patients included in the CD group were attended by six physicians, assisted by three tutors.

2.2. *Phase II – training and testing classifiers*

2.2.1. *Coding attributes*

An experimental model was designed to determine the classifiers that were to be integrated into the web-based system of the CDSS under development. Automated classifiers available from the Waikato Environment for Knowledge Analysis (Weka 3.6.1) software [14] were tested. Weka is a comprehensive set of libraries organized in Java packages of different classes that implement machine learning and data mining algorithms [15].

A total of 35 nominal attributes of signs, symptoms and high-risk groups were listed (Appendix A). All clinical data relating to medical histories were coded and given "y" (yes) or "n" (no) values depending on whether the patient had a certain symptom. The same coding was used for physical examination items, except for subcutaneous tissue (normal, thin or thick) and high-risk groups, s1 (first-degree relatives), s2, s3 and "no" (relatives).

The target variable was the diagnostic impression with "cd" (celiac disease) and "ncd" (non-celiac disease) values. Most were "ndc" (54% of all the patients), thus accounting for 96 clinical cases.

2.2.2. *Artificial intelligence (AI) techniques*

Five AI techniques were tested:

- Decision trees: classification was achieved through performing a sequence of tests on the attributes, using rules of the type "if...then...else". In this study, the ADTree [16], SimpleCart [17] and J48 algorithms were used, which are trees of C4.5 type [18]. To induce the decision trees relating to the ADTree algorithm, the parameter for tree depth (*B*) varied between 10 and 20; in relation to SimpleCart, the minimum number of instances per leaf (*M*) varied between 2 and 9; and in relation to J48, the confidence factor used for pruning (*C*) varied between 0.1 and 0.9 and *M* between 3 and 12.
- Bayesian classifiers: classification was achieved through calculating the probability for each class, assuming conditional independence of the attributes in the NaiveBayes, NaiveBayesSimple and BayesNet algorithms [19]. In the BayesNet algorithm, the alpha parameter (*A*) varied between 0 and 1.0, with increments of 0.25. The algorithms AODE [20] and AODEsr [21] minimize the presupposition of conditional independence, thereby resulting in better precision.
- Artificial Neural Net (ANN): This is a distributed parallel processor composed of simple processing units that have the ability to store knowledge extracted from a database and make it available for classifying new data [22]. In this study, the parameters of learning rate (*L*), varying between 0.3 and 0.5, momentum (*M*), varying between 0.2 and 0.5, and 500 epochs to training (default), were applied to the algorithm Multilayer Perceptron (MLP).
- Support Vector Machines (SVMs): Classification was achieved by separating the data according to the variation in the parameters of a support vector. In this study, different kernels (PolyKernel, Radial Basis Function, sigmoid and linear) were used with different penalty parameters for the error term (*C*) and kernel parameter (*G*). The *C* and *G* parameters varied between 1×10^{-10} and 1×10^{10} [23].
- *K*-nearest neighbors: these classify a given data point according to the distance (frequently Euclidian) to the nearest data point. The algorithm Lazy Bayesian Rules (LBR) [24] is a Bayesian classifier in which lazy learning techniques are applied to induce Bayesian rules. In this study, in the algorithm IBk [25], the number of neighbors (*k*) varied between 1 and 20. In the algorithm KStar [26], o parameter for overall blending (*B*) varied between 10 and 100. The algorithm (LBR) was tested with the standard parameters.

Table 1 – Algorithms used in DB1 training.

Technique	Algorithm	No. of models
Decision trees	ADTree	11
	SimpleCart	8
	J48	30
Bayesian classifier	AODE	3
	AODEsr	5
	NaiveBayes	1
	BayesNet	4
	NaiveBayesSimple	1
ANN	MultilayerPerceptron	18
SVM	LibSVM	158
K-nearest neighbors	IBk	20
	KStar	10
	LBR	1
	Total	270

ANN: artificial neural networks; SVM: support vector machines; AODE: averaged one-dependence estimators; LBR: lazy Bayesian rules.

The variability of the algorithms for each technique and the variations in the parameters resulted in 270 models.

Table 1 shows the algorithms and quantity models used to construct the training models.

No data preprocessing methods were used.

High dimensionality of databases in health care increases the complexity. A preprocessing method can be applied to reduce the number of attributes. Kumar and Srinivas [27] have shown that decreasing the dimensionality of their tuberculosis and hypertension databases increased the performance of the diagnosing method used.

In the next step of this study, the selection wrapper variable [28] was applied to DB1 to reduce the dimension of the feature vector, in order to achieve parameters of greater accuracy.

In Weka 3.6.1, the class WrapperSubsetEval uses a classifier to evaluate the set of attributes as a learning scheme and the cross-validation test to estimate the accuracy rate of a classifier. In this study, the classifier used was NaiveBayes, with the standard parameters and 5-fold cross-validation for accuracy estimation.

Table 2 – Characterization in terms of frequency *F* (%) estimates from training and testing samples, regarding clinical signs and symptoms and high-risk groups.

Type	Input variables	Training			Testing		
		CD ($n = 82$)	NCD ($n = 96$)		CD ($n = 14$)	NCD ($n = 24$)	
		F (%)	*F* (%)	*p*-value	*F* (%)	*F* (%)	*p*-value
General symptoms	Adynamia	9.8	6.3	0.386	0.0	0.0	
	Weight loss	31.7	21.9	0.138	14.3	25.0	0.712
	Anorexia	2.4	14.6	0.005[†]	14.3	12.5	0.734
	Fever	6.1	12.5	0.147	0.0	12.5	
	Irritability	15.9	6.3	0.039[†]	7.1	0.0	
	Apathy	4.9	2.1	0.540	7.1	0.0	
	Anemia	2.4	7.3	0.259	0.0	4.2	
	Edema	8.5	1.0	0.041[†]	7.1	0.0	
Gastrointestinal symptoms	Difficulty swallowing	0.0	2.1		7.1	0.0	
	Nausea	2.4	12.5	0.013[†]	0.0	0.0	
	Regurgitation	2.4	3.1	0.858	0.0	0.0	
	Vomiting	24.4	14.6	0.097	0.0	20.8	
	Upper GI bleeding	0.0	1.0		0.0	0.0	
	Melena	0.0	1.0		0.0	0.0	
	Intestinal bleeding	0.0	9.4		0.0	0.0	
	Acholic stools	0.0	2.1		0.0	0.0	
	Hypocholic stools	0.0	1.0		0.0	0.0	
	Diarrhea	72.0	26.0	0.000[†]	57.1	33.3	0.152
	Bloody diarrhea	3.7	12.5	0.034[†]	0.0	0.0	
	Diarrhea with mucus	7.3	4.2	0.560	0.0	8.3	
	Fatty diarrhea	1.2	0.0		0.0	0.0	
	Constipation	7.3	30.2	0.000[†]	7.1	29.2	0.233
	Abdominal distension	46.3	36.5	0.182	42.9	33.3	0.557
	Abdominal pain	32.9	56.3	0.002[†]	21.4	45.8	0.133
Dental symptoms/ signs	Enamel hypoplasia	2.4	1.2	0.890	7.1	8.3	0.623
Locomotor system	Arthralgia	4.9	3.1	0.831	0.0	16.7	
	Arthritis	1.2	1.0		0.0	4.2	
	Osteoporosis	0.0	0.0		0.0	0.0	
Growth and development	Delayed NPMD	2.4	9.4	0.055	7.1	4.2	
	Underweight	45.1	20.8	0.001[†]	50	16.7	0.070
	Stunting	20.7	15.6	0.377	0.0	12.5	

Skin	Dermatitis herpetiformis	1.2	0.0		0.0	0.0	
High-risk group	Familial CD	8.5[a]	3.1[b]	0.216	0.0	4.2	
Physical examination	Paleness	17.1	16.7	0.942	35.7	0.0	
	SCT						
	Thin	23.2	11.5	0.037[†]	35.7	12.5	0.200
	Normal	40.2	49.0	0.244	35.7	66.7	0.064
	Thick	0.0	6.3	0.0	0.0	0.0	

NPMD: neuropsychomotor development; SCT: subcutaneous tissue

[a] 1st degree relatives

[b] 2nd or 3rd degree relatives.

[†] Values lower than 0.05.

In order to obtain the subset of attributes, a search method that can go across the attribute space to find the best subset for the classifier used by the wrapper needs to be established [15]. One of the most efficient methods is GeneticSearch, which is based on evolutionary algorithms [29]. This was used in this study with the standard parameters of the Weka 3.6.1 software.

2.2.3. *Algorithm selection*

The algorithm with the greatest area under the receiver operating characteristic curve (AUC) was selected for integration with the web-based system. The receiver operating characteristic curve is a graphical plot of sensitivity versus one minus specificity, and is used to evaluate classification and prediction models [30]. In the Weka 3.6.1 software, the AUC is calculated by means of the sensitivity and specificity of each of the folds used in training the classifier. For integration between the algorithm and the web-based system, the Quick middleware [31] was used. Quick enables communication between different applications using the SOAP web service [32], through providing tools for security and fault tolerance.

The CDSS was designed such that it would issue a warning when the probability of CD occurrence was higher than 50%.

2.3. *Phase III – CDSS evaluation*

The CDSS was then evaluated using DB2. The data relating to each clinical case in DB2 were recorded in the CDSS. The probabilities of CD occurrence were recorded for comparison against the diagnostic impression that had been recorded by a medical provider during a consultation.

2.4. *Statistical analysis*

Cross-validation with 10 subsets (10-fold cross-validation) was used for testing. To evaluate and compare algorithms regarding the accuracy of the classification, the parameters used for selecting the algorithm were the highest values for the AUC, sensitivity, specificity and accuracy rate [33].

Comparative analysis between the AUC results for each algorithm before and after variable selection was performed to determine the best parameters for the algorithm to be selected.

Kappa statistics were used to evaluate the agreements between the CDSS and the gold standard; between the physician's impression, i.e. relating to the diagnostic impression recorded by physicians and their tutors during consultations, and the gold standard; and between the CDSS and the physician's impression. In this manner, the diagnostic reliability of the CDSS was assessed.

3. Results

The average System Usability Scale score achieved was 83.5 ± 10.0 (95% CI: [76.3–90.7]), thus indicating that the web system was easy to use and fulfilled the users' expectations.

Table 2 shows the frequencies of symptoms and positive signs presented by patients, which were recorded by the physician provider during the first appointment. A *p*-value was

Table 3 – Classifier analysis measurements, ranked according to AUC, using 35 attributes.

Algorithm	ACC (%)	SEN	SPE	AUC
LBR[a]	76.9	0.79	0.75	0.84
AODE-F1[a]	76.3	0.78	0.75	0.84
AODEsr[a]	75.9	0.78	0.74	0.84
NaiveBayes[a]	76.0	0.77	0.75	0.84
NaiveBayesSimple[a]	76.0	0.77	0.75	0.84
KStar-B60	76.2	0.76	0.77	0.84
BayesNet A0.75	76.0	0.78	0.74	0.83
IBk-k16	73.3	0.80	0.67	0.82
MLP-L0.4-M0.2[b]	73.4	0.70	0.76	0.79
LibSVM-C1.0-G0.0	77.3	0.76	0.78	0.77
J48-C0.7-M9	72.7	0.77	0.69	0.76
ADTree[a]	71.5	0.70	0.72	0.76
SimpleCart[a]	71.9	0.72	0.72	0.73

ACC: accuracy rate; SEN: sensitivity; SPE: specificity; AUC: area under the receiver operating characteristic curve.

[a] Algorithms with default parameters. Kernel used in SVM: sigmoid.

[b] Hidden layers: 22.

estimated to assess the similarity between the proportions of each input variable. If the p-value was greater than 0.05, the proportions were considered similar. These estimates were made using EpiInfo® v.6.

Five AI techniques were tested with 13 variations of algorithms and variations of parameters that resulted in 270 models. The algorithms were run and parameters were obtained for analysis. Table 3 shows the results relating to the best parameters for AUC, sensitivity, specificity and accuracy rate, obtained for each algorithm.

For the database analyzed, the results showed that the LBR algorithm can identify possible CD diagnoses with excellent sensitivity and accuracy rate. The specificity found showed that the classifier was also very reliable for indicating possible negative CD diagnoses. The same analysis could be extended to Bayesian classifiers, particularly AODE-F1, since the same AUC was obtained.

The test wrapper was applied to DB1 (which initially comprised 35 attributes) and the dimensions were reduced to attributes with the highest statistical value for decisions, namely: anorexia, irritability, lethargy, anemia, difficulty in swallowing, nausea, melena, intestinal bleeding, acholic stools, diarrhea, constipation, abdominal pain, delayed neuropsychomotor development, subcutaneous tissue, familial celiac disease, edema and dermatitis herpetiformis.

Table 4 shows the measurements obtained from testing 270 classifiers, using a database with 17 attributes.

Table 4 shows that application of the wrapper did not increase the maximum AUC value of the best classifier. However, in analyzing the results from the set of classifiers, it was found that the mean values had increased significantly. The hypothesis that there was no statistically significant difference in AUC values for DB1 between the use of 35 and the use of 17 attributes was tested using the Wilcoxon test. The p-value estimated from the test was <0.0001 (95% CI), thus showing that there was a statistically significant difference in AUC values for DB1 between the use of 35 attributes and the use of 17 attributes.

Table 4 – Best measurements obtained for each algorithm, ranked according to AUC, using 17 attributes.

Algorithm	ACC (%)	SEN	SPE	AUC
AODE-F1[a]	80.0	0.78	0.80	0.84
NaiveBayes[a]	79.1	0.78	0.80	0.84
NaiveBayesSimple[a]	79.1	0.78	0.80	0.84
KStar-B 40	78.2	0.78	0.77	0.84
BayesNet-A1.0	78.8	0.77	0.80	0.84
LBR[a]	79.2	0.77	0.80	0.84
AODEsr-F1[a]	78.2	0.76	0.80	0.84
Ibk -K 20	74.7	0.82	0.68	0.83
MLP-L 0.3-M0.4[b]	73.3	0.71	0.75	0.79
ADTree[a]	72.8	0.70	0.75	0.79
LibSVM-G0.0-C10.0	76.2	0.75	0.77	0.76
J48-C 0.6-M 2	75.2	0.73	0.77	0.76
SimpleCart[a]	71.4	0.69	0.73	0.71

[a] Algorithms with default parameters. Kernel used in SVM: RBF.

[b] Hidden layers: 13.

Nonetheless, application of the wrapper produced increases in the maximum values of specificity and accuracy.

Considering that the criteria for choosing the most efficient algorithm were the maximum AUC, sensitivity, specificity and accuracy, the classifier selected for the integration to the CDSS was AODE-F1 (default), which is a Bayesian classifier, using DB1 with 17 attributes.

Comparison between the CDSS and the gold standard resulted in sensitivity = 92.9% and specificity = 79.2%. The accuracy of the CDSS was 84.2%. The degree of agreement (kappa) was 0.68 ($p < 0.0001$), thus indicating good reproducibility [34] between the CDSS and the gold standard.

Comparison between the physician's diagnostic impression and the gold standard resulted in sensitivity = 64.3% and specificity = 95.8%. The physicians' accuracy regarding a positive diagnosis of CD was 84.2%, with kappa = 0.64 ($p < 0.0001$), thus showing good agreement.

Correlation between the physician's diagnostic impression and the CDSS gave kappa = 0.46 (p-value = 0.0008), thus indicating moderate agreement.

4. Discussion

It was important to make a preliminary usability evaluation in order to analyze the interaction between the web system and users [35]. The resultant System Usability Scale score (83.5 ± 10.0) showed that the usability was excellent [36], thereby suggesting acceptance among users.

In a review study, Grossi et al. [37] showed that application of AI techniques such as ANN for making diagnoses and prognoses of gastrointestinal conditions is potentially more effective than conventional statistical methods. This is an important finding, given the complexity involved in diagnosing gastrointestinal conditions and the need for invasive tests.

Analysis in Table 2, which characterizes the sample in terms of signs, symptoms and high-risk groups, demonstrated that the most significant common symptoms in the CD group were irritability, edema, diarrhea, thin layer of subcutaneous tissue and weight loss. Anorexia, nausea, bloody diarrhea and constipation were more common in the NCD group. However, most signs and symptoms of CD had similar frequencies in the two groups. This confirms the difficulty in making an initial diagnosis of CD when clinical manifestations are evaluated.

The results from the first phase of the study, in which the AI techniques were tested, suggested that Bayesian classifiers and *k*-nearest neighbors were able to recognize possible diagnoses of CD, with good sensitivity. The specificity found shows that these techniques were also reliable for indicating possible negative diagnoses of CD. Thus, these techniques are potential tools for clinical decision support that should be further tested in other studies. Bayesian classification algorithms are regarded as a reference method within the field of medicine. Because of their efficacy, these algorithms are usually tested before other techniques. They have shown better results in several areas, especially in helping medical diagnosis [38].

Application of the wrapper variable selection technique was efficient for increasing the specificity and accuracy values, although this did not contribute towards increasing the AUC of the best classifier, unlike the findings from other studies within gastroenterology [39,40]. From analysis on the complete set of parameters evaluated, it was seen that an effective improvement occurred.

The findings from the training and test regarding the accuracy of the best classifier showed that the loss of information due to applying the reduction in the number of dimensions of the wrapper did not interfere significantly with the classification task. Further studies on applying variable selection techniques to obtain an attribute vector with smaller dimensions are required in order to establish a minimum data set for significant differentiation, such that this might support development of a protocol with the minimum data for detecting undiagnosed CD cases.

No studies specifically investigating the use of a CDSS supported by AI techniques for diagnosing CD were found in the literature. For other gastrointestinal diseases, related studies mostly described experimental models within gastroenterology that tested and evaluated AI techniques, especially ANN. In a recent study, Pace et al. [41] described the application of ANN and LDA techniques together with a questionnaire for gastroesophageal reflux disease (GERD) as a new model for differentiating between healthy patients and those with GERD. These authors reported accuracy of 99.2% with AUC of 99.1%. In the subset of patients with GERD, the accuracy for differentiating non-erosive gastroesophageal reflux disease (NERD) and erosive esophagitis (EE) was 66.5%, which was an inconclusive result.

Also in relation to GERD, Horowitz et al. [42] conducted a study to identify a set of symptoms that would make it possible to differentiate patients with GERD from those with other types of dyspepsia, using logistic regression, ANN, C5.0 and CART. A comparison of these techniques showed that the best measurements were obtained with ANN, with accuracy of 78.7%. Lahner et al. [40] described the application of ANN to identify cases of atrophic body gastritis (ABG). The best measurements were obtained with ANN, using attribute selection, with an accuracy of 75.8%. In another study, Lahner et al. [39] applied ANN and LDA to detect thyroid disorders in patients with ABG and found that the accuracy was greater than 91%. Sakai et al. [43] reported that a Bayesian network model for diagnosing non-traumatic abdominal pain, consisting of nine variables, had a lower error rate (27%) and greater AUC (0.763) than seen with ANNs and logistic regression.

Compared with other applications within gastroenterology with input variables including a set of clinical manifestations, the present study found measurements that were higher than or close to the Bayesian network model designed for abdominal pain, even when compared with other techniques such as ANN and LDA. The measurements found suggest that the algorithm that was selected can be used with good reliability for identifying patients with CD.

The CDSS evaluation showed an accuracy of 84.2%, which is an excellent result in comparison with the gold standard and is similar to that found in the comparison between the physician's impression and the gold standard. However, the degree of agreement was higher for the CDSS than the agreement found in the comparison between the physician's impression and the gold standard, which suggests that CDSS has good diagnostic reliability. The sensitivity analysis showed that there were more false positives in the CDSS than among the specialists' evaluations. Further studies are required to reduce the rates of false positive and false negative results. However, a suggestion of a possible diagnosis of CD that is not confirmed should not be regarded as an error but rather as an indication to physicians that a case should be further investigated.

The clinical cases used for evaluating the CDSS, although few in number, were considered by the specialists to be sufficient to suggest that the experimental model developed here is valid.

The purpose of the present study was not to replace the procedures for CD confirmation. In effect, a CDSS provides information based on statistical methods to support decision-making for diagnosis and management. In a recent study, Catassi and Fasano [44] pointed out that the use of biopsy has been questioned as the single conclusive test for diagnosing CD in several cases. These authors suggested that diagnoses of CD should be made after a comprehensive evaluation in which clinical presentations, histological and serological tests, HLA analysis and therapeutic response carry equal weight. For a definitive diagnosis, at least four out of the five requirements must be met.

The present study may thus provide a reliable tool for analyzing the clinical manifestations associated with CD. This disease is not considered to be rare, either in Brazil or in European countries [45].

The results and analyses presented are preliminary. The next steps in this study consist of: (i) carrying out an experiment to ascertain the changes to clinical practice that occur through the use of the proposed CDSS; (ii) using the CDSS in a real clinical setting over a long study period, in order to assess the therapeutic gains achieved by physicians and patients [35,46].

5. Conclusions

The preliminary results suggest that CDSS can be reliably used to help in diagnosing CD. The Bayesian classifier AODE can differentiate possible cases of positive and negative diagnoses

Summary table

What is already known on this topic?

- Celiac disease is difficult to diagnose. Medical history and physical examination are the keys in initial investigations on suspected celiac disease.
- Artificial intelligence techniques such as Bayesian networks, artificial neural networks (ANN), decision trees and linear discriminant analysis, among others, have been used with excellent results for diagnoses and prognoses within gastroenterology. There are no studies on diagnosing celiac disease.

What this study has added to our knowledge:

- We proposed a set of attributes relating to signs and symptoms and high-risk groups that was tested and was proved to be efficient for identifying cases of celiac disease.
- Bayesian classifiers seem to improve the accuracy, compared with decision trees, ANNs, support vector machines (SVMs) and *k*-nearest neighbors, for constructing a predictive model for celiac disease with high AUC values. This suggests that these classifiers may be used to construct a clinical decision support system.

of CD with good sensitivity and high accuracy. The measurements obtained in the present study were consistent with those reported in other related studies.

Further studies are required to better analyze the classifiers, including analysis on variable selection techniques, in order to determine a minimum set of clinical data for developing an electronic protocol to be implemented in relation to CDSS and tested within a screening or specialized care setting.

This study may contribute towards development of a computer-assisted system to be used as a tool for uncomplicated, reliable and effective clinical decision-making.

Authors' contributions

- Josceli Maria Tenório: responsible for conception and study design, development of web system and CDSS, data collection, data analysis and the writing of this manuscript.
- Anderson Diniz Hummel: development of a web service, data analysis and critical review.
- Frederico Molina Cohrs: statistical analysis and interpretation of data.
- Vera Lucia Sdepanian: approval of the web system and CDSS, analysis and validation of a data collection and critical review of the manuscript.
- Ivan Torres Pisa: conception and study design, data analysis and critical review of the manuscript.
- Heimar de Fátima Marin: conception and study design, critical review and final approval of the manuscript.

Conflicts of interest

The authors declare that they did not receive any financial support to conduct this research.

Acknowledgements

We thank Dr. Ana Cristina Fontenele Soares for her support of this project and all physicians in the Division of Pediatric Gastroenterology, Universidade Federal de São Paulo.

Dr. Marin was partially supported by grants from NIH (no. D43TW007015-06 BRIGHT) and from CNPq (no. 301735/2009-3).

Appendix A

Attributes analyzed in DB1: signs, symptoms and high-risk groups

Type	Signs/symptoms/ high-risk groups
General symptoms	Adynamia
	Weight loss
	Anorexia
	Fever
	Irritability
	Apathy
	Anemia
	Edema
Gastrointestinal symptoms	Difficulty swallowing
	Nausea
	Regurgitation
	Vomiting
	Upper GI bleeding
	Melena
	Intestinal bleeding
	Acholic stools
	Hypocholic stools
	Diarrhea
	Bloody diarrhea
	Diarrhea with mucus
	Fatty diarrhea
	Constipation
	Abdominal distension
	Abdominal pain
Dental symptoms/signs	Enamel hypoplasia
Locomotor system	Arthralgia
	Arthritis
	Osteoporosis
Growth and development	Delayed neuropsychomotor development
	Underweight
	Stunting
Skin	Dermatitis herpetiformis
High-risk group	Familial CD
Physical examination	Paleness
	Subcutaneous tissue

REFERENCES

[1] V.L. Sdepanian, M.B. Morais, U. Fagundes-Neto, Celiac disease: evolution in knowledge since its original centennial description up to the present day, Arq. Gastroenterol. 36 (1999) 244–257 (in Portuguese).

[2] I.D. Hill, M.H. Dirks, G.S. Liptak, R.B. Colletti, A. Fasano, S. Guandalini, et al., North American Society for Pediatric Gastroenterology, Hepatology and Nutrition, Guideline for the diagnosis and treatment of celiac disease in children: recommendations of the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition, J. Pediatr. Gastroenterol. Nutr. 40 (1) (2005) 1–19.

[3] L.M. Sollid, Celiac disease as a model of gastrointestinal inflammation, J. Pediatr. Gastroenterol. Nutr. 40 (April (Suppl. 1)) (2005) S41–S42.

[4] J. Visser, J. Rozing, A. Sapone, K. Lammers, A. Fasano, Tight junctions, intestinal permeability, and autoimmunity: celiac disease and type 1 diabetes paradigms, Ann. N. Y. Acad. Sci. 1165 (May) (2009) 195–205.

[5] M.I. Torres, M.A. López Casado, A. Ríos, New aspects in celiac disease, World J. Gastroenterol. 13 (8) (2007) 1156–1161.

[6] A. Fasano, C. Catassi, Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum, Gastroenterology 120 (2001) 636–651.

[7] M.N. Marsh, Gluten, major histocompatibility complex and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivitiy ('celiac sprue'), Gastroenterology 102 (January (1)) (1992) 330–354.

[8] M.A. Musen, Y. Shahar, E.H. Shortliffe, Clinical decision–support systems, in: E.H. Shortliffe, J.J. Cimino (Eds.), Biomedical Informatics: Computer Applications in Health Care and Biomedicine, third ed., Springer, New York, 2006, pp. 698–736.

[9] E.S. Berner, T.J. La Lande, Overview of clinical decision support systems, in: Clinical Decision Support Systems: Theory and Practice, second ed., Springer, New York, 2007, pp. 3–22.

[10] Y. Denekamp, Clinical decision support systems for addressing information needs of physicians, Isr. Med. Assoc. J. 9 (11) (2007) 771–776.

[11] A.N. Ramesh, C. Kambhampati, J.R.T. Monson, P.J. Drew, Artificial intelligence in medicine, Ann. R. Coll. Surg. Engl. 86 (2004) 334–338.

[12] P.J. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, Neural Netw. 15 (1) (2002) 11–39.

[13] J. Brooke, SUS – a quick and dirty usability scale, 1986 [text on the Internet] [cited 1.7.2008]. Available from: http://www.usability.gov.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explorations 11 (1) (2009) 10–18.

[15] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.

[16] Y. Freund, L. Mason, The alternating decision tree learning algorithm, Proceeding of the Sixteenth International Conference on Machine Learning, Slovenia, Bled, 1999, pp. 124–133.

[17] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.

[18] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, 1993.

[19] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley-Interscience, New York, 2001.

[20] G. Webb, J. Boughton, Z. Wang, Not so Naive Bayes: aggregating one-dependence estimators, Mach. Learn. 58 (1) (2005) 5–24.

[21] F. Zheng, G.I. Webb, Efficient Lazy Elimination for Averaged-One Dependence Estimators, Proceedings of the Twenty-third International Conference on Machine Learning, 2006, pp. 1113–1120.

[22] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice-Hall, New Jersey, 1999.

[23] C.C. Chang, C.J. Lin, LIBSVM – A Library for Support Vector Machines [text on the Internet], 2001 [cited 5.7.2010]. Available from: http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

[24] Z. Zheng, G. Webb, Lazy learning of bayesian rules, Mach. Learn. 4 (1) (2000) 53–84.

[25] D. Aha, D. Kibler, Instance-based learning algorithms, Mach. Learn. 6 (1991) 37–66.

[26] J.G. Cleary, L.E. Trigg, K*: an instance-based learner using an entropic distance measure [text on the Internet], 1995 [cited 5.7.2010]. Available from: http://www.cs.waikato.ac/nz/ml/publications/1995/Cleary95-KStar.pdf.

[27] C.H.A. Kumar, S. Srinivas, Mining associations in health care data using formal concept analysis and singular value decomposition, J. Biol. Syst. 18 (4) (2010) 787–807.

[28] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.

[29] D.E. Goldberg, Genetic Algorithms in Search Optimization and Machine Learning, 1st ed., Addison-Wesley, Boston, 1989.

[30] T.A. Lasko, J.G. Bhagwat, K.H. Zou, L. Ohno-Machado, The use of receiver operating characteristic curves in biomedical informatics, J. Biomed. Inform. 38 (October (5)) (2005) 404–415.

[31] A.D. Hummel, F.S. Sousa, F. Teixeira, A.E.J. Falcão, F. Mancini, T.M. Costa, et al., Projeto Quick: Desenvolvimento de um middleware para sistemas de apoio a decisão na área da saúde em ambientes distribuídos, in: XII Congresso Brasileiro de Informática em Saúde, Porto de Galinhas, Recife, Brasil. Anais. São Paulo: Sociedade Brasileira de Informática em Saúde, 2010.

[32] E. Cerami, Web Services Essentials: Distributed Applications with XML-RPC, SOAP, UDDI and WSDL, O'Reilly Media, Sebastopol, 2002.

[33] E. Massad, A teoria bayesiana no diagnóstico médico, in: E Massad, RX Menezes, PSP Silveira, NRS. Ortega (Eds.), Métodos quantitativos em medicina, Barueri (SP), Manole, 2004, pp. 189–205.

[34] B. Rosner, Fundamentals of Biostatistics, 6th ed., Duxbury Press, Boston, 2006.

[35] B. Kaplan, Evaluating informatics applications–clinical decision support systems literature review, Int. J. Med. Inform. 64 (November (1)) (2001) 15–37.

[36] A. Bangor, P. Kortum, J. Miller, Determining what individual SUS scores mean: adding an adjective rating scale, J. Usability Stud. 4 (3) (2009) 114–123.

[37] E. Grossi, A. Mancini, M. Buscema, International experience on the use of artificial neural networks in gastroenterology, Dig. Liver Dis. 39 (March (3)) (2007) 278–285.

[38] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artif. Intell. Med. 23 (August (1)) (2001) 89–109.

[39] E. Lahner, E. Grossi, M. Intraligi, M. Buscema, V.D. Corleto, G. Delle Fave, et al., Possible contribution of advanced statistical methods (artificial neural networks and linear discriminant analysis) in the recognition of patients with suspected atrophic body gastritis, World J. Gastroenterol. 11 (2005) 5867–5873.

[40] E. Lahner, M. Intraligi, M. Buscema, M. Centanni, L. Vannella, E. Grossi, B. Annibale, Artificial neural networks in the

recognition of the presence of thyroid disease in patients with atrophic body gastritis, World J. Gastroenterol. 14 (January (4)) (2008) 563–568.

[41] F. Pace, G. Riegler, A. de Leone, M. Pace, R. Cestari, P. Dominici, E. Grossi, The EMERGE Study Group, Is it possible to clinically differentiate erosive from nonerosive reflux disease patients? A study using an artificial neural networks-assisted algorithm, Eur. J. Gastroenterol. Hepatol. 22 (October (10)) (2010) 1163–1168.

[42] N. Horowitz, M. Moshkowitz, Z. Halpern, M. Leshno, Applying data mining techniques in the development of a diagnostics questionnaire for GERD, Dig. Dis. Sci. 52 (August (8)) (2007) 1871–1878.

[43] S. Sakai, K. Kobayashi, J. Nakamura, S. Toyabe, K. Akazawa, Accuracy in the diagnostic prediction of acute appendicitis based on the Bayesian network model, Methods Inf. Med. 46 (2007) 723–726.

[44] C. Catassi, A. Fasano, Celiac disease diagnosis: simple rules are better than complicated algorithms, Am. J. Med. 123 (August (8)) (2010) 691–693.

[45] R.P. Oliveira, V.L. Sdepanian, J.A. Barreto, A.J. Cortez, F.O. Carvalho, J.O. Bordin, M.A. de Camargo Soares, F.R. da Silva Patrício, E. Kawakami, M.B. de Morais, U. Fagundes-Neto, High prevalence of celiac disease in Brazilian blood donor volunteers based on screening by IgA antitissue transglutaminase antibody, Eur. J. Gastroenterol. Hepatol. 19 (1) (2007) 43–49.

[46] G.P. Purcell, What makes a good clinical decision support system, BMJ 330 (April (7494)) (2005) 740–741.