Comparison of Texture Classifier and Human Observer in Diagnosis of Autoimmune Thyroiditis, Observer Variability Evaluation

Š. Holinka[1], **D. Smutek**[1]
[1]3rd Department of Medicine, 1st Medical Faculty, Charles University in Prague, Czech Republic
Supervisor: Doc. MUDr. Ing. Daniel Smutek, Ph.D.

Summary
The objective has been to compare success of the texture classifier and a human observer in diagnosis of the autoimmune thyroiditis from B-mode ultrasound images and to determine inter- and intra-observer variability.
The data set of 161 subjects was classified by four human observers and by the Bayes classifier based on the texture features to three classes (healthy, border state, autoimmune thyroiditis).
Two observers had a higher success rate when classifying the healthy class (74.4% and 83.3%), the other two observers classified better cases with autoimmune thyroiditis (59.0% and 77.4%). The classifier gave the relatively high and balanced success rate for both classes (100,0% for healthy and 87.5% for thyroiditis). The different observers' success rates resulted in the high inter-observer variability, showing only a fair agreement among the human observers. There was no significant difference among human observers in the intra-observer variability.
Due to the fair agreement among observers in the diagnosis of autoimmune thyroiditis from ultrasound images and good results of the classifier, the best way in establishing diagnosis is computer-aided diagnosis combined with observers' clinical experience.

Keywords: thyroid gland, autoimmune thyroiditis, ultrasound image, B-mode ultrasound, texture analysis, computer-aided classification, inter-observer variability, intra-observer variability, Kappa Statistics, weighted Kappa.

1. Introduction
Autoimmune thyroiditis, one of the most frequent diseases of the thyroid gland, is a chronic inflammation of the thyroid parenchyma [1]. The inflammation of the gland causes diffuse changes in the structure of the thyroid tissue. These changes can be detected by ultrasound imaging which is the most widely used diagnostic and monitoring tool for this disease. Thus, the method is suitable as a complementary examination method next to the diagnosis done from clinical examination using immunologic, hormonal and metabolic analyses of blood samples, and from cytological examination using fine-needle aspiration biopsy. However, the assessment of the diffuse processes is difficult [2], [3] and, in daily practice, the diagnosis from B-mode ultrasound images is made qualitatively from the size of the gland being examined, its perfusion, and the structure and echogenicity of its parenchyma. A physician uses clinical experience without giving any quantifiable indices.

This subjective evaluation of the ultrasound image texture is a reason why computer-aided methods for an automatic diagnosis were proposed. In the context, the suitable texture features were founded [4], [5], and a successful classifier for a semi-automatic diagnostic method of autoimmune thyroiditis from B-mode ultrasound images constructed [6].

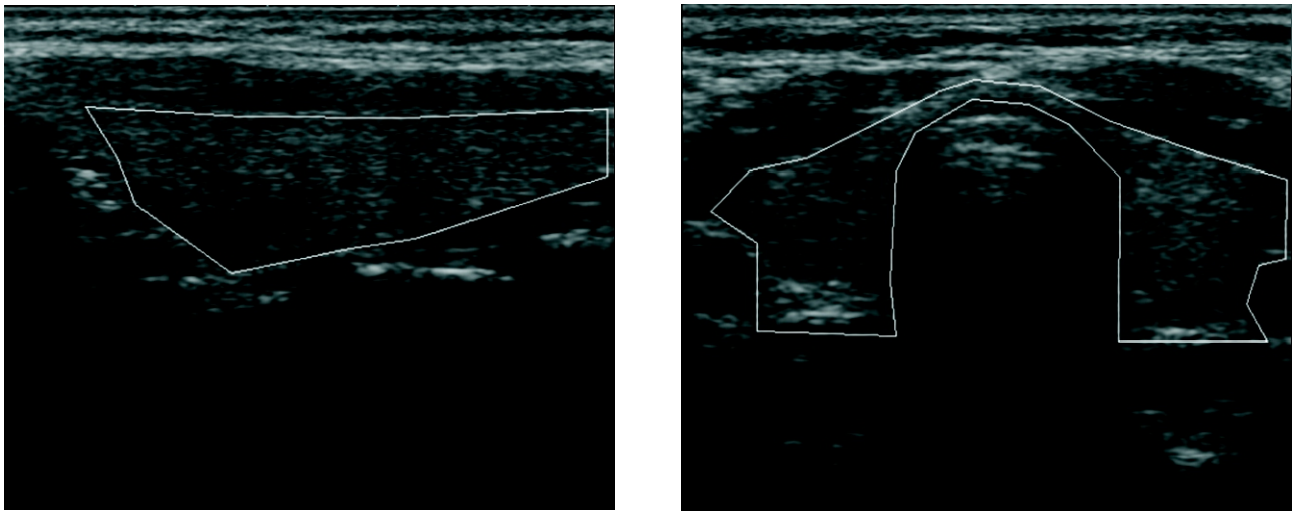
Consequently, the endeavours in comparing computer-aided detection of various disorders with human reading of medical images have appeared. For example the study from the field of screening mammography [7] or the field of pletysmo-graphy [8].

In this paper, we aimed to compare results of human observers (endocrinologists) in a diagnostics of autoimmune thyroiditis from B-mode ultrasound images with results achieved by the classifier described in [6]. Moreover, acquired data were evaluated on inter- and intra-observer variability of human observers.

2. Materials and methods
The data set consists of B-mode ultrasound images of 161 patients (subjects) from the Department of Endocrinology who were referred ultrasound examination of the thyroid gland. The subjects were consecutively screened to the study during three months enrolment period. The subjects with local changes such as nodules were excluded. For each subject from 4 to 8 B-mode ultrasound images were scanned in both longitudinal and transverse planes (see Figure 1) by the ultrasound system EnVisor M2540A equipped with an 8MHz linear probe and were saved in the DICOM format.

The data set were divided to the three classes according to the diagnosis confirmed by a clinical examination: H (normal gland, 26 subjects), BS (borderline state between healthy and inflamed tissue, 14 subjects) and AT (autoimmune thyroiditis, 121 subjects). Moreover a fourth class “Evaluation impossible” was added, which could be used in cases where a human observer was not able to conclude the diagnosis.

Four observers (A, B, C, D) are all endocrinologists with at least 10 years of experience in the ultrasound examination of the thyroid gland. They evaluated the anonymized ultrasound data by using a web-application, developed for this purpose, without any knowledge about the previous diagnosis and clinical results. The observers were instructed to choose one of the classes (H, BS, AT and “Evaluation impossible”) for each subject displayed in the random order in the web application. Time for evaluation was not limited. In general, the observers executed this during three weeks with five second per subject on average. Each subject was evaluated three times (in three rounds) by each of four observers completely.



(a) (b)

Fig. 1. Ultrasound images of thyroid gland scanned for subjects with delineated region of interest: (a) longitudinal scan of inflamed tissue; (b) transverse scan of healthy tissue.

Simultaneously the ultrasound data were evaluated by the classifier (CL) designed and verified in the last project [6]. The classifier is based on the Bayesian decision theory and used the texture features designed by Muzzolini and Haralick [9], [10]. As different ultrasound scanners give different numerical values for same texture features [5], new feature selection was done for the ultrasound scanner used in this study. As the most suitable feature was selected Muzzolini's spatial texture feature, which represented a gradient magnitude of the image texture [9]. The data set was split into a training set and an independent test set (see Table 1) to allow classifier training to the ultrasound image texture acquired from the used scanner.

At first, the success of observers and the classifier in diagnosing to the correct classes were compared. Secondly, inter- and intra-observer variability of human observers were measured by kappa statistic [11]. The kappa values are usually obtained from the interval <0,1> and are interpreted as a poor agreement for $\kappa = 0$ and a perfect agreement for $\kappa = 1$. The inter-observer variability was quantified separately for the each evaluation round by multiple-reader kappa κ_m and weighted kappa κ_w with quadratic weights and computed with 95% confidence limit. Weighted kappa is especially developed to allow evaluate importance of dis-agreements when the observer decision is made over multiple categories (three classes H, BS, AT) [12].

The intra-observer variability was quantified also by using the multiple-reader kappa κ_m, where the calculation was done on three evaluation rounds per observer instead of the calculation done on group of observers.

3. Results

3.1 Comparative study

The success in the evaluation of the whole data set by observers (A, B, C, D) and by the classifier (CL) is shown in Table 2.

Each classification of each subject was supposed to be an independent random experiment. Hence, total evaluations number given by the each observer was 483 in accordance with 161 subjects independently classified three times. The success rate of the classifier was determined on the independent test set including only 52 subjects because remaining subjects were used to train the classifier. And there was no need to do the evaluation three times because of the 100% reproducibility of the classifier results.

In the Table 2, the column "Not classified" is a number of subjects where observers used classification to the class "Evaluation impossible". The classification success is shown as a number of correct evaluations as well as frequencies, which are represented as a ratio of the number of correct evaluations and the number of evaluations in separate classes. For example for the observer A:

H: 33 / (3 * 26) = 0,423
BS: 28 / (3 * 14) = 0,667
AT: 214 / (3 * 121) = 0,590

The major misclassification expresses subjects from the class H classified incorrectly to the class AT and vice versa. The minor misclassification expresses the other incorrect classifications (from BS to H or AT and vice versa).

Tab. 1. Available data.

	class H	class BS	class AT
training set	18	10	81
independent test set	8	4	40

Tab. 2. Success in diagnosis of observers (A,B,C,D) and classifier (CL) in three separate classes H, BS, AT. Success is shown as a number of correct evaluations and frequencies in parentheses.

Observer	Success in separate classes			Overall success	Major misclassif.	Minor misclassif.	Not classified
	H	BS	AT				
A	33 (0.423)	28 (0.667)	214 (0.590)	275 (0.569)	16 (0.033)	190 (0.393)	2
B	58 (0.744)	22 (0.524)	173 (0.477)	253 (0.524)	47 (0.097)	165 (0.342)	18
C	13 (0.167)	25 (0.595)	281 (0.774)	319 (0.660)	13 (0.027)	151 (0.313)	0
D	65 (0.833)	16 (0.381)	139 (0.383)	220 (0.455)	75 (0.155)	169 (0.350)	19
CL	8 (1.000)	4 (1.000)	35 (0.875)	47 (0.904)	2 (0.038)	3 (0.058)	0

Tab. 3. Contingency tables displaying results of the classification done by human observers and the classifier (leave-one-out classification process on the training data set and the classification on the independent test set). Columns of contingency tables represent true classes, each row represents classification results and the main diagonal represents correct classifications.

Observer A

	H	BS	AT	Not classif.
H	33	38	7	0
BS	9	28	5	0
AT	9	138	214	2

Observer B

	H	BS	AT	Not classif.
H	58	18	1	1
BS	16	22	4	0
AT	46	127	173	17

Observer C

	H	BS	AT	Not classif.
H	13	53	12	0
BS	1	25	16	0
AT	1	81	281	0

Observer D

	H	BS	AT	Not classif.
H	65	13	0	0
BS	24	16	1	1
AT	75	131	139	18

Classifier - LOO on training set

	H	BS	AT
H	18	0	0
BS	3	6	1
AT	5	8	68

Classifier – independent test set

	H	BS	AT
H	8	0	0
BS	0	4	0
AT	2	3	35

The minor and the major misclassification frequency are ratios of the number of incorrect classifications and total evaluations per observer (3 * 161) or the classifier (52).

Classification results of human observers and classifier are shown in contingency tables (Table 3).

3.2 Inter- and intra-observer variability

The inter- and intra-observer variability of human observers were examined, see Table 4, 5 and 6. The column N appeared in the following tables denotes the number of all classified subjects except those classified to the class "Evaluation impossible".

Three separate evaluation rounds of 161 subjects per observer were discriminated. The inter-observer variability was quantified separately for each round. In Table 4 multiple-reader kappa κ_m values represented overall agreement among observers. The amount of identically classified subjects by all observers is also mentioned. Weighted kappa κ_w values were evaluated to obtain detailed relation dependencies for pairs of observers.

Tab. 4. Inter-observer (A,B,C,D) agreement, multiple-reader kappa κ_m separately for three evaluation rounds.

Observers	N	All four observers agreed	κ_m
A,B,C,D (r.1)	152	63 (0.416)	0.448
A,B,C,D (r.2)	154	39 (0.253)	0.264
A,B,C,D (r.3)	147	31 (0.211)	0.258

Tab. 5 Inter-observer (A,B,C,D) agreement between paired observers. Class agreement for H, BS and AT and weighted kappa κ_w separately for three evaluation rounds.

;Pair of observers	N	Class agreement			κ_m	Confidence limits
		H	BS	AT		
A&B (r.1)	156	0.593	0.536	0.795	0.670	0.622 - 0.777
A&B (r.2)	158	0.333	0.465	0.707	0.544	0.453 - 0.635
A&B (r.3)	151	0.585	0.565	0.726	0.647	0.667 - 0.737
A&C (r.1)	160	0.533	0.682	0.823	0.698	0.614 - 0.781
A&C (r.2)	161	0.286	0.593	0.790	0.535	0.422 - 0.648
A&C (r.3)	160	0.174	0.474	0.754	0.499	0.400 - 0.598
A&D (r.1)	155	0.480	0.442	0.824	0.673	0.594 - 0.751
A&D (r.2)	155	0.281	0.397	0.720	0.544	0.462 - 0.627
A&D (r.3)	153	0.520	0.450	0.600	0.569	0.485 - 0.653
B&C (r.1)	156	0.429	0.536	0.823	0.675	0.570 - 0.751
B&C (r.2)	158	0.167	0.212	0.671	0.463	0.383 - 0.543
B&C (r.3)	151	0.087	0.226	0.627	0.402	0.320 - 0.484
B&D (r.1)	152	0.706	0.455	0.779	0.717	0.634 - 0.800
B&D (r.2)	154	0.714	0.527	0.720	0.698	0.614 - 0.782
B&D (r.3)	147	0.755	0.605	0.675	0.710	0.629 - 0.792
C&D (r.1)	156	0.286	0.385	0.759	0.521	0.423 - 0.619
C&D (r.2)	155	0.139	0.180	0.618	0.388	0.306 - 0.471
C&D (r.3)	153	0.069	0.175	0.463	0.298	0.225 - 0.370

Tab. 6. Intra-observer (A,B,C,D) agreement, multiple-reader kappa κ_m for individual observers.

Observers	N	All three rounds agreed	κ_m
A	160	94 (0.588)	0.534
B	151	89 (0.589)	0.578
C	161	108 (0.671)	0.545
D	151	80 (0.530)	0.525

Values are shown in Table 5 with class agreement between two observers which is calculated as a number of correct evaluations done by both observers divided by a number of subjects classified to this class by both observers [10], [11].

The intra-observer variance (Table 6) represents agreement in evaluation of each subject in three rounds for each observer. The different values in column N are given by omitting subjects marked as "Evaluation impossible".

4. Discussion and Conclusion

In the comparative study the evaluation success of B-mode ultrasound images by human observers and the classifier were compared. Observers A and C and the pair of observers B and D achieved similar success rates. Thus observers were divided into two groups. Observers A and C achieved the high success rate in the classification to the class AT (59.0% and 77.4% respectively). However, the observer C also classified several healthy subjects (class H) to classes representing the positive disease, i.e. false-positive rate 12 from 13 major misclassifications were done from the class H to the class AT. The observer A major misclassification was balanced 7 from 16 were done from the class H to the class AT (see Table 3).

To the contrary, observers B and D classified more successfully to the healthy class (74.4% and 83.3%) and used more the possibility not to classify unclear subjects (18 and 19 subjects respectively). Further, observers B and D classified with false negative rate. They misclassified 47 and 75 subjects from AT class as healthy (class H) respectively, as shown in Table 3.

The classifier was verified by leave-one-out classification on the training set, where good results (84.4%) were acquired. Classification success on 52 subjects from the independent test set was relatively high and balanced for all three considered classes (100.0% for H, 100.0% for BS and 87.5% for AT). There were 2 major misclassifications both from AT to H (see Table 3), which represent the higher false negative rate. As mentioned in Introduction, the observer's (physician's) decision is based mainly on subjective evaluation of the structure and echogenicity of parenchyma in ultrasound images and their clinical experience, whereas the classifier evaluation is based on quantifiable indices. Thus, the computer-aided diagnosis of autoimmune thyroiditis combined with physician clinical experience would be the best way.

In inter- and intra-observer variability study, it was noted that in the first evaluation round observers agreed more (κ_m = 0.448) than in two following rounds (κ_m = 0.264 and 0.258), see Table 4. The reason could have been that the observers in the first round evaluated data set with considerable interest and higher attention than in two following rounds.

There can be seen in Table 5, that the agreement in pairs of the observer C against B and D, represented by values of κ_w, is lower than in pairs of observers A, B and D against each other. The reason is in the substantial false-positive rating in the data set classification by observer C (see Table 2) against false-negative rating done by observers B and D. Relationships of observers A, B and D are relatively moderate up to substantial. The best agreement is given in pairs A against B and B against D, which is probably caused by the observer B, who made decisions on the data set with the success against to clinical examinations somewhere between the classification success of observers A and D (Table 2).

The intra-observer variability quantified by multi-reader kappa (κ_m) values is compared in Table 6. There is no significant difference among observers (κ_m: 0.534, 0.578, 0.545 and 0.525). When we consider kappa values from the interval <0,1> where κ_m = 0 means the poor agreement and κ_m = 1 means the perfect agreement, all four observers in our study achieved the moderate agreement which also means the moderate intra-observer variability.

Acknowledgement

This work has been supported by the Grant Agency of Charles University in Prague, Czech Republic, grant number 119607 and by the grant SVV-2010-265 513.

The authors wish to thank endocrinologists Prof. Dr. Michal Kršek CSc., Dr. Jan Jiskra PhD. and Dr. Petr Sucharda in the Department of Endocrinology of Charles University in Prague for their effort at evaluation of ultrasound images.

References

[1] Warftsky L., Ingbar S.H. : Disease of the thyroid. In Harrison's Principles of Internal Medicine, page 1712. McGraw-Hill, New York, 12th edition, 1991.

[2] Simeone F.J., Daniel G.H., Müller P.R. et al. : High-resolution real-time sonography. Radiology, 155:431439, 1985.

[3] Solbiati L., Volterrani L., Rizzatto G., et al. : The thyroid gland with low-uptake lesions: Evaluation by ultrasound. Radiology, 155:187196, 1985.

[4] Šára R., Smutek D., Sucharda P., Svačina Š. : Systematic construction of texture features for Hashimoto's lymphocytic thyroiditis recognition from sonographic images. In S. Quaglini, P. Barahona, and S. Andreassen, editors, Artificial Intelligence in Medicine, LNCS, Berlin-Heidelberg, Germany. Springer, 2001.

[5] Smutek D., Šára R., Sucharda P., Tjahjadi T., Švec M. : Image texture analysis of sonograms in chronic inflammations of thyroid gland. Ultrasound in Medicine and Biology, 29(11):1531 1543, 2003.

[6] Holinka Š., Šára R., Smutek D. : Relation Between Structural Changes in B-mode Ultrasound Images of Thyroid Parenchyma and the Presence of Thyroid Antibodies in Blood Sample. Machine Graphics Vision, 18(1):67-82, 2009.

[7] Georgian-Smith D., Moore R.H., Halpern E. : Blinded Comparison of Computer-Aided Detection with Human Second Reading in Screening Mammography. American Journal of Roentgenology, 189:11351141, 2007.

[8] Paez S., Seiden D., Kiel M., Chediak A. : A comparison of human and computer apnea/hypopnea detection from respiratory waveforms derived by respiratory inductive plethysmography. Sleep Research, 25: 521, 1996.

[9] Muzzolini R., Yang Y.-H., Pierson R. : Texture characterization using robust statistic. Pattern Recognition, 27(1):119134, 1994.

[10] Haralick R.M., Saphiro L.G. : Computer and vision. vol. 1. Reading: Addison-Wesley Publishing Company, 1993: 453 506.

[11] Fleiss J.L., Levin B., Paik M.C. : Statistical methods for rates and proportions. John Wiley, New York, third edition, 2003.

[12] Kundel H.L., Polansky M. : Measurement of observer agreement. Radiology, 228(2):303308, August, 2003.

Contact
MUDr. Štěpán Holinka
3rd Department of Medicine,
1st Medical Faculty, Charles University in Prague
U Nemocnice 1
128 08 Prague 2
Czech Republic
e-mail: stepanholinka@gmail.com