Development of a medical expert system: disease staging by a fuzzy classifier

A S Serobabov, L A Denisova

Omsk State Technical University, 11, Mira ave., Omsk, 644050, Russia
E-mail: aserobabow95@mail.ru, denisova@asoiu.com

Abstract. The article considers an approach to the processing and analysis of biomedical data from clinical examination. An automated diagnosis system is presented to be applied in liver disease diagnosis by the results of the initial examination of patients using a classifier based on fuzzy inference. The research is carried out by modeling the proposed system in the environment MATLAB FuzzyLogic. The regularities are revealed making it possible to normalize the results of laboratory tests depending on gender characteristics and to increase the percentage of correctly determined disease stages.

1. Introduction
Medical institutions annually generate large amounts of data on clinical diagnosis of diseases. Since the amount of stored patients' data is growing [2], the workload of doctors is increasing. Thus, making a correct and timely diagnosis requires special information systems that can automate the process of diagnosis and perform the tasks of storage, processing, and normalization of data. Decision support systems (DSS) act as a comprehensive solution to the tasks set [3]. This article proposes DSS for the diagnosis of liver disease. The system includes a step of the disease staging based on a classification algorithm using fuzzy inference [1]. In this case, the decision about assigning a patient to one of the disease stages is formed on the input data of the system, represented as fuzzy variables. This method makes it possible to reduce parameters to single relative units of different biological nature, and also to use linguistic means of determining expert knowledge. Another useful feature of the expert system is the storage of high-quality experience and the ability to transfer it to other doctors, which makes the creation of such systems economically feasible.

2. Problem statement
The article explores an important public health problem - the diagnosis of liver disease, which has become one of the most prominent problems of medicine around the world [4]. The attention of scientists has been focused on noninvasive diagnostic methods since 2003 according to scientific journals [5]. These methods help to avoid any kind of surgery, as well as automate the work of a doctor and identify the disease at an early stage of development.

El-Sappagh et al. [6] used a fuzzy analytical hierarchy process (FAHP) and an adaptive neuro-fuzzy inference system (ANFIS) to study the fibrosis stages in patients with hepatitis, achieving a classification accuracy of 93.3%. Belciug [7] proposed a model that extends a single-hidden layer feedforward neural network trained using logistic regression (LogSLFN) algorithm to the case of multiple classes. Two approaches were used: a parallel LogSLFN, and a cascaded LogSLFN. The model was tested on two medical datasets regarding cancer diagnosis and liver fibrosis staging. According to the test results, the accuracy value for fibrosis was about 81%. Wang et al. [8] present...
feature extraction strategies and classification methods for liver fibrosis using K-nearest neighbors (KNN) and support vector machines (SVM). Their proposed approach showed 99% accuracy but did not take into account the differences between the diagnosis of alcoholic and non-alcoholic fatty liver diseases. Choi et al. [9] developed a deep learning system (DLS) for staging liver fibrosis by using CT images.

All research in this area has significant drawbacks. Thus, El-Sappagh considered non-alcoholic fatty liver disease together with hepatitis. Belciug has a diagnosis accuracy of 80%. Jin Wang does not differentiate between alcoholic and non-alcoholic fatty liver disease. The method proposed in this paper is based on laboratory parameters using an expert system with a fuzzy classifier based on the mathematical fuzzy inference module of Takagi-Sugeno [10]. Following the proposed technique, it is possible to assess the degree of the patient's membership in one of the three stages of liver disease. To do this, three input parameters are used for the disease staging $F_i \ (i=1,2,3)$: $L_{lep}$ is the leptin content, $L_{obr}$ is the number of leptin receptors, $D_{ng}$ is the absence (1) or the presence (2) of non-alcoholic steatohepatitis responsible for inflammatory processes in the patient's body, $C$ is the decision on the staging of the patient's disease. Moreover, the results of cluster analysis underlie the determination of quantitative ranges for lexical variables $L_{lep}, L_{obr}$ based on fuzzification functions.

3. Theory. An algorithm of the expert system for early diagnosis of the disease

To develop and study the course of the disease, a simulation model called ADS has been designed. The structural blocks of the model are shown in Figure 1. In the system under consideration, a doctor enters data about a patient through the graphical user interface. The requests for diagnosis are assumed to contain four values: $L_{lep}, L_{obr}, D_{ng}, G$. Then, the system saves the data to the patient database and preprocessed them. To implement an algorithm for disease staging based on fuzzy inference, normalized input parameters of laboratory tests are fed into the model, the parameters acting as features for a fuzzy classifier. The classifier, in its turn, invokes fuzzy production rules from the rule base. Then, based on the implementation of the Takagi-Sugeno fuzzy inference algorithm [10], the values are transmitted to the diagnostic unit, which converts the output values to graphically display the results to the doctor.
Figure 1. Block diagram of the expert system for early diagnosis of diseases

Step 1. Data for the study go through preprocessing, which includes exclusion of incomplete data, type transformation, normalization of input variables, duplicates removal, feature extraction, exclusion by inclusion criteria, and outlier test. These steps were performed in the previous research work [11].

Step 2. A fuzzy rule base is formed, which contains fuzzy production rules obtained from a preliminary analysis of clinical parameters: \( L_{lep} \), \( L_{obr} \), \( D_{ng} \) and the expert knowledge of hepatologists. The production rules present a set \( R = \{ R_1 \ldots R_n \} \), where each \( R_i \) is a production rule \( R_i \in R \) and is given by the expression \( R_i: \text{IF} \ x_1 \text{is} \ A_1^i \ \text{AND} \ldots \ \text{AND} \ x_M \text{is} \ A_M^i \ \text{THEN} \ y_i = f(i) (x_1 \ldots x_M) \), where \( A = \frac{x}{\mu_A(x)} \), \( \forall x \in X, \mu_A(x) \in [0, 1] \)[16], \( \mu_A \) is the membership degree \( \mu_A \in [0, 1] \), \( x \in X \), \( X \) is the set of input values, \( n \) is the number of rules, \( M \) is the number of terms.

Step 3. The values of parameter \( L_{lep} \) are corrected according to formula 1.

\[
N = \begin{cases} 
1, & \text{if } G = 0; \\
u, & \text{if } G = 1. 
\end{cases}
\]  

(1)

where \( u \) is the constant value of the input parameter correction for gender difference; \( G \) is the gender of the patient (0 is male, 1 is female).

Step 4. Fuzzification of input variables is carried out based on S and Z-functions of membership and the Gaussian function for a set of linguistic terms \( T = \{ S \text{ is a small parameter value, } M \text{ is an average parameter value, } L \text{ is a high parameter value} \} \). The membership functions are represented by the
formulae below. To define the extreme linguistic terms, $S$- and $Z$-functions of membership $\mu_S(x)$ are used, depending on two parameters $a$ and $b$, which determine the extreme points of the inclined part of the membership function. This function has several important properties: the function is non-increasing, its parameters determine the interval $[0; 1]$, within which the function decreases non-linearly. Similarly, the extreme linguistic term $L$ is defined using $S$-functions of membership $\mu_L(x)$ depending on two parameters $a$ and $b$. As the membership function for term $M$, a two-sided Gaussian membership function is adopted, which is combined using two Gaussian membership functions $\mu_M(x)$. Here, parameters $c_1$ and $c_2$ satisfy the following ratio $c_1 < c_2$ and set the values of the fuzzy set kernel, where $c_1$ is the minimum and $c_2$ is the maximum value of the kernel of the fuzzy set. Parameters $\sigma_1$ and $\sigma_2$ set the concentration coefficients of the left and right parts of the membership function. These membership functions for the terms of input parameters are chosen since they have simple formulae with a small number of parameters, which simplifies calculations, time modeling of the system, algorithm modification and simple control of test results.

$$\mu_S(x) = \begin{cases} 
1, & x < a; \\
1 - 2 \left(\frac{x - b}{b - a}\right)^2, & a \leq x \leq \frac{a + b}{2} \\
2 \left(\frac{x - b}{b - a}\right)^2, & \frac{a + b}{2} < x \leq b; \\
0, & x > b 
\end{cases}$$

$$\mu_M(x) = \begin{cases} 
e^{-\left(\frac{(x-c_1)^2}{2\sigma_1^2}\right)}, & x < c_1; \\
1, & \leq x \leq c_2; \\
e^{-\left(\frac{(x-c_2)^2}{2\sigma_2^2}\right)}, & x > c_2; 
\end{cases}$$

$$\mu_L(x) = \begin{cases} 
0, & x < a; \\
2 \left(\frac{x - b}{b - a}\right)^2, & a \leq x \leq \frac{a + b}{2} \\
2 \left(\frac{x - b}{b - a}\right)^2, & \frac{a + b}{2} < x \leq b; \\
1, & x > b 
\end{cases}$$

where $\mu_S, \mu_M, \mu_L$ are the membership functions for terms $S, M, L$.

Step 5. The rule block consists of formulae 2–3. Fuzzy logical operation AND is defined using the $\min$ function.

$$R_i = \min(\mu_A(x), \mu_B(y)) \cdot C_j; \quad (2)$$

$$\mu_i(C) = \min(\mu_A(x), \mu_B(y)). \quad (3)$$

where $C_j$ is the value of the output variable for the $j$-th term with a single value of the membership degree, $\mu_A(x)$ is the grade of membership in term $A$, $x, y$ are fuzzy numbers involved in the production
rule, \( i \) is the rule index, \( \mu(C) \) is the output value for the block of rules, given in the denominator of formula 4, \( R_i \) is the output value of the \( i \)-th block of the rules given in the numerator of formula 4.

**Step** 6. The values obtained from the blocks of rules participate in the operation of reverse transformation of fuzzy variables into crisp ones (defuzzification), and thus the values of output variable \( C \) are formed. This operation is performed by finding the weighted average according to formula 7.

\[
C = \frac{\sum_{i=1,m} R_i}{\sum_{i=1,m} \mu_i(C)} \tag{4}
\]

where \( C \) is the crisp value of output variable \( R_i \) is the output value from the block of rules, \( \mu_i(C) \) is membership in the rule, \( m \) is the number of rules in the system.

To build an expert system, the mathematical package MATLAB with FuzzyLogicToolbox was used [12, 13].

### 4. Results of experimental studies

The study included patients aged 18 to 65 with a cohort of 149 patients with NAFLD (nonalcoholic fatty liver disease). The percentage of men and women is 76.5% and 23.5%, correspondingly. The sample of \( X \) patients is characterized by 129 variables (patient parameters) and is represented by formula 5.

\[
X = \begin{pmatrix}
  x_{11} & \cdots & x_{1n} \\
  \vdots & \ddots & \vdots \\
  x_{m1} & \cdots & x_{mn}
\end{pmatrix}, \tag{5}
\]

where \( i \) is the number of the object (patient); \( j \) is the number of the variable (parameter).

The initial data for the fuzzy classifier are \( L_{lep}, L_{obr}, D_{ng} \), obtained from the previous work [11]. To stage liver fibrosis, an elastometry procedure (a method for diagnosing the fibrosis degree using a fibroscan) was performed for the patients from the study sample in medical institutions in Omsk. This assessment was carried out for 95 patients out of 149. The disease was classified according to the Metavir scale [14], characterizing the degree of liver disease by changes in the internal structure of the organ. Table 1 contains two columns: the stage of the disease and its notation, and the qualitative characteristics of the stage.

| Stage | Qualitative characteristics of the stage |
|-------|-----------------------------------------|
| \( F_0 \) | No fibrosis can be detected |
| \( F_1 \) | Fibrosis exists with expansion of portal zones |
| \( F_2 \) | Fibrosis exists with expansion of most portal zones, and occasional bridging |
| \( F_3 \) | Fibrosis exists with expansion of most portal zones, marked bridging, and occasional modules |
| \( F_4 \) | Presence of cirrhosis |
As a result of step 1, table 2 was built for stage distribution in the sample.

**Table 2. Basic information of the collected data**

| Disease stage | Number of patients | Average patient age | Number of men | Number of women |
|---------------|--------------------|---------------------|---------------|-----------------|
| $F_0$         | 35                 | 49.5                | 24            | 11              |
| $F_1$         | 32                 | 48.5                | 23            | 9               |
| $F_2$         | 19                 | 47.4                | 18            | 1               |
| $F_3$         | 8                  | 49.1                | 7             | 1               |

At step 2, 11 production rules presented in Table 3 have been compiled together with experts in the field of hepatology, based on their practical knowledge and preliminary research. At the intersection of the column and the row of parameter terms $L_{lep}$, $L_{obr}$, the disease stage is determined. At each intersection $L_{lep}$, $L_{obr}$, the values of the disease stage are obtained, but there are two exceptions to this rule. At the intersection of lexical variables $L$ and $M$ for parameters $L_{lep}$ and $L_{obr}$, an additional parameter is introduced, due to the complex distinctness between the second and third stages of the disease. Thus, when the value of parameter $L_{obr}$ corresponds to lexical variable $M$ and the value of $L_{lep}$ corresponds to $L$. At $D_{ng}=1$, the classifier determines stage $F_2$, but at value $D_{ng}=2$, it determines stage $F_3$.

**Table 3. Table of fuzzy rules for the disease staging**

| Parameter value $L_{lep}$ | Parameter value $L_{obr}$ |
|---------------------------|---------------------------|
| $S$ (small parameter value) | $F_1$                     |
| $M$ (average parameter value) | $F_2$                     |
| $L$ (large parameter value) | $F_3$                     |

The resulting graphs of membership functions for parameters $L_{lep}$, $L_{obr}$ and $D_{ng}$ are shown in Figure 2. As seen from the graph, the range of parameter values $L_{lep}$ and $L_{obr}$ is $\in^{+}Q$. Linguistic terms $S, L$ of parameters $L_{lep}$ and $L_{obr}$ are defined by $Z$-functions, $S$-functions at the edges of the graph. Input parameter $L_{obr}$ and its linguistic variable $S$ has $\mu(L_{obr})=1$ in the interval $\{0, 4\}$, due to the biological nature of the parameter. Parameter $D_{ng}$ has a categorical nature since it characterizes the degree of inflammatory processes in the body; therefore, it can only take values $N=1$ or $P=2$, where $N$ means the absence of inflammatory processes in the body, and $P$ means their presence.
The outputs from the blocks of production rules are connected to two adders, thus implementing defuzzification (reverse transformation of fuzzy variables into crisp ones). Based on formula 4 of the weighted average, value $C$ is obtained, being the output value of the expert system, indicating the stage of the patient's disease. The numerator contains the stage values for each block of production rules, and the denominator has the values of the membership degree in this term.

After defuzzification numeric $C$ is formed at the output using a fuzzy classifier; then, it is transmitted to the diagnosis unit and rounded to the nearest integer, which is the resulting stage of the patient's disease. After that, the final decision is shown to the doctor through the graphical interface. The doctor determines the structural changes of the liver due to the Metavir scale and formulates a final diagnosis for the patient.

The system was tested as follows: the data were divided into research and control ones. The research data were used to build membership functions, experts applied them to select ranges of terms, and a correlation matrix was built in [11] to find significant parameters. Control data were not used in the above stages, and experts cannot rely on them when creating fuzzy production rules.

The results of testing on a control sample using the parameter correction subsystem $L_{lep}$ are presented in Table 4.

**Table 4. Testing the expert system with the program correction block $L_{lep}$**

| №  | $L_{lep}$ | $L_{obr}$ | $D_{ng}$ | $G$ | Doctor diagnosis | C value (ADS diagnosis) | Match of system diagnosis and doctor diagnosis |
|----|-----------|-----------|----------|-----|------------------|-------------------------|---------------------------------------------|
| 1  | 86.60     | 4.50      | 1        | 1   | 1                | 1.07 (1)                | +                                           |
| 2  | 11.90     | 12.25     | 2        | 0   | 2                | 1.66 (2)                | +                                           |
The precision of testing the system without a correction module, its results presented in Table 5, is 76%. This difference can be due to the prevalence of men in the sample (33.3% of women and 66.6% of men), and therefore their influence on the overall percentage of diagnosis is greater. Consequently, the correction of parameter $L_{lep}$ leads to a better classification result, because it takes into account inter-gender differences.

| № | $L_{lep}$ | $L_{obr}$ | $D_{ng}$ | $G$ | Doctor's diagnosis | $C$ value (system diagnosis without correction) | Match of ADS diagnosis with doctor's diagnosis | ADS diagnosis with correction with doctor's diagnosis | Difference from the system without correction |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 20.63 | 9.19 | 2 | 0 | 3 | 1.59 (2) | - | - | 0 |
| 4 | 16.60 | 10.16 | 2 | 0 | 2 | 1.54 (2) | + | + | 0.06 |
| 5 | 14.39 | 8.47 | 1 | 0 | 1 | 1.47 (1) | + | + | 0 |
| 6 | 19.13 | 3.82 | 1 | 0 | 1 | 1.02 (1) | + | + | 0.21 |
| 7 | 108.8 | 3.85 | 2 | 1 | 1 | 1.02 (1) | + | + | 0 |
| 8 | 30.66 | 7.86 | 2 | 1 | 1 | 1.55 (2) | + | + | 0.06 |
| 9 | 86.60 | 4.56 | 1 | 1 | 1 | 1.08 (1) | + | + | 0 |
| 10 | 30.14 | 7.03 | 2 | 0 | 2 | 1.68 (2) | + | + | 0.06 |
| 11 | 26.65 | 6.45 | 1 | 0 | 1 | 1.50 (1) | + | + | 0 |
| 12 | 39.57 | 9.67 | 1 | 0 | 1 | 2.06 (2) | - | - | 0 |
| 13 | 38.38 | 7.30 | 2 | 1 | 1 | 1.50 (1) | + | + | 0.06 |
| 14 | 6.551 | 16.52 | 1 | 0 | 2 | 1.96 (2) | + | + | 0 |
| 15 | 33.02 | 5.64 | 1 | 1 | 1 | 1.40 (1) | + | + | 0.06 |
| 16 | 108.80 | 3.85 | 2 | 1 | 1 | 1.06 (1) | + | + | 0 |
| 17 | 23.24 | 7.26 | 1 | 0 | 1 | 1.49 (1) | + | + | 0.06 |
| 18 | 1.97 | 6.62 | 2 | 0 | 2 | 1.45 (1) | - | - | 0 |
| 19 | 22.98 | 7.29 | 1 | 0 | 1 | 1.48 (1) | + | + | 0.06 |
| 20 | 74.3 | 9.23 | 2 | 0 | 3 | 2.75 (3) | + | + | 0 |
| 21 | 64.66 | 8.00 | 2 | 0 | 3 | 2.53 (3) | + | + | 0.06 |

Table 5. Differences between the results obtained with and without the presence of a parameter correction block $L_{LeP}$

The results of the patients' classification in the control sample are shown in Figure 3. The results obtained are divided into three groups depending on the disease stage and are presented as a histogram. Each group contains two columns: doctor's diagnosis, ADS diagnosis. The height of the column is taken as the number of patients diagnosed with a certain disease stage. As seen in Figure 3, the number of cases by stages $F_1$, $F_2$, $F_3$ differ depending on the method of diagnosis. Thus, the diagnoses made by ADS differ from the doctor's diagnosis when determining the second and third stages. This is due to the difference between laboratory indications and direct changes in the structure of the liver, which characterize the degree of the disease.
5. The discussion of the results

The proposed fuzzy system can be used to stage non-alcoholic fatty liver disease and can be applied by a doctor as a preliminary screening stage during the medical examination of the population. Good results have been achieved with the designed system, the percentage of a correctly diagnosed diagnosis exceeding 85% for a 3-categorical sample. The classification system correctly determines the stage between $F_3$ and $F_1$, which is indicated by a 100% correct diagnosis, but since there are little data on stage $F_3$, additional studies are needed. The main errors in staging arise at the boundary between $F_2$ and $F_1$, which can be explained by the undetermined development of the disease. Moreover, there are several detailed aspects of the work to be discussed.

First, the work uses an approach based on laboratory tests. This allows a large amount of patient data to be processed without attracting additional resources such as liver biopsy, MRI (magnetic resonance imaging) and CT (computed tomography) images to detect textural features of the liver.

Second, this approach does not contradict other diagnostic methods. Joint use with already proven methods will allow better results of the disease classification. The block diagram of the expert system for early diagnosis of the disease may include additional modules, thus ensuring the reliability of the results obtained. However, this approach requires additional research to assess the effect of synergistic relationships.

Third, the system can be extended with a special module of correlation substitutions, which will eliminate the system's limitations on the integrity of the input data set. If one of the input parameters of the system is absent, it will not affect the accuracy of the diagnosis significantly. Additional research will result in a better understanding of the complex correlations in such a dynamic system as a human.

Fourth, the designed system has no optimization components, which means that the diagnostic accuracy can be higher. Like any system, it has the potential for improvement. Additional research data will increase the reliability of the study. Implementation of the system in medical institutions approves the results in practice.

6. Conclusion

An expert system with a percentage of correct diagnosis of 85% has been designed. In the course of the study, cross-gender differences in the laboratory parameters have been taken into account. Fuzzy production rules to stage the disease have been formed. As a result of the study, the boundaries of the linguistic terms $S, M, L$ for the parameters $L_{lep}, L_{abr}$ have been obtained. Membership functions have been used to construct a fuzzy classifier.

The proposed method of processing medical data when used with techniques designed for diagnostic purposes makes it possible to create computer software facilitating the clinical diagnosis of diseases.
based on the knowledge and experience of doctors. Such an expert decision support system will significantly increase the efficiency of the hepatologist's work and reduce the time spent on making an initial diagnosis. In addition to the aforementioned beneficial features of the system, it always applies high-quality expertise, which represents the level of thinking of the most qualified experts and allows this knowledge to be transferred to other medical institutions.

References

[1] Sweidan S, El-Sappagh S, El-Bakry H, Sabbeh S, Badria F, Kwak 2019 A Fibrosis Diagnosis Clinical Decision Support System Using Fuzzy Knowledge Arabian Journal for Science and Engineering vol 44 pp 3781–3800

[2] Saleh, E, Valls A, Moreno A, Romero P 2016 Diabetes retinopathy risk estimation using fuzzy rules on electronic health record data Modeling Decision for Artificial Intelligence MDAI Lecture Notes in Computer Science pp 263–274

[3] Nazari S, Fallah M, Kazemipoor H and Salehipour A 2018 A fuzzy inference- fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases Expert Syst. Appl vol 95 pp 261271

[4] Kayaaltý Ö, Aksebzeci B, Karahan İ, Deniz K 2015 Liver fibrosis staging using CT image texture analysis and soft computing Applied Soft Computing vol 25 pp 399–413

[5] Keltch B, Lin Y and Bayrak C 2014 Comparison of AI techniques for prediction of liver fibrosis in hepatitis patients patient facing systems J Med Syst vol 38 p 60

[6] El-Sappagh S, Ali F, Ali A A, Hendawi, Badria F and Suh D 2018 Clinical Decision Support System for Liver Fibrosis Prediction in Hepatitis Patients: A Case Comparison of Two Soft Computing Techniques in IEEE Access vol 6 pp 52911–52929

[7] Belciug S 2021 Parallel versus cascade logistic regression trained single-hidden feedforward neural network for medical data Expert Systems with Applications vol 170

[8] Wang J, Li H, Zhou X 2020 A study of hepatic fibrosis staging methods using diffraction enhanced imaging J Image Video Proc vol 32

[9] Choi K, Jang J, Lee S 2018 Development and validation of a deep learning system for staging liver fibrosis by using contrast agent-enhanced CT images in the liver Radiology vol 289 pp 688–697

[10] Takagi T and Sugeno M 1985 Fuzzy identification of systems and its applications to modeling and control IEEE transactions on systems, man, and cybernetics vol 15 pp 116–132

[11] Serobabov A, Chebanenko E, Denisova L, Crolevec T 2018 Development of an expert system for early diagnosis of diseases: software for primary processing and identification of dependencies The Journal Omsk Scientific Bulletin vol 160

[12] Shtovba S, Designing fuzzy systems using MATLAB. Moscow: Hot line–Telecom, 2007 pp 187–223

[13] Meshcheryakov V and Denisova L 2016 Computer-aided design of the fuzzy control system using the genetic algorithm Dynamics of Systems, Mechanisms and Machines p 7819000

[14] Liao R, Fu Y, Wang T2017 Metavir and FIB-4 scores are associated with patient prognosis after curative hepatectomy in hepatitis B virus-related hepatocellular carcinoma: a retrospective cohort study at two centers in China. Oncotarget. vol 8 pp 1774–1787