The automatization of the medical diagnosis on the basis of an X-ray images of a patient with the restrictions of both possible errors on the desired levels

Kachiashvili K.J.[a,b,c], Kachiashvili J.K.[a,c], Kalandadze R. M.[a], Kvaratskhelia V.V.[a,c]

[a] Faculty of Informatics and Control Systems, Georgian Technical University, 77, st. Kostava, Tbilisi, 0160, Georgia, k.kachiashvili@gtu.edu.ge,
[b] Department of Probability Theory and Mathematical Statistics of the I. Vekua Institute of Applied Mathematics, Tbilisi State University, Tbilisi, Georgia
[c] Department of Informatics of the Muskhelishvili Institute of Computational Mathematics, Georgian Technical University, Tbilisi, Georgia

The International Conference "Distributed Computing and Grid Technologies in Science and Education", the Joint Institute for Nuclear Research (JINR), 3 – 7 July 2023, Dubna, Russia

Presentation structure

1. Introduction
2. Disease data acquisition and preprocessing results
3. The methods used for making a decision
   - 3.1. The Wald's test
   - 3.2. The method of sequential analysis of Bayesian type
4. Investigation of the used methods by simulation and real data
5. Conclusion
6. References

1. Introduction

Making a diagnosis of the disease is the initial and very important stage of the treatment of a sick person, the correctness of which greatly depends on the successful completion of the subsequent stages of treatment. Accurate and timely diagnosis practically (with high probability) ensures the cure of the patient's disease. The diagnosis is made based on the examination of the patient's condition by the doctor. Examination of the condition involves blood, urine and other analyzes of the patient, as well as observation of various organs, which can be done by many different methods, including the use of X-rays and radiation. Based on the results of the observation, the doctor of the relevant profile makes a decision about the presence or absence of the disease. The correctness of the decision depends greatly on the qualification and experience of the doctor. Different doctors can make different decisions on the same data. A misdiagnosis can lead to a disastrous outcome with high probability.

In order to avoid such subjective errors and to improve the quality of diagnosis, in recent decades, attempts have been made to use modern computers for diagnosis through machine learning and artificial intelligence methods (see, for example, (Bishop, 2006)). While diagnosing, as well as when making any decision, two types of errors are possible: mistaking a sick person for healthy, and mistaking a healthy person for sick. The correctness of the decision depends greatly on the qualification and experience of the doctor. The results caused by such errors are diametrically (significantly) different from each other. In the second case, after some stress experienced by the patient, on the basis of additional examinations, the real condition of the patient will be established, and in the first case, the result will be fatally disastrous with a high probability. Based on what has been said, the requirements for automatic diagnosis methods are clearly visible. They should minimize possible errors of both types, especially the possibility of errors of the first type.

Among the diseases that exist today, human lung diseases with pneumonia and cancer occupy an important place. „Pneumonia is a form of acute respiratory infection that affects the lungs. The lungs are made up of small sacs called alveoli, which fill with air when a healthy person breathes. When an individual has pneumonia, the alveoli are filled with pus and fluid, which makes breathing painful and limits oxygen intake that can cause the death“ https://www.who.int/news-room/fact-sheets/detail/pneumonia. „Cancer is a generic term for a large group of diseases that can affect any part of the body and cause the death. One defining feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade adjoining parts of the body and spread to other organs; the latter process is referred to as metastasis. Widespread metastases are the primary cause of death from cancer“ https://www.who.int/news-room/fact-sheets/detail/cancer.

According to the World Health Organization “Pneumonia is the single largest infectious cause of death in children worldwide. Pneumonia killed 740 180 children under the age of 5 in 2019, accounting for 14% of all deaths of children under 5 years old but 22% of all deaths in children aged 1 to 5 years” https://www.who.int/news-room/fact-sheets/ detail/pneumonia. Also “Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths. The most common cancers are breast, lung, colon and rectum and prostate cancers. Each year, approximately 400 000 children develop cancer. Cancer mortality is reduced when cases are detected and treated early” https://www.who.int/news-room/fact-sheets/detail/cancer. Thus, timely correct diagnosis of the presence of the mentioned diseases is a very necessary and important problem.

The article proposes methods of automatic diagnosis of pneumonia and lung cancer, which allow to reduce both types of errors mentioned above to the desired levels. Besides their widespread, these diseases are interrelated as mentioned in the paper (Tanaka S., Inoue M., Yamaji T., et al., 2023): “We found a positive association between incident cancer and risk of death pneumonia in this study. These results imply the possibility that the immunocompromised status and respiratory failure due to antitumor treatment.”

Two types of lung cancer are discussed in the presentation: adenocarcinoma and carcinoma. „Carcinoma is the most common form of cancer. It starts in the epithelial tissue of your skin or internal organs.

Adenocarcinoma is a subtype of carcinoma. It grows in the glands that line the insides of your organs“ https://www.google.com/search?q=What+is+difference+between+Adenocarcinoma+and+carcinoma+diseases%3F%0D%0A&source=hp&ei=LU06ZJyKB--Rxc8Pjouq4Ak&iflsig=AOEireoAAAAAZDpbPTDA8mZaPmgq48pCPr6Vtve EKtYA&ved=0ahUKEwjc1Pbgqav-AhXvSPEDHY6FCpwQ4dUDCAg&uact=5&oq=What+is+difference+between+Adenocarcinoma+and+carcinoma+diseases%3F%0D%0A&gs_lcp=Cgdnd3Mtd2l6EANQAFgAYABoAHAAeACAAQCIAQCSAQCYAQCgAQKgAQE&sclient=gws-wiz.

To make a decision about the diagnosis of the disease, the observation results extracted from the images obtained by radiation irradiation are used, which, like most of the observation results, contain a random component and, therefore, has a random character. Therefore, statistical hypothesis testing methods are used to make decisions, which allow restricting both types of possible errors. Such methods are Wald's sequential analysis method and constrained Bayesian method (Wald, 1947a,b; Kachiashvili, 2010, 2018). It is shown that both methods provide the opportunity to solve the given problem. It is also shown that the constrained Bayesian method, as a rule, requires a relatively small number of observations to make a decision with a given reliability than the Wald method, which is completely consistent with the results obtained earlier by the author of CBM and is its advantage (Kachiashvili, 2013, 2014a, 2015, 2018).

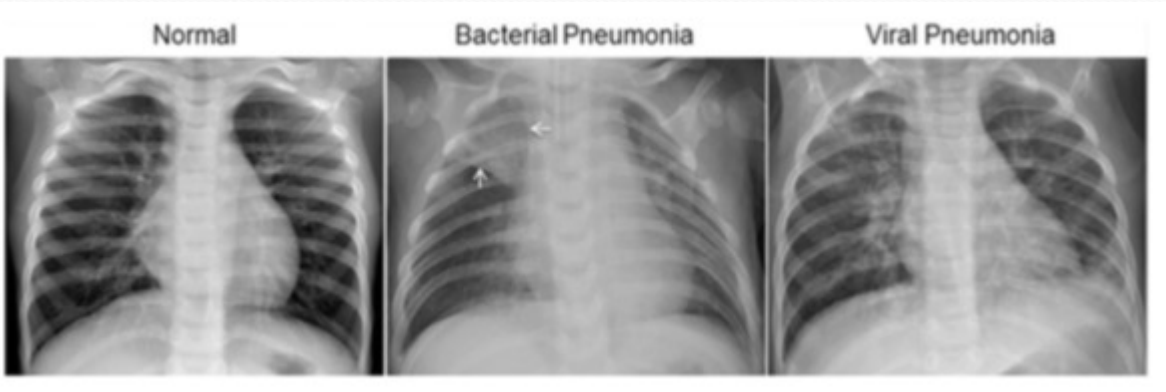
The results of the investigation are distributed in my talk as follows. Disease data acquisition and preprocessing results are presented in the beginning. After the methods used for making decisions are described. Then the results of the investigation of the applied methods using simulation and real data are given with the demonstration of the results of processing of experimental data. In particular, there are shown the results of statistical processing of the data of lung diseases by pneumonia, adenocarcinoma and carcinoma, as well as the results of combined data of both kinds of the cancer. Finally, the results of diagnosis on the basis of the data of pneumonia, adenocarcinoma, carcinoma and the combined data of both kinds of cancer are given.

2. Disease data acquisition and preprocessing results

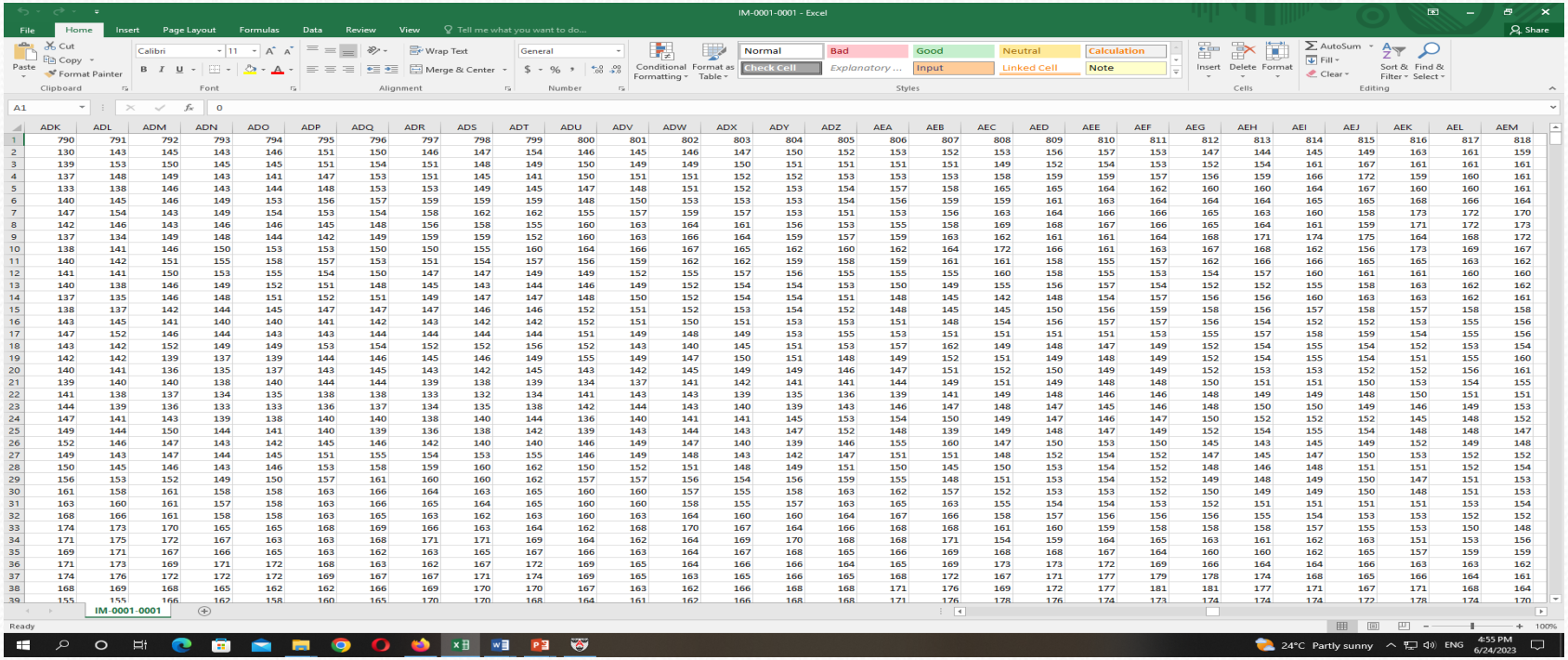
Data from lung pneumonia and lung cancer patients, as well as from healthy patients examined by the same method, were obtained from the Internet at the following web addresses under the appropriate names:

- Chest X-Ray Images (Pneumonia) (https://www.kaggle.com/datasets/paultimothy mooney /chest-xray-pneumonia)
- RSNA Pneumonia Detection Challenge (https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge/overview)
- VinBigData Chest X-ray Abnormalities Detection (https://www.kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection /data)
- Viral Pneumonia, Normal (https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset)
- Chest CT-Scan images Dataset (Cancer) (https://www.kaggle.com/datasets/mohamedhanyyy /chest -ctscan-images)

As it is clear from the indicated addresses, the examination of pneumonia patients was carried out on the basis of X-ray images, and cancer patients - on the basis of CT Scan images. At the mentioned addresses, photographs showing the condition of the lungs of the examined patients obtained by appropriate methods are provided. Photos are in black and white format like shown below.



For the digital representation of visual images, for their further processing, a code was written using the Python programming language, which read the photo using the OpenCV library and displayed the image (information) on it in an Excel file with a certain number of lines and columns, in each cell a number between 0-255 is recorded, which represents the intensity of the corresponding point of the photo image, i.e. pixel intensity value (see image below).



In order to make a decision about the health status of the patient, the data given under the name Chest X-Ray Images (Pneumonia) of patients suffering from pneumonia were processed using a convolutional neural network (CNN) method (Le Cun *et al.*, 1989; Le Cun *et al.*, 1998), which is realized in the programing language Python, under the framework Pytorch. The convolutional neural network is a class of artificial neural networks most commonly applied to analyze visual imagery. CNNs use a mathematical operation called convolution in place of general matrix multiplication in at least one of their layers (Goodfellow et al., 2016). They are specifically designed to process pixel data and are used in image recognition and processing.

For learning the network model, the data was distributed as follows: 3900 photos were used for model training, 300 photos for model validation, and 300 photos for model testing. The dimensions of all the photographs used in the practically acceptable time period to make a decision about the final result or the patient's condition were reduced to the standard size $384\,pixel \times 384\,pixel$. Processing such data (training, validation, testing) took about 25 minutes. 93.91% accuracy was achieved on the validation data, and 95% accuracy on the test data. Using the mentioned method, the processing of photo images with initial sizes (the sizes of which, in the case of pneumonia, are significantly larger than $1000\,pixel \times 1000\,pixel$ for both healthy and diseased patients) is practically impossible without powerful computing resources, because network models take up a lot of space in the computer's RAM, and when a high-dimensional photo is added to it, the problem becomes even worse.

As a rule, the processing of photos larger than $800\,pixel \times 800\,pixel$ dimensions requires quite powerful computing resources, in the absence of which, models of small dimensions are used, which greatly reduces the accuracy of the obtained results.

It is clear that the level of 95% accuracy of the diagnosis is unacceptable for modern medicine. Therefore, in order to increase the reliability of the diagnosis, as well as to develop a simple, fast method that can be implemented in modern, computerized X-ray equipment, it was decided to use statistical hypothesis testing methods for decision-making, which allow to simultaneously restrict both types of possible errors when making a decision. Such methods are Wald's sequential analysis method and CBM, the essence of which is briefly described below.

In order to use the mentioned methods, it is necessary to define a vector representing the state of the objects under investigation (about the state of which a decision must be made), which takes different values depending on the state of the object under investigation. In our case, such a vector turned out to be the dimensions of the patient's photo image represented in pixels, which are equal to the number of rows and columns of the corresponding Excel files. It was found that they take different values for healthy and sick patients and vary randomly from patient to patient.

We demonstrate below the results obtained by processing of the data of pneumonia, adenocarcinoma, carcinoma, and combined data of both types of cancer with the help of statistical package SPSS, as well as the results of statistical processing of lung examination data of healthy patients, obtained by CT Scan method, on the example of pneumonia data.

Results of statistical processing of data showing lung pneumonia.

Descriptive statistics results.

Statistics

	x_nor_R	x_nor_L	x_pne_R	x_pne_L
N Valid	130	130	130	130
Missing	0	0	0	0
Mean	1811.1923	1412.9308	1144.6000	788.1538
Median	1786.0000	1318.5000	1114.0000	744.0000
Mode	1753.00[a]	1125.00	943.00[a]	656.00
Std. Deviation	347.63587	380.56199	226.09779	227.45479
Variance	120850.699	144827.429	51120.211	51735.682
Skewness	.395	.546	.941	1.330
Std. Error of Skewness	.212	.212	.212	.212
Kurtosis	.288	-.220	.783	2.306
Std. Error of Kurtosis	.422	.422	.422	.422

Note 1. The first two columns contain the results of processing the quantities of columns (x_nor_R) and rows (x_nor_L) of the Excel files of healthy patients, and the next two rows contain the same data for patients with pneumonia. The same type of designations are used for other diseases.

The results of correlation analysis.

Correlations

		x_nor_R	x_nor_L	x_pne_R	x_pne_L
x_nor_R	Pearson Correlation	1	.866**	.051	.067
	Sig. (2-tailed)		.000	.563	.448
	N	130	130	130	130
x_nor_L	Pearson Correlation	.866**	1	.020	.035
	Sig. (2-tailed)	.000		.825	.695
	N	130	130	130	130
x_pne_R	Pearson Correlation	.051	.020	1	.891**
	Sig. (2-tailed)	.563	.825		.000
	N	130	130	130	130
x_pne_L	Pearson Correlation	.067	.035	.891**	1
	Sig. (2-tailed)	.448	.695	.000	
	N	130	130	130	130

**. Correlation is significant at the 0.01 level (2-tailed).

The results of testing the normality of the observation results with the χ^2-criterion.

Test Statistics

	x_nor_R	x_nor_L	x_pne_R	x_pne_L
Chi-Square	13.415	9.615	46.231	46.923
df	117	120	78	99
Asymp. Sig.	1.000	1.000	.998	1.000

Based on the results of the research, we conclude that the vector representing the patient's condition is distributed according to the two-dimensional normal distribution law with different values of the parameters of healthy and sick patients. From the results of statistical processing of cancer data, it is clear that the same conclusion is correct for other cases discussed in the article.

Based on these results, we conclude that for each of these diseases, the two-dimensional random vector corresponding to the dimensions of the patient's state photo is a normally distributed vector with correlated parameters. In particular, in the case of pneumonia, the parameters of the normal distribution for healthy patients, the vector of mathematical expectation and the covariance matrix, are given as

$$\mu = (\mu_1, \mu_2) = (1811.1923, 1412.9308),\ W = \begin{pmatrix} w_{11}, w_{12} \\ w_{21}, w_{22} \end{pmatrix} = \begin{pmatrix} 120850.699, 114569.20 \\ 114569.20, 144827.429 \end{pmatrix},$$

and for patients suffering from pneumonia –

$$\mu = (\mu_1, \mu_2) = (1144.60, 788.1538),\ W = \begin{pmatrix} w_{11}, w_{12} \\ w_{21}, w_{22} \end{pmatrix} = \begin{pmatrix} 51120.211, 45821.4796 \\ 45821.4796, 51735.682 \end{pmatrix}.$$

For patients with carcinoma, we have:

$$\mu = (407.4846, 269.6385),\ W = \begin{pmatrix} 667.911, 384.875 \\ 384.875, 1673.861 \end{pmatrix}.$$

For the combined data of both types of cancer, we have:

$$\mu = (401.0615, 265.1538),\ W = \begin{pmatrix} 818.382, 171.3007 \\ 171.3007, 1349.544 \end{pmatrix}.$$

For healthy patients examined by computer tomography method, we have:

$$\mu = (632.6, 476.2857),\ W = \begin{pmatrix} 33007.718, 18500.7969 \\ 18500.7969, 14801.798 \end{pmatrix}.$$

The number of observation results used for computations for each case are given in the corresponding tables of descriptive statistics results (see, for example, the results of pneumonia).

3. The methods used for making a decision

On the basis of the investigation results given in the previous paragraph, the problem of making a decision about the condition of a patient can be formulated as follows. On the basis of the observed values of a random vector $\xi = (\xi_1, \xi_2) \sim N(\mu, W)$, where μ is the vector of mathematical expectation and W is the covariance matrix, must be tested basic hypothesis $H_0 : \mu = \mu_0$, $W = W_0$ vs. alternative one $H_1 : \mu = \mu_A$, $W = W_A$. Here μ_0 and W_0 correspond to the supposition that a patient is healthy while μ_A and W_A correspond to a diseased patient.

Let us consider the set of sequentially obtained i.i.d. observation results $x_1, x_2, ..., x_n, ...$ of a patient concerning of which a decision must be made. A decision must be made in such a way that the probabilities of incorrectly rejected or incorrectly accepted hypotheses, i.e. the Type I and Type II error rates were restricted on the desired levels. For this purpose, let us consider the Wald's test and the method of sequential analysis of Bayesian type (MSABT) (Wald, 1947a,b; Kachiashvili and Hashmi, 2010; Kachiashvili, 2018).

3.1. The Wald's test

The essence of the Wald's sequential test consists in the following (Wald, 1947a,b): compute the likelihood ratio

$$B(x) = p(x_1, x_2, \ldots, x_n \mid H_0) / p(x_1, x_2, \ldots, x_n \mid H_A)$$

for n sequentially obtained observation results, and, if

$$B < B(x) < A,$$

do not make a decision and continue the observation of the random variable. If

$$B(x) \geq A,$$

accept the hypothesis H_0 on the basis of n observation results. If

$$B(x) \leq B,$$

accept the hypothesis H_A on the basis of n observation results.

The thresholds A and B are chosen so that

$$A = \frac{1-\beta}{\alpha} \quad \text{and} \quad B = \frac{\beta}{1-\alpha}.$$

Here α and β are the desirable values of the error probabilities of Types I and II, respectively.

It is proved (Wald, 1947a,b) that in this case the real values of the error probabilities of Types I and II are close enough to the desired values, but still are distinguished from them.

3.2. The method of sequential analysis of Bayesian type

Let us consider a set of hypotheses H_i, $i=1,...,S$ $(S \geq 2)$, involving that the random vector X is distributed by the law $p(x,\theta_i)$, i.e. $H_i : X \sim p(x,\theta_i) \equiv p(x \mid H_i)$; $p(H_i)$ is a priori probability of hypothesis H_i; Γ_i is the region of acceptance of H_i (Γ_i belongs to the observation space of random variable X, i.e. $\Gamma_i \in R^n$, where n is the dimension of the observation vector). The decision is made on the basis of $\mathbf{x}^T = (x_1,...,x_n)$, the measured value of the random vector $\mathbf{X}$. It is possible to formulate different constrained tasks of testing the considered hypotheses (Kachiashvili et al., 2012b; Kachiashvili, 2018). Here we consider only one of them, namely the task with restrictions on the averaged probability of rejection of true hypotheses for stepwise loss function with two possible values 0 and 1.

The essence of this method is the minimization of the averaged probability of incorrect acceptance of hypotheses at restriction of the averaged probability of rejection of true hypotheses, i.e.

$$\min_{\{\Gamma_i\}} \left\{1 - \sum_{i=1}^{S} p(H_i) P(\mathbf{X} \in \Gamma_i \mid H_i)\right\}, \qquad (1)$$

subject to

$$p(H_i) \sum_{j=1, j\neq i}^{S} P(\mathbf{X} \in \Gamma_j \mid H_i) \leq \gamma_i, \; i = 1,\ldots,S. \qquad (2)$$

Solution of task (1) and (2) is (Kachiashvili, 2011; Kachiashvili et al., 2012b)

$$\Gamma_j = \{\mathbf{x} : \lambda_j p(H_j) p(\mathbf{x} \mid H_j) > \sum_{i=1, i\neq j}^{S} p(H_i) p(\mathbf{x} \mid H_i)\}, \; j = 1,\ldots,S. \qquad (3)$$

Coefficient λ_j are determined so that in (2) the equality takes place.

The sequential test developed on the basis of CBM consists in the following (Kachiashvili & Hashmi, 2010; Kachiashvili, 2013, 2018). Let Γ_i^n be the H_i hypothesis acceptance region (3) on the basis of n sequentially obtained repeated observation results; R_n^m is the decision-making space in the sequential method; m is the dimensionality of the observation vector; I_i^n is the population of sub-regions of intersections of hypotheses H_i acceptance regions Γ_i^n $(i=1,...,S)$ with the regions of acceptance of other hypotheses H_j, $j=1,...,S$, $j \neq i$; $E_n^m = R_n^m - \bigcup_{i=1}^{S} \Gamma_i^n$ is the population of regions of space R_n^m which do not belong to any of hypotheses acceptance regions.

The H_i hypotheses acceptance regions for n sequentially obtained observation results in the sequential method are:

$$R_{n,i}^{m} = \Gamma_i^n / I_i^n,\ i = 1,\ldots,S; \tag{4}$$

the no-decision region is:

$$R_{n,S+1}^{m} = \left(\bigcup_{i=1}^{S} I_i^n\right) \bigcup E_n^m, \tag{5}$$

where Γ_i^n, $i = 1,\ldots,n$, are defined by (3) on the basis of n sequentially obtained observation results.

This test is called *the sequential test of Bayesian type* (Kachiashvili & Hashmi, 2010). Such tests could be considered for all constrained Bayesian methods offered in (Kachiashvili et al., 2012a, b; Kachiashvili, 2018) and differing from each other in restrictions.

When the number of hypotheses is equal to two and their a priori probabilities are equal to $1/2$, solution (3) can be rewritten using the Bayes factor:

$$\Gamma_0 : \frac{p(x \mid H_0)}{p(x \mid H_A)} > \frac{1}{\lambda_0}, \; \Gamma_A : \frac{p(x \mid H_0)}{p(x \mid H_A)} < \lambda_A,$$

where λ_0 and λ_A are determined so that in the conditions (2) equalities take place.

It is worth noting the shortcoming of Wald's method: 1) it is optimal for normal distribution in the limit case when $n \to \infty$; 2) it is developed for the case of two hypotheses; 3) its generalization for more than two hypotheses is quite problematic. Although this is done using the Bayes approach, its practical implementation is very difficult.

CBM is free from all these drawbacks. It works for hypotheses of any number and dimension (both continuous and discrete distributions), and the complexity of its implementation practically does not changes.

4. Investigation of the used methods by simulation and real data

By computations on the basis of simulated and real data, we did the diagnoses of lung pneumonia, adenocarcinoma and carcinoma by sequential methods of testing the hypotheses described above. The results will be shown later.

Now let's note that the Bayesian decision-making method for all these diseases used the same values of Lagrange multipliers, which are given in Table 1 and calculated for the pneumonia data for different levels of type I and type II errors.

Table 1. Dependence of Lagrange multipliers on Type 1 and Type II error rates

		CBM			Wald's test	
Type 1 error rate	Type II error rate	Lagrange multipliers			Lagrange multipliers	
α	β	λ_0	$1/\lambda_0$	λ_A	A	B
0.05	0.05	3.4429	0.2904528	3.1893	19	0.052631578947368
0.01	0.01	22.6699	0.0789272	15.1790	99	0.010101010101010
0.001	0.001	65.9375	0.01516587	53.7500	999	0.001001001001001
0.0001	0.0001	355.9375	0.0028095	316.1787	9999	1.000100010001000e-04
0.00001	0.00001	2555.9375	0.00039124587	2416.1787	99,999	1.000010000100001e-05

In this case, the Kullback-Leibler divergence between the hypotheses to be tested is minimal compared to the cancer diseases discussed in the report (see Table 2). The Kullback–Leibler divergence between two multivariate Gaussian distributions is (Kullback, 1978)

$$D_{KL}(p \parallel q) = \frac{1}{2}\left[\log \frac{|W_q|}{|W_p|} - n + (\mu_p - \mu_q)^T W_q^{-1} (\mu_p - \mu_q) + tr\{W_q^{-1} W_p\} \right].$$

Table 2. The Kullback–Leibler divergence between tested hypotheses at the consideration of pneumonia, adenocarcinoma, carcinoma and adenocarcinoma plus carcinoma data.

Type of the disease	The Kullback–Leibler divergence
Pneumonia	2.241213728077075
Adenocarcinoma	3.339872625127654
Carcinoma	3.824174027990074
Adenocarcinoma plus carcinoma	3.187334583538381

It has been proved and shown in papers (Kachiashvili, 2014b, 2015, 2018) that the Lagrange multipliers in CBM that are calculated to ensure decision making with given reliability for hypotheses with minimum Kullback-Leibler distance, ensure making correct decision with higher reliability when the Kullback-Leibler distance between hypotheses increases.

By computations on the basis of simulated and real data, we did the diagnosis of lung pneumonia, adenocarcinoma and carcinoma by sequential methods of testing the hypotheses described above. The Results of diagnosis based on lung pneumonia simulated and real data are given in Appendix below.

Appendix. Results obtained by simulation

Table A1. Decisions made on the basis of simulated data.

Number of experiments	Type 1 error rate	Type II error rate	CBM						Wald's test					
			Average number of observations necessary for making a decision when H_0 is true	Probability of acceptance of basic hypothesis when it is true	Probability of acceptance of alternative hypothesis when basic hypothesis is true	Average number of observations necessary for making a decision when H_A is true	Probability of acceptance of basic hypothesis when alternative hypothesis is true	Probability of acceptance of alternative hypothesis when it is true	Average number of observations necessary for making a decision when H_0 is true	Probability of acceptance of basic hypothesis when it is true	Probability of acceptance of alternative hypothesis when basic hypothesis is true	Average number of observations necessary for making a decision when H_A is true	Probability of acceptance of basic hypothesis when alternative hypothesis is true	Probability of acceptance of alternative hypothesis when it is true
m	α	β	AN	$P(x \in \Gamma_0 \mid H_0)$	$P(x \in \Gamma_A \mid H_0)$	AN	$P(x \in \Gamma_0 \mid H_A)$	$P(x \in \Gamma_A \mid H_A)$	AN	$P(x \in \Gamma_0 \mid H_0)$	$P(x \in \Gamma_A \mid H_0)$	AN	$P(x \in \Gamma_0 \mid H_A)$	$P(x \in \Gamma_A \mid H_A)$
200,000	0.05	0.05	Ex. 1: 1.2605	0.9584	0.0416	1.3427	0.0131	0.9869	1.4785	0.9654	0.0346	1.4468	0.0022	0.9978
			Ex. 2: 1.2617	0.9580	0.0420	1.3454	0.0134	0.9866	1.4770	0.9663	0.0337	1.4480	0.0021	0.9979
			Ex. 3: 1.2620	0.9580	0.0420	1.3430	0.0131	0.9869	1.4774	0.9661	0.0339	1.4476	0.0021	0.9979
			Ex. 4: 1.2608	0.9582	0.0418	1.3424	0.0136	0.9864	1.4775	0.9664	0.0336	1.4487	0.0021	0.9979
			Ex. 5: 1.2611	0.9579	0.0421	1.3431	0.0130	0.9870	1.4796	0.9658	0.0342	1.4492	0.0021	0.9979
200,000	0.01	0.01	Ex. 1: 1.3519	0.9904	0.0096	1.8512	0.0084	0.9916	1.7394	0.9963	0.0037	2.2364	0.0016	0.9984
			Ex. 2: 1.3498	0.9903	0.0097	1.8459	0.0085	0.9915	1.7426	0.9962	0.0038	2.2361	0.0016	0.9984
			Ex. 3: 1.3515	0.9907	0.0093	1.8474	0.0081	0.9919	1.7410	0.9962	0.0038	2.2361	0.0017	0.9983
			Ex. 4: 1.3492	0.9903	0.0097	1.8477	0.0082	0.9918	1.7439	0.9962	0.0038	2.2345	0.0016	0.9984
			Ex. 5: 1.3496	0.9906	0.0094	1.8493	0.0079	0.9921	1.7366	0.9963	0.00375	2.2355	0.0017	0.9983
200,000	0.001	0.001	Ex. 1: 1.6537	0.9992	7.70e-04	2.6844	6.50e-04	0.9993	2.1061	0.9997	2.6500e-04	3.3767	1.1000e-04	0.99989
			Ex. 2: 1.6595	0.9991	9.00e-04	2.6843	7.20e-04	0.9993	2.1021	0.9998	2.400e-04	3.3728	1.5500e-04	0.99984
			Ex. 3: 1.6591	0.9990	0.0010	2.6884	6.00e-04	0.9994	2.1042	0.9998	2.0000e-04	3.3757	1.2500e-04	0.99988
			Ex. 4: 1.6576	0.9991	9.20e-04	2.6850	7.00e-04	0.9993	2.1008	0.9998	2.3500e-04	3.3713	1.4000e-04	0.99986
			Ex. 5: 1.6638	0.9992	8.00e-04	2.6840	6.50e-04	0.9993	2.1049	0.9998	2.3000e-04	3.3739	1.3500e-04	0.99987
200,000	0.0001	0.0001	Ex. 1: 1.9191	0.9999	1.10e-04	3.5608	1.10e-04	0.9999	2.4877	0.99999	1.00e-05	4.5829	1.0000e-05	0.99999
			Ex. 2: 1.9222	0.9999	9.00e-05	3.5623	8.00e-05	0.9999	2.4846	1	0	4.5818	0	1
			Ex. 3: 1.9199	0.9999	1.40e-04	3.5600	6.00e-05	0.9999	2.4872	0.999995	5.0000e-06	4.5831	1.0000e-05	0.99999
			Ex. 4: 1.9201	0.9999	1.00e-04	3.5687	9.00e-05	0.9999	2.4886	1	0	4.5820	0	1
			Ex. 5: 1.9214	0.9999	6.00e-05	3.5638	1.10e-04	0.9999	2.4896	0.99999	1.0000e-05	4.5825	1.0000e-05	0.99999
200,000	0.0000 1	0.00001	Ex. 1: 2.2502	0.999995	5.00e-06	4.5942	1.00e-05	0.999990	2.8869	1	0	5.8102	0	1
			Ex. 2: 2.2437	0.999995	5.00e-06	4.5940	1.00e-05	0.999990	2.8848	1	0	5.8081	0	1
			Ex. 3: 2.2473	0.999995	5.00e-06	4.5934	5.00e-06	0.999995	2.8858	1	0	5.8107	0	1
			Ex. 4: 2.2467	0.999995	5.00e-06	4.5933	0	1	2.8904	1	0	5.8099	0	1
			Ex. 5: 2.2420	0.999995	5.00e-06	4.5935	5.00e-06	0.999995	2.8912	1	0	5.8108	0	1

Table A2. Percentage distribution of the number of observations necessary for making a decision.

Type 1 error rate	Type II error rate	Experiment	Number of observations necessary for making a decision	CBM		Wald's test	
				Hypothesis H_0 is true	Hypothesis H_A is true	Hypothesis H_0 is true	Hypothesis H_A is true
α	β	Ex.	NO	Percentage %	Percentage %	Percentage %	Percentage %
0.05	0.05	Ex. 1	1	76.2888	67.1152	59.8984	57.9160
			2	21.4662	31.5022	32.9562	39.4988
			3	2.1546	1.3806	6.5646	2.5774
			4	0.0870	0.0020	0.5566	0.0078
			5	0.0034	0	0.0238	0
		Ex. 2	1	76.1834	66.9046	60.0284	57.8868
			2	21.5562	31.6580	32.8380	39.4326
			3	2.1666	1.4350	6.5564	2.6736
			4	0.0920	0.0024	0.5602	0.0070
			5	0.0018	0	0.0162	0
		Ex. 3	1	76.1404	67.1030	59.9716	57.8458
			2	21.6138	31.4998	32.9088	39.5508
			3	2.1548	1.3944	6.5470	2.5970
			4	0.0894	0.0028	0.5508	0.0064
			5	0.0016	0	0.0216	0
		Ex. 4	1	76.2454	67.1560	59.9638	57.8040
			2	21.5288	31.4466	32.9088	39.5332
			3	2.1258	1.3952	6.5572	2.6534
			4	0.0986	0.0022	0.5478	0.0094
			5	0.0014	0	0.0216	0
		Ex. 5	1	76.2558	67.0714	59.8239	57.6812
			2	21.4720	31.5508	32.9920	39.7250
			3	2.1832	1.3758	6.6060	2.5876
			4	0.0872	0.0020	0.5584	0.0062
			5	0.0018	0	0.0196	0
0.01	0.01	Ex. 1	1	69.0630	27.8360	43.9655	7.0510
			2	26.9850	59.2880	40.3740	63.3785
			3	3.7480	12.6690	13.5620	28.4515
			4	0.2010	0.2070	1.9615	1.1175
			5	0.0030	0	0.1330	0.0015
		Ex. 2	1	69.1030	28.1320	43.8955	7.0540
			2	27.0670	59.2870	40.2880	63.4360
			3	3.6070	12.4180	13.6320	28.3600
			4	0.2160	0.1630	2.0360	1.1485
			5	0.0070	0	0.1395	0.0015
		Ex. 3	1	69.1120	28.0530	43.9115	7.0985
			2	26.9810	59.2990	40.3990	63.3100
			3	3.6950	12.4600	13.5330	28.4715
			4	0.2100	0.1880	2.0015	1.11850
			5	0.0020	0	0.1485	0.0015
		Ex. 4	1	69.1440	28.0120	43.6445	7.1140

			2	26.9480	59.3430	40.5495	63.4545
			3	3.6850	12.4750	13.7065	28.2980
			4	0.2200	0.1690	1.9705	1.1305
			5	0.0030	0.0010	0.1245	0.0030
		Ex. 5	1	69.3990	27.8880	44.1720	7.1500
			2	26.6880	59.2910	40.2810	63.3055
			3	3.6840	12.6550	13.4040	28.3870
			4	0.2240	0.1660	2.0030	1.1555
			5	0.0050	0	0.1345	0.0020
0.001	0.001	Ex. 1	1	48.6730	0.0650	27.8145	0.0110
			2	38.8260	39.1860	41.5780	5.5910
			3	11.0440	53.0480	23.6305	53.6745
			4	1.3780	7.6480	6.1760	38.1665
			5	0.0750	0.0530	0.7625	2.5510
			6	0	0	0.0375	0.0060
			7	0	0	0.0010	0
		Ex. 2	1	48.3950	0.0710	27.9110	0.0155
			2	38.8380	39.2330	41.6210	5.6215
			3	11.2930	52.9510	23.6300	53.9015
			4	1.3730	7.6880	6.0650	37.9960
			5	0.1000	0.0570	0.7295	2.4580
			6	0	0	0.0410	0.0075
			7	0	0	0.0025	0
		Ex. 3	1	48.4890	0.0600	27.7740	0.0125
			2	38.7670	39.0110	41.8145	5.4970
			3	11.1870	53.0340	23.4910	53.9355
			4	1.4610	7.8200	6.1035	38.0210
			5	0.0900	0.0750	0.7725	2.5265
			6	0	0	0.0430	0.0075
			7	0	0	0.00150	0
		Ex. 4	1	48.4610	0.0700	28.0015	0.0140
			2	38.8850	39.1180	41.6595	5.5635
			3	11.1650	53.1130	23.4555	54.2050
			4	1.4150	7.6410	6.0725	37.7220
			5	0.0720	0.0580	0.7610	2.4885
				0	0	0.0485	0.0070
				0	0	0.0015	0
		Ex. 5	1	48.2900	0.0650	27.9170	0.0135
			2	38.6420	39.2520	41.5440	5.5660
			3	11.5550	52.9550	23.5540	53.8990
			4	1.4310	7.6710	6.1575	38.0640
			5	0.0790	0.0570	0.7740	2.4510
				0	0	0.0530	0.0065
				0	0	0.0005	0
0.0001	0.0001	Ex. 1	1	35.3750	0.0110	16.6810	0.0010
			2	41.8840	1.9950	36.6105	0
			3	18.5910	45.0490	31.1120	2.7285
			4	3.7780	47.8190	12.7700	42.2245
			5	0.3550	5.0980	2.5255	49.1065
		Ex. 2	1	35.3730	0.0080	16.6945	0

			2	41.6180	2.0050	36.7585	0
			3	18.8440	45.1030	31.1295	2.7150
			4	3.7690	47.5550	12.5410	42.2850
			5	0.3770	5.2930	2.5760	49.1370
		Ex. 3	1	35.3430	0.0060	16.6235	0.0010
			2	41.8530	2.0040	36.8840	0
			3	18.6890	45.1460	30.9165	2.6455
			4	3.7290	47.6980	12.6085	42.3175
			5	0.3640	5.1200	2.67100	49.152
		Ex. 4	1	35.1920	0.0090	16.6825	0
			2	41.9930	1.9280	36.6175	0
			3	18.8310	44.5900	31.0655	2.7125
			4	3.6030	48.1510	12.7480	42.2360
			5	0.3620	5.2970	2.5815	49.2250
		Ex. 5	1	35.1230	0.0110	16.5670	0.0010
			2	42.1050	1.9250	36.7045	0
			3	18.6840	44.9790	31.1520	2.7180
			4	3.7010	47.8630	12.6705	42.2545
			5	0.3690	5.2020	2.6075	49.1130
0.00001	0.00001	Ex. 1	1	23.0970	0.0010	9.3650	0
			2	40.3080	0	28.8705	0
			3	26.7710	2.5770	34.1725	0
			4	8.3960	41.5835	20.1370	1.0840
			5	1.3300	49.7085	6.3065	29.3580
			6	0	6.0975	0	57.1675
			7	0	0.0325	0	12.2355
			8	0	0		0.1550
		Ex. 2	1	23.4670	0.0010	9.3355	0
			2	40.1140	0	28.9360	0
			3	26.8110	2.5575	34.2535	0
			4	8.1640	41.5620	20.1430	1.0775
			5	1.3320	49.8220	6.1550	29.4505
			6	0	6.0335	0	57.1860
			7	0	0.0240	0	12.1550
			8	0	0	0	0.1310
		Ex. 3	1	23.0740	0.0005	9.3745	0
			2	40.2450	0	28.9090	0
			3	26.7940	2.5950	34.1910	0
			4	8.4640	41.6630	20.0915	1.0655
			5	1.3190	49.5865	6.2690	29.4165
			6	0	6.1160	0	57.0550
			7	0	0.0390	0	12.3120
			8	0	0	0	0.1510
		Ex. 4	1	23.1030	0	9.3080	0
			2	40.2300	0	28.9475	0
			3	26.9340	2.5540	33.9995	0
			4	8.2460	41.6795	20.1555	1.1030
			5	1.3040	49.6760	6.4250	29.3280
			6	0	6.0660	0	57.1985

			7 8	0 0	0.0245 0	0 0	12.2210 0.1495
		Ex. 5	1 2 3 4 5 6 7 8	23.3780 40.1940 26.7720 8.2890 1.2720 0 0 0	0.0005 0 2.5655 41.6155 49.7470 6.0410 0.0305 0	9.3735 28.7005 34.1695 20.2255 6.3540 0 0 0	0 0 0 1.0775 29.3465 57.1345 12.2970 0.1445

Table A3. Decisions made on the basis of real data of sick patients.

Number of patients	Type I error rate	Type II error rate	CBM				Wald			
			Average number of observations necessary for making a decision when H_A is true	Probability of acceptance of basic hypothesis when alternative hypothesis is true	Probability of acceptance of alternative hypothesis when it is true	Number of made decisions	Average number of observations necessary for making a decision when H_A is true	Probability of acceptance of basic hypothesis when alternative hypothesis is true	Probability of acceptance of alternative hypothesis when it is true	Number of made decisions
m	α	β	AN	$P(x \in \Gamma_0 \mid H_A)$	$P(x \in \Gamma_A \mid H_A)$	N	AN	$P(x \in \Gamma_0 \mid H_A)$	$P(x \in \Gamma_A \mid H_A)$	N
100	0.05	0.05	1	0	1	100	1	0	1	100
	0.01	0.01	1	0	1	100	2	0	1	50
	0.001	0.001	2	0	1	50	3	0	1	33
	0.0001	0.0001	3	0	1	33	3	0	1	33
	0.00001	0.00001	3	0	1	33	4	0	1	25
214	0.05	0.05	1	0	1	214	1	0	1	214
	0.01	0.01	1	0	1	214	2	0	1	107
	0.001	0.001	2	0	1	107	3	0	1	71
	0.0001	0.0001	3	0	1	71	3	0	1	71
	0.00001	0.00001	3	0	1	71	4	0	1	54
214 (the sequence of observations is changed)	0.05	0.05	1	0	1	214	1	0	1	214
	0.01	0.01	1	0	1	214	2	0	1	107
	0.001	0.001	2	0	1	107	3	0	1	71
	0.0001	0.0001	3	0	1	71	3	0	1	71
	0.00001	0.00001	3	0	1	71	4	0	1	53

Table A4. Decisions made on the basis of real data of healthy patients.

Number of patients	Type I error rate	Type II error rate	CBM				Wald			
			Average number of observations necessary for making a decision when H_A is true	Probability of acceptance of basic hypothesis when alternative hypothesis is true	Probability of acceptance of alternative hypothesis when it is true	Number of made decisions	Average number of observations necessary for making a decision when H_A is true	Probability of acceptance of basic hypothesis when alternative hypothesis is true	Probability of acceptance of alternative hypothesis when it is true	Number of made decisions
m	α	β	AN	$P(x \in \Gamma_0 \mid H_0)$	$P(x \in \Gamma_A \mid H_0)$		AN	$P(x \in \Gamma_0 \mid H_0)$	$P(x \in \Gamma_A \mid H_0)$	
90	0.05	0.05	1	1	0	90	1	1	0	90
	0.01	0.01	1	1	0	90	1	1	0	90
	0.001	0.001	1	1	0	90	2	1	0	45
	0.0001	0.0001	1	1	0	90	2	1	0	45
	0.00001	0.00001	2	1	0	45	2	1	0	45
90 (the sequence of observations is changed)	0.05	0.05	2	1	0	45	2	1	0	45
	0.01	0.01	2	1	0	45	2	1	0	45
	0.001	0.001	2	1	0	45	2	1	0	45
	0.0001	0.0001	2	1	0	45	2	1	0	45
	0.00001	0.00001	2	1	0	45	2	1	0	45

The first table show the results obtained with 200,000 data points generated by the distribution parameters corresponding to the observations of healthy and diseased patients given above. The calculation results show that the reliability of diagnosis for healthy and sick patients for each considered case, that is, for each considered restriction of the first and second type error levels, is satisfied both by the Wald criterion and for the decisions made by the MSABT. However, Wald's criterion requires a larger number of observations for each considered case (see Table A2). Tables A3 and A4 of the same appendices, respectively, show the results of decisions made with real data of sick and healthy patients, which completely match the results obtained by modeling and assure us that with the proposed methods it is possible to automatically diagnose the considered diseases with a predetermined reliability.

It is especially important to emphasize the fact that both sequential analysis methods (of Wald and CBM) require practically negligible time (less than one second) and memory for their implementation in modern computerized X-ray equipment (in contrast to the methods based on modern neural networks mentioned above), which allows their widespread implementation in serial equipment.

Conclusion

A method of automatic diagnosis of pneumonia and lung cancer with computerized X-ray equipment is proposed, which requires very little memory and time to make a decision. At the same time, both types of possible errors can be limited to predetermined levels with guarantee. The method is based on the method of sequential analysis of Bayesian type of statistical hypothesis testing. The results of the experimental investigation, both on modeled and real data, showed the ease of implementation, high reliability and accuracy of the proposed method of automatic diagnosis. In our opinion, the implementation of the mentioned method in serially produced relevant equipment will significantly increase the quality of the diagnosis, which in turn will play a decisive role in the final recovery of the patient.

Acknowledgement: Authors were partially supported by European Commission HORIZON EUROPE WIDERA-2021-ACCESS-03 Grant Project GAIN (grant agreement no.101078950)

References

M. Bishop (2006) *Pattern Recognition and Machine Learning*. Springer Verlag.

Goodfellow I., Bengio Y. and Courville A. (2016). *Deep Learning*. MIT Press, 326.

Kachiashvili K.J. (2013) The Methods of Sequential Analysis of Bayesian Type for the Multiple Testing Problem. *Sequential Analysis*, 32: 1-16.

Kachiashvili K.J. (2014a) Comparison of Some Methods of Testing Statistical Hypotheses. Part II. Sequential Methods. *International Journal of Statistics in Medical Research*, 3, 189-197.

Kachiashvili K. J. (2014b). The Methods of Sequential Analysis of Bayesian Type for the Multiple Testing Problem, *Sequential Analysis* 33: 23–38

Kachiashvili K.J. (2015) Constrained Bayesian Method for Testing Multiple Hypotheses in Sequential Experiments. *Sequential Analysis*: *Design Methods and Applications*, 34(2): 171-186.

Kachiashvili K.J. (2018) *Constrained Bayesian Methods of Hypotheses Testing: A New Philosophy of Hypotheses Testing in Parallel and Sequential Experiments*. Nova Science Publishers, Inc., New York, 456 p.

Kachiashvili, K.J. & Hashmi, M.A. (2010). About Using Sequential Analysis Approach for Testing Many Hypotheses. *Bulletin of the Georgian Academy of Sciences*, 4, 2, 20-25.

Kachiashvili K.J., Hashmi M. A. and Mueed A. (2012a) Sensitivity Analysis of Classical and Conditional Bayesian Problems of Many Hypotheses Testing. *Communications in Statistics—Theory and Methods*, 41(4): 591–605.

Kachiashvili K.J., Hashmi M. A. and Mueed A. (2012b) The Statistical Risk Analysis as the Basis of the Sustainable Development. *Int. J. of Innovation and Technol. Management* (World Scientific Publishing Company), 9(3): 1-10. DOI: 10.1142/S0219877012500241

Kullback S. (1978) *Information Theory and Statistics*. Wiley & Sons, Inc.

Le Cun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1(4), 541–551.

Le Cun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 2278–2324.

Tanaka S., Inoue M., Yamaji T., et al. (2023) Increased risk of death from pneumonia among cancer survivors: A propensity score-matched cohort analysis. *Cancer Med.*, 12:6689-6699. doi: 10.1002/cam4.5456

Wald, A. (1947a). *Sequential analysis*. New-York: Wiley.

Wald, A. (1947b). Foundations of a General Theory of Sequential Decision Functions. *Econometrica*, 15, 279-313.

Thank you
for your attention