Institute of Information and Computational Technologies

ISSN : 2788-7677 (Online)
ISSN : 2788-7987 (Print)

# ADVANCED TECHNOLOGIES AND COMPUTER SCIENCE

2020
№3

www.atcs.iict.kz

Institute of Information and Computational Technologies,

# Advanced Technologies and computer science

## №3

Almaty 2020

Institute of Information and Computational Technologies,

Advanced Technologies and computer science



28 Shevchenko str., Almaty,Republic of Kazakhstan
7 (727) 272-37-11
atcs@iict.kz

**About the Journal**

Advance technologies and computer science is a bilingual scientific peer-reviewed, interdisciplinary, electronic journal of open access, including thematic areas:

- Section **"Applied mathematics, computer science and control theory"** includes papers describing modern problems in these areas.
- Section **"Information and telecommunication technologies"** also includes the following topics:
  - Data transmission systems and networks.
  - Internet technologies.
  - Cloud technologies.
  - Parallel computing.
  - Distributed computing.
  - Supercomputer and cluster systems.
  - Big data processing (Big-data).
  - Geographic Information Systems and Technologies.
- In the section **"Artificial intelligence technologies"** in addition to technology, there are works on topics:
  - Intelligent Management Systems.
  - Speech technology and computer linguistics.
  - Pattern Recognition and Image Processing.
  - Bioinformatics and biometric systems.
  - Human-machine interaction.
  - Machine learning.
  - Intelligent Robotic Systems.
- The section **"Information Security and Data Protection"** also covers topics:
  - Software and hardware information protection.
  - Mathematical methods for ensuring information security of complex systems.
- The section **"Modeling and optimization of complex systems and business processes"** may include:
  - Computational mathematics, numerical analysis and programming, mathematical logic.
  - Theory of Statistics.
  - Statistical Methods.

## Editorial Team

## Contents

**IRSTI 50.37.23; 50.41.25**

**UDC 51-76; 004.11**

# BIOMETRIC IDENTIFICATION OF A PERSON USING A MULTI-PARAMETER AUTOMATED SYSTEM

**T. Mazakov[1,2], Erika Lorents (Matsak)[3] M. Aliaskar[1,2], A. Abzhalilova[2], D. Montaeva[2], A. Turlybekova[2]**

[1]RSE Institute of Information and Computational Technologies MES RK CS, Almaty, Kazakhstan

[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan

[3]Tallinn University of Technology IT College, Tallinn, Estonia

[1]tmazakov@mail.ru, [2]87019931011@mail.ru

[1]ORCID ID: https://orcid.org/0000-0001-9345-5167

**Abstract.** The article is devoted to the development of a system for biometric identification of a person by face, fingerprints and voice. Two-dimensional and three-dimensional characteristics of a person's face, taking into account area and volume, were used as informative signs of biometric identification of a person by face. For biometric identification of a person by fingerprints, an FPM10A scanner and an Arduino microcontroller were used.

Another type of signs is local. They are also called minutiae (features or special points) - unique features inherent only in a specific imprint, determining the points of change in the structure of papillary lines (ending, bifurcation, rupture, etc.), the orientation of the papillary lines and coordinates at these points. Each print can contain up to 70 or more minutes. For biometric identification of a person by voice, the MFCC and PLP algorithms are used for digital processing and analysis of audio recordings. Various algorithms are used for acoustic analysis of speech: hidden Markov models, a model of a mixture of Gaussian distributions. The result of determining the tone of speech and the content of speech for the purposes of identification by voice is obtained.

On the Visual FoxPro DBMS, a “Multiparameter automated system for biometric identification of a person” has been developed.

**Keywords:** information security, two-dimensional and three-dimensional image, identification, papillary patterns, voice characteristics, human speech, acoustic modeling.

## Introduction

The problem of information protection and information security is one of the most important aspects of the development of modern society. Currently, the solution to this problem in the development and operation of information systems for various purposes is associated with the development of all sorts of requirements for ensuring their security and the creation of software and hardware from unauthorized access [1-2].

Automatic person recognition for identification has a large number of applications in various fields. Public security problems, the need for remote authentication, the development of human-machine interfaces are causing increased interest in this technology [3].

Methods of biometric identification of a person are increasingly used in access control systems to workplaces, mobile devices, local and global information resources. Since the implementation of the systems does not require specialized equipment, and the biometric feature cannot be lost, forgotten or transferred, the most promising are systems, the principle of which is based on recognition of a person's face.

Authentication methods based on the measurement of human biometric parameters provide 100% identification. At the moment, the following biometric characteristics are successfully used in biometric systems for user authentication: iris, fingerprint, palm print, vascular patterns, face geometry, voice print, signature, DNA comparison, which have properties without which their practical application is impossible [ 4]:

Universality: each person has biometric characteristics.

Uniqueness: there are no two people with exactly the same biometric characteristics.

Consistency: biometric characteristics must be stable over time.

Measurability: Biometric characteristics must be measurable by some kind of physical reading device.

Acceptability is also a very important property. It is least of all associated with any specific biometric parameter, however, without taking it into account, it is impossible to create a complete picture of the effectiveness of the use of biometric systems. The combination of all of the above properties determines the effectiveness of biometric authentication systems.

Currently, there are no biometric parameters that would combine all these properties at the same time, especially when considering acceptability. Therefore, the application of multi-parameter biometric authentication is becoming relevant.

**Implementation. Main part**

For the software implementation of the AS «Biometric Information Security System» was chosen the VisualFoxPro DBMS. Starting with the ninth version of VisualFoxPro, a set of classes GDIPlus and MCI are supplied.

GDIPlus supports both raster (BMP, GIF, PNG, etc.) and vector (WMF, EMF) images. The graphical device interface (GDIPlus) allows the developed applications to use graphics and formatted text for displaying on a monitor screen or printing to a printer [5].

With MCI, you can record, play back audio and video files of various formats.

On the basis of the VisualFoxPro 9 DBMS, the interface part is implemented, which includes the following modes: 1) biological characteristics, 2) characteristics parameters, 3) initial databases, 4) database configuration, 5) simple identification, 6) complex identification, 7) classification [6-8].

After calling the AWP, the main screen of the program appears, shown in Figure 1.

Currently included as biological characteristics are «face video», «fingerprint» and «voice».

Mode - «source databases». Portraits in the following graphic formats can be used as initial data for images: bmp, gif, jpeg, tiff and png. For the «face video» mode, the main information is a volumetric 3d-model, presented as a regular height matrix.

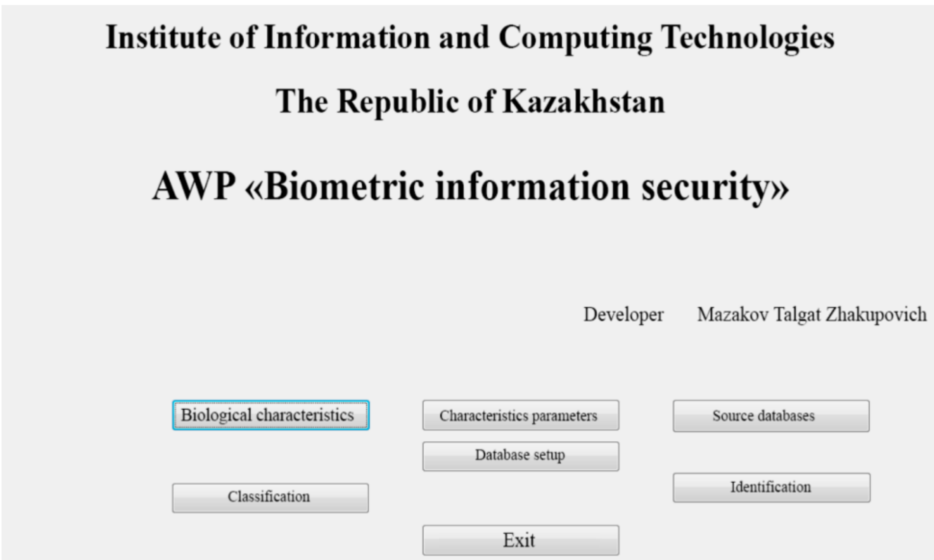


**Figure 1** – AWP main screen

The following types are introduced for characteristics parameters:

1 - point coordinate; 2 - distance (number); 3- area; 4 - volume.

Mode - «classification». The program implements classification according to one or several parameters. In this case, the task of classification is to simplify a data matrix that is too extensive for direct human analysis. There is no single “correct” classification for any dataset. Different numerical strategies usually lead to very different results. Consequently, the help of a numerical consultant is needed to characterize the types of classification available, and it is up to the expert to choose the type that suits him.

Basic algorithm. The initial steps in all agglomerative systems are the same. For n individuals, all $n(n-1)/2$ measures of difference are calculated and the pair of individuals with the smallest measure is combined into one group. It will then be necessary to determine an appropriate measure of difference between this group and the remaining $n-2$ individuals, and in later stages it will obviously be necessary to determine the measure between the individual and the group of any size, as well as between any two groups. At each step of the classification, a union is carried out (between two individuals, between individuals and a group, or between two groups), for which the measure of difference is minimal among all those remaining to this step. The measure must be such that the individual can be regarded as a group of one element. The unification strategy is determined precisely by the measure of differences between the groups. In total, the algorithm calculates $(n-1)^2$ measures. It was shown in [9] that (i, j)-measures can usually be considered from the standpoint of a single linear model. Let there are two groups i and j with ni and nj elements, respectively; the measure of the difference between these groups is denoted by dij. Suppose that dij is the minimum measure of all the remaining ones, so that i and j combine and form a new group k with nk = ni + nj elements. Consider some other group h with nh elements. The values dhi, dhj, dij, nh, ni, and nj are known before joining. Let's put

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma \left| d_{hi} - d_{hj} \right|$$

where parameters $\alpha_i$, $\alpha_j$, $\beta$ and $\gamma$ define the essence of the strategy.

The flexible strategy is applicable for any measure of difference and is determined by four constraints: $\alpha_i + \alpha_j + \beta = 1, \alpha_i = \alpha_j, \beta < 1, \gamma = 0$. The strategy is monotone, and its properties completely depend on $\beta$. If $\beta$ = 0, then the strategy preserves the metric of the space. If $\beta$ is positive, then the strategy shrinks the space, and if negative, it stretches. The program used the value $\beta$ = -10.25 as recommended for practice.

The program implements various classification algorithms, the so-called merging strategies (agglomerative systems): flexible strategy, nearest neighbor strategy, far neighbor strategy, group mean strategy, centroid strategy, sum-of-squares strategy.

Face video

For the characteristic «video image of the face», a number of parameters are defined, which are:

1) point - coordinates of the pupils of the eyes, bridge of the nose, tip of the nose,

2) distance - between the eyes, between the bridge of the nose and the tip of the nose, the base of the nose,

3) perimeter - triangle (pupils of the eyes and tip of the nose), triangle (bridge of the nose and base of the nose),

4) area - isolines of the eye sockets, isolines of the nose,

5) volume - eye sockets, nose.

A volumetric 3d-model, presented as a regular height matrix, is used as the initial data for the «face video».

Algorithms for processing information parameters for identifying a person by face:

1. Coordinates of the center of the pupil of the left eye - (P1x, P1y): determined from a graphic file with a photo of a person;

2. Coordinates of the center of the pupil of the right eye - (P2x, P2y): determined from a graphic file with a photo of a person;

3. The distance between the pupils - P3: calculated using data (P1x, P1y) and (P2x, P2y);

4. Area of the second contour of the left eye socket - P6: a file with 3d data of a person's face is calculated by approximating the contour with an ellipse;

5. Volume of the left eye socket - P8: calculated using the data P5, P6, P7 and the distance (step) between the isolines;

6. Area of the first contour of the right eye socket – P10: calculated similarly to the parameter P5;

7. Coordinates of the left base of the nose – (P15x, P15y): determined from a graphic file with a photo of a person;

8. The height of the tip of the nose – P17: determined from file with 3d data of human face;

9. the Area of the third contour of the left nose – P21: the file with 3d data of the human face is calculated by approximating the contour with a trapezoid;

10. The ratio «distance between pupils» / «Distance between the bridge of the nose and the base of the nose» –P25: calculated using the data (P1x, P1y), (P2x, P2y) and (P14x, P14y).

Fingerprint

The FPM10A module with the Adafruit Arduino library was used to create a block of the biometric fingerprint identification system [10]. Figure 2 shows the elements of the image acquisition and fingerprint identification block. The specified block is implemented on the basis of the Arduino UNO controller. The Arduino is a device based on the ATmega 328 microcontroller. [11] It includes everything you need for convenient operation with the microcontroller: 14 digital inputs/outputs (6 of them can be used as PWM outputs), 6 analog inputs, a 16 MHz quartz resonator, a USB connector, a power connector, a connector for in-circuit programming (ICSP) and a reset button.


**Figure 2** – fingerprint Scanner: 1-FPM10A optical fingerprint scanner; 2 – Arduino UNO; 3 – wires for connecting the scanner to Arduino; 4-USB cable for Arduino

When using a fingerprint sensor, there are two main steps. First, data is recorded in the sensor memory, that is, a unique ID is assigned to each fingerprint, which will be used for comparison in the future. After recording the data, you can proceed to «search», comparing the current image of the fingerprint with those recorded in the sensor memory. With the help of SFGDemo and ArduinoIDE, new fingerprints are loaded, assigning each of them a new ID #. All uploaded fingerprint images are encrypted (figure 3).


**Figure 3** – Uploading a fingerprint to the database

In figure 4, you can see the percentage of matches. Fingerprints that do not match the fingerprints stored in the database are ignored by the scanner.

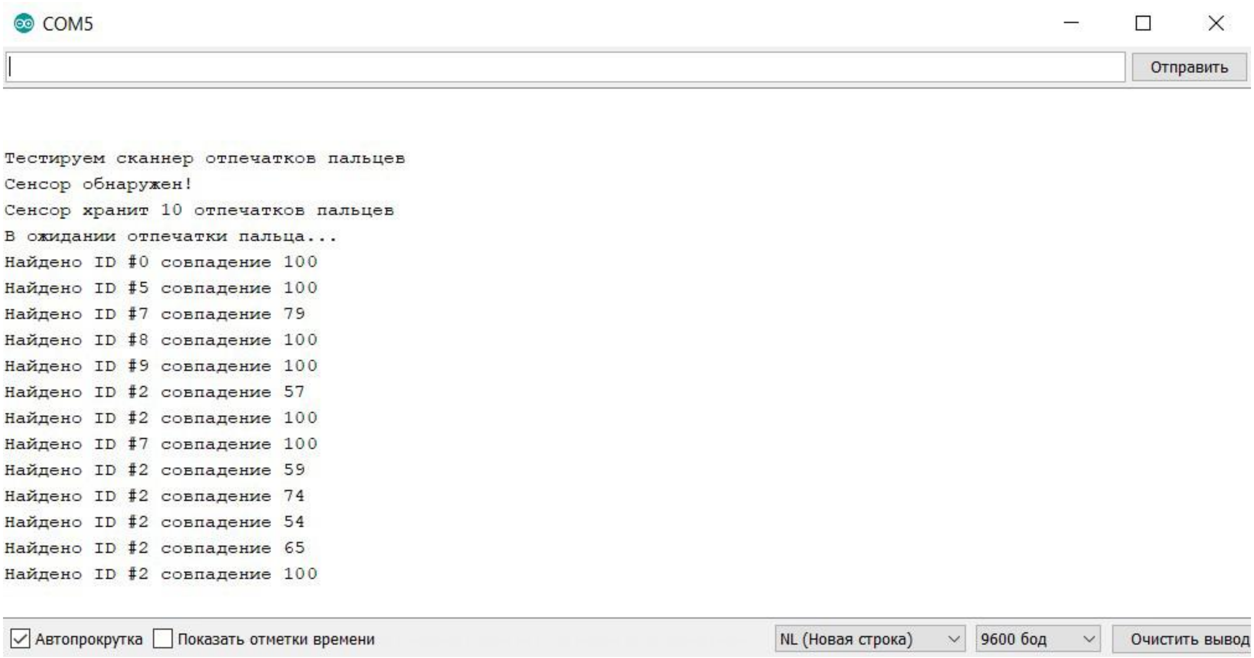


**Figure 4** – Fingerprint recognition

Identification signs of the structure of papillary patterns on the fingers are usually subdivided into global and local signs [12-13].

Global signs include signs that can be seen with the naked eye. These features include: type and type of papillary pattern; direction and steepness of streams of papillary lines; the structure of the central pattern of the pattern; delta structure; the number of papillary lines between the center and the delta and many other signs.

Another type of signs is local. They are also called minutiae (features or special points) - unique features inherent only in a specific imprint, determining the points of change in the structure of papillary lines (ending, bifurcation, rupture, etc.), the orientation of the papillary lines and coordinates at these points. Each print can contain up to 70 or more minutes.

In this study, several types of descriptors were used: SIFT, SURF and ORB [14]. SURF / SIFT algorithms have the best classifying ability when solving everyday search problems on textured images. For fingerprint identification tasks, they have «excess power».

Voice

One of the parameters of biometric identification of a person is the voice, but the voice of a person can change depending on age, emotional state, health or other factors, which makes the identification process more difficult to implement. The oral speech of a person is an ordered system of acoustic signals that are perceived as a sound image, and the oral speech of a person reflects its individual characteristics and features. The acoustic characteristic of the voice is relatively stable over time and remains individual even with pathological changes in the speech organs. The task of voice identification consists in selecting human speech from the input audio stream, its classification and recognition.

The rate of speech is a subjective measure related to the speed of pronunciation of certain segments of speech over time. The tempo can be related to the content, usually the most important words are pronounced slower. The volume and tempo of speech are individual for each person.

The difference in the timbres of different voices is described by different frequency spectra. The mathematical apparatus for analyzing the frequency spectrum is the Fourier transform, as a way to describe a complex sound wave with a spectrogram. Taking into account the peculiarities of human hearing (its non-linear nature in relation to the perception of sound frequencies), the transformation from the Hertz scale to the chalk scale (chalk is a psychophysical unit of sound pitch) is used for this task. The following is the formula for the transition frequency in Hertz (Hz) and the pitch of the sound the Mela

$$m = 1127 * ln(1 + f/700)$$

the calculated spectrum is superimposed with a set of M Mel scale filters, usually M=20 or M=24, usually the more filters the higher the accuracy, while the filters are shifted to those frequencies in which the most in the audio recording:

$$x_i = \sum_{k=0}^{N-1} |X_k| * H_i(f_k), i = 1..M$$

The mel h scale filter has a triangular shape, an example of such a filter is shown below (figure 5).

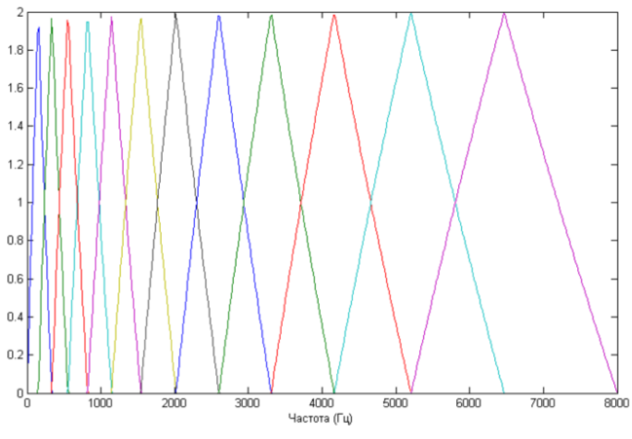


**Figure 5** – example of a mel filter

For acoustic speech analysis, a variety of algorithms are also used, the most common are hidden Markov models (SMM or HMM in the English version), as well as a model of a mixture of Gaussian distributions (SGR or GMM in the English version), neural networks have been actively used in recent years [15].

**Conclusion**

The AWP «Biometric information security system» was developed. Based on the application of multi-criteria optimization methods, different coefficients are calculated for each class, which allow ranking the criteria by importance. The properties of the proposed mathematical algorithm are investigated. For the first time, the human recognition algorithm takes into account such parameters as the volume of the nose, the volume of the eye socket, and other three-dimensional characteristics. A complex identification algorithm has been developed to take into account such phenomena as portrait shift, different photo scale and tilt of the identified face.

Based on the Arduino microcontroller and the FPM10A scanner, a recognition system has been developed for storing data, further processing it, identifying and displaying fingerprint images. The structure of papillary patterns on the fingers was chosen as identification features. The result of matching fingerprints with different rotation through the scanner is obtained. An experimental study of biometric fingerprint identification created on the basis of sift, SURF and ORB descriptors showed that the developed software system has invariance to image rotations. The system has developed three algorithms for analyzing audio recordings to solve the problem of biometric identification by voice.

**Acknowledgements.** The work was carried out at the expense of grant funding of scientific research for 2018-2020 under the project AP05131027 «Development of biometric methods and information protection tools».

## References

[1] Buzov G.A. Practical guide to identifying special technical means of unauthorized receipt of information. – M.: Hotline-Telecom, 2010. 240 p. (in Russian)

[2] Katrin Y.F., Razumovsky, A.V., Spivak A.I. Protection of information by technical means. St. Petersburg: ITMO National Research University, 2012. 416 p. (in Russian)

[3] Bolle R.M., Connel J.H., Pankanti S., Ratha N.K., Senior A.W. Guide to Biometrics, Sprinter-Verlag. – New York, 2004.

[4] Crihalmeanu S., Ross A. Multispectralsclera patternsfor ocularbiometric recognition. – Pattern Recognition. Lett, 2012. 1860–1869.

[5] Klepinin V.B., Agafonova T.P. Visual FoxPro 9.0. The most complete guide. – St. Petersburg: BHV-Petersburg, 2007. 1216 p. (in Russian)

[6] Jomartova Sh.A., Isimov N.T., Bayrbekova G.S., Ziyatbekova G.Z., Abdrazak Zh. Personal identification based on 2D and 3D images. Bulletin of the National Engineering Academy of the Republic of Kazakhstan, 2018. 2 (68). 16-20. (in Russian)

[7] Mazakov T.Zh., Abzhalilova A.A. Authentication and identification of a person by biometric parameter: fingerprint. Bulletin of KazUTB. 2020. 4. 26-33. (in Russian)

[8] Shormanov T.S., Mazakov T.Zh., Ziyatbekova G.Z., Aliaskar M.S. Algorithms for identification of a person by voice. Bulletin of KazUTB. 2020. 1. 58-68. (in Russian)

[9] Aivazyan S.A., Bukhshberger V.M., Enyukov I.S., Meshalkin L.D. Applied statistics. Classification and dimensionality reduction. – M.: Finance and statistics, 1989. 607 p. (in Russian)

[10] Petin V.A. Projects using the Arduino controller. – St. Petersburg: BHV-Petersburg, 2016. 464 p. (in Russian)

[11] A.V. Belov Arduino. From the basics of programming to the creation of practical devices. – SPb.: NiT, 2018. 480 p. (in Russian)

[12] Tan and B. Bhanu. “Robust fingerprint identification” in International Conference on Image Processing, 2002. 1. 1-277.

[13] Polyakov A.V. Algorithms and secure systems for biometric identity authentication // Auth. Cand. Dissertation. Moscow, 2018. 28 p.

[14] Ethan Rublee, Vincent Rabaud, Kurt Konolige, Gary Bradski. ORB: an efficient alternative to SIFT or SURF, Computer Vision (ICCV), IEEE International Conference on. IEEE, (2011). 2564–2571.

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.

**UDC 004.04**

# DATA MINING MODELS FOR HEALTHCARE

**Kuatbayeva A.A[1]., Izteleuov N.E[2]., Kabdoldin A[3]., Abdyzhalilova R.[4]**

IITU university, Almaty, Kazakhstan
[1]Ahamala2017@gmail.com, [2]izt.nurzhan@gmail.com,
[3]abylaikabdoldin@gmail.com,[4]xxx.raihan97@gmail.com

**Abstract.** In the healthcare sector data mining is becoming increasingly popular, because medicine organizations produce and collect large volumes of information day by day. Using data mining and data mining applications can help for each medicine organization and parts in the healthcare industry. For example, clients through data mining can take better services in healthcare, doctors better diagnose patients ' diseases and evolve a course of treatment. Of course, in these examples not the full range of power of data mining in the medicine sector, but main ones. Therefore we have problems like how all this system, also data mining is processed in general and using intelligent data analysis in our country. It is used for storing large amounts of data in different organizations and industries. The objective of this article is to show how the healthcare sector can benefit greatly from data mining and consider possible solution for problems, which described above.

**Keywords:** Data mining, data mining applications, healthcare sector, medicine industry, machine learning, algorithm.

## Introduction

Currently, a large database is very important and the intellectual part is especially valuable. In short, this is all called data mining. It is used for storing large amounts of data in different organizations and industries, as well as for processing and analyzing the same data. This raises the question "Why is this necessary?". The answer to this question is simple, because in the 21st century traditional methods are not relevant. In the time where it requires a minimum of time and effort, as well as money, outdated storage methods are very labor-intensive. Data mining solves these issues. This also applies to the healthcare sector. It is known that new automated systems and data mining allow to evaluate medical and biological indicators of patients ' examination, diagnose diseases and create a treatment algorithm that later has a positive therapeutic effect. Improvements in data technology can improve the efficiency of treatment and diagnostic processes, as well as other parts of the medical industry.

Thus, there are two urgent problems: the problem of how all this system, also data mining is processed. Second problem of using intelligent data analysis in Kazakhstan. In the article possible solutions for these problems.

The purpose of this article is to reveal the possibilities of big data processing. The first part of the article examines the data mining process in the medical sectors and contains answers for the question "How does it work?" and etc. The second and third parts of the article suggest which industries can be used for data mining.

## Data mining. methods and algorithm

At the beginning, it is necessary to understand what is meant by data mining. It is also called Knowledge Discovery In Data and studies the process of finding potentially useful, valid, and new knowledge in databases. In a broader sense, Data Mining refers to the concept of data analysis, which assumes that:

data may be inaccurate, incomplete (contain omissions), contradictory, heterogeneous, indirect, and this is why it is necessary to have huge volumes; therefore, understanding the data in specific applications require significant intellectual effort;

data analysis algorithms themselves may have " elements intelligence", in particular the ability to learn from precedents, that is, to draw General conclusions based on particular observations; the development of such algorithms also requires considerable intellectual effort;

the processes of processing raw data into information and information into knowledge can no longer be performed "manually" in the old way and sometimes require non-trivial automation.

Data mining is a multidisciplinary field that combines several branches of science and the main ones are artificial intelligence, statistics, and a database system (Fig. 1).

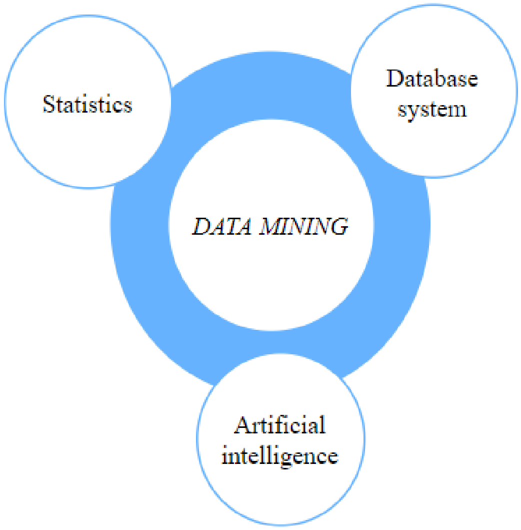


**Figure 1** -.Illustration Data Mining that combines several facets of science

Sometimes there is a different character of multidisciplinary approach. This is when Data Mining is considered a combination of computer science, mathematics,and domain expertise. In this case, computer science describes the environment for creating information products, mathematics builds a theoretical basis for solving problems, and the concept of the subject area allows necessary to understand the reality in which there is a problem situation [28].

There are the following typical steps that accompany the solution of data mining problems:

Analysis of the subject area, formulation of research goals and objectives.

Extracting and saving data.

Pre-processing of data:

cleaning;

integration;

transformation.

Meaningful data analysis using Data Mining methods (establishing General patterns or solving more specific, specific problems).

Interpretation of the results obtained by presenting them in a convenient format (visualization and selection of useful data patterns, generating informative graphs and / or tables).

Using new knowledge to make decisions.

Data Mining methods and algorithms include the following: artificial neural networks, decision trees, symbolic rules, the method of support vectors, linear regression, hierarchical and non-hierarchical methods of cluster analysis, various methods of data visualization and etc. If we consider the Support Vector Machine (SVM classifier) method, the main idea here is to map the source vectors to a higher-dimensional space and search for a dividing hyperplane with a maximum gap in this space [29]. The essence of the standard classifier SVM for the case of two classes can be represented using the following expression:

$f(x, W) = \text{sign}(g(x, W))$, где $g(x, W) = < x, W> + b$,

where are the parameters W (weight vector) and b (free coefficient) defined by the training procedure. Solution boundaries classifiers $g(x, W) = 0$ represent a hyperplane of order L-1 in L-dimensional space (Fig. 2).

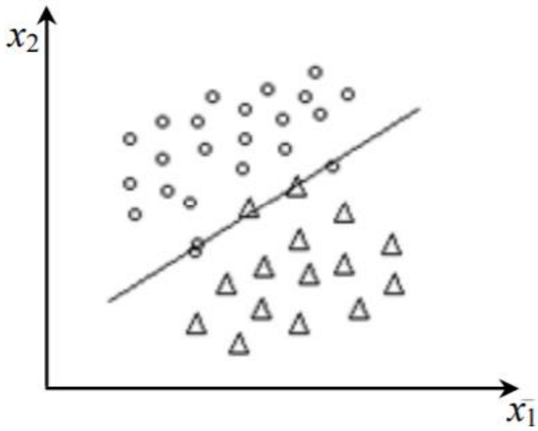


**Figure 2** - Illustration of the SVM classifier for two-dimensional space

Solving the classification problem for the case of several classes can be implemented by multiple pairwise classification and combining results (for example, by the majority rule) [29]. The SVM classifier has a fairly high algorithmic complexity, but it also has high computational efficiency. In addition, it is characterized by high accuracy and robustness of results for various statistical characteristics of training data [28].

**Data mining in healthcare industry**

Healthcare encompasses comprehensive assessment, diagnosis and preventive mechanisms for infections, disabilities and other physical and behavioral impairments in humans [1]. For most nations, the healthcare system is developing at a fast rate. The healthcare field should be seen as a position of rich data because it produces vast volumes of data, including electronic medical documents, administrative reports and other benchmarking findings [2]. However, these safety results remain under-utilized. Data mining, as is recognized, is a non-trivial method of finding real, new, theoretically valuable, and essentially understandable trends in data through mixing, through copious data sets, trends that are too subtle or complicated for humans to recognize. This implies that data mining is able to dig for fresh and useful knowledge from such vast quantities of data. Data mining in health care is primarily used to forecast different illnesses, as well as to support doctors interpret their clinical decisions.

The data mining models can be used in the following healthcare industries: anomaly detection, clustering and classification.

Anomaly Detection

Anomaly analysis is used to find the most important shifts in the data set [3]. Bo Lie et al [4] used three separate forms of anomaly detection, normal support vector data definition, density-induced support vector data definition and Gaussian mixture to test the precision of anomaly detection on unknown dataset of liver disease data collected from UCI. The process is tested using the precision of the AUC. The findings obtained for the healthy data collection were 93.59 per cent on average. Although the total standard deviation from the same data collection is 2.63. The unknown dataset is likely to be present in all databases, anomaly detection will be a reasonable way to address this problem, but as there is only one paper about this approach, we can not say much on the usefulness of the process.

Clustering

Clustering Clustering is a common descriptive method that seeks to define a finite collection of categories or clusters to classify the data [3]. Rui Velosoa[5] used the vector quantization process in the clustering framework to estimate readbacks in intensive medicine. The methods used in the vector quantification process are k-means, k-medoids and x-means. The samples used in this analysis were compiled from the clinical and research findings of the patient. The test for each algorithm is carried out using the Davies-Bouldin Index. The k-means received the strongest outcomes, whilst the x-means received decent results, and the k-medoids got the worse performance. The findings of the study of these studies include a valuable finding in trying to identify the various categories of patients with a greater risk of readmission. A more important comparison of the approach can not be made, since this is the only paper I have addressed in my analysis on vector quantization.

Classification

Classification is the discovery of a predictive learning function that classifies a data item into one of several predefined classes [3]. There are several types of classification: statistical, discriminant analysis, decision tree, swarm intelligence, k-nearest neighbor, logistic regression, Bayesian classifier and support vector.

There is a great deal of scope for data mining technologies in health care. Specifically, they may be categorized as the measurement of medication effectiveness; patient service management; client experience management; and the prevention of fraud and violence. More advanced scientific data mining, such as precision medicine and DNA microarray research, is beyond the reach of this article.

Treatment effectiveness. Data mining software may be developed to determine the effectiveness of medical therapies. Through analyzing and contrasting the triggers, effects and courses of therapy, data mining will include an overview of the courses of action that tend to be effective [6]. Of example, the results of patients groups diagnosed with various medication regimens with the same illness or disorder may be measured to decide which medications perform better and are more cost-effective [7].

Along this line, United HealthCare has used the patient report details to discover opportunities to lower costs and provide quality medicines [8]. Medical profiles have since been created to provide physicians with details about their clinical habits and to equate them with those of other physicians and peer-reviewed professional guidelines.

Certain data mining uses relevant to diagnosis involve comparing the multiple side-effects of medication, gathering specific signs to aid diagnose, determining the most appropriate drug compounds for the diagnosis of sub-populations that react differently from the normal population to such medications, and finding preventive steps that may minimize the incidence of disease [6].

Healthcare management. To help monitor health care, data mining tools may be built to accurately classify and track chronic illness and high-risk patients, plan effective treatments, and minimize hospital visits and claims.

In order to establish effective diagnosis and treatment procedures, for example, the Arkansas Research Network looks at readmittance and resource use and contrasts its findings with existing clinical literature in order to find the right treatment choices, while utilizing facts to justify patient care [7]. The Health Cooperative Group often stratifies consumer demographics through social features and medical problems to determine which communities need the most services, allowing them to establish initiatives to better inform and deter or control such populations [7]. Community Health Alliance has been active in a variety of data analytics projects to improve treatment at reduced prices. Data analysis at the Seton Medical Center is used to shorten patient duration of stay, prevent surgical risks, establish best practices, optimize medical satisfaction, and offer guidance to physicians-all to sustain and increase the standard of health care [9].

Customer relationship management. Although consumer experience management is a critical approach to handling relationships between business organizations-usually banks and retailers and their customers-it is no less relevant in the healthcare sense. Customer experiences can occur via contact centers, medical offices, billing divisions, frustrated environments, and outpatient treatment environments.

As in the case of corporate companies, data mining technologies in the healthcare field may be built to assess the desires, habits of usage and existing and potential needs of consumers to increase their degree of satisfaction [10]. Such methods can also be used to forecast certain items that the health care consumer is willing to buy, whether the individual is willing to comply with recommended medication or whether preventive maintenance is likely to result in a substantial decrease in potential use.

Using data analysis, Patient Capacity Management Corp. has established a Consumer Healthcare Utilization Index that offers an indicator of an individual's willingness to utilize particular clinical facilities, identified by 25 main diagnostic categories, specified diagnostic associated classes or different medical sector areas [11]. The database, focused on millions of healthcare interactions by many million people, is capable of recognizing people who may benefit from the most relevant healthcare programs, enabling patients most in need of particular treatment to seek them, and constantly improving the platforms and messaging used to target effective markets for better safety and long-term patient connections and loyalty. The database was utilized by OSF Saint Joseph Medical Center to deliver the correct information and resources to the most relevant patients at critical times. The end result is more productive and secure contact and improved revenue [11].

**Limitations and perspectives**

Data processing technologies will be a great value to the healthcare sector. We are not without limits, though.

Healthcare data mining may be constrained by data usability, since raw data mining sources frequently occur in various environments and structures, such as management, hospitals, labs and more. Data must also be obtained and incorporated before data mining can take effect. Although some scholars and analysts have recommended that a data center be constructed before an effort is made to mine data, this may be an expensive and time-consuming undertaking. On a good note, Intermountain HealthCare has effectively developed a data archive from five separate sources — a research data server, an intensive care case-mix network, a testing information system, an ambulatory case-mix system, and a treatment program database — and used it to identify and incorporate effective evidence-based healthcare approaches. Oakley [12] recommended a hierarchical network topology instead of a data center for more effective data mining, and Friedman and Pliskin [13] reported a case study of Maccabi Healthcare Services utilizing established repositories to direct future data mining.

Data processing technologies in the health sector may have enormous scope and usefulness. However, the success of healthcare data mining depends on the availability of clean health data. In this context, it is important for the healthcare industry to understand how data can be best collected, processed, packaged and extracted. Potential avenues involve standardization of clinical terminology and exchange of data through organisations to maximize the value of health data mining applications.

Furthermore, since health details are not restricted to textual details, such as patient reports or hospital documents, it is therefore important to expand the usage of text mining in order to extend the complexity and essence of what health data mining will actually achieve. For particular, it is valuable to be able to combine data and text mining.36 This is often helpful to look at how visual imaging photographs can be incorporated with health data mining applications. Throughout these fields, some improvement has been made [14,15].

**Conclusions**

Data mining has played a significant role in the healthcare field, especially in the prediction of different types of diseases. Diagnosis is commonly used in illness detection, which is frequently employed in surgical diagnosis. In addition, there is no data mining tool for solving problems with safety data sets. In order to obtain the highest accuracy among the classifiers that are essential for the medical diagnosis with the characteristics of the data being examined, we need to develop a hybrid model that could solve the problems described. Our potential goal is to improve predictions utilizing hybrid models.

**References**

[1] J.-J. Yang J. Li, J. Mulder, Y. Wang S. Chen, H. Wu Q. Wang, and H. Pan. Emerging information technologies for enhanced healthcare. Comput. Ind., 2015. 69. 3–11.

[2] N. Wickramasinghe, S. K. Sharma, and J. N. D. Gupta. Knowledge Management in Healthcare. 2005. 63. 5–18.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Mag. 1996. 37–54.

[4] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng. SVDD-based outlier detection on uncertain data. Knowl. Inf. Syst. 2013. 3(34). 597–618.

[5] R. Veloso, F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, and J. Machado. A Clustering Approach for Predicting Readmissions in Intensive Medicine. Procedia Technol., 2014. 16. 1307–1316.

[6] Milley, A. Healthcare and data mining. Health Management Technology, 2000. 21(8), 44-47.

[7] Kincade, K. Data mining: digging for healthcare gold. Insurance & Technology, 1998. 23(2), IM2-IM7.

[8] Young, J. & Pitta, J. Wal-Mart or Western Union? United HealthCare Corp. Forbes, 1997. 160(1). 244.

[9] Dakins, D.R. Center takes data tracking to heart. Health Data Management, 2001. 9(1). 32-36.

[10] Biafore, S. Predictive solutions bring more power to decision makers. Health Management Technology, 1999. 20(10). 12-14.

[11] Paddison, N. Index predicts individual service use. Health Management Technology, 2000. 21(2). 14-17.

[12] Oakley, S. Data mining, distributed networks and the laboratory. Health Management Technology, 1999. 20(5). 26-31.

[13] Friedman, N.L. & Pliskin, N. Demonstrating value-added utilization of existing databases for organizational decision-support. Information Resources Management Journal, 2002. 15(4). 1-15.

[14] Ceusters, W. Medical natural language understanding as a supporting technology for data mining in healthcare. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), PhysicaVerlag Heidelberg, New York, 2001. 41-69.

[15] Megalooikonomou, V. & Herskovits, E.H. Mining structurefunction associations in a brain image database. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), Physica-Verlag Heidelberg, New York, 2001. 153-180.

[16] Uskenbayeva R., Moldagulova A., Mukazhanov N.K., Creation of Data Classification System for Local Administration. Advances in Intelligent Systems and Computing. 2020

[17] Uskenbayeva R., Sabina R., Aigerim B., Managing Business Process Based on the Tonality of the Output Information, Advances in Intelligent Systems and Computing

[18] Uskenbayeva R.K., Kuandykov A.A., Rakhmetulayeva S.B., Bolshibayeva A.K. Properties of platforms for the transformation and automation of business processes, International Conference on Control, Automation and Systems, 2019.

[19] Moldagulova A.N., Uskenbayeva R.K., Satybaldiyeva R.Z., (...), Kalpeeva Z.B., Bektemyssova, G.U., On identification of hybrid business processes for effective implementation in the form of cloud services. International Conference on Control, Automation and Systems, 2019.

[20] Uskenbayeva R., Kuandykov A., Kalpeyeva Z., Kassymova A. Formalization of Applications for Processing in the e-Commerce System 2019 Proceedings - 21st IEEE Conference on Business Informatics, CBI 2019

[21] Mikhalevic, I.F., Ryjov A.P. Augmented intelligence framework for protecting against cyberattacks Proceedings - 5th International Conference on Engineering and Telecommunication,

EnT-MIPT 2018.

[22] Ryjov A. Personalization and optimization of information retrieval: Adaptive semantic layer approach. Proceedings - 2016 International Conference on Computational Science and Computational Intelligence, CSCI 2016

[23] Kuatbayeva A.A. Modelling situational room for healthcare. World Applied Sciences Journal, 2014.

[24] Kuatbayeva A.A. Fuzzy logic in healthcare situation room modelling. ACM International Conference Proceeding Series, (2014).

[25] Zamyztin A.V. Data mining. Tomsk, 2016. 120p.

[26] Wang L. (ed.). Support vector machines: theory and applications. Springer Science & Business Media, 2005. 177. 434 p.

[27] Hian Cbye Kob and Gerald Tan, Data Mining Applications in Healthcare, 2005

[28] Neesha Jothia, Nur Aini Abdul Rashidb, Wahidah Husainc. Data Mining in Healthcare – A Review, 2015.

**IRSTI 49.40.01**
**UDC 004.92**

# DEVELOPMENT OF A PERSONALITY RECOGNITION SYSTEM BY VOICE

**O.Zh.Mamyrbayev[1,2], A.S.Kydyrbekova[1,2], D.O.Oralbekova[1,3], B.Zh.Zhumazhanov[1], O.Mohamed[4]**
[1]Institute of information and computational technologies, 050010 Almaty
[2] Al-Farabi Kazakh National University, 050040 Almaty, Kazakhstan
[3]Satbayev University, 050040 Almaty, Kazakhstan
[4]Putra University, Malaysia
(e-mail: kas.aizat@mail.ru https:// https://orcid.org/0000-0001-5740-4100?)
(e-mail: morkenj@mail.ru https://orcid.org/0000-0001-8318-3794)

**Abstract**. the paper discusses the development of a voice recognition system based on cepstral coefficients on the chalk scale (Mel-frequency Cepstral Coefficients, MFCC) and mixed Gaussian models (Gaussian Mixture Models, GMM). Within the framework of this work, the existing methods for solving the problem of automatic speaker identification by voice were investigated. As a result, a complete review of the subject area was carried out and one of the advanced algorithms for solving the problem was implemented, based on the application of the Gaussian mixture model. The components of the Gaussian mixtures simulate the individual characteristics of the voice, which allows highly accurate distinguishing of human voices. It has been experimentally proven that traditional MFCCs using DNN and i-vector classifiers can achieve good results. The aim of the research is to create a simple and convenient automatic speaker recognition system. The description of the created speech database for the Kazakh language is given. Experiments have shown quite good results of the system FAR 8.05%.

**Keywords:** voice recognition system; MFCC; GMM; speech database.

## Introduction

In practice, voice recognition systems are widely used [1]: access to databases, bank accounts, forensics. Automatic voice recognition is a computational task of verifying a user's claimed compliance using characteristics derived from the user's voice. In the case of automatic voice recognition, the speech signal is processed to obtain information specific to the owner of the voice [2,3]. This information is used to create a non-playable speaker ID that is different from the original one. This makes the voice recognition process a secure method of authenticating the user, as opposed to passwords or tokens, which prevent theft, duplication, or forgetting of the voice. Unlike other biometric technologies, which are primarily image-based and require expensive hardware such as a fingerprint sensor or an eye scanner, voice recognition systems are designed for use on any standard public telephone or telephone network. The ability to work with standard telephony equipment allows you to support a wide range of biometric voice applications in a variety of environments.

The Institute of Information and Computer Technologies has been conducting research in the field of voice recognition for several years. As a result of the experience gained, a prototype of a voice recognition system was developed and a speech base for the Kazakh language was compiled. The aim of the study was to create a simple and convenient automatic voice recognition system.

The general scheme of the created system is shown in Fig. 1. The application is built in the MatLab programming language. The system consists of several main subsystems: verification and identification.

User credentials and voice models are stored in a SQLite database. Speech patterns are recorded in the registration subsystem, and after processing these patterns, a speech model is created. New user compatibility is added to the database along with the voice model. The engine also creates a Universal Background Model (UBM) attack model. In the identification subsystem based on the speech model, the search for the nearest users is carried out according to a certain indicator. The result can be a list of users or an empty list. The list is sorted in ascending order of distance from the user being sorted.

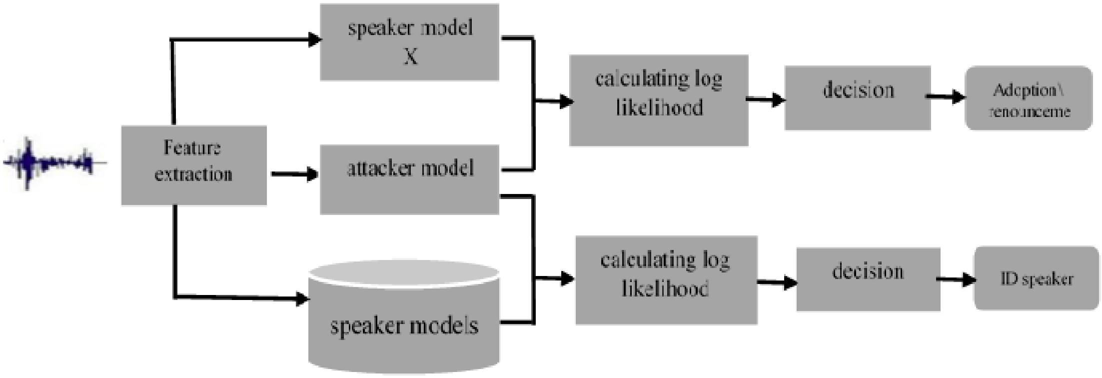


**Figure 1**- General architecture of the speaker verification and identification system

The verification subsystem verifies the identity of the owner of the vote. To do this, the owner of the voice must enter their identification number in the text field and speak the keyword. If authentication is successful, the user's credentials will be displayed and access granted. You can also update users' voice models based on new speech pattern recordings in this system.

Creation of a speech database for the Kazakh language

The performance of an automated test system is highly dependent on the speech database. The operation of the automatic recognition system is influenced by many factors. These include recording conditions, environment, recording devices, duration, speaker gender, age group, and more. It makes no sense to expect a good result from an automatic checking system without knowing the writing conditions. Here are some important reasons why you need a speech impediment:

a) the corpus can display the language of the content appropriate to the geographic environment using audio and voice data;

b) the practical characteristics of harmonized speech, which are not reflected in the textual database, reflect well the individual characteristics of users;

c) unlike text corporations, the speech corpus reflects prosodic information, as well as the chosen pronunciation style of the chosen sociocultural model.

Database description

The speech corpus for the Kazakh language was created by the Institute of Information and Computer Technologies. There are records of 86 speakers (21 men and 65 women). The recording was made in the office using the Cool Edit Pro software. All information includes: single numbers, single words, combinations of numbers and fragments of text.

The database was recorded in .wav format using a microphone with a sampling rate of 11025 Hz and a resolution of 16 bits per session. All native speakers. The total amount of data for each speaker is about 2500KB. The average speaking time for each speaker is about 100 seconds. There are 86 speakers with IDs from 1001 to 1086.

Speech signals output

Research shows that the perception of human speech sounds does not have a linear scale. Therefore, using the MFCC coefficients [4,5], it is possible to more accurately determine the human auditory system. This allows for better data processing.

The speech signal consists of sounds of different frequencies. For each actual audio frequency f, measured in Hz, the subjective pitch is measured on a scale called "chalk". It is a linear frequency below 1000 Hz and logarithmically above 1000 Hz. The following approximation formula is used to calculate the frequency f

$$mel(f) = 2595 * \log_{10}(1 + \frac{f}{700}) . \quad (1)$$

To extract the MFCC coefficients, the signal is divided into frames, to which a window is applied to reduce spectral distortion. As a result, we get a signal of the form

$$y(n)=x(n)w(n),\ 0\le n\le N-1. \tag{2}$$

Applying the Hamming window

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right), 0\le n\le N-1. \tag{3}$$

finally, using the Discrete Cosine Transform (DCT), the logarithmic spectrum is converted back to the time domain, and the MFCC coefficients are calculated

$$C_i=\sqrt{\frac{2}{N}}\sum_{j=1}^{L}\log(m_j)\cos\left(\frac{nj}{L}(i-0.5)\right). \tag{4}$$

**Gaussian mixed models**

Mixed Gaussian models are used to model the distribution of feature vectors received from each user (Mixture Models, GMM) [6,7]. The GMM can be thought of as a nonparametric multivariate Probability Density Function (PDF) that is capable of simulating arbitrary distributions and is the preferred method of speaker modeling.

One of the main advantages of GMM is the ability to create smooth approximations of arbitrary shapes of distributions. GMM has a fast learning phase compared to other approaches, the models can be easily scaled and updated when new speakers are added.

GMM distribution of feature vectors for speaker S is a weighted linear combination of M unimodal densities of Gaussians $b_i^S(x)$, each of which is parameterized by the expectation vector $\mu_i^s$ and the covariance matrix ∑Si. These parameters are collectively represented by the following entry:

$$\lambda_S=\{p_i^S,\mu_i^S,\Sigma_i^S\},\quad i=1,\dots,M \tag{5}$$

where $p_i^s$ – are mixed weights satisfying the condition

$$\sum_{i=1}^{M}p_i^S=1.$$

each speaker has its own model $\lambda_S$ .

For the feature vector x, the mixed density for the speaker S is calculated as

$$p(x|\lambda_S)=\sum_{i=1}^{M}p_i^S b_i^S(x), \tag{6}$$

where

$$b_i^S(x)=\frac{1}{(2\pi)^{D/2}\left|\Sigma_i^S\right|^{1/2}}\exp\left\{\frac{1}{2}(x-\mu_i^S)'\Sigma_i^S(x-\mu_i^S)\right\} \tag{7}$$

For a given sequence of feature vectors X{x1,x2,…,xT}, which are assumed to be independent, the logarithmic likelihood [8] of the speaker model λ_S is presented in the form

$$L_S(X)=\log p(X|\lambda_S)=\frac{1}{T}\sum_{t=1}^{T}\log p(x_t|\lambda_S). \tag{8}$$

To identify a speaker, the last equation is calculated for each speaker registered in the system. This paper uses a GMM with 32 blends for each model.

The speaker's identity is determined by the model with the highest value. Various algorithms are used to find the maximum likelihood of models. One of them is the EM (Expectation-Maximization) algorithm [11].

In this case, for each speaker S, we find the following values:

mixed weights:

$$p_i = \frac{1}{T}\sum_{t=1}^{T} pr(i|x_t,\lambda), \tag{9}$$

mathematical expectation:

$$\mu_i = \frac{\sum_{t=1}^{T} pr(i|x_t,\lambda)x_t}{\sum_{t=1}^{T} pr(i|x_t,\lambda)}, \tag{10}$$

covariance matrix:

$$\Sigma_i = \frac{\sum_{t=1}^{T} pr(i|x_t,\lambda)x_i^2}{\sum_{t=1}^{T} pr(i|x_t,\lambda)} - \mu_i^2, \tag{11}$$

where the posterior probability for component i has the form

$$pr(i|x_t,\lambda) = \frac{p_i b_i(x)}{\sum_{k=1}^{M} p_k b_k(x)}. \tag{12}$$

In modern voice recognition systems, UBM is used to simulate an alternative hypothesis. To simulate the possible space of acoustic possibilities, the standard TV-JFA method was used, which is one of the most effective methods in the field of voice testing [9,10]. The voice model in this approach is as follows:

$$M = m + Ux + Vy + Dz, \tag{13}$$

where M is the supervisor of the mixture of Gaussian distributions (GMM-models) of the speaker's voice, m is the supervisor of the parameters of the universal background model (UBM), U, V, D are their own matrix channels (Eigen Channel), respectively, their own voices and residual variability.

In the space of total variability, the i-vector is obtained using factor analysis, analyzed in the average supervectors of the UBM model and in the T-matrix of total variability. In this case, the voice model is described in the following relationship:

$$M = m + Tw, \tag{14}$$

where w is a low-dimensional vector in the space of possibilities.

In our systems, UBM is a mixture of Gaussian models of the described characteristics. To train the T-matrix and UBM, the features obtained from the training base of the competition were used. The UBM diagonal covariance matrix was studied using the EM-algorithm (Expectation - Maximization) [11]. It has voiceover characteristics and channel independent characteristics. UBM is a set of GMMs trained to distribute functions independently of the speaker.

**Experimental results**

Speech samples specially written for the experiments were taken from the speech database in question. Initially, each speech signal was preprocessed. The Hamming window length was 25 msec and the overlap was 12.5 msec. The following 20 MFCCs were obtained. The training was carried out in 5 samples for each speaker. The duration of the teaching presentation was about 10 seconds. Testing was conducted on 3-second samples.

**Table 1.** Experiment Results

| GMM с 32 смесями | | | | | |
|---|---|---|---|---|---|
| Duration of training | Duration of testing | Accuracy of recognition | FAR | FRR | EER |
| 10 сек | 3 сек | 93,7 % | 8,05% | 7,72% | 7,89% |

The experimental results are presented in Table 1. Recognition accuracy, pseudo-acceptance rate (FAR) and pseudo-rejection rate (FRR), as well as the equal error of type I and II errors (error rate, EER) are given.

**Conclusion**

The work describes the developed prototype of the voice recognition system. The results of testing the system in the Kazakh language on the basis of the collected speech are considered. Experiments have shown high results of his work. In the future, research will continue to develop an automatic voice recognition system to further enhance its security and recognition performance.

**Acknowledgments**

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic Kazakhstan (Grant No. AP08855743).

**References**

[1] J.P. Campbell. Speaker recognition: A tutorial. Proc. IEEE, (1997). 9(85). 1437–1462.

[2] J. Benesty M. Sondhi, Y. Huang, Springer handbook of speech processing, Springer, 2007. 1176 p.

[3] L. Rabiner B.H. Juang, Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 1993. 277 p.

[4] S. Furui, Digital speech processing, synthesis, and recognition. MarcelDekker, 2000. 452 p.

[5] S. Furui. Cepstral analysis techniques for automatic speaker verification. IEEE Tran. acoust., speech, signal processing, 1981. 27. 254-277.

[6] D.A. Reynolds. Speaker identification and verification using. Gaussian mixture speaker models. Speech Communication. 1995. 17. 91-108.

[7] D.A. Reynolds. A Gaussian mixture modeling approach to textindependent speaker identification. Ph.D. Thesis. – Georgia Institute of Technology, September, 1992.

[8] M.N.Kalimoldayev, O.Zh.Mamyrbayev, A.S.Kydyrbekova, N.O.Mekebayev, Algorithms for Detection Gender Using Neural Networks// International journal of circuits, systems and signal processing, ISSN, 2020. 14. 1998-4464.

[9] Mamyrbayev O., M.Turdalyuly, N.Mekebayev, K.Alimhan, T.Turdalykyz., A.Kydyrbekova A. Automatic Recognition of Kazakh Speech Using Deep Neural Networks. Intelligent Information and Database Systems Proceedings, Part II, Indonesia, 2019.

[10] Kydyrbekova A., Othman M., Mamyrbayev O., Akhmediyarova A., B.Zhumazhanov Identification and authentication of user voice using DNN features and i-vector. Cogent Engineering. 2020. 7 (1751557). 1 - 21.

[11] A.P. Dempster, N.M. Laird, D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1977. 39. 1-38.

**IRSTI 81.93.29**
**UDC 004.056.5**

# INFORMATION SECURITY TOOLS IN AN AUTOMATED SYSTEM FOR SECURE INFORMATION EXCHANGE

**Ye. Begimbayeva[1,2], O. Ussatova[1,2], R. Biyashev[1], Andrzej Smolarz[3], A. Abisheva[1,2]**
[1]Institute of Information and Computational Technologies,Almaty, Kazakhstan
[2]al-Farabi Kazakh National University, Almaty, Kazakhstan
[3]Lublin Technical University, Poland.
[1]enlik_89@mail.ru, [2]uoa_olga@mail.ru
[1]ORCID ID: https://orcid.org/0000-0002-4907-3345
[2]ORCID ID: https://orcid.org/0000-0002-5276-6118

**Abstract.** The article presents tools of information security to ensure the secure exchange of information in an automated system. A model of information security for an automated system of the exchange of information is proposed. Proposed model will increase the level of information security in electronic flow of document in the cross border information space. Any cryptographic system is based on the use of cryptographic keys. Key information is understood as the totality of all keys operating in the information network or system. If a sufficiently reliable management of key information is not ensured, then, having taken possession of it, an attacker gains unlimited access to all information in the network or system. Key management includes the implementation of functions such as generating, storing and distributing keys. The article presents requirements for the key management process. The main methods used to distribute keys between users of a computer network are presented.

**Keywords:** cross-border information exchange, electronic digital signature, encryption, information security, access control, cryptography, conflict situations, crypto analysis.

## Introduction

One of the tasks of the correct functioning of any information data transmission system, namely, in cross-border exchange, is to ensure secure data transmission.

The developed model of an automated system for secure cross-border information exchange (AS SCbIE) was developed taking into account the existence of many subsystems [1-5], including:

cryptographic subsystems confidentiality, data integrity and non-repudiation of authorship, by using encryption algorithms and electronic digital signature;

the access control subsystem provides identification and authentication when logging into the system, access of subjects to various objects and resources using specified policies and procedures;

the registration and accounting subsystem provides registration of: entry (exit) of subjects of access to (from) the system; actions of subjects over objects; object status; conflict situations in the system.

The structural general diagram of the developed model of AS SCbIE is shown in the figure 1.

The developed model of the information system uses an encryption algorithm based on non-positional polynomial notations ModNPSS14. This algorithm is a symmetric block cipher that uses blocks of 128 bits, an 8-cycle Feistel network [6-7].

The basis of the developed model of EDS generation and verification is the developed modified (extended) algorithm for EDS generation and verification based on modular notations. The software implementation of the developed modified algorithm for the formation and verification of EDS on the basis of modular notations has been carried out [5].

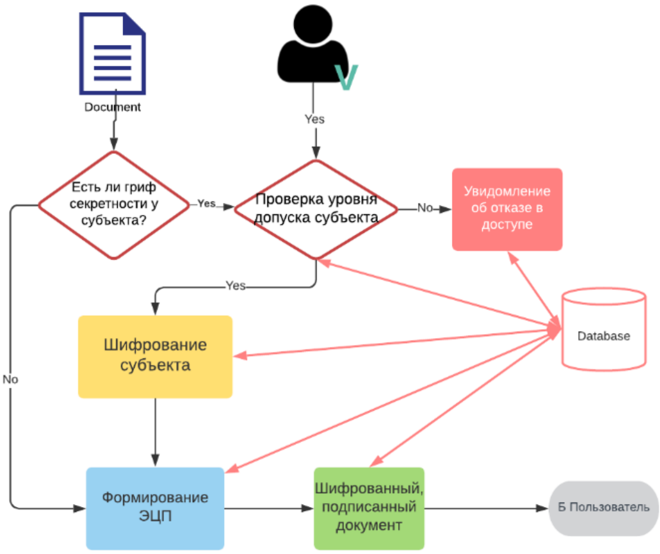


**Figure 1** - Structural general diagram of the information security model in AS SCbIE

Any cryptographic system is based on the use of cryptographic keys. Key information is understood as the totality of all keys operating in the information network or system. If a sufficiently reliable management of key information is not ensured, then, having taken possession of it, an attacker gains unlimited access to all information in the network or system. Key management includes the implementation of functions such as generating, storing and distributing keys.

The scope of work to support and maintain keys significantly exceeds the scope of work on their use for encrypting messages. Keys must be securely distributed to authorized users and kept up to date. They must be reliably protected during their transfer and storage on workstations and servers. Keys must be generated, destroyed and recovered in a secure manner. Key management can be organized manually or through an automated process.

It is imperative that the key management process is properly secured. The key management process has the following requirements:

- the key length must be large enough to provide the required level of security.
- keys must be stored and transferred in a secure manner.
- keys must be completely random, and their generation algorithms must make full use of all available key space.
- the key validity period should correspond to the criticality of the data it protects. Less critical data can be protected with a longer validity key, while critical data requires the use of short validity keys.
- the more often the key is used, the shorter its validity period should be.
- a key must be backed up or a duplicate key must be escrowed to an independent third party in case of an emergency.
- keys must be properly destroyed when they expire [8].

The goal of key management is to neutralize threats such as:

- compromising the confidentiality of private keys;
- compromising the authenticity of private or public keys. In this case, authenticity is understood as the knowledge or ability to verify the identity of the correspondent, to ensure confidential communication with which this key is used;
- unauthorized use of private or public keys, for example the use of a key that has expired.

When using a symmetric cryptosystem, two parties entering into an information exchange must first agree on a secret session key, that is, a key for encrypting all messages transmitted during the exchange. This key must be unknown to everyone else and must be periodically updated at the same time at the sender and receiver. The session key negotiation process is also referred to as key exchange, or key distribution.

An asymmetric cryptosystem assumes the use of two keys - public and private (secret). The public key can be disclosed, but the private key must be kept secret. When exchanging messages, only the public key needs to be sent, ensuring the authenticity of the forwarded public key.

Key distribution is one of the fundamental tasks of cryptography and the most critical process in key management. To understand the scale of the problem, we note that when servicing n users who exchange private information with each other, n (n-1)/2 different secret keys are required. As n grows, the problem of managing a huge number of keys arises. There are several ways to solve this problem. The definition of the most suitable one is chosen depending on the current situation [9]:

- physical distribution;
- issuance of a common key to participants in the interaction by the key issuing center - a "subscriber encryption" scheme;
- provision by a certification authority of access keys to public keys of users and issuance of private keys of a user;
- web of trust. Used in asymmetric cryptosystems;
- key exchange protocols.

The following requirements are imposed on the distribution of keys:

- efficiency and accuracy of distribution;
- confidentiality and integrity of distributed keys.

The following basic methods are used to distribute keys between users of a computer network:

1) using one or more key distribution centers;
2) direct exchange of keys between network users.

The problem with the first approach is that the key distribution center knows to whom and what keys are distributed, and this makes it possible to read all messages transmitted over the network. Potential abuse can significantly compromise network security. In the second approach, the problem is to reliably verify the identity of the network subjects.

Let's take a closer look at the second approach - direct key exchange between network users. When using a symmetric secret key cryptosystem for secure information exchange, two users wishing to exchange cryptographically protected information must have a shared secret key. These users must exchange the shared key over the communication channel in a secure manner. If users change keys frequently enough, key delivery becomes a serious problem.

The task of key distribution is reduced to the construction of such a key distribution protocol that provides:

- mutual confirmation of the authenticity of the session participants;
- confirmation of the validity of the session;
- using the minimum number of messages for key exchange.

It is advisable to base the solution to the key distribution problem on the principle of separating the procedure for confirming the authenticity of partners from the procedure for distributing keys itself. The goal of this approach is to provide a method in which, once authenticated, the participants generate the session key themselves, without the involvement of a key distribution center, so that the key distributor cannot identify the content of the messages.

The model and algorithm of the proposed method of protecting information in an automated control system using a combination of two factors: a permanent and temporary password. The user chooses a permanent password (the first factor) and uses it (account) when registering an account. The developed model is based on two types of two-factor authentication: authenticator applications and login verification using mobile applications are implemented [5,10].

One of the constituent components of the conflict resolution system is the Event Log (logging). Event logs are special files in which the system records significant events, such as user logon events or errors that occur while working with the system. When these types of events occur, an entry is created in the event logs. In detailed descriptions of events, users can find information useful for troubleshooting, finding the cause of problems, conflicts with the system or subsystems. A message in the event log is, first of all, information that can help the administrator and even the user understand what problem or conflict has arisen in the system and how to fix it [11].

Figure 2-3 shows a fragment of the authorization subsystem record and user actions. All user actions in the program are recorded in a separate file.

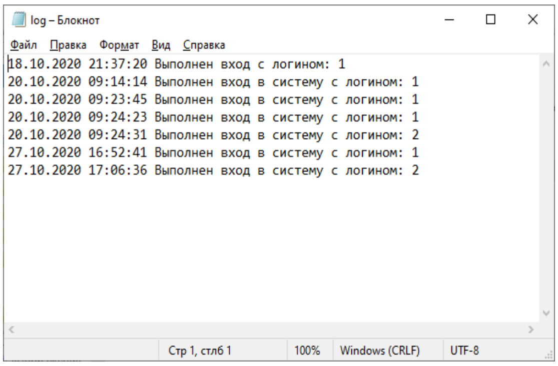


**Figure 2** - Fragment of logging into the system

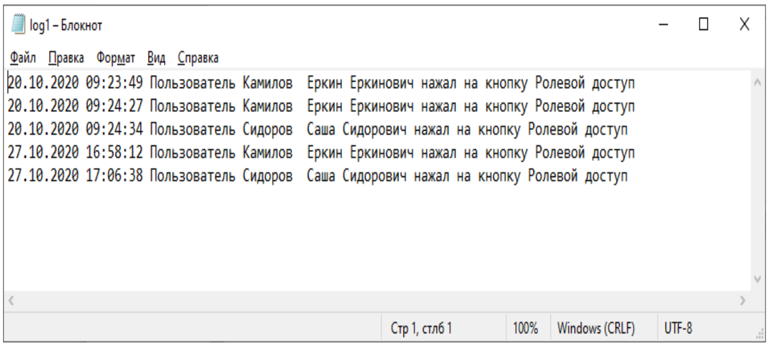


**Figure 3** - Fragment of logging user actions

The proposed structure [11-12] can serve as a basis for building a software complex for detecting conflicts in software systems for protecting information and resolving them. The use of conflict resolution systems for software information security systems will increase the efficiency of the system administrator and reduce the time required to identify the presence of conflict interaction in the information system.

When the program starts, an authorization window will open (Figure 4). To enter the system, you need to enter a username and password (if the user is registered in the database). The password field is masked in order to keep the entered password secret, if necessary, you can see the entered password by clicking on the "Show password" field. If all the entered data are correct, the main program window will open, where all information about the user will appear (Figure 5).

The main window of the program contains information about the employee (user) and about his documents. At the bottom of the main window, there are buttons for demonstrating information access models. "Role-based access" opens a new window with information about the user and available functions and documents (Figure 6).

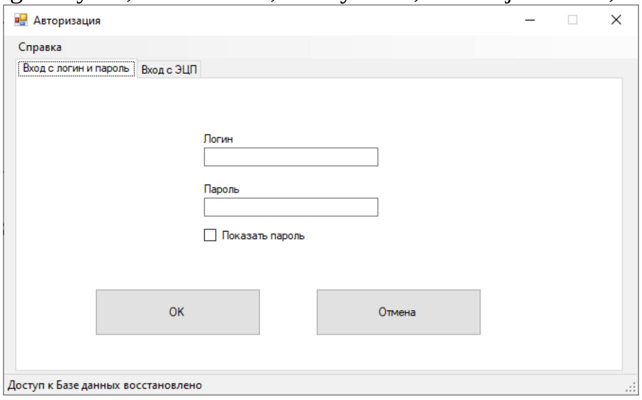


**Figure 4** - Program authorization window

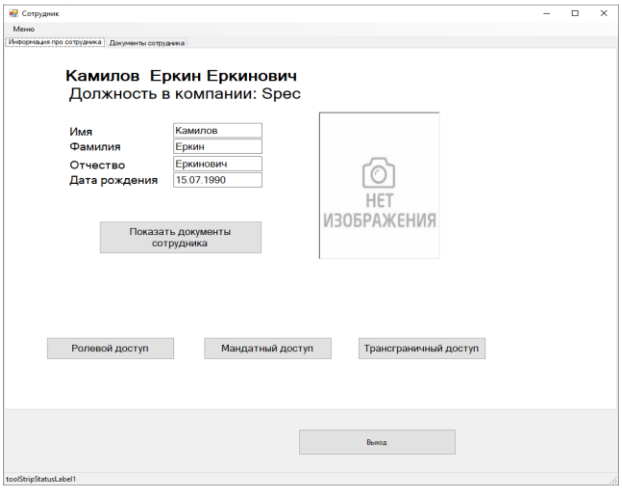


**Figure 5** - The main window of the program

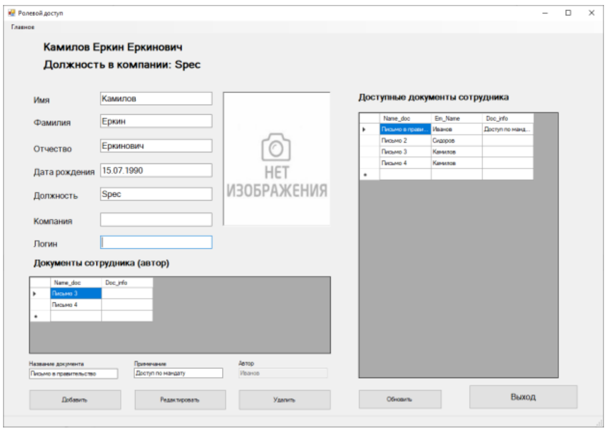


**Figure 6** - Role-based access model window

This user has full access to information (all functions for managing information are available). He can add, edit, delete and view information from the database (DB). The table on the right shows all documents available to the user. The name of the document, the creator of the document and brief information about the document. Figure 7 shows a user with read-only access, the user does not have access to the basic functions of editing information in the database.

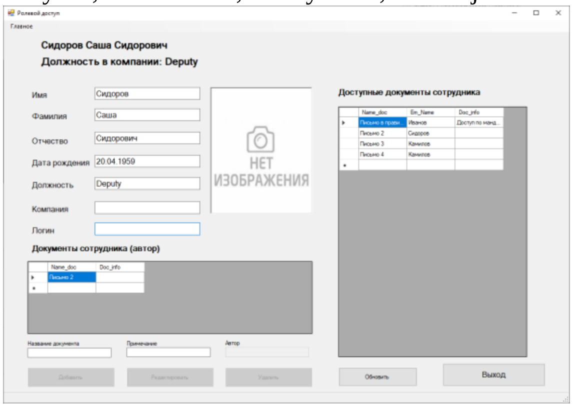


**Figure 7** - An example of a role-based access model (read-only)

## Conclusion

The assembly, debugging and introduction into the experimental, and then into industrial operation of the system is supposed to be carried out together with the team of the integration gateway, whose functions are determined by the EEU documents to ensure the functioning and subsequent development of the integration gateway. Since only they should know the composition of cross-border exchange subscribers on the territory of Kazakhstan and their secrecy labels. Accordingly, encryption and decryption keys must be jointly generated and distributed. This will reduce possible conflicts and optimally resolve them.

## Acknowledgements

Research work was carried out within the framework of the project AP05132584 “Development of Kazakhstan segment of the protected cross border information interaction”, which is being implemented at the Institute of Information and Computer Technologies.

## Reference:

[1] R. G. Biyashev, S. E. Nyssanbayeva, and Ye. Y. Begimbayeva Development of the model of protected cross-border information interaction. Open Engineering, 2016. 6. 199–205.

[2] Biyashev R.G., Begimbayeva Ye.Ye., Rog O.A. Development of an automated information protection system in the process of cross-border exchange. Mater. scientific. conf. "Modern problems of informatics and computing technologies". - Almaty: IICT MES RK, (2019). 148-154 (in Russian).

[3] Biyashev R.G., Begimbayeva Ye.Ye., Ussatova O.A., Rog O.A., Dyussenbaev D.S. Development of means for organizing secure information exchange in the Kazakhstani segment during cross-border interaction. Modern problems of informatics and computational technologies: Mat. scientific. conf. - Almaty: IICTT MES RK, (2020). 105-109 (in Russian).

[4] Begimbayeva Ye.Ye., Biyashev R.G. On the development of a system model for the Kazakh segment of secure cross-border information interaction. Mater. V int. scientific-practical conf. Computer science and applied mathematics. Almaty, (2020). 398-401 (in Russian).

[5] Begimbayeva, Y.Y., Ussatova O.A., Nyssanbayeva, S.E., Biyashev, R.G. Development of an automated system model of information protection in the cross-border exchange. Cogent Engineering. 2020. 7, Is. 1 // DOI: 10.1080/23311916.2020.1724597 (SJR 0.272, Q2, percentile 69).

[6] Dyussenbaev D., Sagan Y., Algazy K., Khompysh A. "MODNPSS14" cipher algorithmin cryptography taldau. Khabarshy KazKKA. 2019. 3. 235-243 (in Russian).

[7] Dyussenbaev D., Ostapenko V., Alkazy K., Sagan K. Cryptanalysis of "MODNPSS14" cipher algorithm. Mater. IV int. scientific-practical conf. "Computer Science and Applied Mathematics". Almaty, (2019). 556-561 (in Kazakh).

[8] Fomina I.A. Key management in cryptographic systems. Bulletin of Nizhegorodskogo universiteta im. N.I. Lobachevsky. Ser. inform. 2010. 4(1). 165-169.

[9] Schneier B. Applied Cryptography. Protocols, algorithms, source texts in the C language. M.: Iz-vo "Triumph", 2003. 816 p.

[10] S. Nyssanbayeva, W. Wojcik, O. Ussatova «Algorithm for generating temporary password based on the two- factor authentication model». Przegląd Elektrotechniczny, Poland. 2019. 5. 101 – 106.

[11] Begimbayeva Ye.Ye. On the model for resolving conflict situations in an automated system of secure information interaction. Actual problems of information security in Kazakhstan: Mater. Int. scientific. - practical. conf. Almaty, 2020. 105-107 (in Russian).

[12] Rustem Biyashev, Saule Nyssanbayeva and Yenlik Begimbayeva Development and Analysis of Possible Conflict Situations Resolution Systems in an Automated System. Proceeding of International Conference on Wireless Communication, Network and Multimedia Engineering. Guilin, China, (2019). 89. 182-184.

**UDC 004.04**

# APPLICATION OF SAS FOR ANALYSIS OF DATA OF MEDICAL INSTITUTIONS FROM DATA.EGOV.KZ PORTAL

**Zh.R. Amanzholova[1], Z.A. Altayeva[2], A.N. Moldagulova[3] L.M. Alimzhanova[4] A.A. Kuatbayeva[5]**
International IT University (IITU), Almaty, Kazakhstan

[1]https://orcid.org/0000-0002-6866-7667, amanzholova.97@mail.ru
[2]zarina_altayeva@mail.ru, [3]a.moldagulova@iitu.kz, [4]dimkim_01@mail.ru,
[5]ahamala2020@gmail.com

**Abstract.** The article discusses the statistical analysis of medical data on the example of infectious diseases using SAS. The analysis is based on a unified method for calculating indicators included in the State Program for the Development of Healthcare of the Republic of Kazakhstan "Densaulyk". The initial data for the research were taken from the open source data.egov.kz, which presents the latest information on infectious diseases in 3 regions and 1 city of the Republic of Kazakhstan. The methodology is based on the use of methods of descriptive statistics, graphical data analysis, building models on the SAS platform and can be used for a more in-depth analysis of infectious diseases in various population groups, as well as for their description in the cities of Kazakhstan. During the research, statistics were obtained using the SAS system for the level of infectious morbidity and the downloaded data was visualized in the form of various diagrams. The information obtained on the level and nature of infectious diseases can be used to assess trends in the health status of the population, plan various types of specialized medical care and treatment and preventive measures, and make rational use of the material and human resources of the healthcare system.

**Keywords:** data analysis, SAS platform, open source, infectious diseases, model.

## Introduction

Health system management is not possible without a detailed statistical analysis of data from medical institutions. Currently, control over the activities of the system of a medical institution is largely based on statistical forms summarizing data on morbidity and mortality of the population, as well as characterizing the activities of medical institutions. Such an approach is generally based on the understanding of health as a state of absence of diseases, which runs counter to the modern definition of World Health Organization (WHO), according to which full physical, mental and social well-being is also necessary for full health [1]. And this entails the need to review the set of indicators necessary for the analysis of the epidemiological picture in the territory of administrative units of any level. Macrodata allows us to identify general trends, but they do not provide an opportunity to track the causes of the changes. Moreover, they do not reflect the relationships that exist between infectious diseases. All this determines the relevance of this research.

Since there is no access to more detailed statistical data, it was decided to analyze the data of patients from medical institutions of Akmola, Aktobe, Almaty regions and the city of Almaty from an open source data.egov.kz portal. Based on the analysis, it is possible to show the potential of microdata and their advantage over traditional macrostatistical forms. It is the characteristics of infectious morbidity and mortality at the municipal level that remain practically outside the interests of medical and demographic statistics, while at this level, the prevention of morbidity can become most successful, as it is as close as possible to individuals with their special problems [2].

The main goal of this research was to study the state of the source of open data from the data.egov.kz portal according to the data of medical institutions and the system of statistical registration of infectious morbidity and mortality at the level of medical institutions for its suitability for statistical and analytical studies. The novelty of the research lies in the development of a model for statistical analysis of data from medical institutions obtained from an open source. The paper shows the methods of working with open data sources and a model for analyzing the epidemiological picture in the territories served by these medical institutions using the Statistical Analysis System (SAS) platform tools.

### Prospects for the application of sas for analysis of data of medical institutions

The progress of modern medical science is based on the analysis of large volumes of fact evidence of clinical, laboratory and other research methods [3]. An objective assessment of these data is greatly facilitated by the methods of statistical analysis, which help to find out not only quantitative, but also qualitative relationships between the studied phenomena [4].

Recently, for the processing, analysis and presentation of data from medical institutions, many statistical packages have been used - one of which is the SAS system. The system consists of modules, each of which performs a certain range of tasks, and it also implements its own programming language. SAS offers analytical solutions that provide effective monitoring. SAS is used for data processing, as well as for educational purposes, to develop skills in students working with this program. Using SAS, it is possible to carry out statistical processing of data of various levels of complexity, in accordance with the tasks: dispersion analysis, multiple linear regression analysis, logistic regression, survival analysis, etc. The SAS system also has great capabilities for presenting data: creating graphs, charts, tables, both for publication and presentation purposes. Interaction with the program is possible both in console mode and through a graphical interface, which is a graphical shell for simplified input of SAS programming language commands [5].

The software market continues to grow. New software packages are developed on the basis of modern computer technologies, which are developing rapidly. The capabilities of the programs are expanding, allowing a person to use them not only as an improvised tool that facilitates work, but also as a full-fledged assistant capable of solving complex medical problems. This research shows the use of the SAS statistical package to analyze data from medical facilities.

### Data processing and methods

Infectious morbidity is one of the main medical and statistical indicators of the health status of the population [6]. An analysis of this indicator over a number of years allows us to draw conclusions about the incidence and dynamics of the incidence, as well as the effectiveness of the complex of socio-hygienic and therapeutic measures aimed at reducing it.

The study was based on data on infectious morbidity and mortality in the Akmola, Aktobe, Almaty regions and Almaty city. The data describing the infectious morbidity in 2020 were selected as initial data. Cases of infectious diseases were recorded monthly. The diseases of the following diagnostic groups are presented: A00-B99 «Some infectious and parasitic diseases». When analyzing the infectious morbidity of the population, three main types of tasks were solved: comparing morbidity indicators in age groups, analyzing morbidity dynamics and analyzing morbidity.

In total, in 2020, 588 cases of infectious diseases were recorded in the Akmola region, in the Aktobe region - 1111, in the Almaty region - 1715, in Almaty - 334.

About 47.2% of patients are women and 52.8% of patients are men. The total number of appeals by territory is presented in Figure 1.

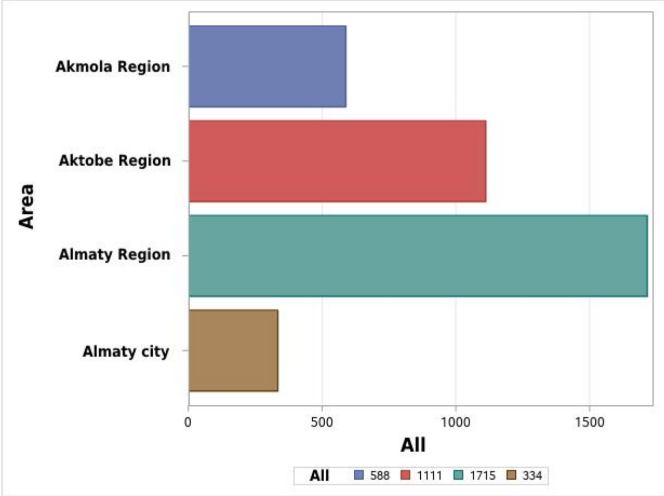


**Figure 1** – Chart of the total number of appeals by territory

**Results and analysis**

The data of infectious diseases among the population in the Republic of Kazakhstan is analyzed and the growth dynamics of the disease by some infectious and parasitic diseases among the population is noted. In addition, indicators were described and analyzed that describe the incidence (in terms of population circulation) of the main age groups of the population (children, teenagers, adults) in the Republic of Kazakhstan with infectious diseases.

For comparison, the age groups in Figure 2 were taken: children (0-14 years old), teenagers (15-17 years old) and adults (18 years and older). The study noted an increase in the incidence rate among adults in early 2020. In children, the incidence rate is slightly lower than in adults. A significant decrease in the incidence rate among teenagers is also noted.

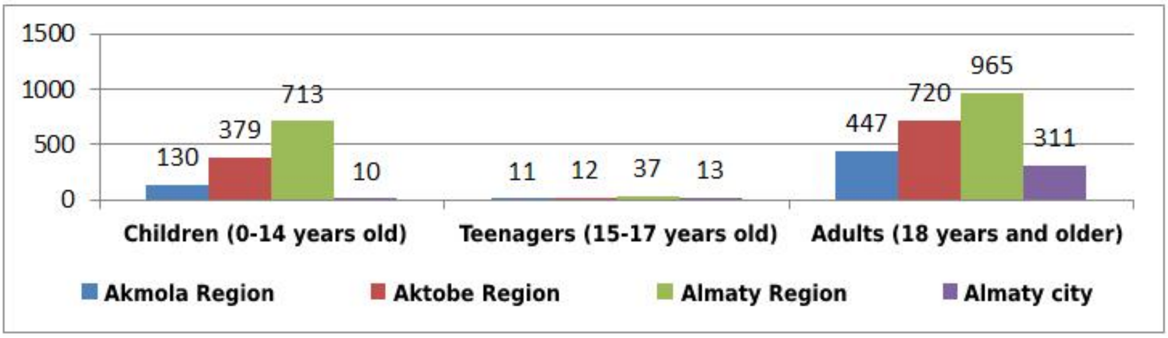


**Figure 2** – Diagram of growth dynamics of infectious diseases of different age groups by region

During the processing of data on infectious diseases, the following parameters were taken to obtain statistics:

Distribution by regions;

The number of all patients with infectious diseases.

The diagram shows the statistics on infectious diseases of all patients in the Akmola, Aktobe and Almaty regions and in Almaty city (Figure 3).

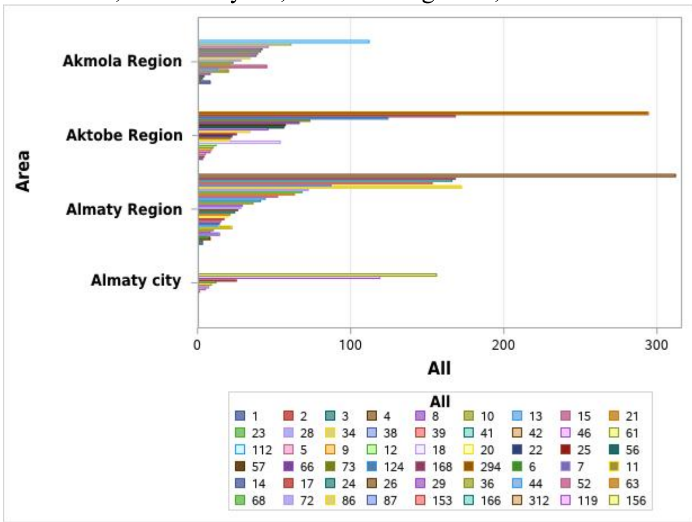


**Figure 3** – Chart on infectious diseases of all patients by regions

The following statistics is given for the number of deaths from infectious diseases in the Akmola, Aktobe and Almaty regions and Almaty city (Figure 4).

Selected options:

Distribution by regions;

The number of deaths from infectious diseases.

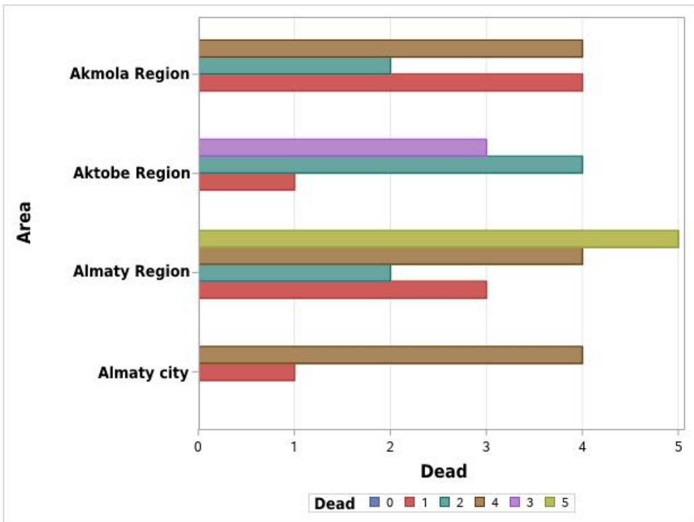


**Figure 4** – Chart by the number of deaths from infectious diseases by regions

The following are statistics on the number of patients with certain infectious and parasitic diseases in the Akmola, Aktobe and Almaty regions and Almaty city (Figure 5).

Selected options:

Distribution by regions;

The number of patients with certain infectious and parasitic diseases.

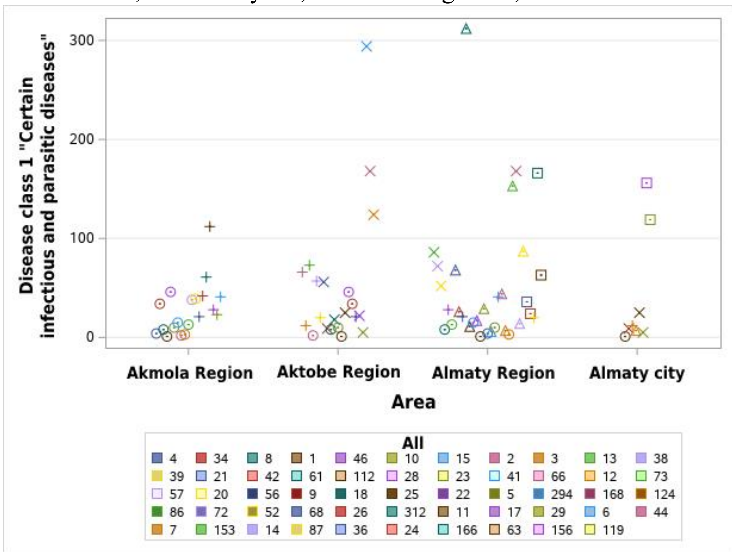


**Figure 5** – Chart by the number of patients with some infectious and parasitic diseases by regions

## Conclusion

In conclusion, it should be noted that a detailed analysis and processing of statistical information is an important component in the management of medical institutions. It is proposed to use statistical packages for processing, analysis and presentation of data from medical institutions - one of which is the SAS system. Using the SAS system, statistical processing of data of various complexity levels is carried out in accordance with the tasks: analysis of variance, logistic regression, survival analysis, etc.

The article uses the SAS system to analyze and process the data of medical institutions from the data.egov.kz portal on the level of infectious morbidity in the population of the Republic of Kazakhstan. Thus, in the course of the study, statistics was obtained using the SAS system for the level of infectious diseases for the beginning of 2020 and the downloaded data was visualized in the form of various diagrams.

As a result, statistics on the level of infectious diseases in the regions (Akmola, Aktobe, Almaty) and Almaty city were obtained:

For different age groups;

By the number of all patients;

By the number of dead;

By the number of patients with certain infectious and parasitic diseases.

The proposed methodology for studying the dynamics of infectious diseases can be applied for a more in-depth analysis of infectious diseases in various population groups, as well as for describing infectious diseases in other regions and cities of the Republic of Kazakhstan, presented in the open source data.egov.kz.

The information obtained on the level and nature of infectious diseases can be used to assess trends in the health status of the population, plan various types of specialized medical care and treatment and preventive measures, and make rational use of the material and human resources of the healthcare system.

## References

[1] Ustav VOZ [Elektronnyy resurs]. Rezhim dostupa: https://www.who.int/ru/about/who-we-are/constitution (7 iyunya 2020)

[2] Jose Bien R. Hernandez, Peggy Y. Kim Epidemiology Morbidity And Mortality [Elektronnyy resurs]. Rezhim dostupa: https://www.ncbi.nlm.nih.gov/books/NBK547668/ (11 yanvarya 2021)

[3] National Academies of Sciences, Engineering, and Medicine. 2015. Improving diagnosis in health care. Washington, DC: The National Academies Press [Elektronnyy resurs]. Rezhim dostupa: https://www.nap.edu/read/21794/chapter/1#iii (11 yanvarya 2021)

[4] M.K. Iztleuov, G.G. Yerimbetova Ispol'zovaniye sistemy SAS dlya analiza rezul'tatov biomeditsinskikh issledovaniy. Batys K̦azak̦stan meditsina zhurnaly. 2012. 3(35). 12-19.

[5] Obzor statisticheskikh programm [Elektronnyy resurs]. Rezhim dostupa: http://www.sciencefiles.ru/section/46/ (7 iyunya 2020)

[6] Morbidity statistics in the EU Report on pilot studies // Publication Office of the European Union, Luxembourg 2014 [Elektronnyy resurs]. Rezhim dostupa: https://ec.europa.eu/eurostat/documents/3888793/5858521/KS-TC-14-003-EN-N.pdf/bd959e6e-10ed-4078-915e-308941c02811 (11 yanvarya 2021)