Diabetes is among the socially significant diseases, which leads to high costs for the diagnosis and treatment of diabetes. Diagnosis and treatment of diabetes is currently one of the important tasks in medicine at the present stage of development of medical services. An important direction in the development of medical services for the population is the development and implementation of various problem-oriented information systems. Similar systems developed earlier did not cover the entire amount of heterogeneous information that is collected when diagnosing and prescribing the course of diabetes treatment, nor did they use technologies and cloud services as tools for Big Data. In this article, let's make use of the predictive analytic to forecast and categorize the type of diabetes which offers an effective method for treating and curing patients at a reduced cost, with improved results such as affordability and availability.

An information system platform has been developed and configured to manage the Hadoop cluster, as well as a non-relational database that uses and processes unstructured data in various formats. All experimental research, development of methods and algorithms, as well as solving computational problems were implemented using software languages for application development. The novelty lies in the research of distributed computing models that provide efficient execution of developed algorithms using the conceptual model of the processes of search, extraction and analysis of unstructured data in large data sets. The practical implementation of algorithms was carried out on the basis of methods of object-oriented programming and object-oriented databases

Keywords: diabetes mellitus, Big Data, Hadoop, MongoDB, information system, Python, database, patient, treatment, platform

UDC 004.4

DOI: 10.15587/1729-4061.2022.266185

DEVELOPING A SYSTEM FOR DIAGNOSING DIABETES MELLITUS USING BIGDATA

Assel Mukasheva
PhD, Associate Professor
Department of Information systems and Cybersecurity
Almaty University of Power Engineering and Telecommunications
Baytursinuli str., 126/1, Almaty, Republic of Kazakhstan, 050013

Sabina Rakhmetulayeva
Corresponding author
PhD, Associate Professor
Department of Information Systems
International Information Technology University
Manas str., 34/1, Almaty, Republic of Kazakhstan, 050000
E-mail: ssrakhmetulayeva@gmail.com

Gulnar Astaubayeva
PhD, Associate professor
Scientific and Educational Department of
Digital Technology and Data Analysis
Narxoz University
Zhandosov str., 55, Almaty, Republic of Kazakhstan, 050035

Sergiy Gnatyuk
Doctor of Technical Science, Professor
Department of Computer Science
Yessenov University
Microdistrict 32, Aktau, Republic of Kazakhstan, 130000

Received date 08.08.2022
Accepted date 14.10.2022
Published date 30.10.2022

***How to Cite**: Mukasheva, A., Rakhmetulayeva, S., Astaubayeva, G., Gnatyuk, S. (2022). Developing a system for diagnosing diabetes mellitus using bigdata. Eastern-European Journal of Enterprise Technologies, 5 (2 (119)), 75–85. doi: https://doi.org/10.15587/1729-4061.2022.266185*

1. Introduction

Diabetes Mellitus (DM) is a socially significant disease, hence the high incidence of the disease among the population, as well as the high costs of diagnosis and treatment of the disease. Despite advances in diabetes mellitus treatment, late complications are still a problem.

Digital healthcare includes advanced medical technology, innovation, and digital communication. The medical revolution is closely linked to the spread of digital health, the use of big data to make better data based decisions [1]. The daily activities of a diabetic specialist are related to digitization, which include electronic medical records, images, laboratory reports, various software for administrative practice, data on glucose levels downloaded from glucose meters, Continuous glucose and insulin pump monitoring sensors. With this data, a diabetic therapist makes therapeutic decisions, and the fuller and more data they have, the more they need information tools that can help them analyze, help them identify specific patterns to detect glycemic abnormalities, understand possible causes and adopt appropriate therapeutic strategies to remedy these disorders [2, 3].

Big data technology offers significant opportunities for diabetes treatment. The most important aspect of the big data revolution is the need to distribute data processing [4]. Thus, important directions can be noted. First, the integration of different sources of information, from primary and secondary care to administrative information, provides a new perspective on patient care processes and individual patient behavior, taking into account the multifaceted nature of the treatment of chronic diseases. Second, the availability of new diabetes treatment technologies capable of collecting large amounts of unstructured real-time data that require distributed platforms for data analysis and decision support.

The rapid growth of data collection in the form of electronic medical records, registers or portable sensors has triggered the big data revolution in the health industry. There is much to be learned from these readily available data, which provide multiple benefits such as improved quality of life, disease diagnosis, treatment and health-care delivery. Other problems associated with big health data include heterogeneous data, a large number of variables and the need to analyze data in real time [5].

Various sectors of the public and private sectors generate, store and analyze big data to improve the services they provide. The available medical data require proper management and analysis to obtain more relevant information [6]. In the modern world, the healthcare field is constantly being improved thanks to the widespread penetration of IoT technologies, where data can be generated from various sources, and then can be applied to support the basic functions of healthcare institutions [7, 8]. In recent years, there has been a great increase in biomedical information, including genomic sequences, electronic medical records, and biomedical signals and images that have taken healthcare to a new era [9–12].

Undoubtedly, new technologies were required to store and retrieve the necessary information from this volume of big data, but other traditional relational databases cannot satisfy the existing needs of users [13]. In the healthcare sector, diabetes is one of the main common types of disease, which form a large amount of medical data and information [14]. Diabetes is a group of metabolic diseases in which a person suffers from elevated blood glucose levels in the body [15].

The introduction of data into health often requires the creation and collection of high-quality data in real time [16]. Big data in healthcare can be used to unite different fields for comprehensive study of the disease [17]. The application of new technologies such as blockchain, cloud computing, and machine learning in healthcare offers promising research opportunities [18]. Big data analysis is of paramount importance in such aspects of health care as patient diagnosis, rapid detection of epidemics and improved patient management [19].

The development and implementation of various problem-oriented information systems (IS) is becoming an important area for the development of public health services. Similar systems that were previously developed did not capture the full amount of heterogeneous information that is collected in the diagnosis and treatment of diabetes mellitus, nor did they use big data technologies and cloud services as tools. Currently, the lack of problem-oriented IT diagnostics and recommendations for DM treatment underlines the relevance. Therefore, the need for research and development of an endocrinologist support system for diabetes diagnosis based on Big Data technology is high. The use of modern information technology in the health-care system will improve the quality of health-care services through technologies for the analysis, processing and visualization of big data. Therefore, the study of these issues determines the relevance of this direction.

2. Literature review and problem statement

Health is always an interesting field for researchers, and the rapidly expanding field of big data analysis has begun to play a key role in the evolution of medical practices and research [20]. Due to its innovative contribution to decision-making and strategic development in the health sector, Big Data Analytics has attracted considerable attention [21]. This paper [22] present a wide interest in big data analytics due to their innovative contribution to decision-making and strategic development in the field of healthcare, as there is an increasing need for understanding trends in massive data sets. Thus, healthcare is one of the promising areas in which big data can be applied. These data have significant potential for improving patient outcomes, reducing financial costs for medical care, and improving the overall quality of life [23]. However, the authors of this study argue that the use of big data in the healthcare field is still in its infancy and the healthcare industry has not adapted quickly enough to the movement of big data when compared with other industries [6]. Thus, Big Data is important in healthcare, and it becomes an important database where the information obtained can be used to treat and manage diseases [24].

In this study, the authors use great data to treat dementia and chronic diseases, which enhances the ability to study diabetes treatment [25]. There are some health challenges in big data processing, and the authors have analyzed the use of three different data processing paradigms [26]. In this paper was proposed extensible architecture of big data, which is based on both streaming and batch computing, in order to further improve the reliability of health systems by generating alerts in real time [27].

In the following studies, the authors compared the diabetes dataset using the Hadoop framework, which is a distribution platform and can be used to analyze large amounts of data [28]. The study discusses key concepts and definitions related to big data, presents significant health efforts and discusses the potential role of big data in the treatment of diabetes, which positively influenced the results of the study [29]. With the development of information technology, data on diabetes patients are increasingly collected digitally. There is no single minimum data set agreed at the national or international level for data collection. The study notes that health service providers generate a wealth of data, including patient data, pathology reports and prescription information, collectively known as “Big Data” [30]. However, the results presented were limited to the collaborating centres that provided data, which influenced the results of the study.

The authors argue that future research should use a systems thinking approach, aided by recent advances in sensor sensors, big data, and related technologies. These opportunities will allow to explore and address all these factors in our quest to develop more targeted and effective public health interventions for overweight, obesity, and diabetes control and prevention [31].

Big data analytics is gaining popularity in medical engineering and healthcare. Stakeholders find that big data analysis reduces health costs and personalizes health care for each individual patient [32, 33]. This research argues that big data analytics can be used in large-scale genetic research, public health, personalized and accurate medicine, and new drug development. There are also studies that present in-depth analytical views of big data in health care. The study presents comprehensive reviews by researchers of different methodologies, technologies and subject areas, including the identification of significant gaps. However, the study has some limitations as it did not have an exact publication date for each article.

In the practice of personalized medicine, in addition to the data provided by patients, data from a variety of sources are used. These data are obtained from patients who have different data subjects and are usually recorded in electronic patient files for clinical purposes. Another source of big data includes computational analysis of these data [18, 34]. It can be observed that research on the application of big data analysis in health care is gaining popularity, especially in

the field of information systems and medical research. All this suggests that it is advisable to conduct a study on the application of big data technology in health care due to high data growth.

3. The aim and objectives of the study

The aim of this research is to make an information system that will assist the healthcare industry in diagnosis of diabetes mellitus, thus, reducing management costs, delivering clinical decision support to physicians, achieving regulatory compliance, improving quality of care, and preventing chronic diseases.

To achieve this aim, the following tasks are accomplished:

– create a mathematical model for diagnosing diabetes mellitus;

– develop an algorithmic system for diagnosing and supporting diabetes treatment based on created mathematical model;

– develop an information system platform using Big Data technology.

4. Materials and methods

4. 1. Object and hypothesis of the study

The purpose of this step is to create a model for diagnosing diabetes mellitus using Big Data. The prospects of using intelligent methods and models are confirmed in systems for determining information risks, measures of information certainty, and other parameters of DM diagnostics. However, the use of information systems (IS) for managing different types of data measures in cloud services is still giving way to classical mathematical algorithms, which is due to the limited use of standard computing technologies. The solution to this problem is seen in the inclusion of an element of expert assessments in the IS based on existing standardized detailed protocols and data from expert domestic diagnostic protocols. To do this, all unstructured formats of medical analyses are entered into the system. After checking all the tests and examinations received, the endocrinologist officially diagnoses the patient's diabetes, and decides on the appointment of treatment and the possibility of hospitalization, if necessary. However, the accuracy of predicting the probability of diagnosis confirmation by these parameters, even with the allowance for parallel calculations, can only be achieved by optimizing unstructured data.

4. 2. Data set used

In this study, statistical data from the registry of patients with diabetes over the past fifteen years in the Republic of Kazakhstan were used. To conduct experimental studies, data on patients with diabetes were provided by the Public Foundation "Kazakhstan Society for the Study of Diabetes" in Almaty. The aim of the study is to build a model for identifying the most accurate experimental method for predicting diabetes. As the date of the sets, medical data provided by the Public Foundation "Kazakhstan Society for the Study of Diabetes" were used as part of the target program "The Burden of Diabetes 2019–2020 for the Republic of Kazakhstan". The data set includes all information about dispensary patients of all citizens of the Republic of Kazakhstan registered with a diagnosis of "Diabetes" from 12.31.2010 to 12.31.2020.

The patient's age, complaints, the length of time they've had the ailment, etc. are all included in the database's extensive list of patient characteristics. On the other hand, not every available sign should be considered significant when constructing models. The elimination of noise and features that are not informative is one of the tasks that are included in the stage of data preparation for modeling.

To implement feature selection, the Elasticsearch method was chosen. For this purpose, key terms will be aggregated. Elasticsearch thereby profiles the generated set by displaying keywords that differentiate it from other data. Therefore, "Filter" chooses all diseases whose names contain the word "Diabetes." The query is filtered by the "name" column, and the value is retained only if it contains the term "Diabetes."

```
from elasticsearch import Elasticsearch
client = Elasticsearch()
indexName = "medical"

docType="diseases"
searchFrom=0
searchSize=3

searchBody={
"fields": ["name"],
"query": {
"filtered":{
"filter": {
        'term': {'name':'diabetes'}
        }
}
        },
"aggregations":{
"DiseaseKeywords":{
"significant_terms": { "field" : "fulltext", "size":30
                                }
}
}
}
client.search(index=indexName,  doc_type=docType,
body=searchBody, from_=searchFrom, size=searchSize)
```

5. Results of experimental studies on the development of a system for the diagnosis and treatment of diabetes mellitus

5. 1. Development of a model for diagnosing diabetes mellitus

The developed system uses sequential execution of stages of mathematical description of the model operation process:

Stage 1 – creating a set of input and output parameters for the model. At this stage it is envisaged the use of the model integration tolerance parameters of the probability of diagnosis the number of criteria, allowing based on analysis of characteristics of the object $H=\{h_1...h_5\}$ and the characteristics of parallel computing on various medical tests $F=\{f_1...f_f\}$ to define a set of tolerance parameters, the number of appeals that will be used as input parameters of the system:

$$X=\left\{x_1,\ldots,x_{N_x}\right\}, \quad (1)$$

where N_x is the number of input parameters for the model. It also defines a set of output parameters of the model that will indicate the presence or absence of parallel calculations of a certain type:

$$Y=\left\{y_1,\ldots,y_{N_y}\right\}, \tag{2}$$

where N_y is the number of output parameters.

Stage 2 – getting statistical data. As a result of analyzing the characteristics of the cloud service in the diagnostics core $M=\{m_1...m_k\}$ the possibility of registering the diagnosis confirmation parameters used to determine X and Y is evaluated. If the assessment is negative, the diagnosis is not confirmed, that is:

$$M=\left\{m_{MPNN}\right\}. \tag{3}$$

Stage 3 – representation of expert knowledge. Using the developed algorithm for applying production rules to represent expert knowledge in plug-ins, the possibility of developing a sub kernel model for diagnosing qualitative results (ultrasound data, etc.) is evaluated. If the score is negative, the sub kernel model is excluded from further consideration:

$$m_{MPNN}\notin M. \tag{4}$$

Otherwise:

$$m_{MPNN}\in M. \tag{5}$$

Stage 4 – checking the set of acceptable types of attributes. The stage is focused on checking the set of acceptable clinical pictures that are not significant for confirming the diagnosis. Thus, a rule is used to define ε:

$$\delta_1\le\delta_2\rightarrow\varepsilon=\delta_1. \tag{6}$$

Otherwise:

$$\varepsilon=\delta_2, \tag{7}$$

where δ_1 is the allowed minimum value of parallel calculations based on signs of deviations from the norm, δ_2 is the acceptable minimum value for false diagnosis recognition, ε is the allowable error.

Stage 5 – determining the minimum size of the training sample. This stage is aimed at determining the minimum acceptable number of evaluation criteria for making a diagnosis. The calculation is as follows:

$$P_{\min}=20N_x, \tag{8}$$

where $P_{\min}$ is the allowed minimum number of examples, N_x is the number of input parameters.

Stage 6 – check. This step involves comparing an acceptable parameter in the sample to identify calculation errors:

$$t_{min}\left(m_j\right)\ge t_d\rightarrow m_j\notin M^{tn}, \tag{9}$$

where M^{tn} is the set of significant indicators with an acceptable deviation, m_j is j^{th} element of the system. Since the cloud service efficiency model uses unstructured data as training examples for this case, the sampling system for the model is equal to the duration of product rules development. For calculation $T_{j,\max}$ the expression is used:

$$T_{j,\max}=T_f-t_j, \tag{10}$$

where t_j is the training period for the j^{th} model $m_j\notin M^{(tn)}$. The result of this stage is the generated set:

$$\left\{T_{1,\max},\ldots T_{L,\max}\right\}, \tag{11}$$

where L is the number of $M^{(tn)}$ elements.

The set of systems that can be used to evaluate the parameters for allowing the number of requests is formed using the expressions (12), (13):

$$T_{j,\max}>T_d\rightarrow m_j^{(tn)}\notin M_z, \tag{12}$$

$$T_{j,\max}<T_d\rightarrow m_j^{(tn)}\in M, \tag{13}$$

where M_z is a set of systems (in this system – types of analyses) that can be used to evaluate the parameters of access tolerance. As a result of the implementation of the algorithm, a set of neural network systems is formed that can be used to evaluate the parameters for allowing the number of requests to recognize parallel calculations. At the same time, the very functioning of the diagnostic system based on a comprehensive data collection model in a cloud service should be connected to a medical treatment Protocol, the choice of which should act as a separate unification module. Therefore, by the stage of application of methodologies for selecting the technology for building IS, it is necessary to study the data types in models with an emphasis on the use of information technology for big data and building a project model for diagnosing diabetes using big data, considering all approaches to the development of such models.

5. 2. Development of the system based on the mathematical model

The Hadoop framework manages data processing and storage for database applications running on clustered servers. Additionally, it can analyze and store both organized and unstructured data. Hadoop can scale to host thousands of hardware nodes and massive volumes of data on typical server clusters.

On the basis of the aforementioned tasks and following a comparative analysis, the following Hadoop tools were selected:

– Hive. It is an open source storage system that queries and parses large datasets stored primarily in Hadoop files;

– MongoDB. It is an open source database management system that uses a document-oriented database model, which in turn supports various forms of data;

– Oozie. It is a server-side workflow scheduling system for managing Hadoop jobs;

– Sqoop. It is a tool designed to transfer data between Hadoop and relational database servers.

This study aims to develop an ecosystem Hadoop [35] to create an IS in its support of the doctor endocrinologist for diagnosis of diabetes mellitus. Access was gained to the technical equipment of Satpayev University, which made it possible to obtain computing nodes for working with BigData technology tools. The platform under development has a total of 4 servers, each with 64 GB of RAM, 600 HD.

To develop and configure the cluster platform, the Apache Ambari software product [36] of the Apache Software Foundation is used. Ambari allows to manage and control the Hadoop cluster [37], as well as integrate Hadoop with your existing enterprise infrastructure. After conducting a comparative analysis, the BigData technology tools were selected and installed in the platform environment of the future ecosystem.

A conceptual diagram of the IS that visually introduces the user to the functioning of the information and simulation system. The relationships between the system modules are also shown in detail here. At the conceptual level, there is an IS for data management or focused on working with data. The relevance of such information systems have increased with the advent of Big Data information technology. Big Data as information technology has such a formal model:

$$BD = \left(Vol_{BD}, I_P, A_{BD}, T_{BD}\right), \quad (14)$$

where Vol_{BD} is the set of different types of volumes; I_p is the set of multiple types of data sources (diagnostic criteria); A_{BD} is the set of methods for analyzing Big Data; T_{BD} is the set of multiple Big Data processing technologies. The following technologies are used at the stages of Big Data processing:

$$T_{BD} = \left(T_{NoSQL}, T_{SQL}, T_{Hadoop}, T_V\right), \quad (15)$$

where T_{NoSQL} – NoSQL database technologies;

T_{Hadoop} – technologies for providing massively parallel processing;

T_{SQL} – structured data processing technologies (SQL databases);

T_v – Big Data visualization technologies. This way, all analyses are loaded into the MongoDB database and stored in the system cluster. The best option that is compatible with a dual-core structure and support (SQL databases) is Hadoop, while all unstructured data (numeric, text, ultrasound images, etc.) is loaded into the MongoDB database. Due to the unstructured nature of data, will pay special attention to technologies for providing massively parallel processing, since the purpose of this work is to study cloud computing and make a project decision on choosing the optimal model for diagnosis and treatment as two parallel cloud services, components of an IS for managing Big Data.

As a result of the studies, an information model for diagnosing diabetes using big data has been developed. Big data was used to develop an information model for the diagnosis of diabetes.

Not only were the boundary diagnostic criteria defined, but the technique of parallel calculations of heterogeneous data to calculate the probability of establishing a diagnosis was also validated. Simultaneously, not only have diagnostic boundary criteria been developed, but also the method of simultaneous calculations of different forms of data to determine the probability of establishing a diagnosis has been validated.

In the Hadoop cluster, a non-relational database that processes unstructured data in multiple formats has been built. In the Hadoop cluster, an unreporting database is generated and utilized to process unstructured data in multiple formats. The HDFS distributed file system has been developed. Created a distributed HDFS file system.

In order to set up the entire cluster, Hadoop provides for the installation of a Unix-like operating system. After the analysis, as well as considering all the required configurations, the choice was stopped on the CentOS 7 distribution [38]. Install Hadoop components using the Apache Ambari server Hortonworks Data Platform (HDP) 3.1 version. Ambari includes a REST API and a browser-based management interface. After starting ambari-server, it is necessary to connect to the server using the http protocol. Then let's perform all the procedures to configure the server.

The main task of all these tools is to work with various types of data, and relational and non-relational database management systems are also used. Thus, the connection of various databases for unstructured and structured data and their interaction with each other [39]. Also, creating on the system platform, in the form of a user-friendly interface, where various operations will be carried out:

– various patient credentials are entered;

– create patient medical records;

– analyzes are introduced where each analysis consists of different numerical and text parameters;

– if the diagnosis is confirmed, then for hospitalization and treatment, an additional 12 types of tests and examinations must be taken;

– connecting various medical databases to hold complete information on all types of diabetes mellitus, as well as to keep the endocrinologist informed of the latest important publications on diabetes. Since the doctor must have all the relevant information about treatment methods.

Our system stores all the patient's medical history so that the doctor can access it remotely. The created database must automatically create replication and be fault-tolerant, and have access to connected medical databases. In addition, students at medical universities can use the system as a simulator to improve their knowledge of diagnosing diabetes.

Node JS was used to build the server [40]. JavaScript is built as an event loop, and Node.js is built from JavaScript [41]. Node.js can handle multiple requests and will act as a client to third-party services, performing only one thread [42]. JavaScript performs the action on the client-side and Node on the server. With Node.js, it is possible to write full-fledged applications. Node can work with external libraries, call commands from JavaScript code, and act as a web server. C Node.js is easier to scale. With thousands of users connecting to the server at the same time, Node.js works asynchronously; that is, it prioritizes and allocates resources more competently.

Our developed product will unite all doctors in Kazakhstan and patients, which will help to better diagnose diseases in patients and help doctors rally to determine the diagnosis and determine the course of treatment. Node.js was chosen because it is necessary users to be able to log into the system and upload their data without difficulty. When building a server on Node.js, it is possible to use ready-made libraries. For example, express [43] is needed to listen to the server, jsonwebtoken [44] is needed to create a token, and MongoDB [45] to create and load unstructured data.

System stores the entire patient's medical history so that the doctor can load unstructured information of various formats into MongoDB, a database connection is required. The request will be next:

```
const objectId=require("mongodb').ObjectID;
const mongodb=require('mongodb')
const MongoClient=mongodb.MongoClient
const connection URL='mongodb://127.0.0.1:27017'
```

When sending analyses, data is downloaded to the server, and data paths are written to the database. The following is an example download that sends three analyses and a username.

```
const fname=req.body.name;
const imagePathUZI=(req.files['image_UZI'][0].destination+»/»+req.files['image_UZI'][0].originalname).substr(9);
const imagePathUZI2=(req.files['image_UZI2'][0].destination+»/»+req.files['image_UZI2'][0].originalname).substr(9);
const imagePathDatchik=(req.files['image_Datchik'][0].destination+»/»+req.files['image_Datchik'][0].originalname).substr(9);

MongoClient.connect(connectionURL, {useNewUrlParser: true}, (error, client)=>{
if (error) {
console.log('Unable to connect with db')
}
const db=client.db(databaseName)
db.collection('user').insertOne({
name: fname,
image_UZI: imagePathUZI,
image_UZI2: imagePathUZI2,
image_Datchik: imagePathDatchik
}, (error, result)=>{
if (error) {
console.log('Error')
}
console.log(result.ops); })
})
})
```

For structured data, user registration was used. It is implemented using mongoose. Then connect to mongoose as follows:

```
mongoose.connect(''mongodb://127.0.0.1:27017/fullstack_2009', {
use New Url Parser: true,
use Unified Topology: true,
use Create Index: true
})
.then(()=>console.log('MongoDB connected.'))
.catch(error=>console.log(error))
```

then register the user where it is necessary to enter email and password

```
const mongoose = require('mongoose')
const Schema = mongoose.Schema
const userSchema = new Schema({
email: {
type: String,
required: true,
unique: true
},
password: {
type: String,
required: true
}
})
module.exports=mongoose.model('users', userSchema)
```

Registration data is entered under the type *String* and each email must be unique, that is, more than one person cannot register with one mail. *Required* is necessary for the user to enter it without fail. After a systematic installation of MongoDB open-source data management system for unstructured data, a data warehouse window appears as shown in Fig. 1.

After the authorization process for the doctor in the main menu of the system, the downloaded data appears in the database as shown in Fig. 2.

Then, after the user authorization process, the loaded data appears in the database in the main menu of the system as shown in Fig. 3.

The list of tests that must be entered by the user after registration in the system for diagnosing diabetes as shown in Fig. 4.

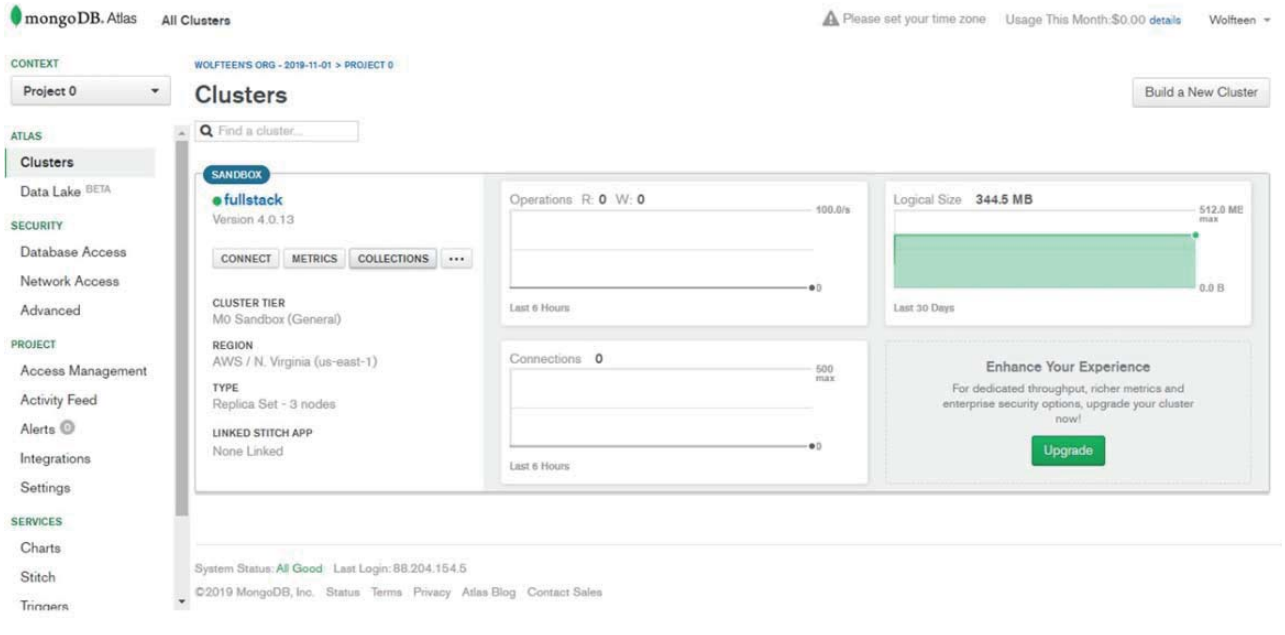

Fig. 1. MongoDB database structure

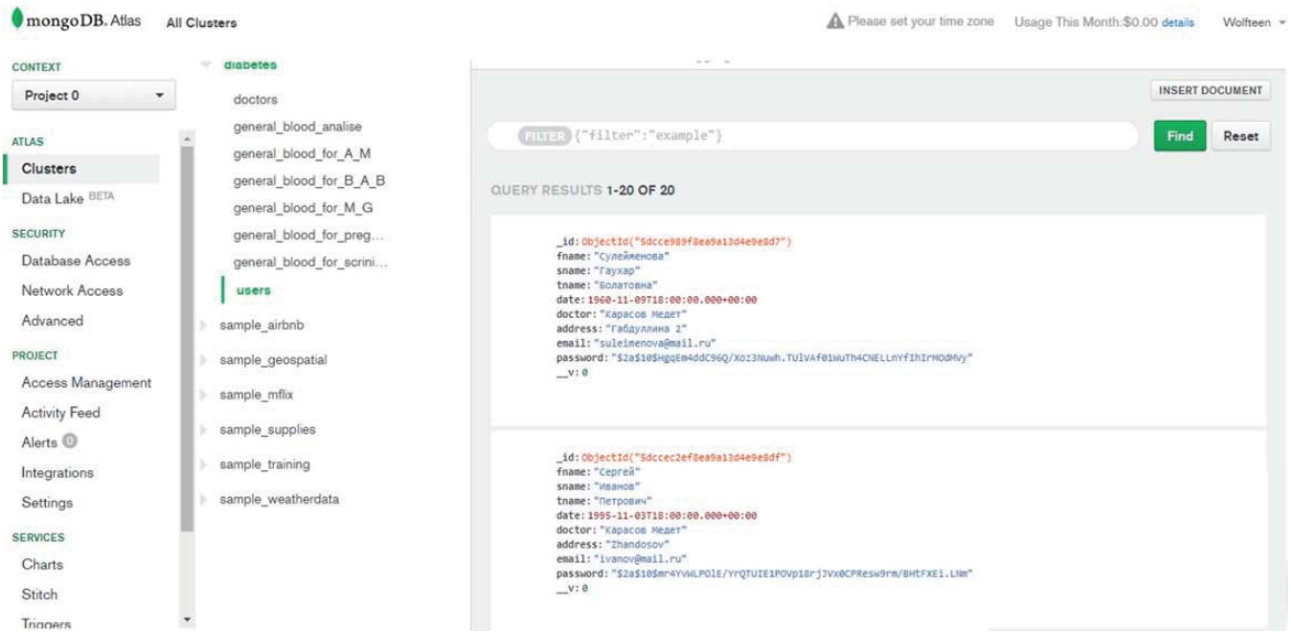

Fig. 2. The registration data of the doctor

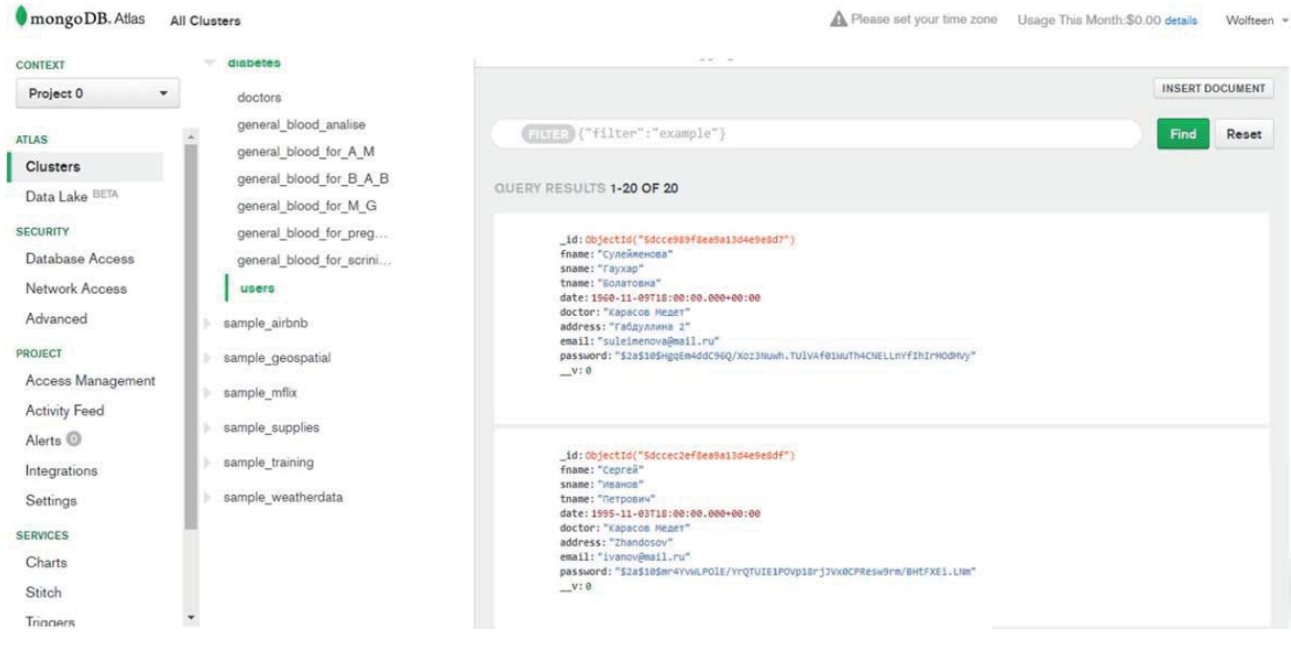

Fig. 3. The registration data of the user

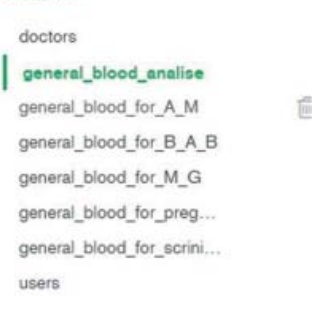

Fig. 4. The list of analyzes in MongoDB

All types of analyzes must be entered and saved. Next, the numerical parameters of the analysis "General blood test" are downloaded, where these data are loaded as shown in Fig. 5.

Then the numerical parameters of the analysis "Biochemical blood analysis" are loaded, where these data are loaded into MongoDB in as shown in Fig. 6.

Next steps contain processes of loading the numerical parameters of the different analysis and unstructured data such as images of ultrasound of lower limb vessels, ultrasound of the abdominal cavity and kidney, fluorography, electrocardiogram.

The downloaded parameters and images of various formats are stored in a database; where after loading the attending doctor can monitor the patient's condition remotely.

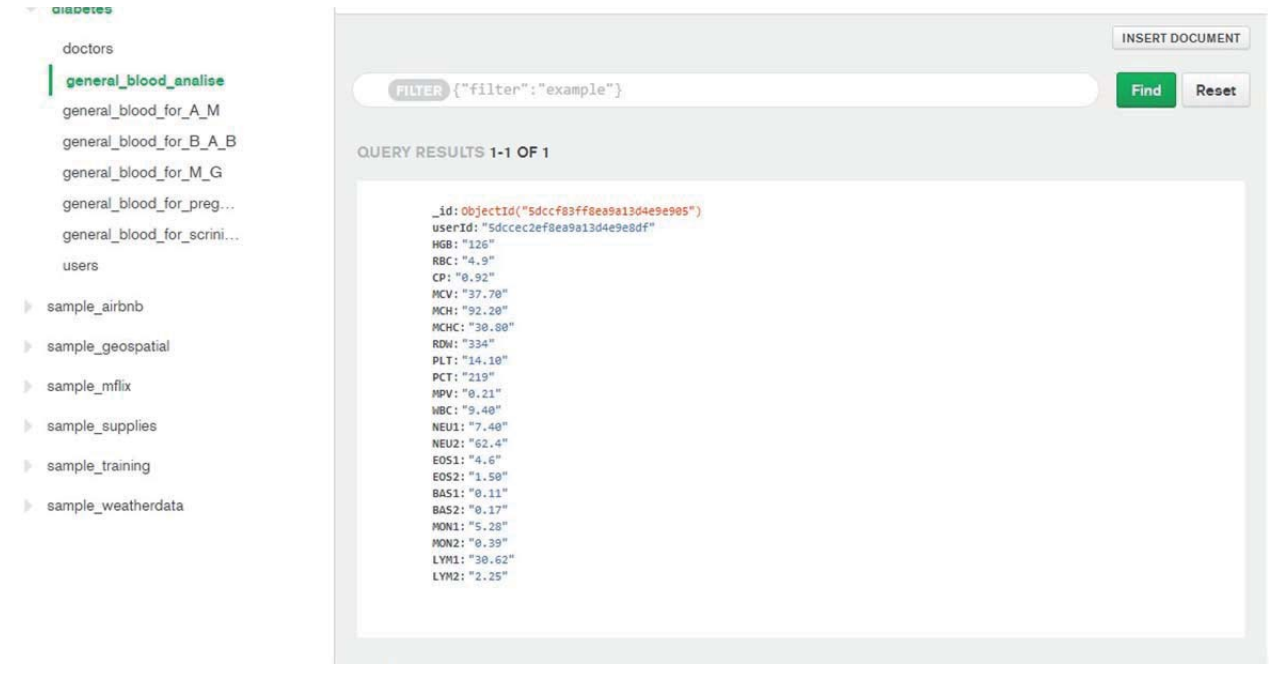

Fig. 5. The numerical data analysis "General blood test"

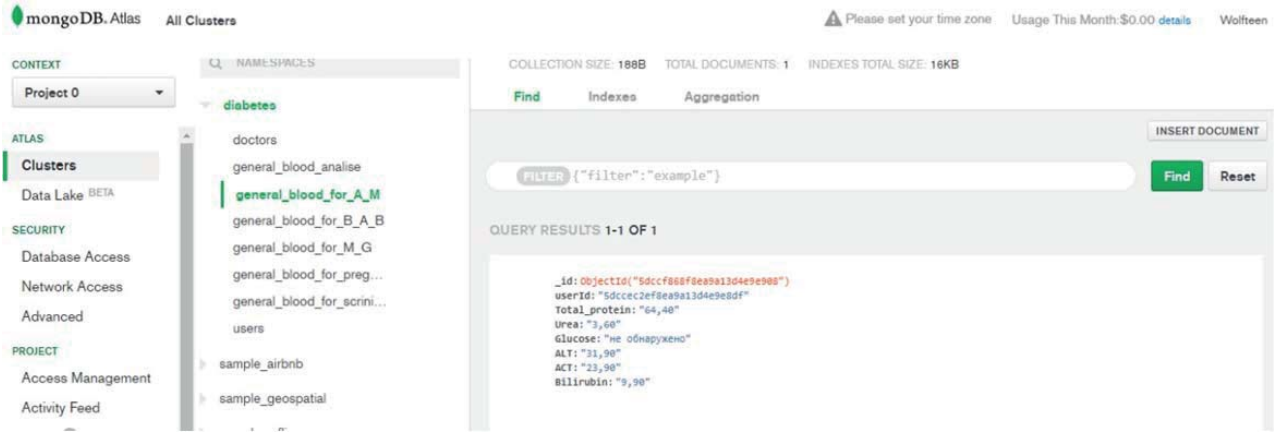

Fig. 6. The numerical data of the analysis "Biochemical analysis of blood"

5. 3. Development of an information system platform

After installing the necessary components of Big Data technology tools, the next main task is to create and connect various databases for unstructured and structured data and their interaction with each other in order to create a platform, in the form of a user-friendly interface, where various operations for diagnosing diabetes will be carried out.

After registration and establishment of an account for the patient, a medical history is created in the database. To authorize employees (doctors) it is necessary to be registered in the technical department of a medical institution, where developers will issue a login and password in advance.

To diagnose diabetes in a patient, it is necessary to direct the patient to be tested in a clinical diagnostic laboratory.

After checking all the received tests and examinations, the endocrinologist officially diagnoses the patient with diabetes and decides on the appointment of treatment and the possibility of hospitalization, if necessary. Downloaded parameters and images of various formats are stored in a database; where after loading the attending physician can monitor the patient's condition remotely.

Electronic files of FreeStyleLibre sensors can be downloaded from the storage card of the carrier, and then downloaded to the database. All downloaded files were locally located in the system repositories and did not cause any difficulties in interacting with them.

This method for diagnosing diabetes was evaluated at the Clinical Diagnostic Center of the International Kazakh-Turkish University named Yassavi in Turkestan, Kazakhstan. The findings of the information system demonstrated the accuracy of the system's chosen operating principle and its performance under real-world settings.

6. Discussion of results of the developed system to assist the endocrinologist in the diagnosis of diabetes

According to the study, the organization of endocrinological care is aimed at reducing the morbidity of patients with endocrine diseases. Using statistical regression anal-

ysis methods, the results published in this study [46] were obtained, which makes it possible to predict the number of patients in the future and to purchase insulin efficiently.

Based on the mathematical apparatus (1)–(13), a list of standard medical analyses with which it is possible to diagnose diabetes is shown. At the same time, not only boundary criteria of diagnostics have been established, but the mechanism of parallel calculations of various types of data for determining the probability of making a diagnosis has been substantiated. An information system platform has been developed and configured to manage and control the Hadoop cluster. The conversion database, which uses and processes unstructured data from various formats, has been effective. The high prevalence, lack of problem-oriented information technology for diagnosis and lack of guidance on diabetes highlight the importance of this work. In this context, there is an urgent need not to stop such research and to continue to develop information systems to assist endocrinologists in diagnosing diabetes. The use of modern information technology in the health system allows to improve the quality of the medical services provided. During the research the authors collaborated with «Kazakhstan Society for the Study of Diabetes» [47]. The main advantage of this kind of research is the development of further health care for the population. Similar systems developed earlier did not cover the full amount of heterogeneous information that is collected in the diagnosis and treatment of diabetes mellitus, nor did they use Big Data technologies and cloud services as tools [48, 49].

As a result of this study, a certificate was obtained on entering information into the state register of rights to objects protected by copyright of the Republic of Kazakhstan No. 9730 dated May 11, 2019.

A significant drawback of the study is the average level of communication with public health institutions for further consultation. Statistics show an increase in diabetes mellitus patients nationwide, which is a matter of great concern. Let's believe that it is necessary to develop the study in the future because the relevance of the direction is very high.

7. Conclusions

1. A mathematical model for diagnosing diabetes mellitus was created. The model is based on an unstructured set of medical analysis data and in order to improve accuracy of prediction, the optimizations of unstructured data took place.

2. Based on the created mathematical model, the system with Big Data tools was developed. The created system allows to store the patients' medical records which helps to conduct diagnosis of diabetes mellitus.

3. In order to interact with the system, the web-platform was developed. It allows users to add and manage patients' medical records in a user-friendly manner. By the end, the data is analyzed in the Big Data system and diagnosis results are provided to the endocrinologist.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Acknowledgements

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP13068032).

References

1. Beam, A. L., Kohane, I. S. (2016). Translating Artificial Intelligence Into Clinical Care. JAMA, 316 (22), 2368–2369. doi: https://doi.org/10.1001/jama.2016.17217
2. Eghbali-Zarch, M., Tavakkoli-Moghaddam, R., Esfahanian, F., Sepehri, M. M., Azaron, A. (2018). Pharmacological therapy selection of type 2 diabetes based on the SWARA and modified MULTIMOORA methods under a fuzzy environment. Artificial Intelligence in Medicine, 87, 20–33. doi: https://doi.org/10.1016/j.artmed.2018.03.003
3. Contreras, I., Vehi, J. (2018). Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. Journal of Medical Internet Research, 20 (5), e10775. doi: https://doi.org/10.2196/10775
4. Fico, G., Arredondo, M. T., Protopappas, V., Georgia, E., Fotiadis, D. (2014). Mining Data When Technology Is Applied to Support Patients and Professional on the Control of Chronic Diseases: The Experience of the METABO Platform for Diabetes Management. Data Mining in Clinical Medicine, 1246, 191–216. doi: https://doi.org/10.1007/978-1-4939-1985-7_13
5. Galetsi, P., Katsaliaki, K. (2019). A review of the literature on big data analytics in healthcare. Journal of the Operational Research Society, 71 (10), 1511–1529. doi: https://doi.org/10.1080/01605682.2019.1630328
6. Dash, S., Shakyawar, S. K., Sharma, M., Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-019-0217-0
7. Dautov, R., Distefano, S., Buyya, R. (2019). Hierarchical data fusion for Smart Healthcare. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-019-0183-6
8. Mazumdar, S., Seybold, D., Kritikos, K., Verginadis, Y. (2019). A survey on data storage and placement methodologies for Cloud-Big Data ecosystem. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-019-0178-3
9. Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W. et al. (2018). Deep Learning and Its Applications in Biomedicine. Genomics, Proteomics & Bioinformatics, 16 (1), 17–32. doi: https://doi.org/10.1016/j.gpb.2017.07.003
10. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19 (6), 1236–1246. doi: https://doi.org/10.1093/bib/bbx044

11. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S. et al. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2 (4), 230–243. doi: https://doi.org/10.1136/svn-2017-000101
12. Bote-Curiel, L., Muñoz-Romero, S., Gerrero-Curieses, A., Rojo- lvarez, J. L. (2019). Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners. Applied Sciences, 9 (11), 2331. doi: https://doi.org/10.3390/app9112331
13. Sabitha, M. S., Vijayalakshmi, S., Sre, R. R. (2015). Big Data-literature survey. International Journal for Research in Applied Science and Engineering Technology, 3, 318–320.
14. Cichosz, S. L., Johansen, M. D., Hejlesen, O. (2015). Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications. Journal of Diabetes Science and Technology, 10 (1), 27–34. doi: https://doi.org/10.1177/1932296815611680
15. Sneha, N., Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-019-0175-6
16. Tang, V., Choy, K. L., Ho, G. T. S., Lam, H. Y., Tsang, Y. P. (2019). An IoMT-based geriatric care management system for achieving smart health in nursing homes. Industrial Management & Data Systems, 119 (8), 1819–1840. doi: https://doi.org/10.1108/imds-01-2019-0024
17. Zhang, R., Simon, G., Yu, F. (2017). Advancing Alzheimer's research: A review of big data promises. International Journal of Medical Informatics, 106, 48–56. doi: https://doi.org/10.1016/j.ijmedinf.2017.07.002
18. Khanra, S., Dhir, A., Islam, A. K. M. N., M ntym ki, M. (2020). Big data analytics in healthcare: a systematic literature review. Enterprise Information Systems, 14 (7), 878–912. doi: https://doi.org/10.1080/17517575.2020.1812005
19. Kamble, S. S., Gunasekaran, A., Goswami, M., Manda, J. (2018). A systematic perspective on the applications of big data analytics in healthcare management. International Journal of Healthcare Management, 12 (3), 226–240. doi: https://doi.org/10.1080/20479700.2018.1531606
20. Soleimani-Roozbahani, F., Rajabzadeh Ghatari, A., Radfar, R. (2019). Knowledge discovery from a more than a decade studies on healthcare Big Data systems: a scientometrics study. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-018-0167-y
21. Shahbaz, M., Gao, C., Zhai, L., Shahzad, F., Hu, Y. (2019). Investigating the adoption of big data analytics in healthcare: the moderating role of resistance to change. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-019-0170-y
22. Hariri, R. H., Fredericks, E. M., Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-019-0206-3
23. Abouelmehdi, K., Beni-Hessane, A., Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. Journal of Big Data, 5 (1). doi: https://doi.org/10.1186/s40537-017-0110-7
24. Fatt, Q. K., Ramadas, A. (2018). The Usefulness and Challenges of Big Data in Healthcare. Journal of Healthcare Communications, 3 (2). doi: https://doi.org/10.4172/2472-1654.100131
25. Kruse, C. S., Goswamy, R., Raval, Y., Marawi, S. (2016). Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Medical Informatics, 4 (4), e38. doi: https://doi.org/10.2196/medinform.5359
26. Landset, S., Khoshgoftaar, T. M., Richter, A. N., Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2 (1). doi: https://doi.org/10.1186/s40537-015-0032-1
27. El aboudi, N., Benhlima, L. (2018). Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. Advances in Bioinformatics, 2018, 1–10. doi: https://doi.org/10.1155/2018/4059018
28. Gajanand, S., Ashutosh, K., Himanshu, S., Ashok, K. S., Priyanka, Dogiwal, S. R. (2020). Diabetes Data Prediction in healthcare Using Hadoop over Big Data. European Journal of Molecular & Clinical Medicine, 7 (4), 1423–1432.
29. Ellaway, R. H., Pusic, M. V., Galbraith, R. M., Cameron, T. (2014). Developing the role of big data and analytics in health professional education. Medical Teacher, 36 (3), 216–222. doi: https://doi.org/10.3109/0142159x.2014.874553
30. Bellazzi, R., Dagliati, A., Sacchi, L., Segagni, D. (2015). Big Data Technologies: New Opportunities for Diabetes Management. Journal of Diabetes Science and Technology, 9 (5), 1119–1125. doi: https://doi.org/10.1177/1932296815583505
31. Kamel Boulos, M. N., Koh, K. (2021). Smart city lifestyle sensing, big data, geo-analytics and intelligence for smarter public health decision-making in overweight, obesity and type 2 diabetes prevention: the research we should be doing. International Journal of Health Geographics, 20 (1). doi: https://doi.org/10.1186/s12942-021-00266-0
32. Rakhmetulayeva, S. B., Duisebekova, K. S., Mamyrbekov, A. M., Kozhamzharova, D. K., Astaubayeva, G. N., Stamkulova, K. (2018). Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis. Procedia Computer Science, 130, 231–238. doi: https://doi.org/10.1016/j.procs.2018.04.034
33. Miah, S. J., Camilleri, E., Vu, H. Q. (2021). Big Data in Healthcare Research: A survey study. Journal of Computer Information Systems, 62 (3), 480–492. doi: https://doi.org/10.1080/08874417.2020.1858727
34. Carnevale, A., Tangari, E. A., Iannone, A., Sartini, E. (2021). Will Big Data and personalized medicine do the gender dimension justice? AI & SOCIETY. doi: https://doi.org/10.1007/s00146-021-01234-9
35. Apache Hadoop 2.7.0 Documentation. Available at: https://hadoop.apache.org/docs/r2.7.0/ Last accessed: 11.05.2020
36. Apache Ambari. Available at: https://ambari.apache.org/ Last accessed: 11.05.2020
37. White, T. (2012). Hadoop: The Definitive Guide. Oreilly & Associates Inc.
38. The CentOS Project. Download CentOS. Available at: https://www.centos.org/download/ Last accessed: 11.05.2020

39. Rakhmetulayeva, S. B., Duisebekova, K. S., Kozhamzharova, D. K., Aitimov, M. Zh. (2021). Pollutant transport modeling using Gaussian approximation for the solution of the semi-empirical equation. Journal of Theoretical and Applied Information Technologythis link is disabled, 99 (8), 1730–1739.
40. About Node.js. Available at: https://nodejs.org/en/about/ Last accessed: 11.05.2020
41. JavaScript.home. Available at: https://www.javascript.com/ Last accessed: 11.05.2020
42. Hezbullah, Sh. (2017). Node.js Challenges in Implementation. Global Journal of Computer Science and Technology: E Network.
43. Fast, unopinionated, minimalist web framework for Node.js. Available at: https://expressjs.com/ Last accessed: 11.05.2020
44. An implementation of JSON Web Tokens. Available at: https://www.npmjs.com/package/jsonwebtoken Last accessed: 11.05.2020
45. Horowitz, E. (2018). Introducing the Best Database for Modern Applications. Available at: https://www.mongodb.com/blog/post/introducing-the-best-database-for-modern-applications Last accessed: 11.05.2020
46. Mukasheva, A., Saparkhojayev, N., Akanov, Z., Apon, A., Kalra, S. (2019). Forecasting the Prevalence of Diabetes Mellitus Using Econometric Models. Diabetes Therapy, 10 (6), 2079–2093. doi: https://doi.org/10.1007/s13300-019-00684-1
47. Kazakhstan Society for the Study of Diabetes. Available at: https://www.kssd.site/
48. Mukasheva, A., Yedilkhan, D., Zimin, I. (2021). Uploading Unstructured Data to MONGODB Using the NoSQLBooster Tool. 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST). doi: https://doi.org/10.1109/sist50301.2021.9465930
49. Mukasheva, A., Iliev, T., Balbayev, G. (2020). Development of the Information System Based on BigData Technology to Support Endocrinologist-Doctors. 2020 7th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE). doi: https://doi.org/10.1109/eeae49144.2020.9278971