3-бөлім **Раздел 3** **Section 3**

Информатика **Информатика** **Computer Science**

IRSTI 27.41.23 DOI: https://doi.org/10.26577/JMMCS1291202610

A.Kh. Nishanov[1*], F.Z. Mengturayev[2], F.F. Ollamberganov[1], U.B. Allayarov[3], M.A. Khasanova[4], G.T. Doniyorova[2]

[1]Tashkent university of information technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

[2]Denau Institute of Entrepreneurship and Pedagogy, Denau, Uzbekistan

[3]Termez Branch of the Tashkent Medical Academy, Termez, Uzbekistan

[4]Tashkent Medical Academy, Tashkent, Uzbekistan

[*]e-mail: nishanov_akram@mail.ru

AN ALGORITHM FOR CREATING A SEMI-SYNTHETIC DATASET FOR DIABETES

Recent advances in the areas of artificial intelligence and machine learning have opened up new avenues for enhancing the practice of medical diagnosis. However, researchers face difficulties in accessing quality datasets because of the sensitive nature of real clinical data related to diabetes mellitus. The main objective of this research is to introduce an algorithm intended to generate a semi-synthetic training dataset aimed at improving classification accuracy for diabetes mellitus, particularly type 1 and type 2 diabetes. An algorithm to generate semi-synthetic diabetes data by statistically analyzing clinical attributes from real patient records. For improving the generation of synthetic samples without altering the properties of the original data, a similarity-based approach focusing on class-object relations was used. The approach used successfully generated synthetic data instances that preserved the inherent structure and distribution typical of real patient data. A similarity-based mechanism ensured the relevance of the created instances, while the study outlined a sequence of steps intended to improve the quality of synthetic datasets. The proposed algorithm creates artificial datasets for diabetes classification with patient data protection. This methodology led to the rise in intra-class similarity from 76.18% to 82.93%, which in turn enhanced the diagnostic accuracy of artificial intelligence-based models.

Key words: diabetes prediction, semi-synthetic dataset, Data augmentation, machine learning algorithms, synthetic medical data, generative model, object similarity.

А.Х. Нишанов[1*], Ф.З. Менгтураев[2], Ф.Ф. Олламберганов[1], У.Б. Аллаяров[3], М.А. Хасанова[4], Г.Т. Дониёрова[2]

[1]Мұхаммед әл-Хорезми атындағы Ташкент ақпараттық технологиялар университеті, Ташкент, Өзбекстан

[2]Денау кәсіпкерлік және педагогика институты, Денау, Өзбекстан

[3]Ташкент медицина академиясының Термез филиалы, Термез, Өзбекстан

[4]Ташкент медицина академиясы, Ташкент, Өзбекстан

[*]e-mail: nishanov_akram@mail.ru

Қант диабетіне арналған жартылай синтетикалық деректер жиынтығын құрастыру алгоритмі

Жасанды интеллект пен машиналық оқыту саласындағы соңғы жетістіктер медициналық диагностика тәжірибесін жетілдірудің жаңа мүмкіндіктерін ашты. Алайда, зерттеушілер қант диабетіне қатысты нақты клиникалық деректердің құпия сипатына байланысты сапалы мәліметтер жиынтығына қол жеткізуде қиындықтарға тап болуда. Бұл зерттеудің басты мақсаты - қант диабетін, әсіресе 1-ші және 2-ші типтегі диабетті жіктеу дәлдігін арттыруға бағытталған жартылай синтетикалық оқу деректер жиынтығын құру алгоритмін ұсыну. Науқастардың нақты жазбаларынан алынған клиникалық көрсеткіштерді статистикалық талдау арқылы жартылай синтетикалық диабет деректерін құру алгоритмі әзірленді.

Бастапқы деректердің қасиеттерін өзгертпей синтетикалық үлгілерді жасауды жақсарту үшін сыныпнысан қатынастарына негізделген ұқсастық тәсілі қолданылды. Бұл тәсіл науқастардың нақты деректеріне тән ішкі құрылымы мен таралуын сақтайтын синтетикалық деректер үлгілерін сәтті жасауға мүмкіндік берді. Ұқсастыққа негізделген механизм жасалған үлгілердің өзектілігін қамтамасыз етті, ал зерттеу синтетикалық деректер жиынтығының сапасын арттыруға арналған қадамдар тізбегін ұсынды. Ұсынылған алгоритм науқастар туралы деректерді қорғай отырып, диабетті жіктеу үшін жасанды деректер жиынтығын құрады. Бұл әдістеме класішілік ұқсастықты 76.18%-дан 82,93%-ға дейін арттыруға әкелді, бұл өз кезегінде жасанды интеллектке негізделген модельдердің диагностикалық дәлдігін жоғарылатты.
Түйін сөздер: қант диабетін болжау, жартылай синтетикалық деректер жинағы, деректерді көбейту, машиналық оқыту алгоритмдері, синтетикалық медициналық деректер, генеративтік модель, объектілердің ұқсастығы

А.Х. Нишанов[1*], Ф.З. Менгтураев[2], Ф.Ф. Олламберганов[1], У.Б. Аллаяров[3], М.А. Хасанова[4], Г.Т. Дониёрова[2]

[1]Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Ташкент, Узбекистан

[2]Денауский институт предпринимательства и педагогики, Денау, Узбекистан

[3]Термезский филиал Ташкентской медицинской академии, Термез, Узбекистан

[4]Ташкентская медицинская академия, Ташкент, Узбекистан

[*]e-mail: nishanov_akram@mail.ru

Алгоритм для создания полусинтетического набора данных по диабету

Последние достижения в области искусственного интеллекта и машинного обучения открыли новые возможности для совершенствования практики медицинской диагностики. Однако исследователи сталкиваются с трудностями в доступе к качественным наборам данных из-за конфиденциальности реальных клинических данных, связанных с сахарным диабетом. Основной целью данного исследования является разработка алгоритма, предназначенного для генерации полусинтетического обучающего набора данных, направленного на повышение точности классификации сахарного диабета, в частности, диабета 1 и 2 типа. Был разработан алгоритм для генерации полусинтетических данных о диабете путем статистического анализа клинических атрибутов из реальных записей пациентов. Для улучшения генерации синтетических выборок без изменения свойств исходных данных был использован подход, основанный на сходстве и ориентированный на отношения между классами и объектами. Этот подход успешно сгенерировал примеры синтетических данных, которые сохранили присущую структуру и распределение, типичные для реальных данных пациентов. Механизм, основанный на сходстве, обеспечил релевантность созданных примеров, в то время как в исследовании была определена последовательность шагов, направленных на повышение качества синтетических наборов данных. Предложенный алгоритм создает искусственные наборы данных для классификации диабета с защитой данных пациентов. Данная методика привела к увеличению внутриклассового сходства с 76.18% до 82.93%, что, в свою очередь, повысило диагностическую точность моделей на основе искусственного интеллекта.
Ключевые слова: Прогнозирование диабета, полусинтетический набор данных, аугментация данных, алгоритмы машинного обучения, синтетические медицинские данные, генеративная модель, сходство объектов.

1 Introduction

In recent years, machine learning architectures have been widely used in various fields, including medicine [1-5]. Artificial intelligence systems enable improvements in medical diagnostics, early disease detection, and personalized patient care. However, the effectiveness of deep learning algorithms is directly related to the volume and quality of data. In particular, since medical data contains personal information, maintaining its confidentiality is a significant challenge.

Diabetes mellitus is a global problem today, with its prevalence increasing year by year. According to the World Health Organization, 108 million people were affected by this disease in 1980, and by 2014 this figure had reached 422 million [6]. The International Diabetes Federation (IDF) projects that this number could rise to 783 million people by 2045. Diabetes is also widespread in Uzbekistan, with 6.3% of the population aged 20-79 suffering from this disease, according to XDF data for 2021.

The use of synthetic data is of great importance in medicine. Synthetic data serves to create a safe environment for analysis and machine learning without disclosing personal data of real patients. In creating synthetic data, statistical models, probabilistic analysis, machine learning, and deep learning methods are employed [7-9]. Synthetic data is divided into three types: fully synthetic, semi-synthetic, and hybrid. Semi-synthetic data is based on generating new data while preserving the statistical characteristics of real patient data. The hybrid approach is used to create large volumes of synthetic data based on small real databases.

This article examines 6 classes and 140 objects for type 1 diabetes, and 16 classes and 340 objects for type 2 diabetes. The small number of objects belonging to certain classes reduces the reliability of the classification process. Therefore, it is crucial to improve the accuracy of the analysis results by increasing the number of objects using synthetic data. This approach aims to enhance the efficiency of artificial intelligence algorithms and improve the diagnosis of diabetes.

2 Related works

Several studies have been conducted on increasing synthetic data for early detection of diabetes mellitus, including P.Sampath, G.Elangovan, K.Ravichandran, and others, who proposed an approach to predicting diabetes using the SMOTE synthetic sampling technique using the ensemble machine learning technique [10]. Among these, P. Sampath, G. Elangovan, K. Ravichandran, and others proposed an approach to predicting diabetes using the SMOTE synthetic sampling technique in combination with ensemble machine learning [10]. Z. Tagmatova, A. Abdusalomov, R. Nasimov, and others proposed a rule-based blending method to represent the quality of the data distribution histogram for assessing the similarity between different classes [11]. Casey Greene suggests early detection of diabetes through data generation on the application of artificial intelligence and machine learning methods in medicine.

There have been a small study on the synthetic multiplication of tabular data presented in the nominal space, reflecting the main symptoms of diabetes mellitus. Moreover, due to the high complexity of hybrid algorithms designed for tabular data existing only in nominal space when classifying diabetes mellitus, this study proposes a relatively simple approach.

3 Problem statement

Suppose that in the N-dimensional space of nominal features, datasets of breast cancer cases $x_{p1}, x_{p2}, \ldots, x_{pm_p} \in X_p$, $p=\overline{1,r}$ is given. Each object in the datasets $x_{pi} = \left(x_{pi}^1, x_{pi}^2, \ldots, x_{pi}^N\right)$, $i = \overline{1, m_p}$, represents a patient in the space of N -dimensional nominal features. Here, x_{pi}, is read as the i-th object of the p - class, N denotes the number of features that

comprise the objects, and m_p denotes the number of objects in the p-class. So, patients $x_{pi} = \left(x_{pi}^1, x_{pi}^2, \ldots, x_{pi}^N\right)$ constitute the $i\left(i = \overline{1, m_p}\right)$ objects of the p -class [12-16].

Problem. It is necessary to generate synthetic training samples using the general sample $x_i \in X, i = \overline{1, M}$ in a given N-dimensional space of nominal features, adding K new objects to each class. In doing so, it is required that the degree of similarity between the objects of classes $x_{p1}, x_{p2}, \ldots, x_{pm_p} \in X_p, p = \overline{1, r}$ be no less than $\delta > 55$. Here, class X_p consists of m_p objects $x_{p1}, \ldots, x_{p_{m_p}}$, and $X = \bigcup_{p=1}^{r} X_p$.

Let the quantity indicating the similarity of objects in the space of nominal symbols be determined by $\rho^j\left(x_{pi,}x_{pq}\right)$ and calculated by (1):

$$\rho_{pi}^j\left(x_{pi,}x_{pq}\right) = \begin{cases} 1, & \text{if} \quad \left(x_{pi}^j - x_{pq}^j\right) = 0; \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $p = \overline{1, r};\ i \neq q = \overline{1, m_p}; j = \overline{1, N}$.

The expressed quantities (1) are the parameters of the vector, which is expressed in the following form $\rho_{pi}\left(x_{pi,}x_{pq}\right) = \left(\rho_{pi}^1\left(x_{pi,}x_{pq}\right), \rho_{pi}^2\left(x_{pi,}x_{pq}\right), \ldots, \rho_{pi}^N\left(x_{pi,}x_{pq}\right)\right.$.

So, for any two objects x_{pi} and x_{pq} of class p in the bool vector space, there exists a bool vector $\rho_{pi}\left(x_{pi,}x_{pq}\right) = \left(\rho_{pi}^1\left(x_{pi,}x_{pq}\right), \rho_{pi}^2\left(x_{pi,}x_{pq}\right), \ldots, \rho_{pi}^N\left(x_{pi,}x_{pq}\right)\right.$. The components of this vector indicate the similarity or difference between the two objects with respect to the considered feature. If $\rho_{pi}^j\left(x_{pi,}x_{pq}\right) = 1$, then the objects x_{pi} and x_{pq} are similar in terms of the j- feature; otherwise, that is, if $\rho_{pi}^j\left(x_{pi,}x_{pq}\right) = 0$, it means they are not similar with respect to the j- feature [8-9].

Let's list the following stages of solving the above-mentioned problem of creating a semi-synthetic datasets:

1. First, the objects of the training set $x_i \in X, i = \overline{1, M}$, are divided into r classes. That is, based on the RCM program, the objects $x_{p1}, x_{p2}, \ldots, x_{pm_p} \in X_p, p = \overline{1, r}$ are separated into classes;

2. Based on the equations (1) given above, all $p = \overline{1, r};\ i \neq q = \overline{1, m_p}; j = \overline{1, N}$ for all parameters of the vector $\rho_{pi}\left(x_{pi,}x_{pq}\right)$. That is $\rho_{pi}\left(x_{pi,}x_{pq}\right) = \left(\rho_{pi}^1\left(x_{pi,}x_{pq}\right), \rho_{pi}^2\left(x_{pi,}x_{pq}\right), \ldots, \rho_{pi}^N\left(x_{pi,}x_{pq}\right)\right.$ vector symbols are calculated for $p = \overline{1, r};\ i \neq q = \overline{1, m_p}; j = \overline{1, N}$;

3. The position of the i-object in the optional p-class in the remaining set of $m_p - 1$ objects of this class is evaluated as follows [10,11]:

$$\Gamma_{pi}\left(x_{pi,}X_p\right) = \frac{1}{m_p - 1} \sum_{q=1}^{m_p - 1} \sum_{j=1}^{N} \rho^j\left(x_{pi,}x_{pq}\right),\ p = \overline{1,r}\ ; i = \overline{1,m_p}; i \neq q.$$

4. The general grade of the arbitrary p-class is calculated based on the criterion $\Gamma_p\left(X_p\right) = \frac{1}{m_p} \sum_{i=1}^{m_p} \Gamma_{pi}\left(x_{pi,}X_p\right)$, $p = \overline{1,r}$. The degree of similarity of their objects is evaluated as follows:

$$\nu_p\left(X_p\right) = = \frac{\Gamma_p\left(X_p\right) * 100\%}{N},\ p = \overline{1,r}$$

5. When synthetically creating a new object $\overline{x}$ for classes $X_p, p = \overline{1,r}$, objects created with the preservation of essential features that ensure an increase in the average similarity level $\nu_p(X_p)$ of objects in this class are added to this class.

6. If the number of objects in class p is $m_p = 1$, then when creating a synthetic object, it is created while preserving the features that distinguish the object of the class from objects of other classes. It is required to satisfy $\nu_p(X_p) > \delta$.

These calculations are performed for all objects of the semi-synthetic datasets $\overline{x}_{pi}$, $p = \overline{1,r}$ $i = \overline{1,K}$; and the obtained new classes $x_{p1}, x_{p2}, \ldots, x_{pm_p} \in X_p, p = \overline{1,r}$, produce a semi-synthetic datasets. Here, $m_p = m_p + K$.

4 Diabetes symptoms and real dataset

In this study, 1106 medical records of patients with diabetes mellitus were examined, of which 480 were deemed suitable in collaboration with doctors from the regional endocrinology clinic. Among these, 340 were classified as type 2 diabetes, and 140 as type 1. For type 2 diabetes, 50 susceptible signs of 19 symptoms were identified. Based on these objects and symptoms, they were categorized into 16 classes depending on the type of disease. For type 1 diabetes mellitus, 10 symptoms were divided into 12 perceivable features and 7 classes. When classifying diabetes mellitus using the RCM program, the degree of object similarity was divided into classes up to 65%. Symptoms and signs for type 1 and type 2 diabetes are presented in Tables 1-2.

Table 1

Symptoms and signs for Type 2

T/r	**Symptoms of diabetes**	**Signs of possible symptom manifestation**
s^1	Age (Age at the onset of illness)	1- Up to 30 years old 2- Over 30 years old
s^2	Mouth dryness	1- No dry mouth 2- Mild or moderate dry mouth 3- Severe dry mouth
s^3	Feeling of thirst	1- No feeling of thirst 2- Mild or moderate feeling of thirst 3- Strong feeling of thirst
s^4	Drinking water	1- Normal 2- Moderate 3- Frequently and often
s^5	Frequent and excessive urination	1- Normal 2- Moderate 3- Frequent and increased
s^6	By descent	1- Yes 2- No
s^7	Mental and emotional stress	1- Yes 2- No
s^8	Pancreatic injuries	1- Yes 2- No
s^{10}	Appetite	1- Normal 2- Increased appetite
s^{11}	Rate of disease progression	1- 10 days to 6 months 2- Longer than 6 months 3- Unknown
s^{12}	Pain in the legs	1- None 2- Light 3- Strong
s^{13}	Frostbite on the hand	1- None 2- Light 3- Strong
s^{13}	Persistent sores on palms/fingers	1- No 2- Yes
s^{14}	Tingling in the feet/toes	1- None 2- Light 3- Strong
s^{15}	Frostbite in the feet/toes	1- None 2- Light 3- Strong
s^{16}	Warmth in the feet/toes	1- None 2- Light 3- Strong
s^{17}	Non-healing wounds on feet/toes	1- None 2- Light 3- Strong
s^{18}	Difficulty controlling urination	1- Can hold 2- Can't hold
s^{19}	Getting hungry quickly	1- None 2- Light 3- Strong

Table 2
Symptoms and their recognizable signs for Type 1

T/r	**Symptoms of diabetes**	**Signs of possible symptom manifestation**
s^1	Age (Age at the onset of illness)	1- Up to 30 years old 2- 30 to 35 years old
s^2	Mouth dryness	1- No dry mouth 2- Mild or moderate dry mouth 3- Severe dry mouth
s^3	Feeling of thirst	1- No feeling of thirst 2- Mild or moderate feeling of thirst 3- Strong feeling of thirst
s^4	Frequent and excessive urination	1- Normal urination 2- Moderate urination 3- Frequent and increased urination
s^5	General fatigue	1- No general weakness 2- Moderate general weakness 3- Severe general weakness
s^6	Losing 10 kilograms in a month	1-yes 2-no
s^7	By descent	1-yes 2-no
s^8	Pancreatic injuries (infectious and non-infectious)	1-yes 2-no
s^9	Mental and emotional stress	1-yes 2-no
s^{10}	Rate of disease progression	1- From 10 days to 1 month. 2- From 1 month to 3 months. 3- Longer than 6 months

The first column of the table presents the designation of symptoms, the second column

contains the names of symptoms of diabetes mellitus, and the third column lists the possible signs that these symptoms may exhibit.

The dataset, created based on the aforementioned symptoms and their indicators, has been converted to a nominal space, which is crucial for the early diagnosis of diabetes mellitus (Table 3-4).

Appearance of the dataset for Type 2

Table 3

	s^1	s^2	s^3	s^4	s^5	s^6	s^7	s^8	s^9	s^{10}	s^{11}	s^{12}	s^{13}	s^{14}	s^{15}	s^{16}	s^{17}	s^{18}	s^{19}	**Class**
x^1	2	3	2	3	2	1	1	2	1	1	3	3	1	3	3	3	2	2	2	1
x^2	2	3	2	3	3	2	2	2	1	2	3	3	2	3	3	3	2	2	3	1
x^3	2	3	2	2	3	2	2	2	1	2	3	2	1	3	3	3	2	2	2	1
x^4	2	3	3	3	3	2	1	2	1	2	3	1	1	3	3	3	1	2	2	1
x^5	2	3	3	3	3	2	1	2	1	2	3	3	1	3	3	3	3	2	1	1
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
x^{336}	2	3	3	3	2	2	1	2	2	2	2	1	1	1	2	1	1	1	1	16
x^{337}	2	3	3	3	3	2	1	2	1	2	2	1	1	2	2	1	1	1	1	16
x^{338}	2	3	3	3	3	2	1	2	1	2	2	1	1	2	2	1	1	1	1	16
x^{339}	2	3	3	3	3	2	1	2	1	2	2	2	1	1	2	2	1	1	1	16
x^{340}	2	3	3	3	3	2	1	2	1	2	2	2	1	1	2	2	1	1	1	16

The first training dataset consists of 340 objects, which have been categorized into 16 classes based on 19 features.

Appearance of the dataset for Type 1

Table 4

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	x^9	x^{10}	Class
x^1	1	3	3	3	3	2	2	2	2	1	1
x^2	1	3	3	3	3	2	2	2	2	1	1
x^3	1	3	3	3	3	1	2	2	2	1	1
x^4	1	3	2	3	3	2	2	2	2	1	1
x^5	1	3	3	3	3	2	2	2	2	1	1
	...	...	...	...	...	...	...	...	...	...	
x^{136}	1	3	3	3	2	2	1	2	2	1	7
x^{137}	1	3	3	3	2	2	1	2	2	1	7
x^{138}	1	3	3	3	3	2	1	2	2	1	7
x^{139}	1	3	3	3	1	2	1	2	1	1	7
x^{140}	1	3	3	3	1	2	1	2	2	1	7

The database for type 1 diabetes contains 140 objects, which are categorized into 7 classes based on 10 characteristics.

(Fig. 1) shows a diagram of the existing classes in the dataset for type 2 diabetes, broken down by objects.

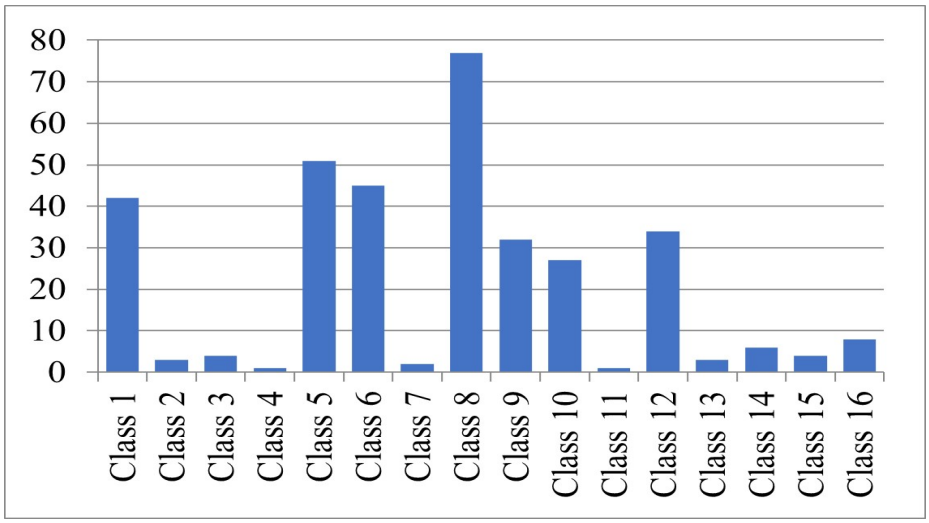

Figure 1. Objects count in the classes of type 1 diabetes mellitus

This training sample consists of 340 objects, which have been divided into 16 classes based on 19 features using a clustering algorithm focused on a set of informative characteristics. The distribution of objects across classes is as follows: class 1 contains 42 objects, class 2 has 3 objects, class 3 has 4 objects, class 4 has 1 object, class 5 has 51 objects, class 6 has 45 objects, class 7 has 2 objects, class 8 has 77 objects, class 9 has 32 objects, class 10 has 27 objects, class 11 has 1 object, class 12 has 34 objects, class 13 has 3 objects, class 14 has 6 objects, class 15 has 4 objects, and class 16 has 8 objects.

The diagram of the existing classes in the educational sample for the type 1, respectively, in the section of objects, is shown in Fig. 2.

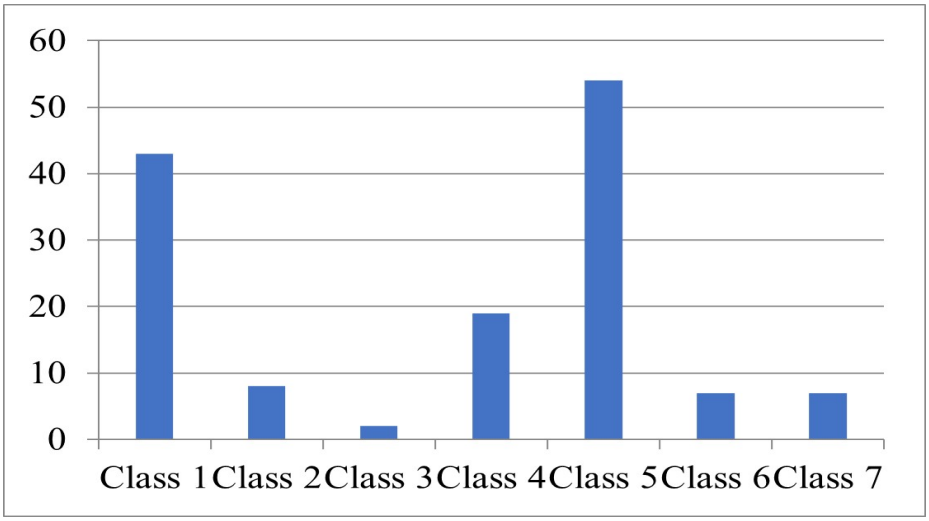

Figure 2. Objects count in the classes of type 1 diabetes mellitus

In the educational sample for type 1 diabetes mellitus, 140 objects are provided. Using an algorithm for selecting a set of informative features, these objects are divided into 7 classes based on 10 characteristics. The distribution of objects across classes is as follows: 43 objects in the 1st class, 8 objects in the 2nd class, 3 objects in the 3rd class, 19 objects in the 4th class, 54 objects in the 5th class, 7 objects in the 6th class, and 7 objects in the 7th class.

These dataset include subjective and objective signs of types of diabetes mellitus, genetic predisposition, and the body's various reactions to the disease. The possible values of the symptoms represent the severity of the disease and the rate of its progression. The dataset also serves to improve disease diagnosis, identify the illness in its early stages, and develop individual therapeutic approaches. However, the insufficient number of objects in all existing classes affects the complete and successful implementation of the classification process. As a solution to this problem, an algorithm for creating a new synthetic object based on existing objects in the class was developed.

5 Algorithm for creating a semi-synthetic dataset on diabetes

This algorithm aims to create a new expanded dataset by evaluating the similarity of synthetic objects generated for a nominal feature space.

The algorithm consists of the following steps.

Step 1. The objects of the general training sample are entered into the database. The initial database is formed in the context of all objects $x_i \in X, i = \overline{1, M}$;

Step 2. The process of preliminary data processing is carried out. This stage involves data cleaning, which includes eliminating errors, restoring lost values, removing duplicates, and clarifying ambiguous information in the collected data. After these adjustments, data normalization is performed.

Step 3. The general training dataset is preliminarily divided into classes $x_{p1}, x_{p2}, \ldots, x_{pm_p} \in X_p,\ p = \overline{1, r}$.

Step 4. For each formed class, the average similarity values of objects $\nu_p(X_p)$, $p=\overline{1,r}$ are determined.

Step 5. A new object $\overline{x}$ is created for classes X_p, where $p = \overline{1, r}$, while preserving essential features that ensure an increase in the average similarity level $\nu_p(X_p)$ of objects within this class.

Step 6. If a class p contains only one object ($m_p = 1$), then when creating a synthetic object, it is formed while preserving the features that distinguish the object of this class from objects of other classes.

Step 7. Steps 5 and 6 are repeated until K new synthetic objects are added to each class. As a result, classes consisting of hybrid objects are formed.

The problem mention in Section 3 is solved using the proposed algorithm.

6 Experimental results

In the experimental trials, classes with fewer than 20 objects were selected from the general educational sample, based on which synthetic objects were created (Fig. 1).

The results obtained using the algorithm for creating a semi-synthetic dataset based on the average similarity of objects within a class are presented in the following (Table 4). The aforementioned table was populated according to the procedure described above.

The results of this experiment are aimed at evaluating the effectiveness of the algorithm for creating a semi-synthetic dataset. The study describes the number of objects belonging to different classes, their degree of similarity, and the changes observed after introducing synthetic objects. As evident from the table, the number of real objects in classes varies, and

the degree of similarity between them differs. After the introduction of synthetic objects, these similarities increased correspondingly.

Table 4

Analytical results of the algorithm for creating a semi-synthetic dataset (K=50)

Class	**2**	**3**	**4**	**7**	**11**	**13**	**14**	**15**	**16**
Object count	3	4	1	2	1	3	6	4	8
Average object similarity (%)	71.93	76.32	-	68.42	-	75.44	77.54	79.82	83.83
Average similarity between classes including synthetic objects (%)	86.01	87.18	69.59	89.36	70.51	86.68	81.93	88.27	86.88
Difference(%)	**14,08**	**10,86**	**-**	**20,94**	**-**	**11,24**	**4,39**	**8,45**	**3,05**

The data in the table indicate that initially the average degree of similarity between classes was 76.18 percent. However, after implementing the semi-synthetic dataset algorithm, the average degree of similarity between classes increased to 82.93 percent (Fig. 3).

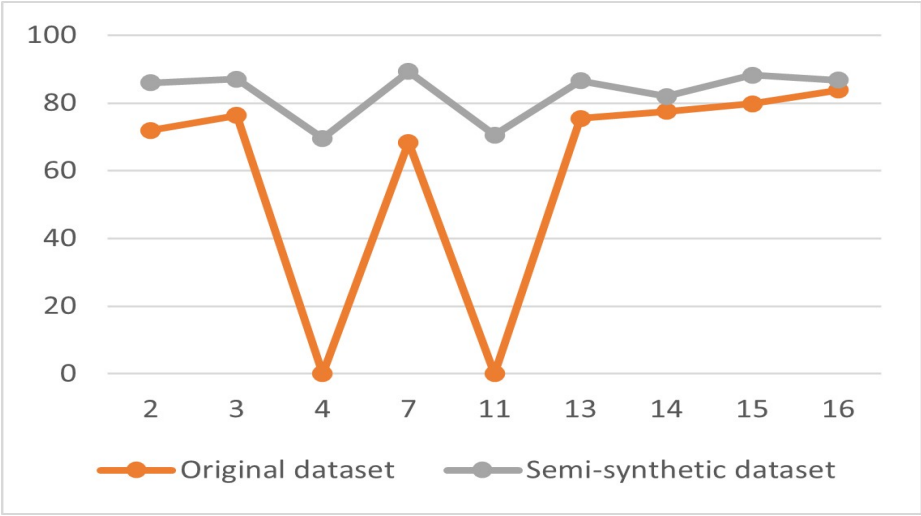

Figure 3. Change in average similarity levels between class objects

To comparatively assess the reliability of the developed algorithm, a classification of the created hybrid training sample was conducted using popular machine learning methods. The recognition rates for each class using three different methods are shown in detail in Table 5.

The results of the experimental trials demonstrated that the dataset obtained through the algorithm developed for creating a semi-synthetic dataset has high reliability.

Class	**Decision Tree (Precision)**	**KNN (Precision)**	**Proposed algorithm (Precision)**
2	93%	100%	100%
3	100%	100%	100%
4	100%	100%	100%
7	100%	100%	100%
11	90%	100%	100%
13	100%	100%	100%
14	94%	100%	100%
15	91%	91%	91%
16	93%	74%	80%
Accuracy	**95,7%**	**94%**	**96,6%**

7 Conclusion

This article proposes the use of artificial intelligence and machine learning technologies for the classification of diabetes mellitus. Due to data confidentiality issues and the problematic nature of directly accessing real patient information, an algorithm for generating synthetic training samples has been developed. This algorithm ensures the creation of new synthetic objects while preserving the statistical characteristics of real data.

The work developed a mechanism for generating synthetic data based on the clinical signs of patients with type 1 and type 2 diabetes mellitus. For type 2 diabetes, 16 classes and 340 objects were examined, while for type 1 diabetes, 7 classes and 140 objects were analyzed. This approach helps to expand the database and improve the accuracy of artificial intelligence models in diagnosing the disease.

The proposed algorithm aims to calculate the similarity degree of class objects and optimize their distribution. After the addition of synthetic objects, the average similarity degree of the classes increased from 76.18% to 82.93%.

In conclusion, it should be emphasized that in this study, an algorithm for generating synthetic training samples was developed to improve the classification and diagnosis of diabetes mellitus, while maintaining patient data confidentiality. This algorithm ensures the creation of new objects while preserving the statistical properties of the original data.

References

[1] Gonzales, A., G. Guruswamy, and S. R. Smith. “Synthetic Data in Health Care: A Narrative Review.” *PLoS Digital Health* 2 (2023): e0000082.

[2] Kokosi, T., and K. Harron. “Synthetic Data in Medical Research.” *BMJ Medicine* 1 (2022): e000167.

[3] Turimov Mustapoevich, D., D. Muhamediyeva Tulkunovna, L. Safarova Ulmasovna, H. Primova, and W. Kim. “Improved Cattle Disease Diagnosis Based on Fuzzy Logic Algorithms.” *Sensors* 23 (2023): 2107.

[4] McDuff, D., T. Curran, and A. Kadambi. "Synthetic Data in Healthcare." *arXiv* (2023). arXiv:2304.03243.

[5] Surendra, H., and H. Mohan. "A Review of Synthetic Data Generation Methods for Privacy Preserving Data Publishing." *Journal of Scientific and Technology Research* 6 (2017): 95–101.

[6] Aljohani, A., and N. Alharbe. "Generating Synthetic Images for Healthcare with Novel Deep Pix2Pix GAN." *Electronics* 11 (2022): 3470.

[7] Kaur, D., M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon. "Application of Bayesian Networks to Generate Synthetic Health Data." *Journal of the American Medical Informatics Association* 28 (2021): 801–811.

[8] Reiter, J. "Using CART to Generate Partially Synthetic Public Use Microdata." *Journal of Official Statistics* 21 (2005): 441–462.

[9] Noguer, J., I. Contreras, O. Mujahid, A. Beneyto, and J. Vehi. "Generation of Individualized Synthetic Data for Augmentation of the Type 1 Diabetes Data Sets Using Deep Learning Models." *Sensors* 22, no. 13 (2022): 4944. https://doi.org/10.3390/s22134944.

[10] Sampath, P., G. Elangovan, K. Ravichandran, et al. "Robust Diabetic Prediction Using Ensemble Machine Learning Models with Synthetic Minority Over-Sampling Technique." *Scientific Reports* 14 (2024): 28984. https://doi.org/10.1038/s41598-024-78519-8.

[11] Tagmatova, Z., A. Abdusalomov, R. Nasimov, N. Nasimova, A. H. Dogru, and Y. I. Cho. "New Approach for Generating Synthetic Medical Data to Predict Type 2 Diabetes." *Bioengineering* 10, no. 9 (2023): 1031. https://doi.org/10.3390/bioengineering10091031.

[12] Nishanov, A., O. Ruzibaev, and N. Tran. "Modification of Decision Rules 'Ball Apolonia' in the Problem of Classification." In *2016 International Conference on Information Science and Communications Technologies (ICISCT)*, 2016. https://doi.org/10.1109/ICISCT.2016.7777382.

[13] Nishanov, A., Sh. Saidrasulov, and E. Babadjanov. "Analysis of Methodology of Rating Evaluation of Digital Economy and E-Government Development in Uzbekistan." *International Journal of Early Childhood Special Education* 14, no. 2 (2022): 2447–2452. https://doi.org/10.9756/INT-JECSE/V14I2.230.

[14] Nishanov, A., O. Ruzibaev, J. C. Chedjou, K. Kyamakya, K. Abhiram, D. De Silva, G. Djurayev, and M. Khasanova. "Algorithm for the Selection of Informative Symptoms in the Classification of Medical Data." In *Developments of Artificial Intelligence Technologies in Computation and Robotics*, vol. 12 (2020): 647–658. https://doi.org/10.1142/9789811223334_0078.

[15] Nishanov, A., M. Akbarova, A. Tursunov, F. Ollamberganov, and D. Rashidova. "Clustering Algorithm Based on Object Similarity." *Journal of Mathematics, Mechanics and Computer Science* 123, no. 3 (2024): 108–120. https://doi.org/10.26577/JMMCS2024-v123-i3-4.

[16] Nishanov, A., F. Zaripov, B. Akbaraliev, E. Babadjanov, and B. Geldibayev. "Improved Deep Learning Model for Cattle Identification Using Muzzle Images." *Journal of Mathematics, Mechanics and Computer Science* 125, no. 1 (2025). https://doi.org/10.26577/JMMCS2025125102.

[17] Nishanov, A., A. Tursunov, F. Ollamberganov, and D. Rashidova. "Algorithm for Clustering Different Types of Drugs Affecting Blood Pressure." *Journal of Mathematics, Mechanics and Computer Science* 125, no. 1 (2025). https://doi.org/10.26577/JMMCS2025125104.

Авторлар туралы мәлімет:

Нишанов Ахрам Хасанович – доктор технических наук, Мұхаммед әл-Хорезми атындағы Ташкент ақпараттық технологиялар университетінің программалық инженерия факультетінің профессоры (Ташкент, Өзбекстан, электрондық почта: nishanov_ akram@mail.ru).

Менгтураев Фарход Зиятович – Денау кәсіпкерлік және педагогика институтының Ақпараттық технологиялар кафедрасының ага оқытушысы (Денау, Өзбекстан, электрондық почта: f.mengtoraev@dtpi.uz).

Олламберганов Файзулла Фарход угли – Мұхаммед әл-Хорезми атындағы Ташкент ақпараттық технологиялар университетінің жүйелік және қолданбалы программалау кафедрасының докторанты (Ташкент, Өзбекстан, электрондық почта: fayzulla0804@gmail.com).

Аллаяров Уктамжон Бекташович – Ташкент медицина академиясының Термез филиалының Ішкі аурулар пропедевтикасы, оңалту, этногылым және эндокринология кафедрасының оқытушысы (Термез, Өзбекстан, электрондық почта: criptolione7777@gmail.com).

Хасанова Малика Ахрамовна – Ташкент медицина академиясының No2 факультетінің госпитальдық терапия кафедрасының оқытушы ассистенті (Ташкент, Өзбекстан, электрондық почта: malikabonuxasanova@gmail.com).

Дониёрова Гулшан Тошмирзаевна – Денау кәсіпкерлік және педагогика институтының Ақпараттық технологиялар кафедрасының оқытушысы (Денау, Өзбекстан, электрондық почта: gulshandoniyorova68@gmail.com).

Сведения об авторах:

Нишанов Ахрам Хасанович (ответственный автор) – доктор технических наук, профессор факультета программной инженерии Ташкентского университета информационных технологий имени Мухаммада Ал-Хоразми (Ташкент, Узбекистан, электронная почта: nishanov_ akram@mail.ru);

Менгтураев Фарход Зиятович – старший преподаватель кафедры информационных технологий Денауского института предпринимательства и педагогики (Денау, Узбекистан, электронная почта: f.mengtoraev@dtpi.uz);

Олламберганов Файзулла Фарход угли – докторант кафедры системного и прикладного программирования Ташкентского университета информационных технологий имени Мухаммада аль-Хорезми (Ташкент, Узбекистан, электронная почта: fayzulla0804@gmail.com);

Аллаяров Уктамжон Бекташович – кафедра пропедевтики внутренних болезней, реабилитации, этномедицины и эндокринологии Термезского филиала Ташкентской медицинской академии (Термез, Узбекистан, электронная почта: criptolione7777@gmail.com);

Хасанова Малика Ахрамовна – Преподаватель-ассистент кафедры госпитальной терапии факультета No 2 Ташкентской медицинской академии (Ташкент, Узбекистан, электронная почта: malikabonuxasanova@gmail.com);

Дониёрова Гулшан Тошмирзаевна – преподаватель кафедры информационных технологий Денауского института предпринимательства и педагогики (Денау, Узбекистан, электронная почта: gulshandoniyorova68@gmail.com).

Information about authors:

Nishanov Akhram Khasanovich (corresponding author) – DSc, professor of the Faculty of Software engineering of Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi (Tashkent, Uzbekistan, email: nishanov_ akram@mail.ru);

Mengturayev Farhod Ziyatovich – Senior lecturer of the Department of Information Technology Denau Institute of Entrepreneurship and Pedagogy (Denau, Uzbekistan, f.mengtoraev@dtpi.uz);

Ollamberganov Fayzulla Farxod o'g'li – PhD student Department of System and applied programming of Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi (Tashkent, Uzbekistan, email: fayzulla0804@gmail.com);

Allayarov Uktamjon Bektashovich – Department of propedeutics of internal diseases, rehabilitation, Ethnoscience and Endocrinology, Termez Branch of the Tashkent Medical Academy (Termez, Uzbekistan, email: criptolione7777@gmail.com);

Khasanova Malika Akhramovna – Teaching Assistant of the Department of Hospital Therapy of Faculty No. 2 of Tashkent Medical Academy (Tashkent, Uzbekistan, email: malikabonuxasanova@gmail.com);

Doniyorova Gulshan Toshmirzayevna – Teacher of of the Department of Information Technology Denau Institute of Entrepreneurship and Pedagogy (Denau, Uzbekistan, email: gulshandoniyorova68@gmail.com).

Received: July 05, 2025
Accepted: December 02, 2025