MOSCOW STATE UNIVERSITY

The Faculty of Computational Mathematics and Cybernetics

Machine learning (ML) for biomedicine
Research of diabetes mellitus using neural networks

Yudina Diana

217 group

Moscow

2025

Table of contents

Introduction

Machine learning (ML) is a set of artificial intelligence methods that can be used to create self - learning computer systems (in particular, neural networks).

Machine learning is one way to make solving various tasks easier. This report will examine aspects of machine learning to solve various medical problems.

Relevance

The urgency of this problem is very high. Currently, there are very few medical centers, hospitals, and scientific laboratories in Russia that use medical technology. More than 10,000 patients come to hospitals every day to facilitate the work of doctors. which is incomprehensibly complex. Programmers, mathematicians, and physicists invent new technologies, programs, applications, and medical equipment to make the diagnosis and further decision faster.

Introduction to Machine learning

As we said in the introduction, machine learning is a tool that programmers use to "train" a model to solve problems. The main task of this approach is to select a special algorithm and then reproduce it.

There are a lot of types of subtasks in machine learning:

- **clustering** (the distribution of objects by attributes, when it is not known how many objects there are in total)
- **classification** (when objects and attributes need to be categorized, the number of objects is known)
- **regression** (by various parameters - objects, attributes, analyze statistics and make some predictions for the future)

Machine learning consists of three main aspects: **data, functions, and algorithms.** It is not difficult to guess that each of them is important and without data we will not be able to implement the program, the same as without algorithms we will not be able to solve the problem. There are a lot of algorithms, each of them is applied to a specific task, for example, to image segmentation or to static forecasts, etc.

Biological base

Biomedicine is a branch of medicine that studies the human body from a theoretical perspective, its structure and function in norm and pathology, pathological conditions, methods of their diagnosis, correction and treatment.

For machine learning in biomedicine, it is naturally necessary to know biology well enough, especially anatomy, genetics, histology, etc., depending on which field the programmer will work in.

Mathematical base

Mathematics is one of the main directions in machine learning and, in principle, in programming. Not many people understand, but any medical research, especially a blood test, can be described using differential and integral equations. Therefore, we will introduce topics of mathematical analysis that are necessary to know for machine learning:

- integral equations
- differential equations
- mathematical statistics, thanks to which the learning model will be able to predict the patient's future condition.

Diabetes research using machine learning

It is best to consider machine learning by example. So our task is to "Study diabetes mellitus". Let's try to figure out what algorithms and data sets we need to solve this problem.

Task statement:

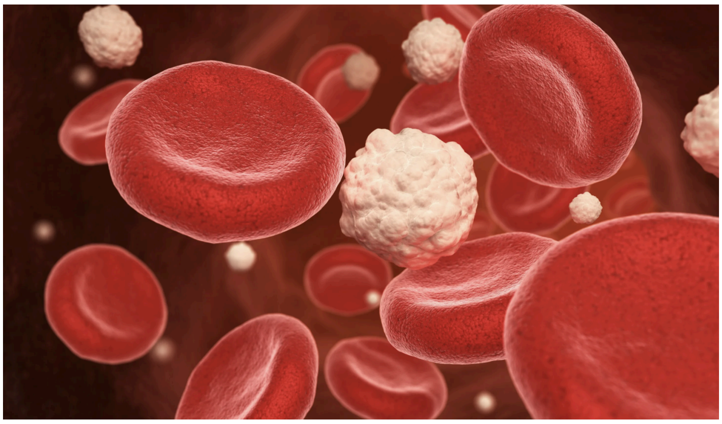
To conduct a study based on ready-made data for the study of diabetes mellitus.

Tasks:

- making a diagnosis (whether a person has diabetes or not)
- investigate the tests that need to be taken for diabetes
- further research and prediction of the development of the disease

Biological basis:

For research on diabetes mellitus, which can be of two types, different studies are used, in our case we will examine the blood for glucose content. For this study, we will use integral and differential equations describing the movements of various blood components (in our case, glucose). Looking at the photos, it can be seen that glucose is characterized by its white color and rounded shape.



Software implementation:

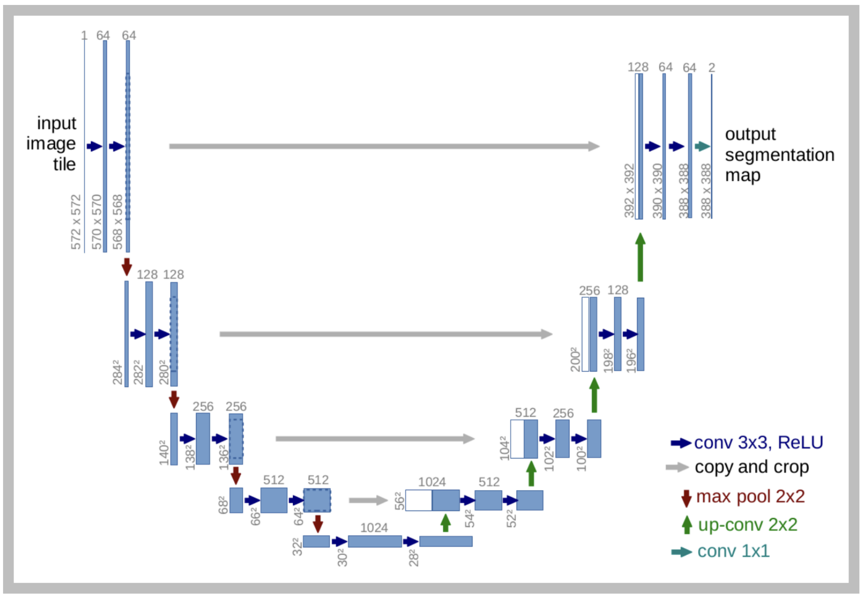
To implement our project, it is necessary to create two neural networks for different studies. The doctor provides the system with a blood smear using neural network No. 1, which will segment images and examine the ratio of glucose to other blood components. And in the future, to identify the diagnosis (whether a person is sick or not). Neural network No. 2 will predict how long it will take for a person to recover, as well as identify the patient's diagnosis on a daily basis throughout his treatment.

To write the code, we will use the Python language, with the connection of various libraries. Consideration of the code as we will write it will not be considered in this report, but we will consider the algorithms that we will use to write this system.

U-Net Neural Network

U-Net is a convolutional neural network designed for fast and accurate image segmentation. It is used in solving machine learning problems. This neural network is used for image processing tasks (in our case, doctors upload a blood smear to the system). Our main task is to segment the image and isolate glucose.

The architecture of the U-Net convolutional neural network



U-Net architecture as proposed by Ronneberger et al. (2015).

Encoder (Contracting Path)

The main task of the Encoder (encoder, compressing paths) is to reduce the size of photos with the allocation of important features, which allows you to reduce the spatial resolution. In turn, the compression path also consists of several parts:

- Convolutional Layers (convolutional layers) Each block contains two consecutive 3×3 convolutional layers with the ReLU activation function
- Pooling Layers (sub discretization) Pooling reduces the size of images by a factor of 2 and helps to concretize information and reduce the computational complexity of the model.
- Increasing the number of filter features

As the model moves, the number of filters doubles (64 -> 128 -> 256, etc.). This helps to further refine the features and work with more complex data.

Decoder (Expanding Path)

The main task of the Decoder (the restoring part, the decoder) is to restore the original size using information from the compressing path, i.e.

- Upsampling, Transposed (Convolution Increases the size of images by 2 times)
- Skip Connections (Skip Connections is one of the key steps in using the U-Net neural network. This is the stage of connecting the Encoder and Decoder, which allows you to restore small details that could have been lost during the operation of the model.)
- Convolutional Layers (After Skip Connections, two convolutional layers with ReLU are executed, this helps to apply the features that we highlighted earlier with spatial expansion.)

To write such a neural network, you need a ready-made dataset. At the moment, there is no ready-made dataset that would be in an open dossier, so the programmer's task is to create a dataset with photos. Which we can connect to the program. Various libraries and processing functions are used for image processing, which are included in their interfaces. You can use the **petroscope library**, which was created by the Krylov laboratory, Department of Mathematical Physics at Moscow State University. This library is adapted to handle various segmentation methods.

Let's look at an example of neural network code.

```
class UNet(nn.Module):
    def __init__(self, enc_chs=(3,64,128,256,512,1024), dec_chs=(1024, 512, 256, 128, 64)
        super().__init__()
        self.encoder     = Encoder(enc_chs)
        self.decoder     = Decoder(dec_chs)
        self.head        = nn.Conv2d(dec_chs[-1], num_class, 1)
        self.retain_dim  = retain_dim

    def forward(self, x):
        enc_ftrs = self.encoder(x)
        out      = self.decoder(enc_ftrs[::-1][0], enc_ftrs[::-1][1:])
        out      = self.head(out)
        if self.retain_dim:
            out = F.interpolate(out, out_sz)
        return out
```

When we write this neural network, we will be able to process images without a mathematical base. The future neural network will use a mathematical and biological base.

Neural network for further diagnosis of the disease

The next neural network that needs to be written is a neural network for diagnosis and further investigation of the patient's diagnosis. The main questions that need to be asked before writing a neural network.

- What will the neural network do?
- How will we diagnose it?
- What mathematical and biological terms will we use?

Let's answer our questions and give a complete description of this neural network, which we will write from scratch.

The previous neural network that we used to process the blood smear at the output will give us the percentage of glucose, red blood cells, white blood cells, platelets and plasma. For our second neural network, we will use a table of norms to establish a diagnosis. If the diagnosis is established, the system automatically adds the patient to the database and begins working on further diagnosis. The diagnostics that the neural network will produce will be based on mathematical statistics, as well as its basic concepts (mathematical expectation, variance, correlation, etc.)

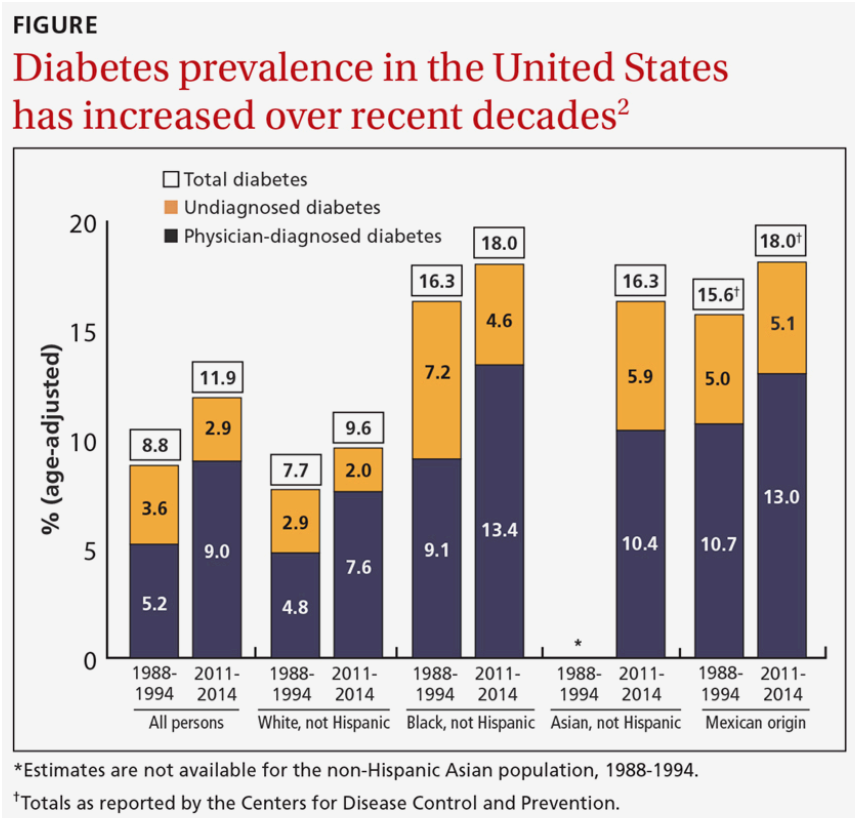
Table 1. Prediabetes diagnostic criteria according to health authorities.

Diagnostic criterion	WHO	ADA	NICE
HbA_{1c}	Not recommended for diagnosis	39–47 mmol/mol (5.7–6.4%)	42–47 mmol/mol (6.0–6.4%)
2-hour OGTT	7.8–11.0 mmol/L	7.8–11.0 mmol/L	7.8–11.0 mmol/L
Fasting plasma glucose	6.1–6.9 mmol/L	5.6–6.9 mmol/L	6.1–6.9 mmol/L

ADA=American Diabetes Association; NICE=National Institute for Health and Care Excellence; OGTT=oral glucose tolerance test; WHO=World Health Organization.

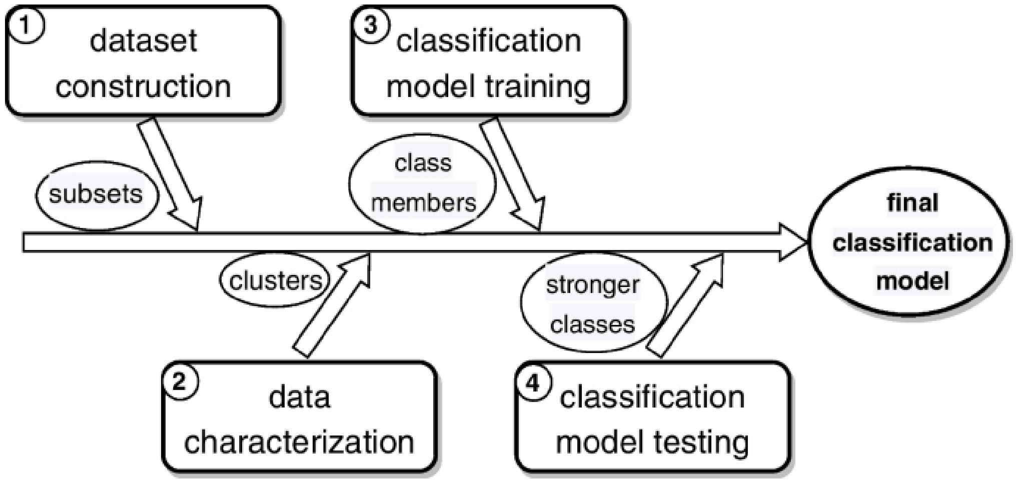
Studying diabetes, we know that there are two types of diabetes. **Type 1** diabetes is an autoimmune disease that causes the immune system to attack the beta cells of the pancreas, which are responsible for producing insulin, and as a result, the body has an insulin deficiency. **Type 2** diabetes is the more common form of this disease and often occurs against a background of overweight, unhealthy lifestyle and low physical activity. At the same time, it is diagnosed both in overweight people and in people without deviations in body weight. For the study of what type of diabetes, the neural network is not yet ready for such processing.

Having received the analyses, from a mathematical point of view, we will apply probability theory and mathematical statistics to demonstrate the future forecast. Since we do not have specific datasets, we will take a public dataset that will contain statistics for the last 30 years in the United States.



From the software implementation side, namely machine learning and its types, which we considered at the very beginning, we recommend using classification and regression. The classification will identify which category of diabetes this snapshot belongs to based on the previous neural network. The classification method implies that we select objects (in our case, photographs) based on their characteristics. (blood glucose levels, and other signs that the doctor deals with), this code will output the probability of a certain type of illness. The second type is regression, as we said earlier, it will output the average value of statistics, with what probability the patient will be cured, when approximately this happens.

The diagram below shows how the study takes place.



Overview of the general procedure applied in the study

Thus, summing up the above, we can conclude that this study, using neural networks, is an innovation in biomedicine and in the IT sphere, then we will consider the pros and cons of using ML.

Pros and cons of biomedical research using machine learning

Prons

1. High Relevance and Societal Impact

- Diabetes is a global health challenge, with rising prevalence. Early and accurate diagnosis can reduce complications and healthcare costs.

2. Multi-modal Data Utilization

- *Combining medical images (via U-Net) and clinical/statistical data (via a custom neural network) allows for deeper insight into a patient's condition..*

3. Personalized Treatment Prediction

- Enables the development of individualized treatment plans based on patient-specific data.

4. Automation and Decision Support

- Reduces burden on healthcare professionals.

- Acts as a decision support system for more accurate diagnosis and faster treatment planning.

5. Educational and Research Value

- Provides a strong foundation for further studies in AI, medicine, and data science.
- Can contribute to publications, open datasets, or clinical tool development.

Cons

1. Data Quality and Availability

- Medical datasets are often small, imbalanced, or incomplete.
- Getting access to labeled, reliable, and diverse data is difficult due to privacy and legal concerns (HIPAA, GDPR, etc.).

2. Interpretability Issues

- Deep learning models, especially neural networks, are often black boxes.
- In the medical field, explainability is crucial for trust and legal compliance.

3. Clinical Validation Required

- Models may work well in theory but need extensive clinical trials to prove their reliability in practice.
- Gaining acceptance among medical professionals can be slow.

4. Risk of Bias

- If training data is biased (e.g., not representative of all populations), predictions may be inaccurate or unfair.
- Potential ethical issues regarding equity in diagnosis or treatment.

5. Integration Complexity

- Implementing ML models into real healthcare systems involves technical, legal, and organizational hurdles.
- Requires collaboration with clinicians, IT departments, and administrators.

Conclusion

This research project explores the integration of machine learning techniques, particularly deep neural networks, into the field of biomedical diagnostics, with a focus on diabetes mellitus types 1 and 2. By leveraging two specialized models — U-Net for image processing and a custom-built statistical neural network for patient diagnosis and treatment recommendations — the project demonstrates a powerful, multi-modal approach to enhancing medical decision-making.

The strengths of the study lie in its relevance to global health challenges, its use of diverse data sources, and its potential to support personalized medicine. At the same time, it acknowledges critical challenges, including data availability, model interpretability, and the necessity of clinical validation to ensure safe deployment in real-world settings.

Ultimately, this project contributes to the ongoing transformation of healthcare through AI-driven solutions. It showcases how artificial intelligence, when carefully developed and responsibly applied, can improve diagnostic accuracy, optimize treatment plans, and ultimately lead to better patient outcomes. Future work should focus on expanding datasets, improving model transparency, and collaborating with healthcare professionals to bring such innovative tools into clinical practice.

References

1. Goodfellow I., Pouget-Abadie J., Mirza M., et al. Generative Adversarial Networks // Advances in Neural Information Processing Systems. 2014.
2. Ledig C., Theis L., Huszár F., et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network // CVPR. 2017.
3. Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019

4. https://biodatamining.biomedcentral.com/articles/10.1186/s13040-023-00340-2 , Machine learning based study for the classification of Type 2 diabetes mellitus subtypes