ABSTRACT
for the dissertation of Ainur Yerkos
submitted for the degree of Doctor of Philosophy (PhD)
in the educational program 8D06102 – Computer Engineering
on the topic: "Development of a Method and a Software Package for Big Medical Data Processing"

Relevance of the Research Topic. In the context of global digitalization of healthcare systems, the volume of medical data accumulated during diagnosis, treatment, and long-term patient monitoring grows exponentially. This array of data, characterized by the speed of accumulation, diversity of structures, and complexity of interpretation, is a classic example of Big Data. The effective use of such data opens unprecedented opportunities for the development of predictive medicine and the creation of medical decision support systems. These opportunities allow clinicians, taking into account the individual characteristics of patients, to identify risks in advance and optimize treatment trajectories, as a result of which medical practice rises to a new level.

One of the most important and resource-intensive areas of modern medicine is transplantology, in particular kidney transplantation. The shortage of donor organs and the risk of transplant rejection require the introduction of high-accuracy survival prediction tools. Despite the availability of classical statistical approaches, they often do not account for nonlinear dependencies of risk factors in the long term and their dynamic nature. The limitations of these approaches reduce the quality of decisions in transplantology and complicate doctors' work. However, in this area, there are complex technological barriers to the application of intelligent data analysis methods (Data Mining).

First, medical samples are imbalanced by nature: cases with unfavorable outcomes (organ rejection, death) are significantly rarer than stable conditions. This leads to the fact that standard machine learning algorithms, showing high overall accuracy, prove ineffective in predicting rare but critically important clinical events. Existing balancing methods often introduce noise or lead to the loss of valuable information.

Second, traditional software solutions are insufficiently flexible for processing heterogeneous data and do not account for the dynamics of patient condition changes at different stages after surgery. They rely on static models, which does not allow fully reflecting the temporal evolution of real clinical processes.

Thus, the development of a special cluster balancing method that effectively works with noise and complex data geometry (for example, the NAUS method), as well as the creation of an integrated software package based on it for multi-stage forecasting, is a relevant scientific and technical task. The implementation of such tools not only increases the accuracy of medical forecasts by an average of 16%, but also allows optimizing the distribution of

donor resources, which has direct socio-economic significance and contributes to improving the quality of life of patients. In addition, these innovations serve as the basis for updating clinical protocols, strengthening interdisciplinary interaction, and increasing the overall efficiency of transplantology. Modern approaches like NAUS address these challenges by preserving critical patterns in imbalanced datasets, enabling more reliable models that adapt to evolving patient profiles over time. Such advancements foster collaboration between data scientists and clinicians, paving the way for personalized post-transplant care strategies. Ultimately, they empower healthcare providers to make proactive interventions, enhancing patient safety and long-term outcomes in transplant programs worldwide. The integration of these methods also supports continuous learning systems that evolve with new clinical insights, ensuring sustained relevance in dynamic medical environments.

Research Objective: To develop a data balancing method and a software package based on machine learning models to improve the accuracy of predicting clinical outcomes in kidney transplantation.

Research Tasks:

- Develop a class balancing method for processing medical data, including data on major disease forms.
- Implement five machine learning models trained on specific patient cohorts based on graft survival duration.
- Develop a software package and integrate the trained models for practical graft survival prediction after surgery.

Methods of Research:

Proven data mining methods (machine learning, Data Mining), statistical analysis methods, and numerical experiments.

Object of Research:

Big medical data in the field of nephrology.

Subject of Research:

The NAUS (Noise-Aware Undersampling with Subsampling) class balancing algorithm and software tools for risk prediction based on medical indicators.

Scientific Novelty:

A class balancing method has been developed, applicable to various types of medical data, ensuring minimization of prediction errors, as confirmed by processing data on key diseases in transplantology.

Provision for the Defense:

The developed class balancing method ensures the minimization of prediction errors, confirmed by an average 16% increase in accuracy metrics compared to existing analogs.

Theoretical and Practical Significance: The theoretical significance lies in the study of imbalanced big medical data processing and the development of the NAUS algorithm, which offers a multi-stage approach to data preparation

with noise filtering and radial undersampling. The practical significance lies in the application of the software package for predicting terminal renal failure risks.

Reliability of Scientific Results

The reliability of the results is ensured through the use of rigorous methods of intelligent data analysis and statistical analysis. The results are confirmed by numerical experiments on stable class balance, as well as successful comparison with existing undersampling methods.

Relevance to research programs

The dissertation work was carried out under grant funding for scientific research:

- «A machine learning approach for personalized prediction of graft survival in kidney transplant recipients before transplantation» (State registration №0122PK00958, grant funding, 2022-2024).
- «Development of a multimodal artificial intelligence-based application for kidney transplant data preprocessing and analysis» (State registration №0124PK00417, grant funding, Al-Farabi Kazakh National University, 2024-2026).

Structure and Content of the Dissertation

The dissertation consists of an introduction, four main chapters, a conclusion, and a list of references.

The introduction, the relevance of the topic is substantiated, the goal and objectives of the research are clearly defined, the object and subject of the research are determined, the scientific novelty and practical significance of the work are revealed, and the provisions put forward for defense are also proposed. In this section, the necessity of modern medical research is especially emphasized, since the volume of big data is growing rapidly and their effective analysis improves clinical decisions.

The first chapter, a comprehensive analytical review of modern methods for processing big medical data was conducted. In biomedicine, problems of class imbalance and data heterogeneity were investigated. Existing software solutions for monitoring the condition of patients after kidney transplantation were analyzed and the limitations of classical statistical methods compared to machine learning algorithms were identified. During the review, the strengths and weaknesses of the methods were comprehensively considered, which strengthened the basis of the research.

The second chapter, the theoretical justification and implementation of the NAUS (Noise-Aware Undersampling with Subsampling) data balancing method were proposed. The results of verifying the method on control medical data sets were presented, confirming its advantage in preserving information in the sample. This approach, taking into account noise, is oriented towards improving data quality, therefore enhancing the stability of models.

The third chapter, the process of developing five predictive models was

described. Algorithms for various time horizons of graft survival were analyzed. Cross-validation and evaluation of models by AUC and F1-score metrics were proposed, which showed high predictive accuracy in conditions of dynamically changing patient indicators. The models are intended for preliminary identification of transplantation risk, which facilitates the work of doctors.

The fourth chapter, the architecture and software implementation of the developed package were proposed. The results of experimental tests of the system in conditions close to clinical practice were provided, and the effectiveness of integrating models into the process of making medical decisions was demonstrated. The package, thanks to its modular structure, is easily adaptable and ready for widespread use in hospitals.

Publications based on the research results

Seven printed works have been published on the dissertation topic, including 2 in publications recommended by he Science and Higher Education Quality Assurance Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, 3 works in peer-reviewed journals included in the international citation databases «Web of Science» and/or «SCOPUS», and 2 works in international conference proceedings.

- Salybekov A., Yerkos A., Kunikeyev A.D., Buribayev Z., Mochida Y., Ishioka K., Hidaka S., Kobayashi S. Predictive Modeling of Graft Failure Risk in Deceased Donor Kidney Transplants: Leveraging Machine Learning for Improved Outcomes and Data-Driven Insights: TH-PO036 // Journal of the American Society of Nephrology. - 2023. - Т. 34. - №. 11S. - С. 99 (ASN Kidney Week, November 02, 2023)
- Imanbek R., Buribayev Z., Yerkos A., Shaikalamova S., Abdirazak Z., Dauletiya D. HRBwSD: Hybrid radial based with standard deviation // 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST). - IEEE, 2024. - С. 487-491.
- Buribayev Z., Shaikalamova S., Yerkos A., Imanbek R. EKMGS: A HYBRID CLASS BALANCING METHOD FOR MEDICAL DATA PROCESSING //Scientific Journal of Astana IT University. - 2024. - С. 5-16.
- Buribayev Z., Yerkos A., Zhetpisbay Z. NEW APPROACH TO ADDRESSING CLASS IMBALANCE IN MEDICAL DATASETS CONSIDERING SPECIFICS //Scientific Journal of Astana IT University. - 2025. - С. 78-87.
- Salybekov A. A., Yerkos A., Sedlmayr M., Wolfien M. Ethics and algorithms to navigate AI’s emerging role in organ transplantation //Journal of Clinical Medicine. - 2025. - Т. 14. - №. 8. - С. 2775 (CiteScore: 5.2 / Impact Factor: 2.9; Q1)
- Salybekov A.A., Wolfien, M., Yerkos, A., Buribayev, Z., Hidaka, S., & Kobayashi, S. Phase-Specific Kidney Graft Failure Prediction with Machine Learning Model //Frontiers in Artificial Intelligence. - Т. 8. - С. 1682639 (CiteScore: 7.3 / Impact Factor: 4.7; Q2)

- Buribayev Z., Yerkos A., Zhetpisbay Z., Wolfien M. Noise-Aware Undersampling for imbalanced medical data (NAUS) //Informatics in Medicine Unlocked. - 2026. - C. 101731 (CiteScore: 8.7 / Impact Factor: 4.3; Q2)