**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**
**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**
**ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ**
**Кафедра системного анализа и компьютерного моделирования**

ГРУДОВИК Ксения Ильинична

# РАЗРАБОТКА ПРОГРАММНОГО ПРИЛОЖЕНИЯ ДЛЯ ОЦЕНКИ БИОЛОГИЧЕСКОГО ВОЗРАСТА ВЗРОСЛОГО НАСЕЛЕНИЯ ПО ГЕНОМНЫМ И КЛИНИЧЕСКИМ ДАННЫМ

Аннотация (реферат) дипломной работы

Научный руководитель:
кандидат физико-математических наук,
доцент Н.Н. Яцков

Консультант:
кандидат физико-математических наук,
П.В. Назаров

Допущена к защите
«___» _____ 2024 г.
Зав. кафедрой системного анализа и компьютерного моделирования

кандидат физико-математических наук, доцент
_____ Н.Н. Яцков

Минск, 2024

# РЕФЕРАТ

В дипломной работе 48 страниц, 10 рисунков, 12 таблиц, 32 источника, 2 приложения.

Ключевые слова: сокращение размерности данных, классификация, геномные данные, экспрессия генов, биоинформатика, анализ данных, машинное обучение, предсказание возраста.

Темпы старения различаются среди людей, а биологический возраст служит более надёжным показателем, нежели хронологический возраст. Применение методов интеллектуального анализа данных и машинного обучения для разделения людей на возрастные группы мало изучено, тем не менее, оно может улучшить понимание механизмов старения и эффективность лечения возрастных заболеваний.

Цель работы – разработать модели машинного обучения и реализовать программное приложение для предсказания биологического возраста по геномным и/или клиническим данным, а также разработать систему защиты и безопасности приложения.

Объектом исследования являются биофизические характеристики биологического возраста взрослого населения, изучаемые по геномным и клиническим экспериментальным данным об экспрессии генов.

Предметом исследования являются методы предсказания биологического возраста по геномным и/или клиническим данным.

Моделирование данных выполнено с помощью небольших искажений исходных данных. При сокращении размерности данных применены методы главных компонент и t-SNE. Разделения людей на возрастные группы по данным экспрессии генов проведено с помощью методов Random Forest (RF), Elastic-Net, eXtreme Gradient Boosting (XGBoost) и опорных векторов (SVM).

Наивысшая точность разделения людей на возрастные группы по данным экспрессии генов получена для метода XGBoost 70,35%. Выделены информативные гены, которые могут являться биомаркерами старения.

Разработано веб-приложение, позволяющее предсказывать биологический возраст пользователя по его геномным данным.

Разработана система защиты и безопасности веб-приложения.

# РЭФЕРАТ

У дыпломнай рабоце 48 старонак, 10 ілюстрацый, 12 табліц, 32 крыніцы, 2 пракладання.

Ключавыя словы: скарачэнне памернасці даных, класіфікацыя, геномныя даныя, экспрэсія генаў, біяінфарматыка, аналіз даных, машыннае абучэнне, прадказанне ўзросту.

Тэмпы старэння адрозніваюцца сярод людзей, а біялагічны ўзрост служыць больш надзейным паказчыкам, чым храналагічны ўзрост. Ужыванне метадаў інтэлектуальнага аналізу даных і машыннага абучэння для раздзялення людзей на ўзроставыя групы мала вывучана, тым не менш, яно можа палепшыць разуменне механізмаў старэння і эфектыўнасць лячэння ўзроставых захворванняў.

Мэта працы – распрацаваць мадэлі машыннага абучэння і рэалізаваць праграму для прадказання біялагічнага ўзросту па геномных і/або клінічных даным, а таксама распрацаваць сістэму абароны і бяспекі праграмы.

Аб'ектам даследавання з'яўляюцца біяфізічныя характарыстыкі біялагічнага ўзросту дарослага насельніцтва, якія вывучаюцца па геномных і клінічных эксперыментальных даных аб экспрэсіі генаў.

Прадметам даследавання з'яўляюцца метады прадказання біялагічнага ўзросту па геномных і/або клінічных дадзеных.

Мадэліраванне даных выканана з дапамогай невялікіх скажэнняў зыходных даных. Пры скарачэнні памернасці даных ужытыя метады галоўных кампанентаў і t-SNE. Раздзялення людзей на ўзроставыя групы па даных экспрэсіі генаў праведзены з дапамогай метадаў Random Forest (RF), Elastic-Net, eXtreme Gradient Boosting (XGBoost) і апорных вектараў (SVM).

Найвышэйшая дакладнасць падзелу людзей на ўзроставыя групы па даных экспрэсіі генаў атрымана для метаду XGBoost 70,35%. Вылучаны інфарматыўныя гены, якія могуць з'яўляцца біямаркерамі старэння.

Распрацавана вэб-праграма, якая дазваляе прадказваць біялагічны ўзрост карыстальніка па яго геномных даных.

Распрацавана сістэма абароны і бяспекі вэб-прыкладанні.

# STRUCTURAL ABSTRACT

The thesis contains 48 pages, 10 illustrations, 12 tables, 32 sources, 2 appendices.

Key words: data dimensionality reduction, classification, genomic data, gene expression, bioinformatics, data mining, machine learning, age prediction.

Aging rates vary among people, and biological age is a more reliable indicator than chronological age. The usage of data mining and machine learning methods for age group classification is understudied. However, it can improve the understanding of aging mechanisms and the effectiveness of age-related disease treatment.

The goal of the work is to develop machine learning models and implement a software application to predict biological age from genomic and/or clinical data, as well as to develop a protection and security system for the application.

The object of the study is the biophysical characteristics of the biological age of the adult population, studied using experimental genomic and clinical gene expression data.

The subject of the study is methods for predicting biological age from genomic and/or clinical data.

Data modeling is performed using small distortions of the experimental data. For data dimensionality reduction principal component analysis and t-SNE were applied. Age group classification based on gene expression data was carried out using Random Forest (RF), Elastic-Net, eXtreme Gradient Boosting (XGBoost) and support vector machines (SVM) methods.

The highest accuracy of 70.35% of age group classification based on gene expression data was obtained using the XGBoost method. Informative genes that may serve as aging biomarkers have were identified.

A web application has been developed. It allows predict the biological age of a user using their genomic data.

A protection and security system for the web application were developed.