DOI: https://doi.org/10.15276/aait.06.2023.29
UDC 004.8

Machine learning for human biological age estimation based on clinical blood analysis

Volodymyr H. Slipchenko[1)]
ORCID: https://orcid.org/0000-0002-3405-0781; ddpolytechnic2016@gmail.com
Liubov H. Poliahushko[1)]
ORCID: https://orcid.org/0000-0003-3287-8523; liubovpoliagushko@gmail.com. Scopus Author ID 58246840200
Vladyslav V. Shatylo[1)]
ORCID: https://orcid.org/0000-0001-5395-2097; v.shatylo@kpi.ua. Scopus Author ID 58247671300
Volodymyr I. Rudyk[1)]
ORCID: https://orcid.org/0009-0004-4774-6579; rudykviv@gmail.com
[1)] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine

ABSTRACT

This article explores the issue of estimating the biological age of a person using machine learning techniques. Biological age is a statistical indicator that reflects the degree of aging of an organism compared to other individuals in a specific population to which the organism belongs. This indicator aids medical professionals in diagnosing and treating diseases and assists researchers in studying the aging process in humans. There is no definitive correct formula for its determination because it is a statistical indicator and its value may vary depending on the dataset (population) and the selected set of indicators. **The study aims** is to create neural networks and choose a set of biomarkers that is both informative and easily accessible to the majority of individuals for evaluating biological age, ensuring both high recognition accuracy and operational speed. **The object of study** is the determination of the biological age of a person using information technology methods. **The subject of study** is the application of neural networks for determining the biological age of a person based on a clinical analysis of the human body's condition. Biomarkers correlating most with biological age were selected using the Pearson statistical method. The first neural network took selected biomarker values and previously calculated biological age as input and returned a predicted biological age as output. The second neural network took the predicted biological age and chronological age as input and returned a corrected predicted biological age as output. Accuracy assessment used the Pearson correlation coefficient, as well as classic error metrics such as coefficient of determination mean absolute error, and mean squared error. **As a result of the research**, were studied the dataset to identify biomarkers with the highest correlation coefficient values. Neural network architectures were selected and implemented to calculate biological age through general blood analysis. The best hyper parameters were selected experimentally and neural networks were trained. The obtained results conclude that a set of biomarkers for effective biological age recognition based on a comprehensive blood analysis has been developed and processed. Four neural networks were developed to realize the aim of the research (two for each gender). The Pearson correlation coefficient between the determined corrected biological age and chronological age for men is 0.9946, and for women is 0.9978, which is an indicator of high recognition accuracy. **The scientific novelty** of the conducted research lies in the application of an approach to assess human biological age based on the use of two neural networks and a set of biomarkers included in standard blood analysis packages. The proposed approach has allowed for an increase in the accuracy of biological age assessment and its usability in medical practice. This approach has been successfully applied to analyze the biological age of Ukrainian citizens, contributing significantly to the advancement of research in the field of biological age for medical professionals.

Keywords: Biological age; biomarkers; blood analysis; neural networks; machine learning; deep learning

***For citation*: Slipchenko V. H., Poliahushko L. H., Shatylo V. V., Rudyk V. I. "Machine learning for human biological age estimation based on clinical blood analysis". *Applied Aspects of Information Technology.* 2023; Vol.6 No.4: 431–442. DOI: https://doi.org/10.15276/aait.06.2023.29**

INTRODUCTION

Chronological age (CA) indicates the amount of time a person has lived since birth. It is often associated with the aging of the human body. However, individuals of the same age may have different appearances and health conditions, casting doubt on the relevance of such a connection.

Biological age (BA), on the other hand, characterizes the physical condition of the human body and the degree of its damage, estimated based on various biomarkers (weight, height, blood pressure, blood tests, data about physical activity, etc.) [1, 2]. Knowing BA allows for various conclusions about a person's lifestyle and its impact on their health. Moreover, unlike CA, this indicator can be altered.

The estimate of biological age has several advantages, including the ability to diagnose numerous diseases, assess the individual's organism's condition, and researches the aging process's pace.

Overall, the research on biological age has been ongoing since the last century; however, there is still no definitive method for its determination. Most researchers use mathematical and statistical methods such as multiple linear regressions, principal component analysis, and the Klemera-Doubal method.

With the development and popularization of machine learning algorithms, scientists have started applying them to determine BA. This approach allows for a much deeper exploration of data and dependencies between them without using complex formulas.

Most studies focused on determining biological age employ a similar algorithm to calculate its value:

1) selection of a set of biomarkers;
2) pre-processing of biomarker values;
3) utilization of an existing method or introducing a new one to calculate BA based on these biomarker values.

A biomarker (biological marker) is an objective indicator that unequivocally captures some aspect of the organism's state at a specific moment. To estimate BA, commonly used biomarkers include blood test samples, physical activity data, various medical examinations, etc.

The effectiveness of estimating biological age is typically assessed by the correlation coefficient between CA and BA, also considering the coefficient of determination, means absolute error (MAE), and mean squared error (MSE).

ANALYSIS OF EXISTING RESEARCH AND PUBLICATIONS

For more than 60 years, scientists and researchers have been working on determining biological age and have developed three main statistical methods to address this task.

The article [2] discusses the use of the multiple linear regression method to determine the BA of residents (437 people, including 169 men and 268 women) of Hiroshima, Japan, selecting 9 key biomarkers (skin elasticity, systolic blood pressure, pulmonary vital capacity, hand grip strength, light extinction time, vibratory perception in the ankle, visual acuity, auditory function, serum cholesterol) using the correlation coefficient method.

As a result, they obtained a mean square error of 6.8 for men and 5.9 for women. This approach showed satisfactory results at the time, although it had vulnerabilities to multicollinearity and the biomarker paradox. In the [3] discusses a method for addressing the issue of multicollinearity through the use of the Z-score.

Another method, also discussed in the article [3], namely the principal component analysis method, was employed for the same task. To determine a set of biomarkers that were later orthogonally transformed into linearly uncorrelated data, several methods such as correlation analysis, stability analysis, and dimensionality analysis were utilized. A drawback of this approach is the distortion of data at both ends of the regression. Additionally, the computed value is a relative indicator that is challenging to convert into years.

Later Klemera and Doubal formed one of the fundamental statistical methods still used today, namely the Klemera-Doubal method (KDM) [4]. While working with this method, many researchers made their adjustments and improvements to address existing drawbacks or simplify calculations.

For example, Cho et al. developed a new method, KDM2, combining KDM and principal component analysis, which simplified calculations [5]. This paper doesn't primarily address the selection of biomarkers, although it acknowledges its significance. Instead, the main emphasis is on assessing the accuracy of the estimated BA using different computational algorithms.

Recently, Kwon and Belsky presented a software implementation of the KDM method as one of the modules in their BioAge application [6]. Other implemented methods of determining BA are PhenoAge [7], and homeostatic dysregulation [8].

In the article [9], a stepwise multiple regression method with systematic error correction was employed to assess the biological age of the cardiovascular, respiratory, musculoskeletal systems, autonomic regulation, and metabolic age. Anthropometric results, echocardiography with Doppler, spirometry, ECG with heart rate variability analysis, two-photon X-ray absorptiometry, and clinical laboratory biochemical tests served as biomarkers. As a result, a relatively high accuracy (error of 4-5 years) was achieved in estimating the aging rates of physiological organ systems.

Furthermore, there is a large body of research focused on estimating BA using machine learning. Most of them utilize a branch of machine learning called neural networks, which get biomarker values are input, and return BA values are output.

Mamoshina et al. applied deep learning to determine biological age [10]. They developed and tested over 40 different deep learning models, each with a distinct architecture. The models received input from over 60.000 samples of classical biochemical blood analysis, and the output was the predicted BA. All data were pre-normalized, and

only samples within normal values were chosen. As a result, 21 machine learning models with the best results were selected and combined into an ensemble. That ensemble demonstrated a recognition accuracy of about 83.5 %. The main drawback of their study is the use of samples only from healthy individuals, leading to distortion of BA values for people whose biomarker values are out of range of the normal values.

Another popular solution is the application of Convolutional Neural Network (CNN) models for recognizing biological age based on a person's photos. They are often combined with other types of layers, such as Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and others.

To determine BA, Pyrkov et al. used 1D CNN based on data on a person's physical activity [11], Cole and colleagues chose a 3D architecture to process MRI data [12], Rahman and Adjeroh developed their variations of 2D and 3D architectures for the analysis of various medical biomarkers from the NHANES dataset [13].

Lima et al. employed deep learning to determine a person's biological age based on raw results of 12-channel electrocardiogram [14]. They used a convolutional neural network, similar to the proposed residual network for image classification [15], adapting it for processing one-dimensional signals.

Bortz, Guariglia et al. implement a technique for estimating biological age using an Elastic-Net neural network based on 25 selected circulating blood biomarkers [16]. Although the paper focused on mortality risk research, their method of determining biological age showed a higher accuracy than [7].

All the solutions discussed above use various variations of human biomaterial (blood analyses, etc.) or the results of medical examinations as biomarkers. Another branch of biological age research involves assessing its value using human images.

Later, the FineGrained method was proposed, which combines CNN, RNN, and LSTM architectures for age recognition from a person's photo [17]. The photo is processed and formatted to the size of 224 by 244 pixels before it enters the model. Then, it passes through a certain number of convolutional and LSTM layers. The model returns the predicted BA value. Despite high accuracy, this approach is highly sensitive to the gradient vanishing problem due to the use of RNN.

Tang et al. introduced another architecture based on Generative Adversarial Networks (GANs) called Conditional Generative Adversarial Networks with Identity Preservation [18], which also performs age recognition from photos. It consists of 3 modules: CGANs generate images with a synthesized face and a specific target age; the identity module ensures the transfer of basic facial features from the original image to the synthesized one; the age classifier module additionally checks the synthesized image for adherence to the target age.

However, most of these solutions use biomarkers whose values are difficult or impossible to obtain for most people. Not everyone can afford expensive medical examinations or the use of specialized devices, making biological age inaccessible to numerous individuals.

Since the values of many biomarkers require special examinations and research, it is more practical to use readily available indicators, such as a complete blood count.

Today, rapid and efficient determination of biological age can provide doctors with information for the early diagnosis of many diseases and the identification of problems with specific systems in the human body. Also, it can help researchers with the study of the rate of aging of people.

THE AIM AND OBJECTIVES OF THE RESEARCH

The object of study is the determination of the biological age of a person using information technology methods. The subject of study is the application of neural networks for determining the biological age of a person based on a clinical analysis of the human body's condition.

The work aims to develop neural networks and select an informative and conveniently accessible to most individuals set of biomarkers for the assessment of biological age, which will guarantee the high accuracy of recognition and operational speed.

To achieve the goal, it is necessary to solve the following tasks:

1) select an easily accessible set of biomarkers for effective estimation of biological age;

2) pre-process the chosen input data according to modern machine learning methods to prepare them for training neural networks;

3) design the optimal architecture of neural networks for estimating biological age;

4) implementation of the chosen architectures and conduct their training;

5) analyze the accuracy of estimations based on typical biological age metrics.

MATERIALS AND RESEARCH METHODS FOR ESTIMATE BIOLOGICAL AGE

The training data for neural networks were provided by the "D.F. Chebotarev Institute of Gerontology of the National Academy of Medical Sciences of Ukraine". The dataset includes the results of 928 clinical blood tests of healthy people whose biological age corresponds to the chronological age. This dataset was divided by gender (408 males and 520 females), considering physiological differences that significantly influence blood test indicators (Fig. 1) [19]. The average CA for males was 64.8 years and for females was 61.9 years.

To process and scale the data, the StandardScaler standardization method [20] from the sklearn library [21] was utilized. Its main idea is to bring all values into a distribution from -1 to 1 with a mean value of 0. This is done according to the formulas:

$$z = \frac{x - \mu}{\sigma}, \quad (1)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i), \quad (2)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad (3)$$

where z is scale value; x is current value; μ is mean value; N is number of elements; σ is standard deviation.

To determine the target feature, the biological age of each individual was calculated using a modification of the Klemera-Doubal method by Morgan Levin [22]. Since the range of calculated values for biological age using this method differs from the usual chronological age range, a correction algorithm, similar to Levin's work, was used to adjust the obtained biological age.

Additionally, based on the obtained data, biomarkers with the best correlation indicators to biological age were selected using the Pearson correlation method [23].

The Pearson correlation coefficient is a data characteristic that describes the strength and direction of the linear relationship between two quantitative values and it is calculated by the formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \quad (4)$$

where r is a Pearson correlation coefficient; x_i and y_i are data points; $\bar{x}$ is the mean of the x-values; $\bar{y}$ is the mean of the y-values.

Its values range from -1 to 1, where 0 indicates no correlation, negative values indicate a negative correlation (with an increase in one variable, the other variable linearly decreases), and positive values indicate a positive correlation (both variables increase linearly).

However, the correlation coefficient alone does not provide enough information to claim that the metric is informative for predicting the target variable. Another important indicator is the p-value, which indicates the statistical significance of this parameter. The correlation is considered statistically significant if the p-value is not less than the significance level (most often the α value is 0.05).

To calculate it, first of all, it was necessary to compute the value of the t-statistics according to the formula:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}, \quad (5)$$

	Age	Leukocytes	Erythrocytes	Hemoglobin	Hematocrit	Platelet	PCT	MCV	MCH	MCHC	RDW	MPV	PDW	Lymphocytes	Monocytes	Granulocytes	ESR
0	72.994	6.2	4.85	130.0	0.403	240.0	0.222	83.0	26.9	323.0	14.3	9.2	13.8	47.1	6.40	46.5	11.0
1	59.617	4.2	4.25	124.0	0.378	176.0	0.169	89.0	29.2	329.0	12.9	9.6	15.9	36.8	5.60	57.6	31.0
2	44.144	5.5	4.58	125.0	0.398	235.0	0.197	87.0	27.4	315.0	13.3	8.4	14.1	32.8	11.10	56.1	18.0
3	67.425	7.9	4.40	132.0	0.378	248.0	0.224	86.0	30.1	350.0	13.6	9.0	14.2	33.0	4.90	62.1	11.0
4	65.733	8.4	4.46	132.0	0.379	305.0	0.248	85.0	29.6	349.0	14.7	8.1	13.7	23.4	4.60	72.0	17.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
403	59.422	6.2	5.71	149.0	0.483	201.0	0.165	85.0	26.1	309.0	13.3	8.2	15.1	28.9	8.60	62.5	4.0
404	84.022	6.7	4.16	124.0	0.366	175.0	0.160	88.0	29.7	339.0	12.9	9.1	15.3	30.3	3.81	65.9	30.0
405	65.131	5.1	4.76	132.0	0.411	218.0	0.187	86.0	27.7	321.0	13.0	8.6	13.8	26.0	9.40	64.6	27.0
406	68.022	4.7	4.25	128.0	0.380	203.0	0.188	90.0	30.2	337.0	14.6	9.2	15.4	32.7	5.70	61.6	6.0

Fig 1. **Dataset for men (similarly for women)**
***Source*: compiled by the authors**

After obtaining this value and considering that degrees of freedom is $N - 2$, the p-value can be obtained using the t-distribution.

During the work, the decision was made to split the task of determining biological age into two parts: recognizing biological age and correcting its value. Separate neural networks were created for each part.

To design the architecture and develop neural networks, the programming language Python [24] and its library TensorFlow [25] were used.

The selection of the number of layers, neurons, and hyperparameters was done experimentally. The accuracy and prediction efficiency were assessed by calculating the Pearson correlation coefficient between biological age and chronological age, the determination coefficient, and the calculation of mean absolute error and mean square error.

The coefficient of determination, a statistical metric, evaluates the extent to which variations in one variable can account for variances in another variable when forecasting the result of a specific event. This coefficient gauges the strength of the linear correlation between two variables.

MAE assesses the average size of errors within a prediction set, regardless of their direction. It calculates the mean over the test sample of the absolute variances between predictions and actual observations, assigning equal weight to all individual differences.

MSE calculates the mean of the squared variances between predicted values and actual target values. The squaring process assigns greater importance to larger errors, rendering the MSE responsive to outliers.

ReLU (Rectified Linear Unit) and PReLU (Parametric Rectified Linear Unit) were used as activation layers for neural networks.

ReLU is one of the most popular activation functions, primarily used in regression tasks [26]. It is a non-linear function that transforms input values to a range from 0 to positive infinity.

PReLU is an activation function that introduces learnable parameters to the classic ReLU [27]. This helps overcome the vanishing gradient problem with minimal training cost increase.

As the optimization function, AdaGrad (Adaptive Gradient) was utilized [28]. It is a stochastic optimization method that minimizes the expected value of a stochastic objective function concerning a set of parameters, considering a sequence of function realizations.

RESULTS OF THE DEVELOPMENT OF NEURAL NETWORKS FOR DETERMINING BIOLOGICAL AGE

Table 1 shows the mean values of each biomarker, standard error (SE), the results of calculating the Pearson correlation coefficients for chronological age (4), as well as the t-statistics (5) and p-value for the male part of the dataset.

Despite the fact that the most common choice of α is 0.05, due to a relatively small data volume, the decision was made to use $\alpha = 0.1$ for male and $\alpha = 0.15$ for female.

For men, the minimum absolute correlation value for the selection of statistically significant biomarkers was chosen equal to 0.75, and for women it was equal to 0.65.

Therefore, for training neural networks to estimate biological age, the following significant biomarkers were selected from the male dataset, which have the highest correlation: hemoglobin, hematocrit, platelets, MCV, MCH, RDW, and ESR.

Table 2 shows the mean values of each biomarker, standard error (SE), the results of calculating the Pearson correlation coefficients for chronological age (4), as well as the t-statistics (5) and p-value for the female part of the dataset.

For training neural networks to estimate biological age, the following significant biomarkers were selected from the female dataset, which have the highest correlation: leukocytes, platelets, MCHC, RDW, lymphocytes, monocytes, and ESR.

Before training, both samples were divided with an 80% training set (326 for males, 416 for females) and a 20% test set (82 for males, 104 for females).

After that, all data were standardized using the StandardScaler method (1)-(3). As a result, two sets of data were obtained, the values of which are in the range from -1 to 1 with a middle value of 0 (Fig. 2).

The neural networks developed will predict biological age in years with an accuracy of 3 decimal places, chosen as the units of measurement. As the output is not constrained to integers, it can later be interpreted into the desired format (years, months, days, etc.).

Following tests, the architecture of the neural network for predicting the biological age of males was selected (Fig 3a). The input layer receives 7 values (according to the selected biomarker set) and consists of 32 neurons, followed by a hidden layer with 16 neurons, then a layer with 8 neurons, and finally, an output layer with 1 neuron returning the predicted value.

Table 1. **Information about the male part of selected dataset**

Biomarker name	Unit	Mean	SE	Correlation	t-statistics	p-value
Leukocytes	10^9 /L	6.614	0.132	-0.001	-0.029	0.977
Erythrocytes	10^{12} /L	4.634	0.025	-0.061	-1.225	0.221
Hemoglobin	g/L	139.241	0.692	-0.109	-2.204	0.028
Hematocrit	%	0.412	0.002	-0.086	-1.732	0.084
Platelets	10^9 /L	223.254	2.831	-0.087	-1.764	0.078
Platelet crit (PCT)	%	0.199	0.004	-0.009	-0.180	0.857
Mean Corpuscular Volume (MCV)	fL	90.414	0.575	-0.127	-2.572	0.010
Mean Corpuscular Hemoglobin (MCH)	pg	30.250	0.198	-0.095	-1.921	0.055
Mean Corpuscular Hemoglobin Concentration (MCHC)	g/L	337.577	0.841	-0.006	-0.122	0.903
Red Cell Distribution Width (RDW)	fL	13.575	0.047	0.121	2.453	0.015
Mean Platelet Volume (MPV)	fL	8.643	0.040	-0.015	-0.297	0.766
Platelet Distribution Width (PWD)	%	13.632	0.086	0.019	0.388	0.698
Lymphocytes	%	31.364	0.404	0.064	1.285	0.199
Monocytes	%	5.911	0.108	0.006	0.117	0.907
Granulocytes	%	62.970	0.422	-0.051	-1.026	0.305
Erythrocyte Sedimentation Rate (ESR)	mm/hr	13.127	0.449	0.272	5.694	0

***Source*: compiled by the authors**

Table 2. **Information about the female part of selected dataset**

Biomarker name	Unit	Mean	SE	Correlation	t-statistics	p-value
Leukocytes	10^9 /L	6.614	0.132	-0.001	-0.029	0.977
Erythrocytes	10^{12} /L	4.634	0.025	-0.061	-1.225	0.221
Hemoglobin	g/L	139.241	0.692	-0.109	-2.204	0.028
Hematocrit	%	0.412	0.002	-0.086	-1.732	0.084
Platelets	10^9 /L	223.254	2.831	-0.087	-1.764	0.078
Platelet crit (PCT)	%	0.199	0.004	-0.009	-0.180	0.857
Mean Corpuscular Volume (MCV)	fL	90.414	0.575	-0.127	-2.572	0.010
Mean Corpuscular Hemoglobin (MCH)	pg	30.250	0.198	-0.095	-1.921	0.055
Mean Corpuscular Hemoglobin Concentration (MCHC)	g/L	337.577	0.841	-0.006	-0.122	0.903
Red Cell Distribution Width (RDW)	fL	13.575	0.047	0.121	2.453	0.015
Mean Platelet Volume (MPV)	fL	8.643	0.040	-0.015	-0.297	0.766
Platelet Distribution Width (PWD)	%	13.632	0.086	0.019	0.388	0.698
Lymphocytes	%	31.364	0.404	0.064	1.285	0.199
Monocytes	%	5.911	0.108	0.006	0.117	0.907
Granulocytes	%	62.970	0.422	-0.051	-1.026	0.305
Erythrocyte Sedimentation Rate (ESR)	mm/hr	13.127	0.449	0.272	5.694	0

***Source*: compiled by the authors**

	Leukocytes	Erythrocytes	Hemoglobin	Hematocrit	Platelet	PCT	MCV	MCH	MCHC	RDW	MPV	PDW	Lymphocytes	Monocytes	Granulocytes	ESR
0	-1.055	0.289	0.413	0.068	-2.071	9.285	0.223	0.539	0.732	0.760	0.805	-0.249	0.384	-0.554	-0.255	1.201
1	-0.530	1.120	1.845	2.318	0.013	0.116	0.740	0.364	-1.036	-1.439	0.805	1.889	0.556	1.230	-0.877	-1.229
2	-0.793	0.209	-0.232	0.272	-0.390	-0.433	-0.036	-0.364	-0.977	-0.706	-0.542	1.195	0.764	1.184	-1.088	-0.345
3	-0.043	-0.582	-0.734	-0.614	0.696	0.592	-0.122	-0.163	-0.211	0.236	0.928	1.195	-0.498	-0.554	0.590	1.974
4	2.244	1.674	2.132	2.227	0.486	0.092	0.223	0.138	-0.388	-0.287	-0.542	2.236	-0.559	-1.148	0.802	-1.339
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
403	-0.830	-1.235	-2.023	-1.386	-0.565	-0.531	-0.381	-0.815	-1.331	1.074	-0.542	-1.579	0.433	0.178	-0.489	-0.566
404	0.894	0.427	0.842	0.841	-0.250	0.226	0.223	0.238	-0.034	-0.183	2.153	0.964	-0.253	-0.325	0.297	-0.566

Fig. 2. **Standardized dataset for men (similar for women)**
***Source*: compiled by the authors**

Between all layers, there are additional activation layers with the PReLU function, which is an improved version of the classic ReLU.

Hyperparameters for training were experimentally chosen: AdaGrad optimizer with a learning rate of 0.1, MSE loss function, and a batch size of 80. The training lasted for 95 epochs and took 2.2 seconds.

The neural network for determining the biological age of females had the same hyperparameters (except for the learning rate, which in this network was set to 0.03), but a different architecture (Fig. 3b). The first input layer has 18 neurons and accepts 7 input values (according to the selected biomarker set).

After that, there are 3 hidden layers with 16, 12, and 8 neurons, respectively, and finally an output layer with 1 neuron. Training took 4.9 seconds and 235 epochs.

The next neural network (Fig. 3c) was designed to correct the biological age values of males to the chronological range. The first layer with 16 neurons takes 2 values as input: the predicted biological age and the chronological age. The next layer is a single hidden layer with 8 neurons, followed by an output layer. The network completed training in 8.2 seconds and 495 epochs.

For correcting the biological age values of females, the most complex of the developed neural networks was created (Fig. 3d). The input layer, similar to the male network, takes 2 parameters but consists of 32 neurons. After that, there are 3 hidden layers with 16, 12, and 8 neurons, respectively. The output layer with 1 neuron also includes the ReLU activation function. The training was completed in 95 epochs in 2.4 seconds.

As a result, four neural networks were obtained, demonstrating high prediction accuracy across various accuracy metrics (MSE, MAE, Coefficient of determination, Minimal error, Maximum error), as shown in Table 3.

The Pearson correlation coefficient between the corrected biological age and chronological age for men is 0.9946, and for women, it is 0.9978 (Fig. 4).

Thus, during the study, a method for determining a person's biological age based on indicators of their complete blood count using neural networks was presented, and its implementation was developed.

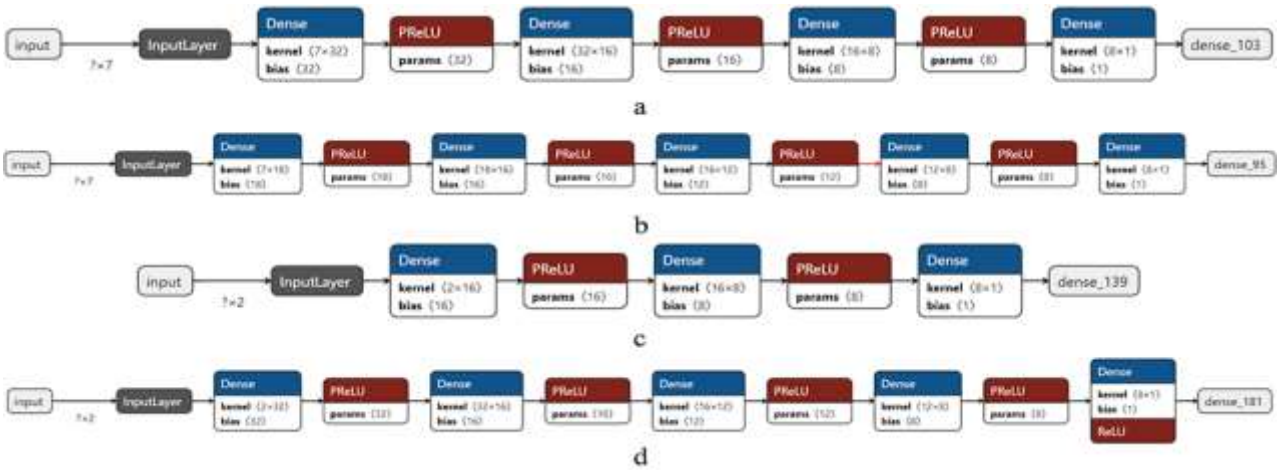

Fig. 3. **Architectures of the developed neural networks:**
a – for BA men; b – for BA women; c – for correction of men's BA;
d - for correction of women's BA
***Source*: compiled by the authors**

Table 3. **Results of testing the developed neural networks**

Accuracy metric	Men		Women	
	NN for estimating BA	NN for estimating corrected BA	NN for estimating BA	NN for estimating corrected BA
MSE	2.925	1.265	3.151	0.266
MAE	0.931	0.649	0.707	0.371
Coefficient of determination	0.998	0.992	0.998	0.998
Minimal error	0.013	0.001	0.003	0.001
Maximum error	8.936	11.138	15.843	4.669
Pearson correlation coefficient ($p < 3.063 \times 10^{-140}$)	–	0.9946	–	0.9978

***Source*: compiled by the authors**

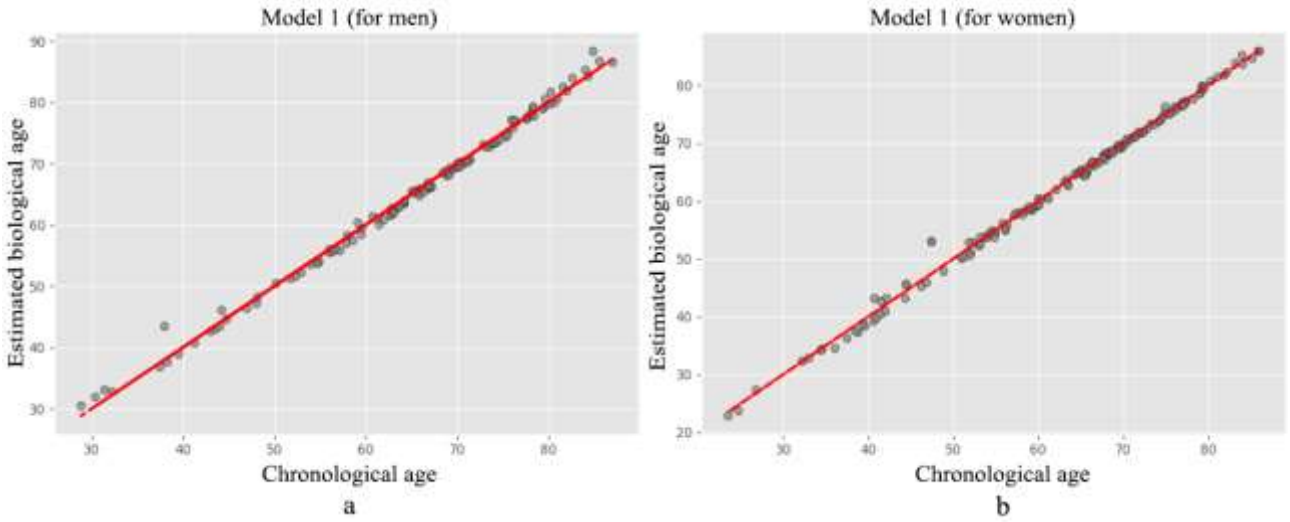

Fig. 4. **Correlation of estimated biological age and chronological age: a – for men; b - for women**
***Source*: compiled by the authors**

DISCUSSION OF THE CREATED MODELS

A significant amount of research has been focused on determining biological age using statistical and computational methods. This work introduces a novel method based on the development of two neural networks. The first network performs biological age recognition without using explicit formulas. This approach allows for more flexible determination and avoids various unpredictable situations. However, the calculated value may fall outside the typical chronological age range (from 1 to 80-100 years). The second network addresses this issue and transforms the obtained value into a normalized form.

Earlier, at the "D. F. Chebotarev Institute of Gerontology of the National Academy of Medical Sciences of Ukraine" [29, 30], research was conducted using neural networks. They utilized the same dataset as in our work, but different sets of biomarkers were chosen, and biological age was assessed using only one neural network. That approach yielded the following accuracy indicators: a correlation coefficient of 0.92 for males and 0.72 for females, as well as an MAE of 3.68 for males and 6.55 for females.

As a result of our research using the proposed approach, the accuracy significantly surpassed these values: the correlation coefficient for males is 0.9946 and for females – 0.9978; MAE for males is 1.265, and for females – 0.266. This demonstrates a substantial advantage of the developed method and its implementation.

CONCLUSIONS

The results of the conducted research allow for the following conclusions:

– A new approach to assess human biological age based on the use of two neural networks and a set of biomarkers included in standard blood analysis packages was applied.

– The proposed approach has allowed for an increase in the accuracy of biological age assessment and its usability in medical practice.

– This approach has been successfully applied to analyze the biological age of Ukrainian citizens, contributing significantly to the advancement of research in the field of biological age for medical professionals.

– The developed method and its implementation can be used by researchers and scientists for determining human biological age or for further modification and improvement of the developed method.

ACKNOWLEDGMENTS

Thanks to the “D.F. Chebotarev Institute of Gerontology of the National Academy of Medical Sciences of Ukraine” for the provided data.

REFERENCES

1. Diebel, L. & Rockwood, K. “Determination of biological age: Geriatric assessment vs biological biomarkers”. *Current Oncology Reports*. 2021; 23 (9): 104, https://www.scopus.com/authid/detail.uri?authorId=7103175498. DOI: https://doi.org/10.1007/s11912-021-01097-9.

2. Li, Z., Zhang, W., Duan, Y. et al. “Progress in biological age research”. *Front Public Health*. 2023; 11: 1074274, https://www.scopus.com/authid/detail.uri?authorId=57225184212. DOI: https://doi.org/10.3389/fpubh.2023.1074274.

3. Bafei, S. & Shen, C. “Biomarkers selection and mathematical modeling in biological age estimation”. *NPJ Aging*. 2023; 9 (13): 1–13, https://www.scopus.com/authid/detail.uri?authorId=57787946200. DOI: https://doi.org/10.1038/s41514-023-00110-8.

4. Jia, L., Zhang, W., & Chen, X. “Common methods of biological age estimation”. *Clinical interventions in aging*. 2017; 12: 759–772, https://www.scopus.com/authid/detail.uri?authorId=56547962000. DOI: https://doi.org/10.2147/CIA.S134921.

5. Cho, I., Park, K. & Lim, C. “An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI)”. *Mechanisms of Ageing and Development*. 2010; 131 (2): 69–78, https://www.scopus.com/authid/detail.uri?authorId=23395997900.
DOI: https://doi.org/10.1016/j.mad.2009.12.001.

6. Kwon, D. & Belsky, D. “A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge”. *GeroScience*. 2021; 43 (6): 2795–2808, https://www.scopus.com/authid/detail.uri?authorId=57217082861. DOI: https://doi.org/10.1007/s11357-021-00480-5.

7. Levine, M., Lu, A., Quach, A., Chen, B. et al. “An epigenetic biomarker of aging for lifespan and healthspan”. *Aging (Albany NY)*. 2018; 10 (4): 573–591, https://www.scopus.com/authid/detail.uri?authorId=7404034758. DOI: https://doi.org/10.18632/aging.101414.

8. Cohen, A., Milot, E., Yong, J., Seplaki, C., Fülöp, T., Bandeen-Roche, K. & Fried, L. “A novel statistical approach shows evidence for multi-system physiological dysregulation during aging”. *Mechanisms of Ageing and Development*. 2013; 134 (3–4): 110–117. https://www.scopus.com/authid/detail.uri?authorId=26643237900. DOI: https://doi.org/10.1016/j.mad.2013.01.004.

9. Pisaruk, A., Shatylo, V., Grygorieva, N. et al. “Biological age of physiological systems of the organism and profile of human aging”. *Journal of the National Academy of Medical Sciences of Ukraine* (in Ukrainian). 2022; 28 (1): 504–527. DOI: https://doi.org/10.37621/jnamsu-2022-4-2-2.

10. Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., Ostrovskiy, A., Cantor, C., Vijg, J. & Zhavoronkov, A. “Deep biomarkers of human aging: Application of deep neural networks to biomarker development”. *Aging*. 2017; 8 (5): 1021–1033, https://www.scopus.com/authid/detail.uri?authorId=56893719500. DOI: https://doi.org/10.18632/aging.100968.

11. Pyrkov, T., Slipensky, K., Barg, M., Kondrashin, A., Zhurov, B., Zenin, A., Pyatnitskiy, M., Menshikov, L., Markov, S. & Fedichev, P. “Extracting biological age from biomedical data via deep learning: too much of a good thing?”. *Scientific Reports*. 2018; 8: 5210, https://www.scopus.com/authid/detail.uri?authorId=15758226100. DOI: https://doi.org/10.1038/s41598-018-

23534-9.

12. Cole, J., Poudel, R., Tsagkrasoulis, D., Caan, M., Steves, C., Spector, T. & Montana, G. "Predicting brain age with deep learningm from raw imaging data results in a reliable and heritable biomarker". *NeuroImage*. 2017; 163: 115–124, https://www.scopus.com/authid/detail.uri?authorId=7403376270. DOI: https://doi.org/10.1016/j.neuroimage.2017.07.059.

13. Rahman, S. & Adjeroh, D. "Estimating biological age from physical activity using deep learning with 3D CNN". *IEEE International Conference on Bioinformatics and Biomedicine*. San Diego: USA. 2019. p. 1100–1103, https://www.scopus.com/authid/detail.uri?authorId=56564697600. DOI: https://doi.org/10.1109/BIBM47256.2019.8983251.

14. Lima, E., Ribeiro, A., Paixao, G. et al. "Deep neural network-estimated electrocardiographic age as a mortality predictor". *Nature Communications*. 2021; 12 (5117), https://www.scopus.com/authid/detail.uri?authorId=57205652239. DOI: https://doi.org/10.1038/s41467-021-25351-7.

15. He, K., Zhang, X., Ren, S. & Sun, J. "Deep residual learning for image recognition". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: USA. 2016. p. 770–778, https://www.scopus.com/authid/detail.uri?authorId=57209052101.
DOI: https://doi.org/10.1109/CVPR.2016.90.

16. Bortz, J., Guariglia, A., Klaric, L., Tang, D., Ward, P., Geer, M., Chadeau-Hyam, M., Vuckovic, D. & Joshi, P. "Biological age estimation using circulating blood biomarkers". *Communications Biology*. 2023; 6 (1): 1–10, https://www.scopus.com/authid/detail.uri?authorId=57415099100. DOI: https://doi.org/10.1038/s42003-023-05456-z.

17. Zhang, K., Liu, N., Yuan, X., Guo, X., Gao, C., Zhao, Z. & Ma, Z. "Fine-Grained age estimation in the wild with attention LSTM networks". *IEEE Transactions on Circuits and Systems for Video Technology*. 2020; 30 (9): 3140–3152, https://www.scopus.com/authid/detail.uri?authorId=56359224500. DOI: https://doi.org/10.1109/TCSVT.2019.2936410.

18. Tang, X., Wang, Z., Luo, W. & Gao, S. "Face aging with identity-preserved conditional generative adversarial networks". *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: USA. 2018. p. 7939–7947, https://www.scopus.com/authid/detail.uri? authorId=57020306700. DOI: https://doi.org/10.1109/CVPR.2018.00828.

19. Colineaux, H., Neufcourt, L., Delpierre, C., Kelly-Irving, M. & Lepage, B. "Explaining biological differences between men and women by gendered mechanisms". *Emerging Themes in Epidemiology*. 2023; 20 (1): 1–17, https://www.scopus.com/authid/detail.uri?authorId=56059169600.
DOI: https://doi.org/10.1186/s12982-023-00121-6.

20. "Sklearn.preprocessing. StandardScaler – scikit-learn 0.3.2 documentation". – Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing. StandardScaler.html. – [Accessed: July, 2023].

21. "Scikit-learn: machine learning in Python — scikit-learn 0.3.2 documentation". – Available from: https://scikit-learn.org/stable/index.html. – [Accessed: July, 2023].

22. Levine, M. "Modeling the rate of senescence: Can estimated biological age predict mortality more accurately than chronological Age?". *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*. 2013; 68 (6): 667–674, https://www.scopus.com/authid/detail.uri?authorId=7404034758. DOI: https://doi.org/10.1093/gerona/gls233.

23. Rodgers J. & Nicewander, A. "Thirteen ways to look at the correlation coefficient". *The American Statistician*. 1988; 42 (1): 59–66. DOI: https://doi.org/10.1080/00031305.1988.10475524.

24. "Welcome to Python.org". – Available from: https://www.python.org. – [Accessed: June, 2023].

25. "TensorFlow". – Available from: https://www.tensorflow.org. – [Accessed: June, 2023].

26. Brownlee J. "A gentle introduction to the Rectified Linear Unit (ReLU)". – Available from: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. – [Accessed: June, 2023].

27. "PreLU layer". – Available from: https://keras.io/api/layers/activation_layers/prelu. – [Accessed: June, 2023].

28. Duchi, J., Edu, J., Technion, E. & Singer, Y. "Adaptive subgradient methods for online learning and stochastic optimization". *Journal of Machine Learning Research*. 2011; 12: 2121–2159, https://www.scopus.com/authid/detail.uri?authorId=26221757400. – Available from:

https://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf. – [Accessed: June, 2023].

29. Pisaruk, A. & Mekhova, L. “Estimating biological age by hematological blood parameters”. *Ageing and Longevity*. 2021; 2 (3): 14–21. DOI: https://doi.org/10.47855/jal9020-2021-3-2.

30. Pisaruk, A., Shatilo, V., Antoniuk-Shcheglova, I., Naskalova, S., Bondarenko, O., Chyzhova, V., Shatilo, V. & Polyagushko. L. “Human biological age: Regression and neural network models”. *Fiziologichnyi Zhurnal*. 2023; 69 (2): 3–10, https://www.scopus.com/authid/detail.uri?authorId=6701801366. DOI: https://doi.org/10.15407/f69.02.003.

Conflicts of Interest: The authors declare no conflict of interest

Received 18.09.2023
Received after revision 11.12.2023
Accepted 15.12.2023

DOI: https://doi.org/10.15276/aait.06.2023.29
УДК 004.8

Машинне навчання для визначення біологічного віку людини на основі клінічного аналізу крові

Сліпченко Володимир Георгійович [1)]
ORCID: https://orcid.org/0000-0002-3405-0781; ddpolytechnic2016@gmail.com
Полягушко Любов Григорівна[1)]
ORCID: https://orcid.org/0000-0003-3287-8523; liubovpoliagushko@gmail.com. Scopus Author ID 58246840200
Шатило Владислав Валерійович [1)]
ORCID: https://orcid.org/0000-0001-5395-2097; v.shatylo@kpi.ua. Scopus Author ID 58247671300
Рудик Володимир Іванович[1)]
ORCID: https://orcid.org/0009-0004-4774-6579; rudykviv@gmail.com
[1)] Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, пр. Берестейський, 37. Київ, 03056, Україна

АНОТАЦІЯ

Стаття розглядає проблему визначення біологічного віку людини за допомогою методів машинного навчання. Біологічний вік – це статистичний показник, який показує ступінь старіння організму у порівнянні з іншими представниками певної популяції, до якої належить організм. Цей показник допомагає медикам у діагностиці і лікуванні захворювань, а також науковцям у вивченні темпів старіння людини. Не існує однозначно правильної формули для його визначення, оскільки значення біологічного віку для однакових вхідних даних буде відрізнятися в залежності від обраної популяції а також обраного набору показників. **Метою цього** дослідження є розробка нейронних мереж для обчислення біологічного віку людини, які забезпечать високу точність і швидкість роботи, а також вибір набору біомаркерів, який одночасно буде інформативним та доступним більшості людей. **Об’єктом дослідження** є визначення біологічного віку людини методами інформаційних технологій. У якості **предмету дослідження** обрано використання нейронних мереж для визначення біологічного віку людини на підставі клінічного аналізу стану організму людини. Пошук біомаркерів з найбільшим показником кореляції до біологічного віку відбувався за допомогою статистичного методу Пірсона. На вхід першій нейронній мережі подавали значення обраних біомаркерів та обчислений біологічний вік, на виході отримували прогнозований біологічний вік. На вхід другій нейронній мережі подавали прогнозований біологічний вік та хронологічний вік, на виході отримували скореговане значення прогнозованого біологічного віку. Для визначення точності розпізнавання використовувалися коефіцієнт кореляції Пірсона, а також класичні похибки: коефіцієнт детермінації, середня абсолютна похибка, середня квадратична похибка. **У результаті роботи** виконано дослідження отриманого набору даних, що дозволило визначити набір біомаркерів, які показали найвищі значення коефіцієнту кореляції, отже найкраще підходять для визначення біологічного віку. Обрано архітектури нейронних мереж, розроблено їх програмні реалізації для обчислення біологічного віку на основі загального аналізу крові. Обрано найкращі гіперпараметри та проведено навчання нейронних мереж. **З отриманих результатів** можна зробити висновок, що було отримано та оброблено набір біомаркерів для ефективного розпізнавання біологічного віку на основі загального аналізу крові та розроблено чотири нейронні мережі для виконання цієї задачі (по дві на кожну стать). Коефіцієнт кореляції Пірсона між визначеним скорегованим біологічним та хронологічним віком для чоловіків рівний 0.9946, а для жінок 0.9978, що є показником високої точності розпізнавання. **Науковою новизною** проведеного дослідження є застосування нового підходу оцінки біологічного віку людини заснованого на використанні двох нейронних мережах та набору біомаркерів, що входять в стандартні пакети аналізів крові. Запропонований підхід дозволив підвищити точність оцінки біологічного віку та доступність його використання в медичній практиці. Запропонований підхід успішно застосовано для аналізу біологічного віку громадян України, що дозволить значно просунути дослідження в сфері біологічного віку медичним фахівцям

Ключові слова: біологічний вік; біомаркери; аналіз крові; нейронні мережі; машинне навчання, глибоке навчання

ABOUT THE AUTHORS

Volodymyr H. Slipchenko - Doctor of Engineering Sciences, Professor, Professor at the Department of Digital Technologies in Energy. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine.
ORCID: https://orcid.org/0000-0002-3405-0781; ddpolytechnic2016@gmail.com
Research field: Development of integrated monitoring systems for ecological, energy, and economic factors; modeling for medical and biological systems; development of automated hardware and software systems; machine learning; neural networks

Сліпченко Володимир Георгійович - доктор технічних наук, професор, професор кафедри Цифрових технологій в енергетиці. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна

Liubov H. Poliahushko – PhD, Associate Professor at the Department of Digital Technologies in Energy. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine.
ORCID: https://orcid.org/0000-0003-3287-8523; liubovpoliagushko@gmail.com. Scopus Author ID 58246840200
Research field: Modeling for medical and biological systems; development and implementation of automated hardware and software systems; machine learning; neural networks

Полягушко Любов Григорівна – кандидат технічних наук, доцент кафедри Цифрових технологій в енергетиці. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна

Vladyslav V. Shatylo - PhD Student of the Department of Digital Technologies in Energy. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine.
ORCID: https://orcid.org/0000-0001-5395-2097; v.shatylo@kpi.ua. Scopus Author ID 58247671300
***Research field**:* Biological age; fractal analysis; machine learning; neural networks

Шатило Владислав Валерійович - аспірант кафедри Цифрових технологій в енергетиці. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна

Volodymyr I. Rudyk - Master at the Department of Digital Technologies in Energy. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine.
ORCID: https://orcid.org/0009-0004-4774-6579; rudykviv@gmail.com
Research field: Biological age; machine learning; neural networks

Рудик Володимир Іванович - магістр кафедри Цифрових технологій в енергетиці. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна