Deep Learning vs Gradient Boosting in age prediction on immunology profile

Alena Kalyakulina[1,2*†]
Igor Yusipov[1,2†]
Elena Kondakova[3]
Maria Giulia Bacalini[4]
Claudio Franceschi[2]
Maria Vedunova[3]
Mikhail Ivanchenko[2]

[1] Research Center for Trusted Artificial Intelligence, The Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia, 109004
[2] Institute of Information Technologies, Mathematics and Mechanics, Lobachevsky State University, Nizhny Novgorod, Russia, 603022
[3] Institute of Biology and Biomedicine, Lobachevsky State University, Nizhny Novgorod, Russia, 603022
[4] IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna, Italy, 40139
[*] Corresponding author: Alena Kalyakulina, kalyakulina.alena@gmail.com
[†] co-first authorship

Abstract

Background

The aging process affects all systems of the human body, and the observed increase in inflammatory components affecting the immune system in old age can lead to the development of age-associated diseases and systemic inflammation.

Results

We propose a small clock model SImAge using a limited number of immunological biomarkers. To solve the problem of regressing chronological age from cytokine data, we first use a baseline Elastic Net model, gradient-boosted decision trees models, and several deep neural network architectures. On the full dataset for 46 immunological parameters, LightGBM, DANet, and TabNet models showed the best results. Dimensionality reduction of these 3 models with SHAP values revealed the 10 most age-associated immunological parameters, which formed the basis of the SImAge small immunological clock. The best result of the SImAge model has mean absolute error of 6.28 years, it was shown by the DANet deep neural network model. Explicable artificial intelligence methods were used to explain the model solution for each individual participant.

Conclusions

We proposed an approach to construct a small model of immunological age, SImAge, using the DANet deep neural network model, which showed the smallest error on the 10 immunological parameters. The resulting model shows the highest result among all published studies on immunological profiles. Since gradient-boosted decision trees and neural networks show similar results in this case, we can consider parity between these types of models for immunological profiles.

Keywords: Immunological Profile, Deep Neural Network, Tree-based Model, Explainable Artificial Intelligence

1. Introduction

1.1. Background

The immune system plays an important role in protecting the human organism from various infections. An increase in inflammatory components affecting the innate and adaptive arms of the immune system [1–4] is observed in old age, which may be associated with the development of various age-associated diseases [5–8] and systemic inflammation [9–14]. This chronic inflammation, caused by the metabolic products of damaged cells and environmental influences, has been associated with aging and leads to progressive damage of many tissues [5,15]. This phenomenon of increased levels of circulating inflammatory mediators has been called inflammaging [16–18]. However, there are no standard biomarkers to fully characterize inflammaging [19,20].

The aging process is extremely complex and involves all body systems, so there have been many attempts to characterize it and estimate its rate from different perspectives: combining different clinical parameters, predicting the risk of mortality or cognitive deterioration [21–25]. A broad class of aging biomarkers are various chronological age predictors, or clocks. Epigenetic clocks, which attempt to predict a person's age based on DNA methylation data [26–29], are the most common. Recently, the inflammatory clocks have also developed [30–32]. Both genetic factors involved in the formation of immune composition [33–35] and environmental factors [36,37] can lead to significant individual differences in immune characteristics. An individual's immune status continuously changes over time; therefore, immune age can vary even among healthy individuals of the same chronological age [38]. Understanding the relationship between biological aging and an individual's immunological profile and predisposition to age-associated diseases may facilitate the development of tools to slow aging and improve longevity [39].

One of the first immunological age biomarkers is Inflammatory Biologic Age, which uses 9 inflammatory markers to construct an aggregate age estimate based on data from over 3,000 subjects using the Klemera-Doubal method [21]. This inflammatory age acceleration has been shown to be associated with increased risks of cardiovascular disease and mortality. The original work compares mean values with standard deviations of clinical and inflammatory ages; no error values are given, so an explicit comparison with other approaches is difficult. Another notable aggregated biomarker of immunological aging is IMM-AGE, which demonstrates multidimensional changes in immune status with age and is significantly

associated with mortality. The association of IMM-AGE score with cardiovascular disease has also been shown [30]. Aggregating information on cellular phenotyping and cytokine responses for a fairly wide age range of individuals (40-90 years), IMM-AGE demonstrates a P value of 10^-60 for linear regression with age. However, this metric is only an estimate, and it is difficult to obtain a predicted age value in years. A better known recently developed metric for human immunological age is the iAge model [31]. This inflammatory aging clock uses information on cytokines, chemokines and growth factors from a thousand of healthy and diseased subjects ranging from age 8 to 96 years to construct an aggregated age estimate using a guided autoencoder. This nonlinear method, a type of deep neural network, eliminates the noise and redundancy of the raw data while preserving important biological information. Nevertheless, this model requires the entire input data to be fed to the input, so this model cannot be called fully compact. The average error of this model is 15.2 years, and the association of iAge with multimorbidity, immune aging, frailty, and cardiovascular aging has been shown. There are also immunological clock models that take into account changes in inflammatory/immunological status in certain diseases. One example of such age predictors is the ipAGE model, which is sensitive to age-associated acceleration in chronic kidney disease [32]. Constructed using the classic Elastic Net approach for 38 immunological biomarkers, ipAGE demonstrates a mean absolute error of 7.27 years and root mean squared error of 8.33 years for the test subset of the control group and detects immunological biomarkers significantly differentially expressed between cases and controls. The model better assesses the phenotype of accelerated aging in patients with end-stage renal disease by quantifying inflammaging than various epigenetic clock models. The smaller error relative to the other models may be due to the small training sample size and the lack of cross-validation. Another class of aggregated immunological biomarkers constructs age estimates based on peptides or plasma proteins data [40,41]. Such models are characterized by a large dimensionality of input data: from hundreds or thousands of human plasma proteins and up to several million peptide features. A Lasso model clock based on nearly three thousand plasma proteins showed the best mean absolute error of 2.44 years on a test dataset of 3 thousand people with an age range of 18-76 years among 12 different models [40]. This model also confirmed the hypothesis that people who regularly engage in aerobic activity have a lower predicted age than people committed to a sedentary lifestyle. Another model, antibody repertoire immune age, was constructed using the Elastic Net method for 125,000 antibody bindings [41]. As for some other models, no error values are given for this model, but only the correlation with chronological age is characterized (Pearson's correlation coefficient 0.75, P-value < 10^-107). The huge dimensionality of the input data for such models, already mentioned earlier, is a drawback that can lead to difficulties in clinical application.

All of the described immunological age predictors use different quantitative measures, so the input data for all of the models described are tabular: the rows contain samples (individual participants) and the columns contain features (immunological measures). Traditionally, one of the most common methods for constructing age predictor models (clocks) for different types of data is Elastic Net. This model is used as the basis for multiple age predictor models based on epigenetic [26–29], immunological [32,41], transcriptomic [42], metabolomic [43–45], microRNA [46], and proteomic [47] data. More advanced are the methods of gradient boosting, which are also actively used to construct age predictors, particularly from epigenetic [48] or gut microbiome [49]

data. In the last few years, deep learning methods, including neural networks, autoencoders, and other approaches have been developed. These approaches are also actively used to construct various clock models. There are models for epigenetic [50], immunological [31], transcriptomic [51], gut microbiome [52] and even hematological [53–55] data.

Elastic Net is an example of a linear model, therefore, despite its widespread use not only in age prediction tasks and ease of interpretation, it has a number of drawbacks: it assumes linear relationships between dependent and independent variables, does not take into account possible nonlinear relationships, and is sensitive to outliers [56]. The gradient-boosted decision trees (GBDT) and neural network approaches have great potential, in particular for tabular data. For classification and regression problems for tabular data, GBDT models such as XGBoost, LightGBM and CatBoost are the dominant group of methods [57,58]. All of them train several decision trees to predict the final result, although each of the mentioned methods has its own features: XGBoost iteratively trains an ensemble of shallow trees [59]; LightGBM splits the tree by leaf rather than by depth or level [60]; CatBoost builds symmetric trees [61]. In an attempt to compete with GBDT models, several neural network models of different types have been created specifically for tabular data. With the increasing number of heterogeneous data estimation tasks with both continuous and categorical features, for which traditional approaches are not always well applied, the need for flexible solutions, in particular based on deep neural networks that can capture complex nonlinear dependencies in the data, has arisen [62]. Recent research has seen regularization as a way to improve the performance of deep neural networks on tabular data [63]. Due to the success of attention-based approaches for textual and visual data, it also became applicable to tabular data [64]. The development of transformer architectures for tabular data made it possible to work with high-dimensional datasets [65–67]. Hybrid methods have also been developed that attempt to combine the advantages of tree and neural network models [68]. Despite the many tree, neural network, and hybrid approaches that have been developed, there is still no consensus on which methods perform best for heterogeneous tabular datasets [57,58,62,67].

Simple linear models are easy to interpret, with the importance of individual features determined by the values of the coefficients in front of them; the majority of the tree models can also be interpreted. For neural network models, the situation is more complicated: not all existing models allow obtaining a ranking of the importance of individual features for model prediction. However, this is where explainable artificial intelligence (XAI) approaches can help, which can determine the contribution of individual features to the final prediction of almost any model [69–75]. One categorization divides explainability into global and local [76–78]. Global explainability helps to interpret the behavior of the model as a whole - which features have the highest impact on the prediction of certain classes (for classification) or specific values (for regression). When analyzing biomedical data, global explainability helps to identify biomarkers and their combinations that are specific to a particular task or studied phenomenon (e.g., the association of a biomarker with a pathological condition, disease, or a clinical characteristic, age, or gender). Explainability can also be used to reduce the dimensionality of models. Discarding unimportant, noisy features and keeping only those that affect the model often allows to reduce the input dimensionality of the model, to increase efficiency, and even to improve the result. Local explainability helps to determine why the model made its prediction for a particular sample, and how the feature values for that sample affected it. When

solving problems for biomedical data, this type of model explainability is crucial in questions of personalized medicine.

1.2. Study design and novelty

The primary goal of this work is to develop a small clock using a limited number of immunological biomarkers. We use the classic Elastic Net model as the baseline, which is the most common choice for constructing age predictors on various types of biological data. To solve the problem of chronological age regression on tabular data we also use tree structures, namely, gradient-boosted decision trees (GBDT), one of the most well-proven approaches specifically for tabular data. When comparing GBDT models with deep learning models, in many problems GBDTs showed similar or better results, but required less tuning and showed better performance [57]. Nevertheless, we also consider several neural network architectures designed to solve problems on tabular data. Linear models, in particular Elastic Net, are easy to interpret, while tree and neural network models are highly nonlinear, are much more complex, and require methods of explicable artificial intelligence. We first apply XAI methods to determine the most important features for age prediction models based on immunological data. We also use this ranking to build portable models consisting of a small number of immunological biomarkers. Such small models are much easier to implement, can have widespread practical applications, and are potentially cheaper. Further, XAI methods are used to explain the model decision for each particular sample (person), which provides a significant contribution to the development of a personalized medicine approach.

2. Results

2.1. Data description

Participants

The study involving a control group of volunteers recruited in the Nizhny Novgorod region in 2019-2021 was performed at Lobachevsky State University of Nizhny Novgorod. Exclusion criteria were chronic diseases in the acute phase, oncological diseases, acute respiratory viral infections, and pregnancy in women. The total number of participants was 260, including 173 women and 87 men, i.e., the sex distribution was not balanced with a significant predominance of women. The age distribution is also not uniform, with an overall age range from 19 to 96 years, participants from the age group of 30-60 years prevail (Figure 1(a)).

Features

Figure 1(b) lists all of the 46 immunological biomarkers considered in this study. They were obtained using Luminex xMAP technology for blood plasma samples (see Methods Section 4.1 for details). To examine possible correlations between these immunological parameters and their correlation with age, Pearson correlation coefficients and p-values for testing non-correlation (Benjamini-Hochberg FDR-corrected p-values<0.05) were calculated for each pair of features. It can be seen that CXCL9 is most correlated with age, which was also found in earlier works [31,32,79,80]. Recently, it has been increasingly evidenced that CXCL9 is an important

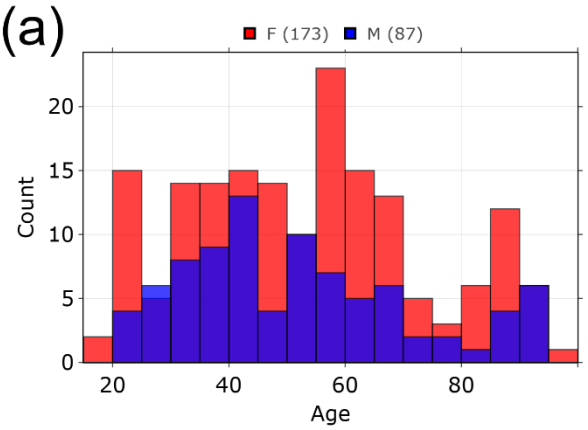
marker of inflammation and plays a key role in the development of age-associated diseases, such as neurodegeneration [81], chronic kidney disease [32], glaucoma [82] and various inflammatory diseases [83–85]. In addition, there is a whole group of interleukin biomarkers that are significantly correlated with each other (the most strongly correlated pairs are IL2 vs IL17A, IL2 vs IL15, IL17A vs IL15 with Pearson correlation coefficients almost equal to one). The relationship between IL2 and IL15 has been demonstrated [86], including cancer therapy [87,88], as well as the relationship between IL2 and IL17A [89,90], and between IL15 and IL17A [91].

All data used in the present work, such as immunological biomarker values, sex and age of participants, are presented in Supplementary Table 1.

Participants:
260 healthy volunteers from 19 to 96 years, including 173 women and 87 men, recruited in Nizhny Novgorod, Russia

Features:
46 immunological biomarkers obtained using Luminex xMAP technology for blood plasma samples.

(a)



(b)

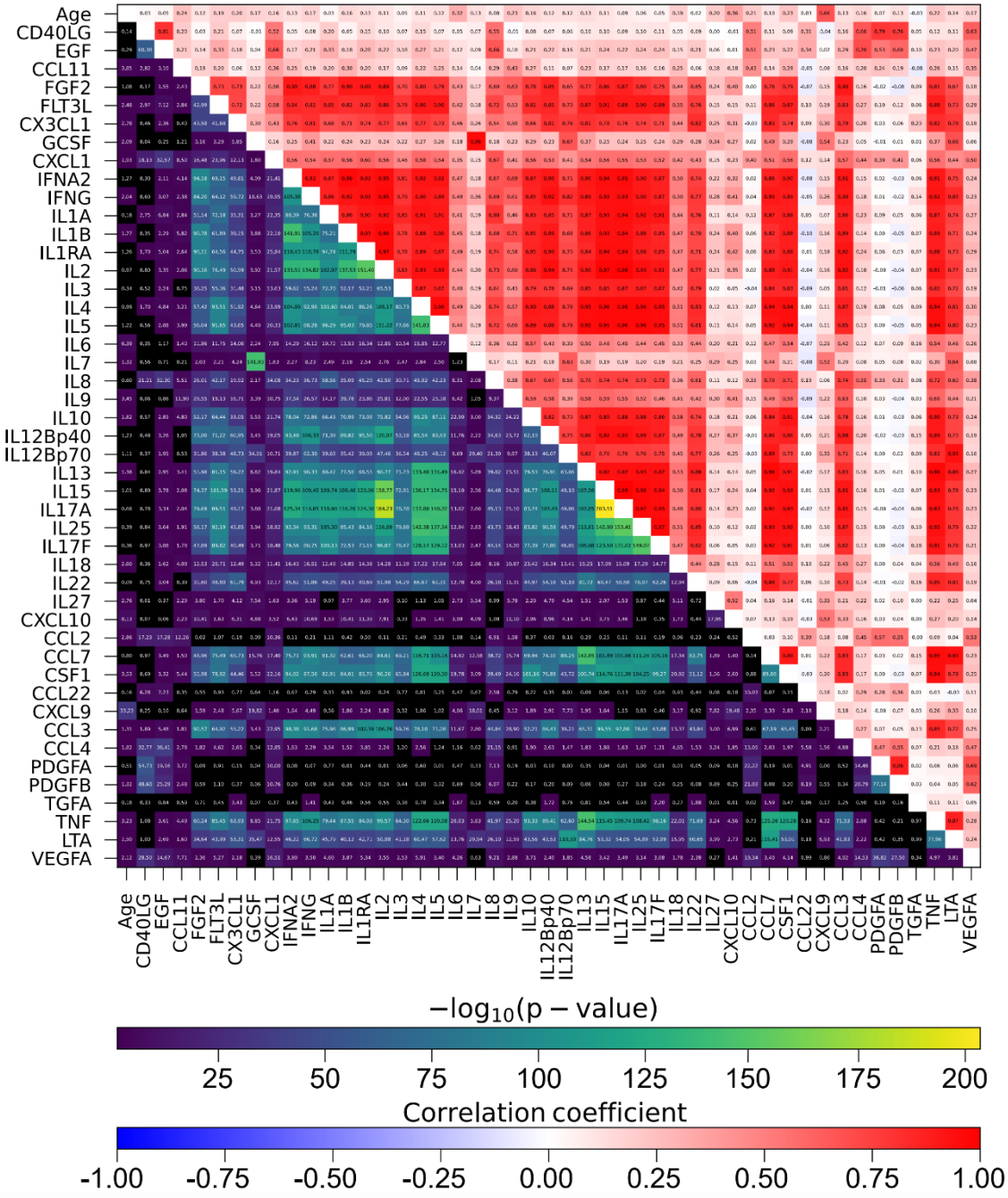


Figure 1. Information about the data analyzed in this study. (a) Distribution of samples by age and sex. (b) Relationships between all immunological biomarkers and age: Pearson correlation coefficient (top triangle) and the p-value for testing non-correlation (bottom triangle) were calculated for each pair of features. All p-values were adjusted using Benjamini-Hochberg FDR-correction (p-value<0.05).

To illustrate the distribution of individual immunological biomarkers depending on age, Figure 2 shows probability density functions. The plot for CXCL9 shows that there is a linear dependence between the values of this biomarker and age.

Figure 2. Probability density functions of sample distributions for all immunological biomarkers (Y axis) and chronological age (X axis). The values of chronological age and all biomarkers are presented in Supplementary Table 1.

2.2. Experiment design

The main goal of this study is to develop a machine learning model that solves the problem of regression of chronological age (in other words, clocks) on a reduced set of the most significant immunological biomarkers (Small Immuno Age - SImAge). Initially, there are 46 immunological markers available, but the goal is to effectively reduce their number so that the accuracy of the age predictor model does not decrease significantly. One of the most important aspects that makes it reasonable to reduce the number of features used in the model is cost-effectiveness. There are manufacturers (i.e. Merck KGaA, Bio-Rad) who are able to produce custom panels with a pre-selected set of biomarkers (for example, specific 10 immunological parameters instead of 46), which can reduce the total cost of the panel and, therefore, of the analysis by several times.

The workflow used for our study is shown in Figure 3. The first step is to train the models using all available immunological biomarkers (Figure 3(a)). We used a wide variety of machine learning models, ranging from classical linear regression with Elastic Net regularization (a popular choice in age regression tasks [26,27,32,42,44,47]) to GBDT and various DNN (deep neural network) architectures specifically designed for tabular data. The GBDT and DNN models tested in this paper are state-of-the-art models that compete for better results in many studies [57,62,66,67]. One of the goals of this study is to find out which class of models is the best in our special problem setting with a quite limited number of samples. A detailed description of the used models is given in Section 4.2. Multifold cross-validation (Section 4.3) was used to obtain the most robust result, and hyperparametric optimization for the best combination of parameters for each model was used to obtain the best result (Section 4.3).

After the best baseline models were identified, the selection of the most important features among them was performed (Figure 3(b)) based on global explainability functionality of SHAP values (Section 4.4). For this purpose, the Top-3 best models with the lowest averaged mean absolute errors (MAE) across all cross-validation splits were first identified. For these models we ranked features by frequency of occurrence in Top-10 in each cross-validation splits. As a result, 10 features were selected, the most frequent in the Top-10 features of all the best models. To build the best portable model for predicting human immunological age (SImAge), we used the same types of models that provided the best result in the previous step (Top-3 models). However, now the models do not take all 46 immunological parameters as input, but only the 10 best ones determined earlier. Multifold cross validation for robustness (Section 4.3) and hyperparametric search for the best result (Section 4.3) were also used in the building of portable models. As a result, one model with the best MAE on the test subset was chosen, which was defined as SImAge (Figure 3(c)).

Then, for the SImAge model the analysis of the predictions produced by the model was performed (Figure 3(d)). Using SHAP values (Section 4.4), the contribution of certain immunological biomarker values to the model predictions for individual participants was estimated. The cumulative contribution of the individual features to the predictions with positive and negative age acceleration was also analyzed.

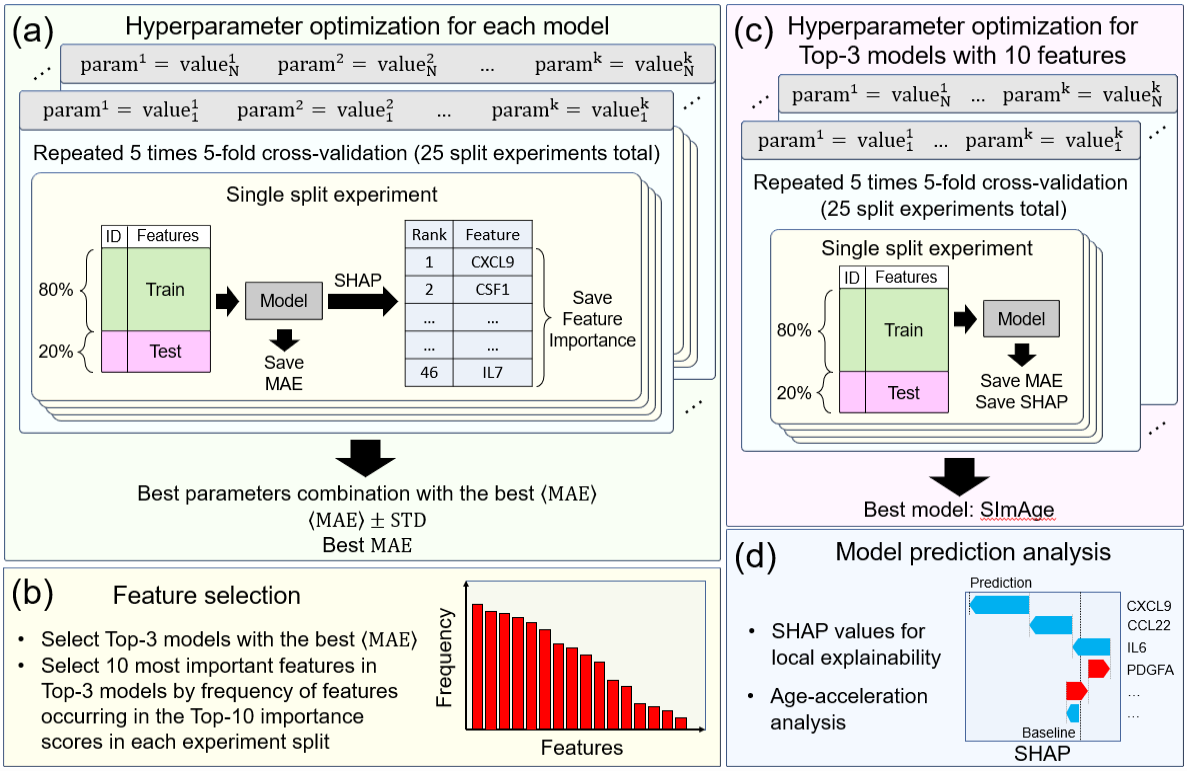


Figure 3. The main steps of the performed analysis. (a) Building baseline machine learning models on a complete list of 46 immunological biomarkers. Each individual experiment was a multifold cross-validation. Multiple runs of such experiments were performed to achieve the best result with optimal values of the model parameters. (b) Identification of the Top-10 most important features from the Top-3 baseline models. (c) Building the SImAge model - the best model on the Top-10 selected biomarkers. Cross validation and hyperparametric search are also used here. (d) Analysis of age predictions produced by SImAge - the contribution of individual immunological parameter values to individual participant predictions (using local explainability and SHAP values), as well as the importance of features in groups of people with positive and negative age acceleration.

2.3. Baseline results

Models from three conceptually different classes are used to solve the problem of regression of chronological age from immunological profile data: (i) the classical linear Elastic Net model, (ii) Gradient Boosted Decision Trees: XGBoost [59], LightGBM [60], CatBoost [61] and (iii) Deep Neural Networks: simple Multilayer Perceptron (MLP), Neural Additive Models (NAM) [92], Neural Oblivious Decision Ensembles (NODE) [68], Deep Abstract Networks (DANet) [93], TabNet [64], Automatic Feature Interaction Learning via Self-Attentive Neural Networks (AutoInt) [65], Self-Attention and Intersample Attention Transformer (SAINT) [66], Feature Tokenizer and Transformer (FT-Transformer) [67]. A detailed description of the used models is given in Section 4.2.

For each model, a single experiment with a particular configuration of model parameters consisted of 25 cross-validation splits (details in Methods, Section 4.3). Within each split, an individual model was trained on the training dataset (80%) and validated on the test dataset

(20%). After 25 splits, the mean and standard deviation of MAE on the test set were calculated. This experiment was repeated many times with different combinations of model parameters during hyperparametric search, whose main aim is to determine the optimal combination of parameters with the minimal average MAE (more details in the methods, Section 4.3). Besides this, the minimal MAE value for the best split was saved. This step of the analysis is represented schematically within the overall study design in Figure 3(a).

Table 1 shows the results of solving the age prediction problem for all baseline models trained on the full set of 46 immunological biomarkers. Among the Top-3 best models in terms of mean MAE, there is one GBDT model (LightGBM) as well as two neural network architectures (DANet and TabNet). All these models have similar MAE values, so for our study we can rather speak about GBDT and DNN parity, than about any obvious superiority of one of the model types. Compared to the other models, the Top-3 models have a mean MAE less than 9 years and the MAE for the best model is less than 7 years. This result is significantly superior to that of the classical linear regression model with Elastic Net regularization, which is one of the most popular choices when building age estimation models for different types of data. All applied GBDT and DNN models (even not included in the Top-3) also surpassed Elastic Net. The result is also better than previously published works in the field of chronological age prediction by immunological biomarkers. The average reconstruction error on the test dataset for the iAge model is 15.2 years [31], and the MAE for the ipAGE model is 7.27 years on the control group [32].

Table 1. Results of age prediction by baseline models using all 46 immunological biomarkers. The Top-3 models with the best mean MAE values are highlighted.

Type	Model	$\langle \mathrm{MAE} \rangle \pm \mathrm{STD}$	Best MAE
Linear	Elastic Net	11.99 ± 1.71	9.81
GBDT	XGBoost	9.31 ± 0.96	7.75
	LightGBM	$\mathbf{8.88 \pm 0.97}$	$\mathbf{6.90}$
	CatBoost	9.02 ± 0.90	7.26
DNN	MLP	9.66 ± 1.83	7.46
	NAM	10.49 ± 1.06	8.31
	NODE	9.48 ± 0.82	8.30
	DANet	$\mathbf{8.59 \pm 1.01}$	$\mathbf{6.82}$
	TabNet	$\mathbf{8.81 \pm 1.26}$	$\mathbf{6.41}$
	AutoInt	10.69 ± 1.64	8.78
	SAINT	9.66 ± 1.57	7.83
	FT-Transformer	9.24 ± 1.06	7.28

2.4. Feature selection and dimensionality reduction

Linear machine learning models, like Elastic Net, are easy to interpret: the greater the absolute value of the coefficient before the feature, the more important this feature is. The GBDT models considered here have a built-in functionality to determine the feature importance, unlike most of the considered neural network architectures (only NAM [92] and TabNet [64] architectures have such built-in functionality). The same model-agnostic approach, based on the calculation of SHAP (SHapley Additive exPlanations) values, is used to obtain the feature importance in all the analyzed models [94]. This is a game theoretic approach to explaining the results of any machine learning model, which links optimal credit allocation to local explanations using classical Shapley values from game theory and related extensions (details in Methods, Section 4.4) [95].

SHAP values can show how each individual feature either positively or negatively affects the final prediction of the model (age estimation in our case). Using this approach, feature importance values were calculated in each of the 25 splits for the Top-3 best models (75 splits in total). In each individual split, the Top-10 most relevant features were determined according to their ranking by importance value. Figure 4(a) shows the number of times each of the features appeared in the Top-10 in all the splits. In particular, CXCL9 was in the Top-10 in each of the splits, so it takes the maximum frequency value of 75 in the figure. As a result, the 10 immunological parameters with maximum frequencies were selected (highlighted in red in Figure 4(a)). These are the most robust biomarkers and were selected based on the results of all baseline models. These biomarkers are then used to construct a portable immunological clock. This step of the study is shown schematically as part of the general flowchart in Figure 3(b). Such Top-10 biomarkers as IL6 [96,97], CSF1 [98], CXCL10 [99], PDGFA [100], CCL2 [101] are associated with aging, and the other CXCL9 [31], CCL22 [102], PDGFB [103], CCL4 [104], IL1RA [105] are associated with different age-associated and inflammatory pathological conditions.

Figure 4(b) shows the relative distribution of the 10 selected immunological biomarker values and chronological age. The diagonal elements illustrate the distribution of each individual feature, the scatter plots in the upper triangle and the probability density functions in the lower triangle illustrate the relationship between each pair of features. It can be observed that there is a clear proportional relationship between PDGFA and PDGFB, which is expected because both of these biomarkers belong to the group of platelet-derived growth factor (PDGF), growth factors for fibroblasts, smooth muscle cells and glia cells [106]. The values of the Pearson correlation coefficients and the corresponding p-values for these biomarkers can be found in Figure 1(b). It should also be noted that there are no biomarkers from the above-mentioned set of mutually correlated interleukins among the 10 selected features.

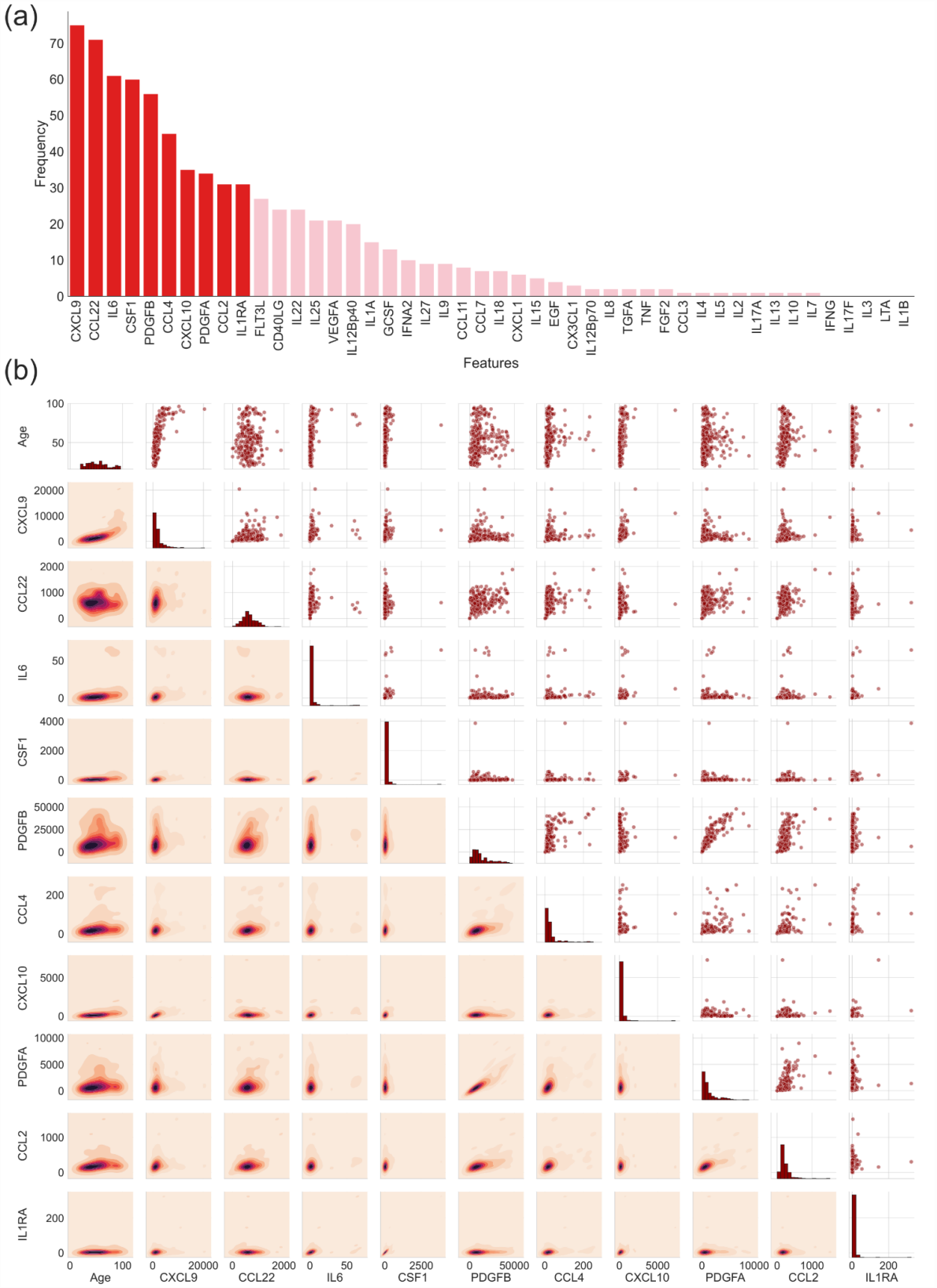


Figure 4. The 10 most important immunological features that were selected for the construction of the small immunological clocks. (a) Ranking of the features according to their frequency of occurrence in the Top 10 by importance values for all 75 splits (25 splits each for the top 3 models). The 10 selected biomarkers with the highest frequency values are

highlighted in red. (b) Relative distribution of 10 most important immunological biomarker values and chronological age. The diagonal elements illustrate the distribution of each individual feature, the scatter plots in the upper triangle and the probability density functions in the lower triangle illustrate the relationships between each pair of features.

2.5. Small immunological clocks (SImAge)

Portable chronological age regression models are constructed for the selected 10 most significant immunological features. Models ranked in the Top-3 best baseline results are considered: LightGBM [60], DANet [93], TabNet [64].

Like in the baseline experiments, here for each model with a certain combination of parameters 5 times repeated 5-fold cross-validation was performed (total of 25 splits). Each split consisted in splitting the dataset into a training subset (80% of samples) and a test subset (remaining 20% of samples). The results of model validation on the test subset were saved for all splits, then the average MAE was calculated and the best MAE achieved for one particular split was determined (more details in the Methods, Section 4.3). A hyperparametric search was also performed to find the best combination of parameters that provided the smallest average MAE value (more details in the Methods, Section 4.3). This step of the study is shown schematically as part of the general flowchart in Figure 3(c).

Table 2 shows the results for the LightGBM, DANet, and TabNet models, which received 10 selected biomarkers as input. The best result obtained for the DANet model is highlighted in the table. As a result it turned out that both average MAE and MAE for the best model are superior to the corresponding ones for all baseline models. This allows us to conclude that 10 immunological parameters are sufficient to obtain a valid age estimation. It is also worth noting that a panel with 10 immunological biomarkers is several times cheaper than a full panel of 46 biomarkers, which also makes the results of the analysis more affordable for a particular participant.

Table 2. Results of models based on 10 selected immunological biomarkers. The DANet model with the best results is highlighted.

Model	$\langle \text{MAE} \rangle \pm \text{STD}$	Best MAE
LightGBM	8.33 ± 0.73	6.67
DANet	$\mathbf{8.06 \pm 0.90}$	$\mathbf{6.28}$
TabNet	8.39 ± 0.95	6.36

The final best model predicting age on a small number of immunological biomarkers (SImAge) has a MAE=6.28 years on the test set, which outperforms both the iAge and ipAGE models. Figure 5(a) illustrates the age distribution for the training and test sets for the SImAge model (see Methods, Section 4.3 for details on stratification and cross-validation). Figure 5(b) shows the prediction result of this model with ground truth values of chronological age.

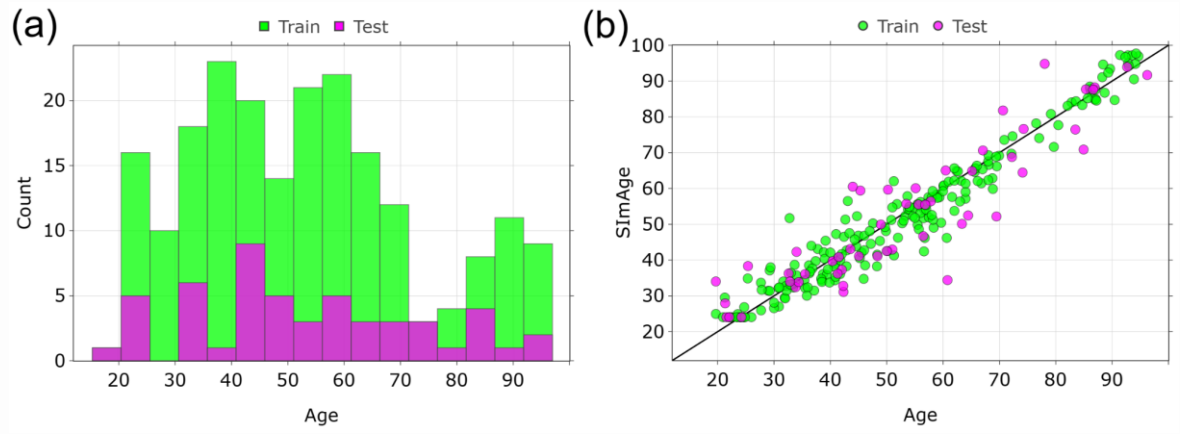


Figure 5. Results for the best model predicting age on a small number of immunological biomarkers (SImAge). (a) Histogram of the samples' age distribution on the training and test sets. (b) Result of SImAge prediction for training and test sets versus chronological age. The black solid line is the bisector of SImAge=Age.

2.6. Model predictions analysis

In addition to obtaining information about the importance of features, SHAP values can also be used for local explainability. As one of the most common XAI methods, SHAP values allow to increase model transparency for a particular prediction. SHAP values also help to explain why a particular sample gets its prediction, and to determine the contribution of individual feature values to the prediction.

For each sample and each feature, SHAP values are calculated to show how a particular value of a selected feature for a certain sample changes the basic prediction of the model (the average prediction of the model in the training set). This approach is most clearly illustrated with waterfall plots, which show the process of how SHAP values are added to the base value and arrive at an individual prediction (Figure 6). This representation allows us to observe which features affected the change in model predictions relative to the mean value in each case, and to analyze the cumulative positive and negative contributions of all the features (which can lead to a rather large model error). Figure 6(a) shows the correct model prediction for a sample with a small age, while Figure 6(b) shows it for a sample with a large age. In both cases we can see that the most important and age-correlated biomarker CXCL9 significantly shifts the model prediction (small values shift negatively, large values shift positively). Figure 6(c, d) shows examples of model predictions with large negative and large positive errors, respectively. Interestingly, in these particular cases, the highest influence on the erroneous model prediction was caused by features values, which are not the most important among the Top-10 (see Figure 4).

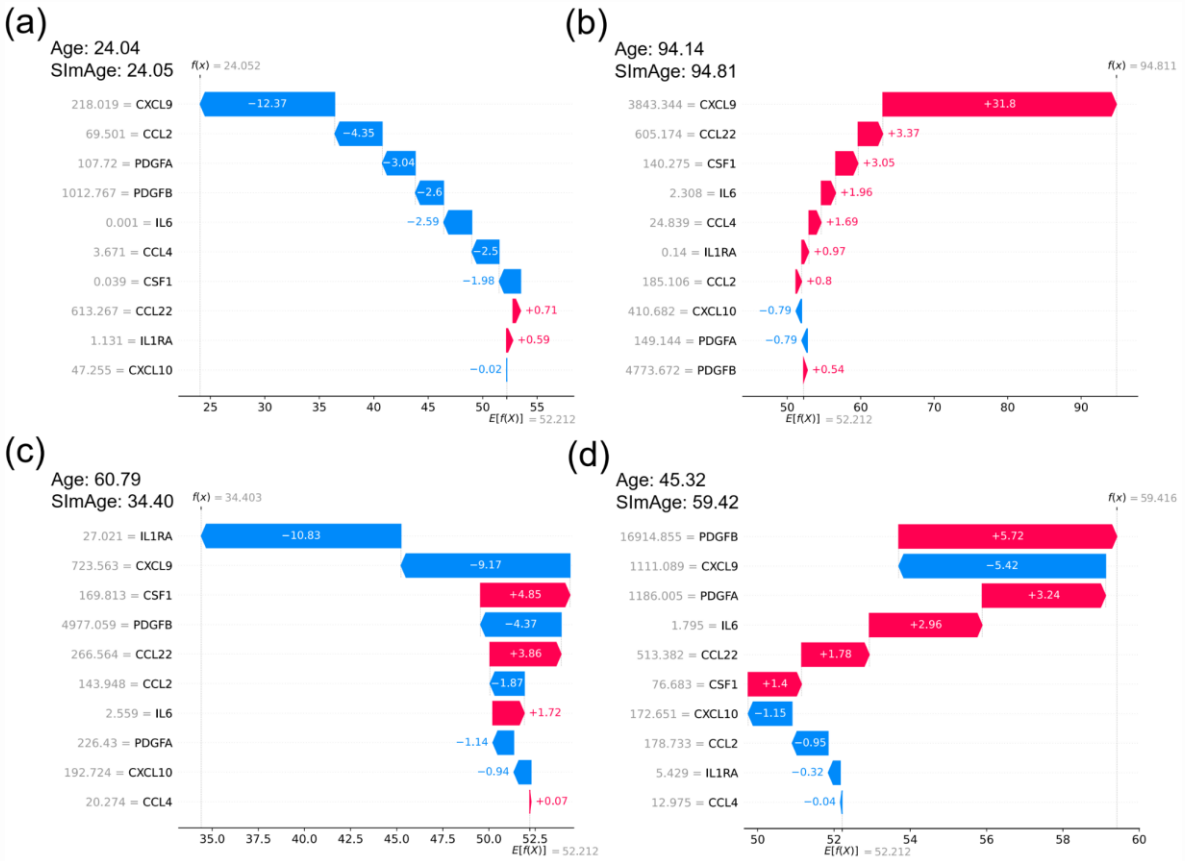


Figure 6. The local explainability of the SImAge model based on SHAP values is illustrated by waterfall plots. The bottom part of each waterfall plot starts with the expected value of the model output (the average prediction of this model on the training set). Each row shows by how much in the positive (red) or negative (blue) direction each feature shifts the prediction relative to the expected value to the final model prediction for that sample. (a) Example of a sample with low model error and small age. (b) Example of a sample with low model error and large age. (c) Example of a sample with a high negative model error. (d) Example of a sample with high positive model error.

The question of the cumulative effect of individual features on the predictions of samples with positive and negative age acceleration is of separate interest. All participants were divided into 2 groups with positive and negative age acceleration (acceleration was calculated as the difference between SImAge and chronological age: Acceleration = SImAge – Age); for each group, the total contribution of individual immunological parameters to the final prediction was analyzed (the average value of absolute SHAP values was calculated). In general, the most important features in both groups are similar (Figure 7), except for differences in the order of individual features: CXCL9 and CCL22 have the highest contribution in both groups (with CXCL9 being more dominant in absolute SHAP value in the group with positive age acceleration, whereas in the group with negative acceleration the difference in absolute values between them is much smaller), PDGFB is ranked significantly higher in the group with negative age acceleration. The distributions of SHAP values and features values in both groups are similar (right beeswarm parts of Figure 7).

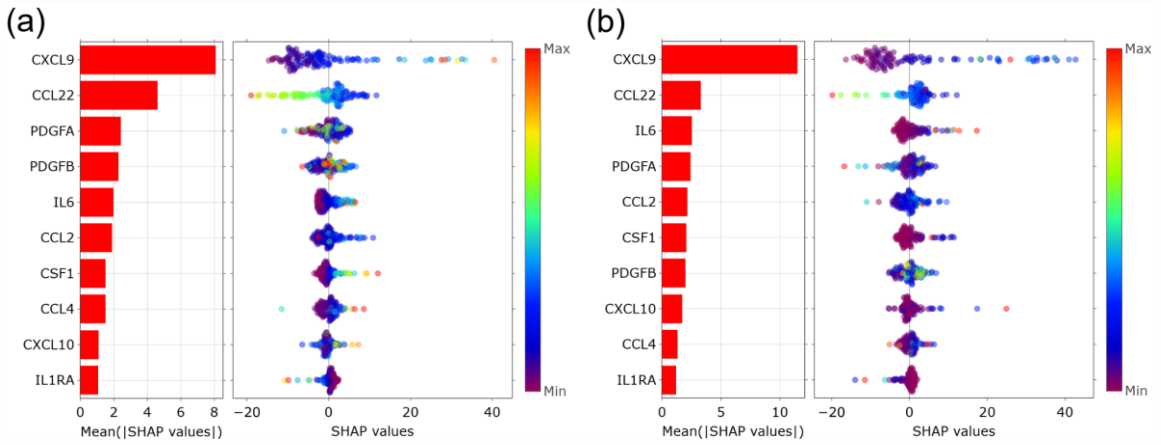


Figure 7. Feature importance for samples with (a) negative and (b) positive age acceleration. The left bar plots illustrate the global importance of each feature, which is calculated as the average absolute value for a given feature across all participating samples. The right beeswarm plots show an information-dense summary of how the top features affect the output of the model. Each sample is represented by a separate dot on each feature row. The X coordinate of each dot is determined by the SHAP value of that feature, the dots pile up along each feature row to show density. The color shows the original relative value of the feature.

3. Discussion

3.1. Conclusion

In this paper, we propose a small SImAge immunological clock that predicts a person's age based on a limited set of immunological biomarkers. The best known studies in this field include such models as IMM-AGE [30], iAge [31], ipAge [32], which are based on different cytokine datasets and show quite good prediction results.

As a baseline model, we consider Elastic Net which is one of the most common approaches in constructing clock models based on biomedical data of different types. We then compare this model with various GBDT and DNN models, which have proven themselves for various machine learning tasks on tabular data. It was shown earlier that deep neural networks are not suitable for all types of tabular data, and GBDT models often outperform deep models on tabular data. Moreover, deep neural networks are often more difficult to optimize than GBDT models [57,58,62,67]. Twelve different models with cross-validation and hyperparametric search were built for 46 immunological parameters. Both types of models, GBDT and DNN, were in the Top-3, and both types performed very well, while Elastic Net significantly underperformed them. As a result, the errors of the best models were lower than those of iAge and ipAGE. The next step was to reduce the model dimensionality. Since not all models are able to calculate feature importance, a unified approach based on SHAP values was used. For the Top-3 models and all cross-validation splits, we ranked the features according to their importance value. According to the frequency of occurrence at the top of the ranking, 10 immunological features that are most robust for different splits within the same model as well as for different models and even different types of models were selected. In particular, CXCL9 has been shown to play

a key role in age-related chronic inflammation [31], CCL22 increases expression of pro-inflammatory mediators and decreases expression of anti-inflammatory mediators [102], IL-6 is associated with mortality risk and physical and cognitive performance [96,107]. Based on the data of 10 selected immunological parameters, new models (3 types of models that showed the best results for the full data set - LightGBM, DANet, TabNet) are built and their results are evaluated. The DANet model showed the smallest error with a result of 8.06 ± 0.90 years (6.28 years for the best split), which is even better than for the complete data. We can say that dimensionality reduction filters out the least significant noisy features, leaving only the most significant ones for age prediction. Another advantage is the potential cheapness of the small custom panel to produce data compared to the full panel, which is applicable not only to immunological data, but also to, for example, epigenetic data [108] and others. Local explainability methods based on SHAP values can be applied to the resulting model to obtain an individual trajectory for predicting a certain age value for each individual participant. As a result, abnormal values of certain parameters can be observed, particularly those contributing to the increase in predicted age, which may indicate the need for an in-depth medical examination. Also, subgroups of participants with positive and negative age acceleration were analyzed separately, and it was shown that CXCL9 has a greater contribution in the group with positive age acceleration. Previously, its contribution to accelerated cardiovascular aging [31], age acceleration in Chagas disease patients [80] was shown. Also PDGFB is ranked higher in the group with negative age acceleration.

Thus, we proposed an approach to build a small model of immunological age - SImAge - using the DANet DNN model, which showed the highest result among all the tested models based on 10 immunological parameters. The obtained model shows the lowest error among all the published studies on immunological profiles. Further work may include expanding the list of models tested, expanding the analyzed data through open sources, and testing the SImAge model for participants with different diseases (including age-associated and/or immunological diseases).

3.2. Limitations

The proposed analysis has several limitations at its various stages. The first limitation is the limited sample size of the data, since machine learning methods typically perform better on large amounts of data. Nevertheless, it is interesting how machine learning techniques focused on big data perform on relatively limited data sets. The second limitation concerns the obtained parameters of machine learning models - they are only locally optimal within the proposed limited hyperparametric search. It is possible that more optimal parameters exist for certain models, which may lead to a completely different ranking of the best models. Third, there are various dimensionality reduction and feature selection algorithms, only one of which is presented in this paper. Finally, the field of neural network architectures is actively developing and this paper considers only a limited list of some of the most popular architectures. Some of the architectures, such as TabTransformer [109] and its modifications, are not presented in the paper, because they are focused on working with categorical features, and when working only with continuous data (like our immunological biomarkers) are reduced to relatively simple MLPs, whose variations are presented in the paper.

4. Methods

4.1. Data details

All possible inconveniences and risks were explained to each participant, as well as the details of the procedure. Each participant signed an informed consent and filled out a personal data processing consent, taking into account the principle of confidentiality, which implies the accessibility of personal data only to the research group. The study was approved by the local ethical committee of Lobachevsky State University of Nizhny Novgorod. All research procedures were in accordance with the 1964 Declaration of Helsinki and its later amendments. The analysis was performed on plasma using the K3-EDTA anticoagulant, without hemolysis and lipemia. Plasma was thawed, spun (3000 rpm, 10 min) to remove debris, and 25 µl was collected in duplicate. Plasma with antibody-immobilized beads was incubated with agitation on a plate shaker overnight (16–18 h) at 2–8 °C. The Luminex® assay was run according to the manufacturer's instructions, using a custom human cytokine 46-plex panel (EMD Millipore Corporation, HCYTA-60 K-PX48). Assay plates were measured using a Magpix (Milliplex MAP). Data acquisition and analysis were done using a standard set of programs MAGPIX®. Data quality was examined based on the following criteria: standard curve for each analyte has a 5P R2 value > 0.95. To pass assay technical quality control, the results for two controls in the kit needed to be within the 95% of CI (confidence interval) provided by the vendor for > 40 of the tested analytes. No further tests were done on samples with results out of range low (< OOR). Samples with results out of range high (> OOR) or greater than the standard curve maximum value (SC max) were not tested at higher dilutions.

4.2. Age estimation models

This study considers various machine learning models for solving the problem of regression of chronological age using immunological profile data. All of these models focus on tabular data, which have several features: (i) features are already extracted, they can be heterogeneous and represent a mixture of continuous, categorical, and ordinal values, which can be independent or correlated; (ii) there is no any inherent position information, which means that the column ordering is arbitrary. This is very different both from textual data, where tokens are always discrete, and ordering impacts semantic meaning, and from images, where the mutual positioning of features (pixels) is important. Models for tabular data must discover correlations without relying on the positional information. A peculiarity of our data representation is that all considered immunological features are continuous (no categorical and ordinal features).

Elastic Net is a relatively simple model and a popular choice for constructing various kinds of biological clocks on tabular data such as epigenetic [26–29], immunological [32,41], transcriptomic [42], metabolomic [43–45], microRNA [46], and proteomic [47] data.

In the last few years a number of papers have been published comparing the efficiency of applying GBDT and DNN to different tabular data. In [58] it was concluded that the superiority of these two classes of models depends on the size of the dataset. For mid-sized and small datasets, GBDT algorithms lead, while parity is observed for large datasets. It was found that uninformative features have a much stronger negative impact on DNN compared to GBDT. The papers [57,62] also found that GBDTs generally perform better than various DNN

architectures. At the same time, there are papers proposing completely new DNN architectures, which for a large list of tabular datasets give better results than GBDT [66,67].
The following subsections present the basic concepts of the models used to predict age from immunological data.

Linear Model: Elastic Net

Elastic net is an extension of linear regression that adds L1 (Lasso) and L2 (Ridge) regularization penalties to the loss function during training. The Ridge penalty shrinks the coefficients of correlated predictors towards each other while the Lasso tends to pick one of them and discard the others. Linear regression assumes a linear relationship between input immunological features (x_1, x_2, ..., x_N) and target variable – chronological age (y are ground truth values and $\hat{y}$ are predictions):

$$\hat{y_i} = \beta^0 + \sum_{j=1}^{M} x_i{}^j \beta^j = \beta^0 + x_i^T \beta,$$

where i is the sample index ($i = 1, \ldots, N$). b_i ($j = 0, \ldots, M$) are coefficients in the linear model, which are found via optimization process that seeks to minimize loss function:

$$loss = \frac{1}{2N} \sum_{i=1}^{N} (y_i - (\beta^0 + x_i^T \beta))^2 + \lambda[\alpha ||\beta||_1 Penalty_{L1} + (1-\alpha) Penalty_{L2}] \ ,$$
$$Penalty_{L1} = \sum_{j=1}^{M} |\beta^j| \ ,$$
$$Penalty_{L2} = \frac{1}{2} \sum_{j=1}^{M} (\beta^j)^2 \ ,$$

where hyperparameter $\lambda \geq 0$ controls the overall strength of both penalties to the loss function and $0 \leq \alpha \leq 1$ is a compromise between Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$). In the performed experiments, equal contribution of both penalty types is set ($\alpha = 0.5$) and only the parameter λ is varied. The implementation of the algorithm was taken from the sklearn library version 1.1.2 [110]. The range of varying parameters is presented in Supplementary Table 2 (sheet "ElasticNet").

Gradient Boosted Decision Trees: XGBoost, LightGBM, and CatBoost

XGBoost (eXtreme Gradient Boosting) [59], LightGBM (Light Gradient Boosting Machine) [60], and CatBoost (Categorical Boosting) [61] are all GBDT (Gradient Boosted Decision Tree) algorithms. These models are based on ensemble learning, a technique that combines predictions from multiple models to produce predictions that are potentially more stable and better generalizable. By averaging individual model errors, the risk of overtraining is reduced while maintaining strong prediction performance. There are two main classes of ensemble learning methods: bagging (a method that builds models in parallel using a random subset of data and then aggregates the predictions from all the models), and boosting (a method that builds models sequentially using all data, with each model improving upon the error of the previous one). XGBoost, LightGBM, and CatBoost are all variations of gradient boosting algorithms. The differences between the three models lie in the splitting method and type of tree growth.

Splitting Method.

XGBoost uses a pre-sorted algorithm which takes into account all features and sorts them by value. After that, a linear scan is done to select the best split with the highest information gain.

The histogram-based modification instead of considering all values groups them into discrete bins and finds the split point based on these bins, which is more efficient than the pre-sorted algorithm.

LightGBM offers gradient-based one-side sampling (GOSS) which selects the split using all instances with large gradients (higher errors) and random samples of instances with small gradients (smaller errors). To keep the same data distribution when computing information gain, GOSS introduces a constant multiplier for data instances with small gradients. As a result, GOSS achieves a good balance between increasing the speed of reducing the number of data points and preserving accuracy for learned decision trees.

Catboost offers Minimal Variance Sampling (MVS) technique, which is a weighted sampling version of Stochastic Gradient Boosting (SGB) [111]. The weighted sampling occurs at tree-level, not at split-level. The observations for each boosting tree are sampled to maximize the accuracy of the split scoring.

Tree Growth.

XGBoost splits trees up to a certain maximum depth (specified by a hyperparameter) and then starts pruning the tree backwards and removes splits beyond which there is no positive gain. It uses this approach since a split with no loss reduction can be followed by a split with loss reduction.

LightGBM uses leaf-wise tree growth: it chooses to grow a leaf that minimizes losses, allowing an unbalanced tree to grow.

Catboost builds a balanced tree: at each level of such a tree, it chooses the feature-split pair that leads to the lowest loss, and it is used for all nodes in the level.

Besides the described properties, these models differ in algorithms for imputation of missing values, processing of categorical features and methods for computing feature importance, which is not relevant in this study: in the immunology dataset there are no missing values, no categorical features, and the feature importance will be computed in a universal way for all models using SHAP values. Software implementations of these models are used from the corresponding packages: XGBoost version 1.6.2, LightGBM version 3.3.2, CatBoost version 1.1. The range of varied parameters as well as their description is presented in Supplementary Table 2 (sheets “XGboost”, “LightGBM”, “CatBoost”). In all GBDT models, the maximum number of rounds of the training process was set to 1000, and the number of early stopping rounds was set to 100 (models stopped the training process if the evaluation metric in the test dataset was not improving for 100 rounds).

Deep Neural Networks

Several implementations of neural network architectures focused on tabular data are used in this work. Many of the considered networks have similar blocks and layers, which have identical concepts and functions, but may slightly differ in the implementation details. The concepts and basic elements of the deep neural network architectures used in this work, taking into account the basic details of their software implementations, are presented in Figure 8. For training all neural network models, the Adam optimization algorithm was used [112], whose main hyperparameters (learning rate and weight decay) were selected for each model individually during the hyperparametric search. Their specific values are presented in Supplementary Table

2. For all models, the maximum number of epochs has been set to 1000, and for early stopping the number of epochs has been set to 100 (models stopped the training process if the evaluation metric on the test dataset was not improving for 100 epochs). All models were implemented in the PyTorch Lightning framework [113], which is a lightweight PyTorch [114] wrapper for high-performance AI research. Feature importance for all models were calculated using SHAP values and corresponding global explainability methods [94] (see Section 4.4 for details).

Common Elements.

Element-wise addiction, multiplication and concatenation (of output vectors) are used in many neural network architectures to variously accumulate outputs of several previous layers at once.

Dropout is a special layer that provides regularization and prevents co-adaptation of neurons by zeroing out randomly selected neurons [115]. It works only during training and is turned off in evaluation mode.

Mask layers are used for instance-wise feature selection in several considered models (TabNet and DANet).

Linear (or fully-connected, or dense) layer applies a linear transformation to the input data with an optional bias. It is often used in the final stages of neural networks (but not only in them) to change the dimensionality of the output of the preceding layer, so that the model can easily determine the relationship between the data values in which the model operates.

Normalization layers bring all input values to the same scale with a mean equal to zero and variance equal to one. This improves productivity and stabilizes neural networks.

Embedding layers transform input continuous features into a new embedding space of a different dimensionality, using both linear and nonlinear transformations. During the neural network training, similar samples can be clustered in this space.

Activation functions are used to activate neurons from hidden layers and to introduce different nonlinearities in the decision boundary of the neural networks. In many used models the type of activation function is a hyperparameter.

EntMax [116] and SparseMax [117] are modifications of activation functions that transform continuous vectors into a discrete probability distribution, resulting in a sparse output. They are used to implement the attention mechanism in some models, allowing to take into account the influence of only the most important features.

The designations of the described general blocks in the general scheme of the neural network concepts used in this work are shown in Figure 8(a).

Multilayer Perceptron (MLP).

A simple neural network architecture, consisting of several dense blocks, which in turn consist of linear, dropout, and activation layers, as shown in Figure 8(b). The software implementation of the neural network is adapted from the pytorch-widedeep library version 1.2.1. The main hyperparameters are the architecture of dense blocks, activation function type, and probability of dropout. Exact values of the varied parameters are presented in the Supplementary Table 2 (sheet “MLP”).

Neural Additive Model (NAM).

In this architecture [118], a separate dense block consisting of linear layers and activation functions is constructed for each input feature. All these independent dense blocks are trained jointly and can learn arbitrarily complex relations between their input and output features. The main advantage of this method is its ease of interpretability, as each input function is processed

independently by a different neural network. The contribution of each individual function to the overall result can be easily assessed. However, in this paper we will use the interpretability via SHAP values method. A scheme of the architecture is shown in Figure 8(c). The software implementation of the neural network is adapted from the NAM library version 0.0.3 implementing this architecture for the PyTorch framework. The hyperparameters of the network are the architecture of dense blocks, activation function type, probability of dropout and additional regularization parameters [92]. Exact values of the varying parameters are given in the Supplementary Table 2 (sheet "NAM").

Neural Oblivious Decision Ensembles (NODE).

This architecture [68] combines decision trees and deep neural networks, so that they can be trained (via gradient-based optimization) in an end-to-end manner. This method is based on so-called oblivious decision trees (ODTs), a special type of decision tree that uses the same splitting function and splitting threshold in all internal nodes of the same depth (as in CatBoost [61]). Besides using oblivious decision tables, another important design choice is the recent EntMax transformation [116], which effectively performs soft splitting feature selection in decision trees within the NODE architecture. A scheme of the architecture is shown in Figure 8(d). The software implementation of the neural network is adapted from the PyTorch Tabular package version 0.7.0 [119]. The hyperparameters of the network are number of NODE layers, numbers of ODTs in each NODE layer, the depth of ODTs, sparse activation function type. Exact values of varying parameters are presented in Supplementary Table 2 (sheet "NODE").

Deep Abstract Network (DANet).

The architecture [93] is focused around abstract layers, the main idea of which is to group correlated features (through a sparse learnable mask) and create higher level abstract features from them. The DANet architecture consists of stucking such abstract layers into blocks. The blocks are combined one by one, and each block has a shortcut connection that adds raw features back to each block. A scheme of the architecture is shown in Figure 8(e). The software implementation of the neural network is adapted from the supplementary code repository of the corresponding article [93]. The hyperparameters of the network are number of abstract layers to stack, number of masks, the output feature dimension in the abstract layer, and dropout rate in the shortcut module. Exact values of the varying parameters are presented in the Supplementary Table 2 (sheet "DANet").

TabNet.

The architecture [64] consists of sequential modules (steps), each of which implements a sequential attention mechanism that selects the most significant features. Attentive transformers learn the relationship between the relevant features and decide which features to pass using the SparseMax (or EntMax) functions. In the feature transformer block, all the selected features are processed to generate the final output. Each feature transformer is composed of several normalization layers, linear layers and several Gated Linear Unit (GLU) blocks [120], that control which information should pass through the network. One of the main features of the model is the possibility to obtain feature importance. However, in this paper we will use the method of interpretability via SHAP values. The scheme of the architecture is shown in Figure 8(f). The software implementation of the neural network is adapted from the pytorch-widedeep library version 1.2.1. The hyperparameters of the network are width of the decision prediction layer, number of decision steps, number of GLU blocks, parameters of

batch normalization, and relaxation parameters. Exact values of the varying parameters are presented in Supplementary Table 2 (sheet “TabNet”).

AutoInt.

This model tries to automatically learn the interactions between features and create a better representation, and then use that representation in downstream tasks. The model first transforms the features into embeddings, and then applies a series of attention-based transformations to the embeddings. The output of the model is the sum of the outputs of the multi-head self-attention mechanism and the residual connection block. A scheme of the architecture is shown in Figure 8(e). The software implementation of the neural network is adapted from the PyTorch Tabular package version 0.7.0 [119]. The hyperparameters of the network are different characteristics of multi-headed attention layers, dropout rates, and linear layers configuration in MLP. Exact values of the varied parameters are presented in the Supplementary Table 2 (sheet “AutoInt”).

Self-Attention and Intersample Attention Transformer (SAINT).

This hybrid architecture is based on self-attention, which applies attention to both rows and columns, and includes an enhanced embedding method [66]. SAINT projects all features into a combined dense vector space. These projected values are passed as tokens into the transformer encoder, which performs a special attention mechanism. A scheme of the architecture is shown in Figure 8(e). The software implementation of the neural network is adapted from the pytorch-widedeep library version 1.2.1 and uses Einstein notation for operations on tensors (rearrange from einops python package [121]). The hyperparameters of the network are the number of attention heads per transformer block, the number of SAINT-transformer blocks, dropout rates, and linear layers configuration in MLP. Exact values of the varying parameters are presented in Supplementary Table 2 (sheet “SAINT”).

Feature Tokenizer and Transformer (FT-Transformer).

Like the previous two models, it is an adaptation of the transformer architecture [122] for the tabular domain. First, in this architecture, the function tokenizer transforms features into embeddings. Then the embeddings are processed by the Transformer layer stack with special multi-head self-attention blocks. A scheme of the architecture is shown in Figure 8(e). The software implementation of the neural network is adapted from the pytorch-widedeep library version 1.2.1. The hyperparameters of the network are number of attention heads per transformer block, number of FT-Transformer blocks, dropout rates, and linear layers configuration in MLP. Exact values of the varying parameters are presented in Supplementary Table 2 (sheet “FT-Transformer”).

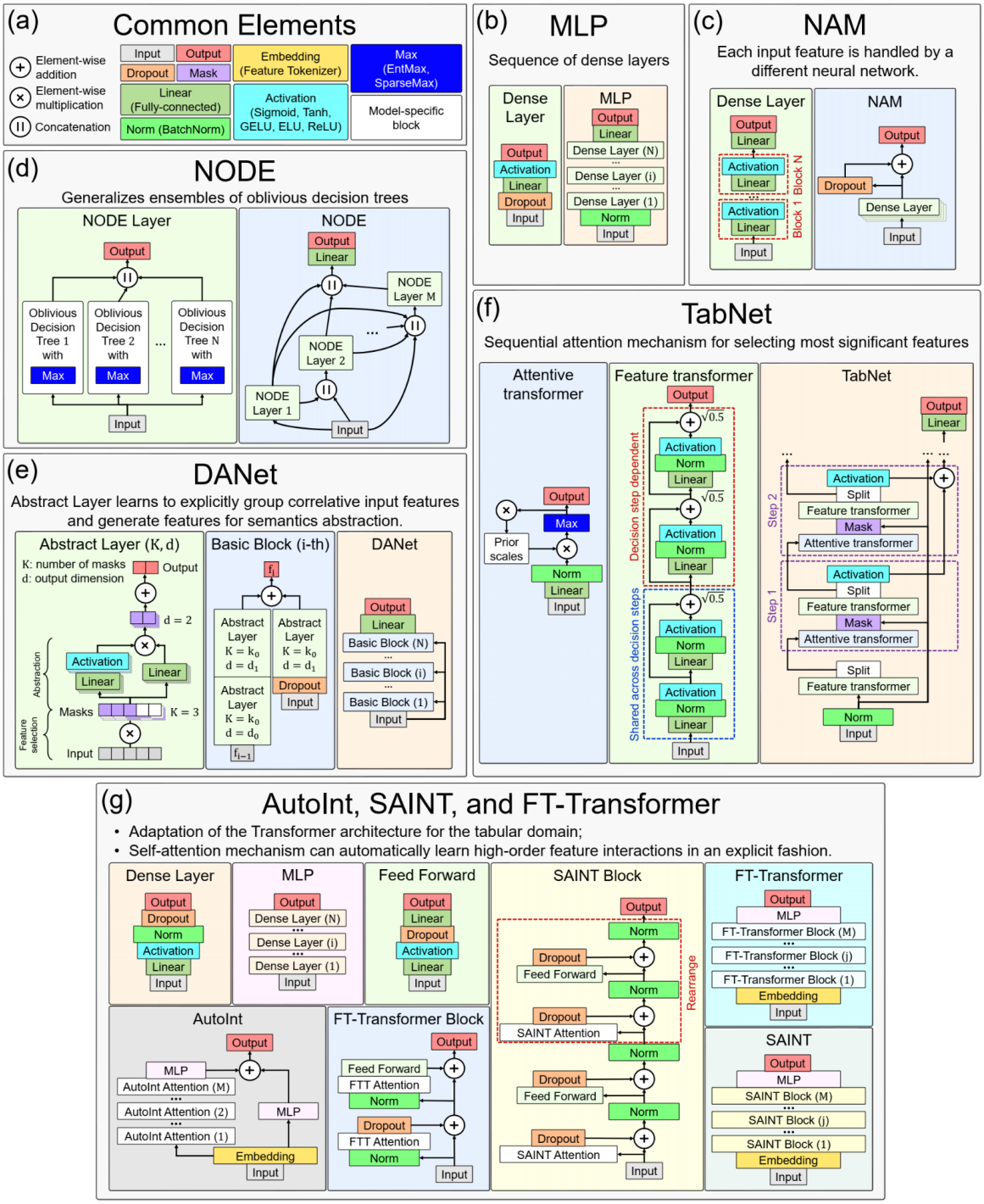


Figure 8. Main concepts of neural network architectures used for chronological age regression on immunological parameters. (a) Blocks and layers used in many architectures. (b) Multilayer Perceptron (MLP) - the simplest neural network architecture composed of a sequence of dense layers. (c) Neural Additive Model (NAM) - the architecture where a separate neural network is built for each input feature. (d) Neural Oblivious Decision Ensembles (NODE) - the architecture that combines ensembles of oblivious decision trees with the benefits of both end-to-end gradient-based optimization and the power of multi-layer hierarchical representation learning. (e) Deep Abstract Network (DANet), based on special abstract layers that learn to explicitly group correlative input features and generate features

for semantics abstraction. (f) TabNet architecture, which implements a sequential attention mechanism to select the most important features. (g) AutoInt, SAINT, and FT-Transformer are conceptually similar neural network architectures, based on adaptation of Transformer architecture for the tabular domain.

4.3. Experiment details

Cross-validation and metrics

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited amount of data. In this paper, we used a repeated k-fold cross-validation procedure with 5 repeats and k=5 (25 split experiments in total). A single run of the k-fold cross-validation procedure can lead to a noisy estimate of model performance. Different splits of the data can produce very different results. Repeated k-fold cross-validation can improve the estimated performance of the machine learning model. This involves simply repeating the cross-validation procedure several times and reporting the mean result over all folds from all runs. This mean result is expected to be a more accurate estimate of the true unknown underlying mean performance of the model in the dataset, calculated using standard error.

The software implementation of the repeated k-fold cross-validation procedure was used from the the sklearn library version 1.1.2 [110] with the stratification parameter set to make the distribution of samples by age approximately the same in the train and test sets.

Various metrics such as MAE [32,42,44,45,51,52,52–55], RMSE [26,44], correlation between chronological age and predicted age [28,45,47], median error [27,50], and others are used to evaluate the efficiency of machine learning models in solving age estimation problems. The most popular among them is MAE. In this work, MAE was also chosen as the main observable metric. It is calculated by the formula:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|,$$

where y_i – ground truth values of target metric (chronological age), $\hat{y}_i$ – model predictions, N – number of samples. For this metric two values will be given for each model: (i) the main metric by which models are sorted – average MAE ($\pm$STD) for all 25 cross-validation splits; (ii) the secondary metric – MAE for the best model (for one particular split, where the best results were obtained).

Hyperparameters optimization

Each machine learning model considered in this paper has certain hyperparameters. To obtain the best result, hyperparameter search of the best combination was performed, which yielded the minimum MAE value averaged over all the split cross-validations.

The software implementation of hyperparameter optimization was taken from the optuna python package version 3.0.2 [123], using the Tree-structured Parzen Estimator (TPE) algorithm [124]. TPE is an iterative process that uses the history of evaluated hyperparameters to create a probabilistic model that is used to propose the next set of hyperparameters for evaluation. The total number of optimization trials for each considered model was set to 200. The number of random sampled startup trials was set to 50. The number of candidate samples used to calculate

the expected improvement was set to 5. Supplementary Table 2 lists the varied hyperparameters for each model with a description and range of variation.

4.4. SHAP values

SHAP (Shapley Additive ExPlanations) values are a game theory approach to explain model predictions. The algorithm for calculating SHAP values starts with some base value (averaged model output) to predict based on prior knowledge, and then sequentially tests the features of the data to explain the effect of the introduction of that feature on the base value to make the final prediction. It takes into account the order in which the features are introduced, as well as the interaction between features, which helps to understand the model performance.

SHAP values for a regression problem have the same dimensionality as the original dataset - they are computed for each sample and for each feature. SHAP values show how a particular value of a selected feature for a particular sample changed the baseline model prediction (the average model prediction on the train dataset).

The model output can be interpreted as the reward that will be distributed among the group of features that helped obtain the reward. SHAP values determine how the model output, the reward, should be distributed among the features. Given the prediction of model f on sample x_i the method estimates SHAP values $\phi(x_i, f)$ by finding an approximation to the model, $g_i(x_i)$, one model per sample x_i. This model $g_i(x_i)$ is locally accurate: its output converges to $f(x_i)$ for this sample by summing up M attributions $\phi_j(x_i, f)$ for Mfeatures. The model $g_i(x_i)$ is consistent: features that are truly important to one model's predictions versus another are always assigned higher importance. The summarization of effects:

$$f(x_i) \approx g_i(x_i) = E[f(x_i)] + \sum_{j=1}^{M} \phi_j(x_i, f),$$

where $E[f(x_i)]$ is the expected reward across the training samples [125]. SHAP values can be computed as a weighted sum of the contribution of a feature to the model prediction, taking into account all possible permutations of other features introduced into the model. This method requires significant computational resources, so various estimators of SHAP values (explainers) are implemented in the shap python package version 0.41.0 [94].

For GBDT and DNN models, [94] presents special types of explainers, TreeExplainer and DeepExplainer, respectively. DeepExplainer does not support specific layers of some architectures (for example, SparseMax and EntMax in TabNet and Node). For this reason, the KernelExplainer is used as unified as possible for all model types.

Data availability statement

All data generated during this study are included in this published article and its supplementary files.

Abbreviations

AI: Artificial Intelligence; AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Network; CatBoost: Categorical Boosting; CI: Confidence Interval; DANet: Deep

Abstract Network; DNN: Deep Neural Network; FDR: False Discovery Rate; FT-Transformer: Feature Tokenizer and Transformer; GBDT: Gradient-Boosted Decision Tree; GLU: Gated Linear Unit; GOSS: Gradient-based One-Side Sampling; LightGBM: Light Gradient Boosting Machine; MAE: Mean Absolute Error; MLP: Multilayer Perceptron; MVS: Minimal Variance Sampling; NAM: Neural Additive Model; NODE: Neural Oblivious Decision Ensemble; ODT: Oblivious Decision Tree; OOR: Out Of Range; RMSE: Root Mean Squared Error; SAINT: Self-Attention and Intersample Attention Transformer; SGB: Stochastic Gradient Boosting; SHAP: Shapley Additive Explanations; STD: Standard Deviation; TPE: Tree-structured Parzen Estimator; XAI: Explainable Artificial Intelligence; XGBoost: eXtreme Gradient Boosting.

Competing Interests

The authors declare that they have no competing interests.

Author's Contributions

Conceptualization: A.K., I.Y., M.V., M.I.; Investigation: E.K.; Formal analysis: A.K., I.Y.; Methodology: A.K., I.Y., M.I.; Software: A.K., I.Y.; Resources: E.K., M.V.; Supervision: M.G.B., C.F., M.V., M.I.; Visualization: A.K., I.Y.; Writing – original draft: A.K., I.Y.; Writing – review & editing: A.K., I.Y., E.K., M.G.B., C.F., M.V., M.I.

Acknowledgements

The authors acknowledge the use of computational resources provided by the "Lobachevsky" supercomputer.

References

1. Dorshkind, K., Montecino-Rodriguez, E. & Signer, R. A. J. The ageing immune system: is it ever too old to become young again? *Nat Rev Immunol* **9**, 57–62 (2009).
2. Gruver, A. L., Hudson, L. L. & Sempowski, G. D. Immunosenescence of ageing. *J Pathol* **211**, 144–156 (2007).
3. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
4. Nikolich-Žugich, J. The twilight of immunity: emerging concepts in aging of the immune system. *Nat Immunol* **19**, 10–19 (2018).
5. Kotas, M. E. & Medzhitov, R. Homeostasis, inflammation, and disease susceptibility. *Cell* **160**, 816–827 (2015).
6. Liu, F. *et al.* Mechanosignaling through YAP and TAZ drives fibroblast activation and fibrosis. *Am J Physiol Lung Cell Mol Physiol* **308**, L344-357 (2015).
7. Crusz, S. M. & Balkwill, F. R. Inflammation and cancer: advances and new agents. *Nat Rev Clin Oncol* **12**, 584–596 (2015).

8. Franceschi, C. & Campisi, J. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *J Gerontol A Biol Sci Med Sci* **69 Suppl 1**, S4-9 (2014).
9. Efeyan, A., Comb, W. C. & Sabatini, D. M. Nutrient-sensing mechanisms and pathways. *Nature* **517**, 302–310 (2015).
10. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, Inflammation, and Cancer. *Cell* **140**, 883–899 (2010).
11. Hunter, R. L. *et al.* Inflammation induces mitochondrial dysfunction and dopaminergic neurodegeneration in the nigrostriatal system: LPS effects nigrostriatal mitochondria. *Journal of Neurochemistry* **100**, 1375–1386 (2007).
12. Nathan, C. & Cunningham-Bussel, A. Beyond oxidative stress: an immunologist's guide to reactive oxygen species. *Nat Rev Immunol* **13**, 349–361 (2013).
13. Oh, J., Lee, Y. D. & Wagers, A. J. Stem cell aging: mechanisms, regulators and therapeutic opportunities. *Nat Med* **20**, 870–880 (2014).
14. Thevaranjan, N. *et al.* Age-Associated Microbial Dysbiosis Promotes Intestinal Permeability, Systemic Inflammation, and Macrophage Dysfunction. *Cell Host & Microbe* **21**, 455-466.e4 (2017).
15. Goldberg, E. L. & Dixit, V. D. Drivers of age-related inflammation and strategies for healthspan extension. *Immunol Rev* **265**, 63–74 (2015).
16. Franceschi, C. *et al.* Inflamm-aging. An evolutionary perspective on immunosenescence. *Ann N Y Acad Sci* **908**, 244–254 (2000).
17. Franceschi, C. *et al.* Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. *Mech Ageing Dev* **128**, 92–105 (2007).
18. Franceschi, C., Garagnani, P., Parini, P., Giuliani, C. & Santoro, A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol* **14**, 576–590 (2018).
19. Franceschi, C., Garagnani, P., Vitale, G., Capri, M. & Salvioli, S. Inflammaging and 'Garb-aging'. *Trends Endocrinol Metab* **28**, 199–212 (2017).
20. Morrisette-Thomas, V. *et al.* Inflamm-aging does not simply reflect increases in pro-inflammatory markers. *Mechanisms of Ageing and Development* **139**, 49–57 (2014).
21. Murabito, J. M. *et al.* Measures of Biologic Age in a Community Sample Predict Mortality and Age-Related Disease: The Framingham Offspring Study. *The Journals of Gerontology: Series A* **73**, 757–762 (2018).
22. Levine, M. E. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *J Gerontol A Biol Sci Med Sci* **68**, 667–674 (2013).
23. Belsky, D. W. *et al.* Quantification of biological aging in young adults. *Proc Natl Acad Sci U S A* **112**, E4104-4110 (2015).
24. Swindell, W. R. *et al.* Indicators of 'healthy aging' in older women (65-69 years of age). A data-mining approach based on prediction of long-term survival. *BMC Geriatr* **10**, 55 (2010).

25. Newman, A. B., Boudreau, R. M., Naydeck, B. L., Fried, L. F. & Harris, T. B. A physiologic index of comorbidity: relationship to mortality and disability. *J Gerontol A Biol Sci Med Sci* **63**, 603–609 (2008).
26. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* **49**, 359–367 (2013).
27. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).
28. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* **10**, 573–591 (2018).
29. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303–327 (2019).
30. Alpert, A. *et al.* A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nat Med* **25**, 487–495 (2019).
31. Sayed, N. *et al.* An inflammatory aging clock (iAge) based on deep learning tracks multimorbidity, immunosenescence, frailty and cardiovascular aging. *Nat Aging* **1**, 598–615 (2021).
32. Yusipov, I. *et al.* Accelerated epigenetic aging and inflammatory/immunological profile (ipAGE) in patients with chronic kidney disease. *Geroscience* **44**, 817–834 (2022).
33. Orrù, V. *et al.* Genetic variants regulating immune cell levels in health and disease. *Cell* **155**, 242–256 (2013).
34. Patin, E. *et al.* Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nat Immunol* **19**, 302–314 (2018).
35. Roederer, M. *et al.* The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. *Cell* **161**, 387–403 (2015).
36. Tsang, J. S. Utilizing population variation, vaccination, and systems biology to study human immunology. *Trends Immunol* **36**, 479–493 (2015).
37. Shen-Orr, S. S. & Furman, D. Variability in the immune system: of vaccine responses and immune states. *Curr Opin Immunol* **25**, 542–547 (2013).
38. Kaczorowski, K. J. *et al.* Continuous immunotypes describe human immune variation and predict diverse responses. *Proc Natl Acad Sci U S A* **114**, E6097–E6106 (2017).
39. Burch, J. B. *et al.* Advances in geroscience: impact on healthspan and chronic disease. *J Gerontol A Biol Sci Med Sci* **69 Suppl 1**, S1-3 (2014).
40. Lehallier, B., Shokhirev, M. N., Wyss-Coray, T. & Johnson, A. A. Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging. *Aging Cell* **19**, e13256 (2020).
41. Arvey, A. *et al.* Age-associated changes in the circulating human antibody repertoire are upregulated in autoimmunity. *Immunity & Ageing* **17**, 28 (2020).
42. Meyer, D. H. & Schumacher, B. BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell* **20**, (2021).
43. Ravera, S. *et al.* Discrete Changes in Glucose Metabolism Define Aging. *Sci Rep* **9**, 10347 (2019).
44. Hwangbo, N. *et al.* A Metabolomic Aging Clock Using Human Cerebrospinal Fluid. *The Journals of Gerontology: Series A* **77**, 744–754 (2022).

45. Robinson, O. *et al.* Determinants of accelerated metabolomic and epigenetic aging in a UK cohort. *Aging Cell* **19**, e13149 (2020).
46. Huan, T. *et al.* Age-associated microRNA expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell* **17**, (2018).
47. Sathyan, S. *et al.* Plasma proteomic profile of age, health span, and all-cause mortality in older adults. *Aging Cell* **19**, e13250 (2020).
48. Li, X., Li, W. & Xu, Y. Human Age Prediction Based on DNA Methylation Using a Gradient Boosting Regressor. *Genes* **9**, 424 (2018).
49. Chen, Y. *et al.* Human gut microbiome aging clocks based on taxonomic and functional signatures through multi-view learning. *null* **14**, 2025016 (2022).
50. Galkin, F., Mamoshina, P., Kochetov, K., Sidorenko, D. & Zhavoronkov, A. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. *Aging Dis* **12**, 1252–1262 (2021).
51. Mamoshina, P. *et al.* Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. *Frontiers in Genetics* **9**, (2018).
52. Galkin, F. *et al.* Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. *iScience* **23**, 101199 (2020).
53. Putin, E. *et al.* Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging (Albany NY)* **8**, 1021–1030 (2016).
54. Mamoshina, P. *et al.* Population Specific Biomarkers of Human Aging: A Big Data Study Using South Korean, Canadian, and Eastern European Patient Populations. *The Journals of Gerontology: Series A* **73**, 1482–1490 (2018).
55. Mamoshina, P. *et al.* Blood Biochemistry Analysis to Detect Smoking Status and Quantify Accelerated Aging in Smokers. *Sci Rep* **9**, 142 (2019).
56. Yan, X. & Su, X. *Linear regression analysis: theory and computing*. (World Scientific Pub. Co., 2009).
57. Shwartz-Ziv, R. & Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion* **81**, 84–90 (2022).
58. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? Preprint at https://doi.org/10.48550/arXiv.2207.08815 (2022).
59. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785.
60. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
61. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *arXiv:1706.09516 [cs]* (2019).
62. Borisov, V. *et al.* Deep Neural Networks and Tabular Data: A Survey. *arXiv:2110.01889 [cs]* (2022).
63. Kadra, A., Lindauer, M., Hutter, F. & Grabocka, J. Well-tuned Simple Nets Excel on Tabular Datasets. Preprint at https://doi.org/10.48550/arXiv.2106.11189 (2021).

64. Arik, S. O. & Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *arXiv:1908.07442 [cs, stat]* (2020).
65. Song, W. *et al.* AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks. in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* 1161–1170 (2019). doi:10.1145/3357384.3357925.
66. Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B. & Goldstein, T. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. Preprint at https://doi.org/10.48550/arXiv.2106.01342 (2021).
67. Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting Deep Learning Models for Tabular Data. in *Advances in Neural Information Processing Systems* vol. 34 18932–18943 (Curran Associates, Inc., 2021).
68. Popov, S., Morozov, S. & Babenko, A. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. Preprint at https://doi.org/10.48550/arXiv.1909.06312 (2019).
69. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. in *Computer Vision – ECCV 2014* (eds. Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) vol. 8689 818–833 (Springer International Publishing, 2014).
70. Bach, S. *et al.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **10**, e0130140 (2015).
71. Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv:1605.01713 [cs]* (2017).
72. Lipton, Z. C. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]* (2017).
73. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]* (2016).
74. Doshi-Velez, F. & Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]* (2017).
75. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* **73**, 1–15 (2018).
76. Baehrens, D. *et al.* How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.* **11**, 1803–1831 (2010).
77. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]* (2014).
78. Samek, W., Wiegand, T. & Müller, K.-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv:1708.08296 [cs, stat]* (2017).
79. Bolomsky, A., Schreder, M., Zojer, N. & Ludwig, H. The chemokine CXCL9 (MIG) is an independent predictor of overall survival in newly diagnosed multiple myeloma. *Clinical Lymphoma Myeloma and Leukemia* **15**, e237–e238 (2015).
80. de Araújo, F. F. *et al.* CXCL9 and CXCL10 display an age-dependent profile in Chagas patients: a cohort study of aging in Bambui, Brazil. *Infect Dis Poverty* **9**, 51 (2020).
81. Koper, O., Kamińska, J., Sawicki, K. & Kemona, H. CXCL9, CXCL10, CXCL11, and their receptor (CXCR3) in neuroinflammation and neurodegeneration. *Adv Clin Exp Med* **27**, 849–856 (2018).

82. Chua, J. *et al.* Expression profile of inflammatory cytokines in aqueous from glaucomatous eyes. *Mol Vis* **18**, 431–438 (2012).
83. Romagnani, P. *et al.* Expression of IP-10/CXCL10 and MIG/CXCL9 in the thyroid and increased levels of IP-10/CXCL10 in the serum of patients with recent-onset Graves' disease. *Am J Pathol* **161**, 195–206 (2002).
84. Egesten, A. *et al.* The CXC Chemokine MIG/CXCL9 Is Important in Innate Immunity against *Streptococcus pyogenes*. *J INFECT DIS* **195**, 684–693 (2007).
85. Faé, K. C. *et al.* CXCL9/Mig Mediates T cells Recruitment to Valvular Tissue Lesions of Chronic Rheumatic Heart Disease Patients. *Inflammation* **36**, 800–811 (2013).
86. Ikemizu, S., Chirifu, M. & Davis, S. J. IL-2 and IL-15 signaling complexes: different but the same. *Nat Immunol* **13**, 1141–1142 (2012).
87. Waldmann, T. A. The shared and contrasting roles of IL2 and IL15 in the life and death of normal and neoplastic lymphocytes: implications for cancer therapy. *Cancer Immunol Res* **3**, 219–227 (2015).
88. Yang, Y. & Lundqvist, A. Immunomodulatory Effects of IL-2 and IL-15; Implications for Cancer Immunotherapy. *Cancers* **12**, 3586 (2020).
89. Vokaer, B., Charbonnier, L.-M., Lemaître, P. H., Spilleboudt, C. & Le Moine, A. IL-17A and IL-2-Expanded Regulatory T Cells Cooperate to Inhibit Th1-Mediated Rejection of MHC II Disparate Skin Grafts. *PLoS ONE* **8**, e76040 (2013).
90. Aljabali, M. A. & Kuts, L. SERUM LEVELS OF IL-2 AND IL-17A ARE RELATED TO CLINICAL TYPE AND SEVERITY OF ALOPECIA AREATA. *Wiad Lek* **75**, 263–267 (2022).
91. Pandiyan, P. *et al.* The role of IL-15 in activating STAT5 and fine-tuning IL-17A production in CD4 T lymphocytes. *J Immunol* **189**, 4237–4246 (2012).
92. Agarwal, R. *et al.* Neural Additive Models: Interpretable Machine Learning with Neural Nets. (2020) doi:10.48550/ARXIV.2004.13912.
93. Chen, J., Liao, K., Wan, Y., Chen, D. Z. & Wu, J. DANets: Deep Abstract Networks for Tabular Data Classification and Regression. (2021) doi:10.48550/ARXIV.2112.02962.
94. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. (2017) doi:10.48550/ARXIV.1705.07874.
95. Lipovetsky, S. & Conklin, M. Analysis of regression in game theory approach. *Appl. Stochastic Models Bus. Ind.* **17**, 319–330 (2001).
96. Maggio, M., Guralnik, J. M., Longo, D. L. & Ferrucci, L. Interleukin-6 in aging and chronic disease: a magnificent pathway. *J Gerontol A Biol Sci Med Sci* **61**, 575–584 (2006).
97. Rea, I. M. *et al.* Age and Age-Related Diseases: Role of Inflammation Triggers and Cytokines. *Front. Immunol.* **9**, 586 (2018).
98. Lira-Junior, R., Åkerman, S., Gustafsson, A., Klinge, B. & Boström, E. A. Colony stimulating factor-1 in saliva in relation to age, smoking, and oral and systemic diseases. *Sci Rep* **7**, 7280 (2017).
99. Bradburn, S. *et al.* Dysregulation of C-X-C motif ligand 10 during aging and association with cognitive performance. *Neurobiol Aging* **63**, 54–64 (2018).
100. Karlsson, C. & Paulsson, Y. Age related induction of platelet-derived growth factor A-chain mRNA in normal human fibroblasts. *J Cell Physiol* **158**, 256–262 (1994).

101. Song, Y. *et al.* Aging enhances the basal production of IL-6 and CCL2 in vascular smooth muscle cells. *Arterioscler Thromb Vasc Biol* **32**, 103–109 (2012).
102. Ren, G., Al-Jezani, N., Railton, P., Powell, J. N. & Krawetz, R. J. CCL22 induces pro-inflammatory changes in fibroblast-like synoviocytes. *iScience* **24**, 101943 (2021).
103. Keller, A. *et al.* Mutations in the gene encoding PDGF-B cause brain calcifications in humans and mice. *Nat Genet* **45**, 1077–1082 (2013).
104. Estevao, C. *et al.* CCL4 induces inflammatory signalling and barrier disruption in the neurovascular endothelium. *Brain, Behavior, & Immunity - Health* **18**, 100370 (2021).
105. Blandford, S. N. *et al.* Interleukin-1 receptor antagonist: An exploratory plasma biomarker that correlates with disability and provides pathophysiological insights in relapsing-remitting multiple sclerosis. *Multiple Sclerosis and Related Disorders* **52**, 103006 (2021).
106. Andrae, J., Gallini, R. & Betsholtz, C. Role of platelet-derived growth factors in physiology and medicine. *Genes Dev.* **22**, 1276–1312 (2008).
107. Puzianowska-Kuźnicka, M. *et al.* Interleukin-6 and C-reactive protein, successful aging, and mortality: the PolSenior study. *Immun Ageing* **13**, 21 (2016).
108. Bell, C. G. *et al.* DNA methylation aging clocks: challenges and recommendations. *Genome Biol* **20**, 249 (2019).
109. Huang, X., Khetan, A., Cvitkovic, M. & Karnin, Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. Preprint at https://doi.org/10.48550/arXiv.2012.06678 (2020).
110. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
111. Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38**, 367–378 (2002).
112. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. (2014) doi:10.48550/ARXIV.1412.6980.
113. Falcon, W. & The PyTorch Lightning team. PyTorch Lightning. (2019) doi:10.5281/ZENODO.3828935.
114. Paszke, A. *et al.* PyTorch: an imperative style, high-performance deep learning library. in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* 8026–8037 (Curran Associates Inc., 2019).
115. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).
116. Peters, B., Niculae, V. & Martins, A. F. T. Sparse Sequence-to-Sequence Models. (2019) doi:10.48550/ARXIV.1905.05702.
117. Martins, A. F. T. & Astudillo, R. F. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. (2016) doi:10.48550/ARXIV.1602.02068.
118. Radenovic, F., Dubey, A. & Mahajan, D. Neural Basis Models for Interpretability. Preprint at https://doi.org/10.48550/arXiv.2205.14120 (2022).
119. Joseph, M. PyTorch Tabular: A Framework for Deep Learning with Tabular Data. (2021) doi:10.48550/ARXIV.2104.13638.

120. Dauphin, Y. N., Fan, A., Auli, M. & Grangier, D. Language Modeling with Gated Convolutional Networks. (2016) doi:10.48550/ARXIV.1612.08083.
121. Rogozhnikov, A. Einops: Clear and Reliable Tensor Manipulations with Einstein-like Notation. in *International Conference on Learning Representations (ICLR 2022)* (2022).
122. Vaswani, A. *et al.* Attention Is All You Need. (2017) doi:10.48550/ARXIV.1706.03762.
123. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. (2019) doi:10.48550/ARXIV.1907.10902.
124. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for Hyper-Parameter Optimization. in *Advances in Neural Information Processing Systems* vol. 24 (Curran Associates, Inc., 2011).
125. Molnar, C. *Interpretable machine learning: a guide for making Black Box Models interpretable*. (Lulu, 2019).