Risk forecasting tools based on the collected information for two types of occupational diseases

Marc Deminov[1,2]*, Petr Kuztetsov[2,3], Alexander Melerzanov[4,5]

[1]"Soyuz Sport i Zdorov'e" (LLC), Skolkovo Innovation Center, Moscow 143026, Russia

[2]National Association of Medical Informatics, Verkhnyaya Krasnoselskaya, 20C1, Moscow 107140, Russia

[3]Project office "Digital Transformation in Occupational Medicine" of the Research Institute of Occupational Medicine. 31 Budennogo Ave., Moscow 105275, Russia

[4]“Applied Genetics” Center Moscow Institute of Physics and Technology (National Research University) 9 Institutsky Ln, Dolgoprudny, Moscow Region 141700, Russia

[5]“Laboratory of innovative Technologies and Artificial Intelligence in Pubblic Health”
Semashko National Research Institute of Russian Academy of Science
12 Vorontsovo Pole, Moscow 105064, Russia

Correspondence: Marc Deminov md@ibs-m.ru

Abstract

The article represents the construction of algorithms for monitoring and predicting the risk of occupational diseases (sensorineural hearing loss and vibration disease from exposure to local and general vibration) and the use of data from clinical and instrumental examination of patients.

Key words: labor health, employee health risk estimation, sensorineural hearing loss, vibration disease, effects of vibration on health, health risk forecasting, information and statistical methods of forecasting

The health problems of workers caused by acoustic and vibrational influences are noted as important and require study and specific quantitative methods of assessment and forecasting by the World Health Organization (WHO 2021). Measurements and research in this area are carried out in many countries (Sliwinska-Kowalska 2020; Roberts et al. 2018; Ntlhakana et al. 2020; Dobie 2008; Mahbub et al. 2020; Ekman 2021), with the main starting point being the industry standards (ISO 1999:2013; ISO 2631-5:2018) for predicting health risks from acoustic and vibration impacts.

Risk assessment and prognosis of occupational diseases, as well as the formation of therapeutic and preventive measures are carried out in order to protect health and preserve the ability to work, prevent and timely detect occupational diseases of workers engaged in work with harmful and (or) dangerous production factors, as well as in cases provided for by the legislation of the Russian Federation (GOST R ISO 1999-2017). Employees engaged in certain types of work undergo mandatory medical examinations. The assessment when applying for a job is carried out in order to achieve:

- determining the compliance of the health status of a person entering a job with the work assigned to him;
- early detection and prevention of diseases (Periodic medical examinations);
- dynamic health monitoring;
- timely detection of diseases, initial forms of occupational diseases;
- early features of the impact of harmful and (or) hazardous production factors on the health of workers;
- formation of risk groups for the development of occupational diseases.

More than 15 million people annually fall under these requirements throughout the Russian Federation, which in turn entails significant requirements for the availability of a highly qualified staff of occupational pathologists evenly distributed throughout the territory of the Russian Federation, the costs of their education and maintenance.

Development of a mathematical model based on risk calculation methodology using artificial intelligence tools

In order to optimize and objectify the process of passing medical examinations and decision-making, it is extremely necessary to have an automated system to support the adoption of medical decisions by a professional pathologist to assess the risk and forecast the onset of occupational disease, including the formation of therapeutic and preventive measures aimed at minimizing the likelihood of its occurrence. At the moment, there is no such system in the Russian Federation, while the medical community expresses interest in its creation and dissemination, including for the objectification of the assessment of quality control of medical examinations.

Based on the above, the purpose of this project is to develop a decision support system for a professional pathologist (hereinafter referred to as SPP) based on a software product that includes an automated model for determining risk and predicting the future state of hearing (risk group) using Artificial Intelligence technologies (hereinafter referred to as AI) based on the results of the analysis of medical indicators of a doctor's study obtained by conducting an employee questionnaire and instrumental diagnostics, including depending on the level and type of acoustic impact in the workplace, as well as depending on other factors and parameters.

The solution of a particular problem of automated determination of the risk of sensorineural hearing loss (hearing loss), as well as the risk of vibration disease, will form the basis for the development of similar tools for other nosological profiles.

To solve the task, the scientific group of the project has determined the sequence of actions:

Step 1. Formulation and preliminary steps of the problem solution.

1. Collect a structured set of samples of employees in the number of more than 500 people with data on medical indicators obtained by conducting an employee questionnaire and instrumental diagnostics and other data on each of the employees in machine-readable form (for example, MS Excel tables);
2. Bring the data to a format acceptable for further automated processing, remove typos, conduct format-logical control for each column with data, identifying anomalies.
3. To classify employees into 5 risk groups (negligent low risk; low risk; medium; high; very high) of the onset of the disease under study in accordance with the current recommendations of the occupational pathologist to identify risk groups by key parameters. We will carry out such a classification with the help of an expert doctor, repeating the logic of his actions in everyday practice, we will put the risk group line by line opposite each employee.
4. Thus, we will get a machine-expert-trained sample at the output for its subsequent processing using AI technologies. A similar data preparation operation can be implemented in a software tool or using Excel.

Step 2. Building an automated data analysis model

5. Further, to solve the problems of automated data research and automated model construction, powerful and, at the same time, accessible software tools were used: scikitlearn, pandas, numpy, seaborn, streamlit (python programming language) and other open-source libraries for working with data. The choice of programming language library data is determined based on the following criteria.

All the basic methods and functions necessary for an exhaustive research in the field of data analysis are implemented in the available software libraries (Figure 1)

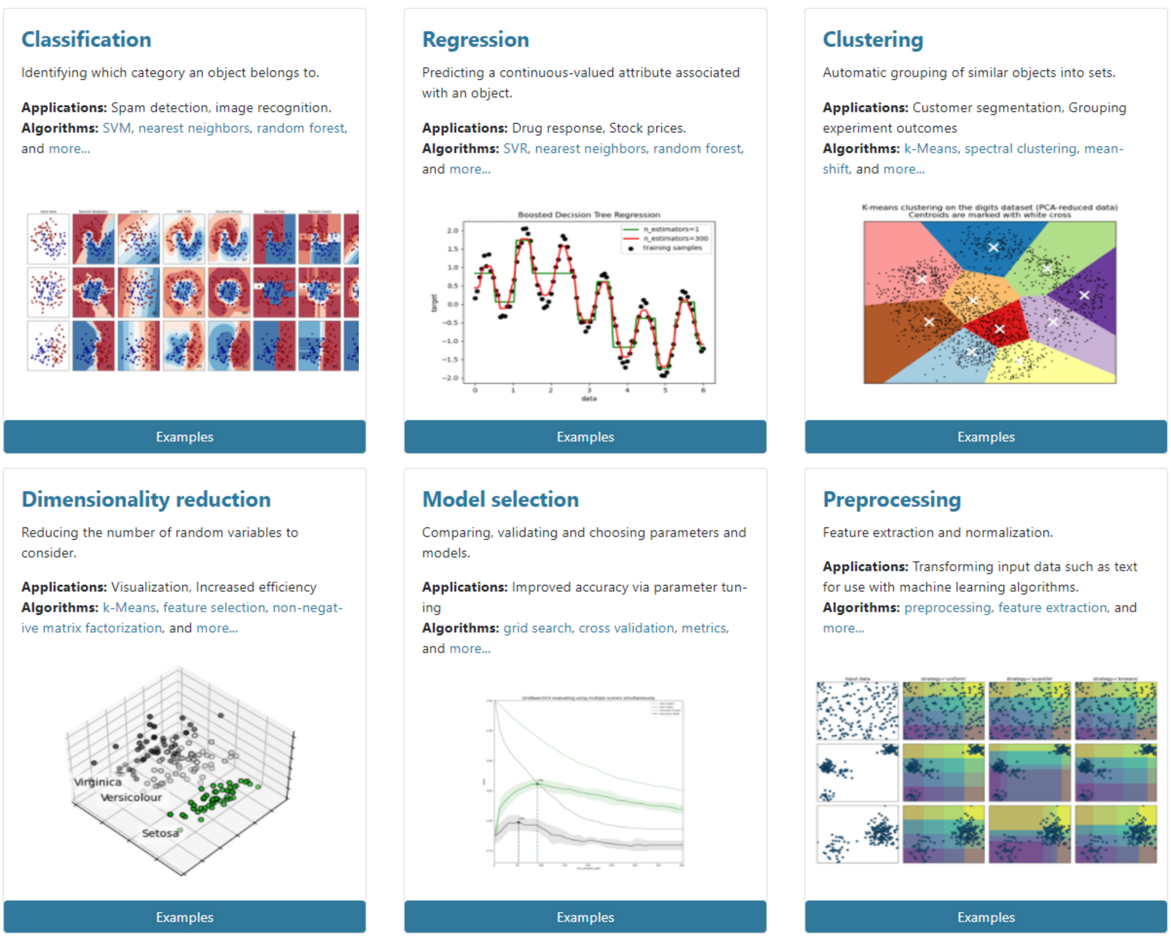

Figure 1. Examples of visualization of results obtained using analytical methods implemented in the SciKitLearn software library package. It can be seen that almost all the most effective modern analytical technologies and methods are presented in this library package

Python libraries have absorbed almost all modern analytical mathematical capabilities and provide researchers and developers with a wide range of convenient, ready-to-assemble functional software elements: a quick (so-called “seamless”) transition from model development to its implementation in prototype, and then in industrial code; powerful visualization tools (see, for example, the seaborn library) and debugging (Figure 2).

Figure 2. Visualization of the capabilities of libraries in Python

6. Uploading the specified data to the test development environment

7. For further work on the data, built-in tools of factor analysis, clustering, regressions, categorical analysis using the methods of solution trees and "random forests" were used to identify the most information-relevant features that determine the classification by risk groups. There was a need to conduct an experiment with different techniques. The following methods were used:

a. The principal component method, including varying the number of leading linear combinations of principal components, that is, the proportion of information left for analysis
b. Comparison of regression models for the f2 metric
c. Method for selective enumeration of combinations of data to build prediction models of the group according to the worker
d. Methods of classifiers based on continuous variables (multilayer neural networks, convolutional networks, etc.)
e. Methods of classifiers based on categorical and continuous variables (random trees,

random forest, etc.)

f. Validation of the constructed model in the metric AUC of the ROC curve by repeated random split into training (70% of the data), validation (15%) and test (15%) of the sample.

g. Based on the results and methods of the Bali stages described above, the most accurate (that is, giving the least error) numerical and analytical models of automated risk group determination are selected; they are reflected in detail in the section of the mathematical report.

9. Analytical prognostic models of health groups are constructed based on forecasts of medical indicators that are input to the risk group definition model; an assessment of the accuracy of the prognostic model is provided, as well as a justification for the accuracy assessment.

10. The expert evaluation of the obtained predictive models was carried out; the optimal ones were determined.

The applied methods belong to the field that is commonly referred to by the collective term "Artificial Intelligence"; in this work, AI methods such as "teaching without a teacher" (cluster analysis), factor analysis, regression analysis, other information-statistical, analytical methods that are significantly dependent on computational algorithms that are used.

This approach can be illustrated by the first definition of artificial intelligence given by John McCarthy in 1956 at a conference at Dartmouth University. According to McCarthy, "AI researchers are free to use methods that are not observed in humans, if necessary to solve specific problems."

A leading researcher in the subject area, Gennady Osipov (President of the Russian Association of Artificial Intelligence, permanent member of the European Coordinating Committee on Artificial Intelligence (ECCAI), PhD, professor) gives the following description of AI: "artificial intelligence is an experimental science. Experimentalism of artificial intelligence is that by creating certain computer concepts and models, the researcher compares their behavior with each other and with examples of the solution of the same tasks by the specialist, modifies them based on this comparison, trying to achieve the best match results."

This work was carried out in full compliance with the above mentioned paradigm of artificial intelligence.

The following is a description of the features of objects with their detailed properties and characteristics, an analysis of the types of tasks to be solved and the available target variables, a description of the models and the result of their work with the data obtained, as well as conclusions.

The work was carried out on the data of the FGBNU Research Institute of Labour Health, presented in the form of a table with 892 rows and 132 columns. Each line is a description of the subject of the study - the patient. The first three columns are the target features that we want to automatically predict.

Overview of the methods used

In recent years, the use of computing systems and personal computers has been widely spread and implemented. A digitalization program is being actively implemented in the country. This leads to the fact that sufficient amounts of data are accumulated to build effective and modern decision-making systems that help employees solve problems, sometimes completely excluding human participation.

Such a process inevitably affects medicine – an area in which measurements have been carried out throughout history, on the basis of which the necessary result was obtained. In recent years, methods of modern data analysis have become increasingly used for medicine. This applies both to specialized studies with the analysis of images, video streams and audio tracks, and general issues related to the processing of arrays of tabular data.

These algorithms can be used to analyze tabular data and obtain estimated results based on them. The capabilities of these algorithms are demonstrated in this paper. Any research begins with an analysis of the existing features that describe objects – this is the first part of this work. Next, we consider the specifics of the task and choose ways to solve it. After that, we consistently analyze various machine learning algorithms and analyze their results. After discussing the results, we attempted a long-term prediction of changes in risks for a particular employee. The final part presents conclusions and observations.

Feature description of objects

Each object corresponds to 129 features that make up its medical description. Before building a model, you need to check the presented data for correctness. In this study, it was found that there are missing values for some of the features. So there are empty cells in the columns describing the Vibration sensitivity at different frequencies. In all cases, the missing values were replaced by the average values for the attribute in question. In addition, all categorical features are presented in the form of letter classes. In this form, directly, in the presented format, these features cannot be applied as input data for mathematical models, therefore, the presented letter designations must be changed to numerical ones. In this paper, category "a" was replaced by the number 0, category "b" by 1. If there are other classes in the attribute, then they were replaced by numbers accordingly.

Features are naturally divided into groups. Examples of such a division and an analysis of the available features are given below.

Separation of features by the type of their values

Most of the features in the data under consideration have a categorical nature. Categorical features describe the belonging of the object of research to a certain type or class. In our data, most of the categorical features are binary, representing two classes. Most often, this may be the presence of any complaints, the fact of observation of a certain property or behavior of the patient. Examples of such features are shown in Table 1.

№	*Feature name*
1	Sex
2	Pain in the feet/and shins
3	Pain in the lumbar spine
4	Hyperhidrosis of the hands
5	Hyperhidrosis of the feet
6	Perforation Mt-2
7	Traumatic brain injuries
8	Smoking
9	Harmful factors
10	1.1.3.
11	1.1.4.3.1.
12	1.1.4.3.2.
13	1.1.4.3.3.
14	1.1.4.5.
15	1.1.4.8.2.
16	1.2.1.

17	1.2.10.
18	1.2.14.1.
19	1.2.2.
20	1.2.21.1.
21	1.2.25.
22	1.2.30.1.

Table 1. Examples of features: the Python scikit learn library allows the researchers to monitor categorical, numerical, textual and other features during debugging and testing of the program

Numerical features represent the description of objects in the form of a numerical expression. At the same time, these features can take both a fixed set of values (such features as year of birth, work experience, etc.) and an infinitely large set of values of real numbers. In our task, we work with medical data, so that the acceptable range of accepted values for almost every attribute is known. According to the general meaning of the attribute, it is possible to understand in which range its values lie. When receiving data from a source, according to these criteria, it is possible to determine the correctness of the data provided or to find anomalies in the data.

As with any set of numerical values, various functions can be applied to numerical ones, finding important statistical parameters of the data provided. Thus, it is possible to estimate the mean value, variance and other possible parameters of distributions that can give a general idea of the feature under consideration.

Table 2 represents a list of numerical features as an example.

№	*Numerical feature name*
1	Risk of sensorineural hearing loss
2	Risk local vibration
3	Risk general vibration
4	Year of Birth
5	Work experience
6	The year of the current employment
7	Thermometry of the right hand, degrees Celsius
8	Thermometry of the left hand, degrees Celsius
9	Thermometry of the right foot, degrees Celsius
10	Thermometry of the left foot, degrees Celsius
11	Vibration sensitivity at 125 Hz on the right, dB
12	Vibration sensitivity at 125 Hz on the left, dB
13	Vibration sensitivity at a frequency of 32 Hz on the right
14	Vibration sensitivity at 63 Hz on the right

15	Vibration sensitivity at 250 Hz on the right
16	Vibration sensitivity at a frequency of 32 Hz on the left
17	Vibration sensitivity at 63 Hz on the left
18	Vibration sensitivity at 250 Hz on the left
19	Hearing threshold at 250 Hz, (dB)
20	Hearing threshold at the frequency of 500 Hz (dB)
21	Hearing threshold at the frequency of 1000 Hz(dB)
22	Hearing threshold at the frequency of 2000 Hz(dB)
23	Hearing threshold at a frequency of 3000 Hz(dB)
24	Hearing threshold at the frequency of 4000 Hz(dB)
25	Hearing threshold at the frequency of 6000 Hz(dB)
26	Hearing threshold at the frequency of 250 Hz (dB)
27	Hearing threshold at a frequency of 500 Hz, (dB)
28	Hearing threshold at a frequency of 1000 Hz, (dB)
29	Hearing threshold at a frequency of 2000 Hz, (dB)
30	Hearing threshold at a frequency of 3000 Hz, (dB)
31	Hearing threshold at 4000 Hz,(dB)
32	Hearing threshold at a frequency of 6000 Hz, (dB)

Table 2. Examples of numerical features

Next, we will consider the division of features by meaning, where we will consider in more detail the descriptions that were presented above.

Separation of features by meaning

It is more convenient to present a general description of objects by dividing their features into groups by meaning. Then, considering each group separately, it is possible to understand in more detail what data the developed models will have to work with. This stage is also very useful, because by doing it, the researchers can find "faces" and "insights" in the data that can significantly improve the result.

Examples of the division of features by meaning:

- general information (socio-demographic and formal information about work),
- pain and complaints,
- thermometry of the hands,
- vibration sensitivity,
- medical factors (binary), hearing

— threshold at various frequencies,

— harmful factors,

— harmful work.

Figures 3-8 show distributions and histograms of patients by features.

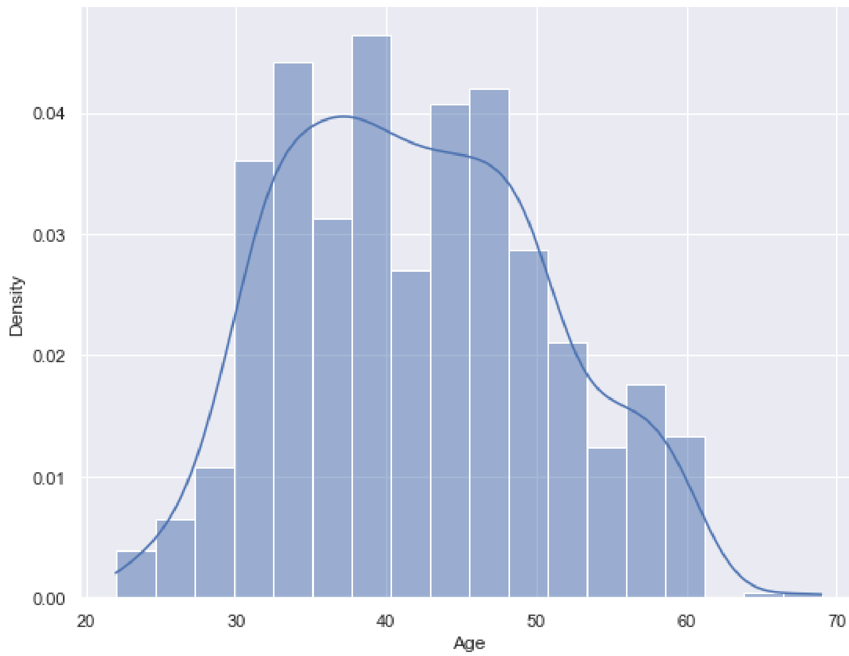

Figure 3 - Distribution of the studied objects by age

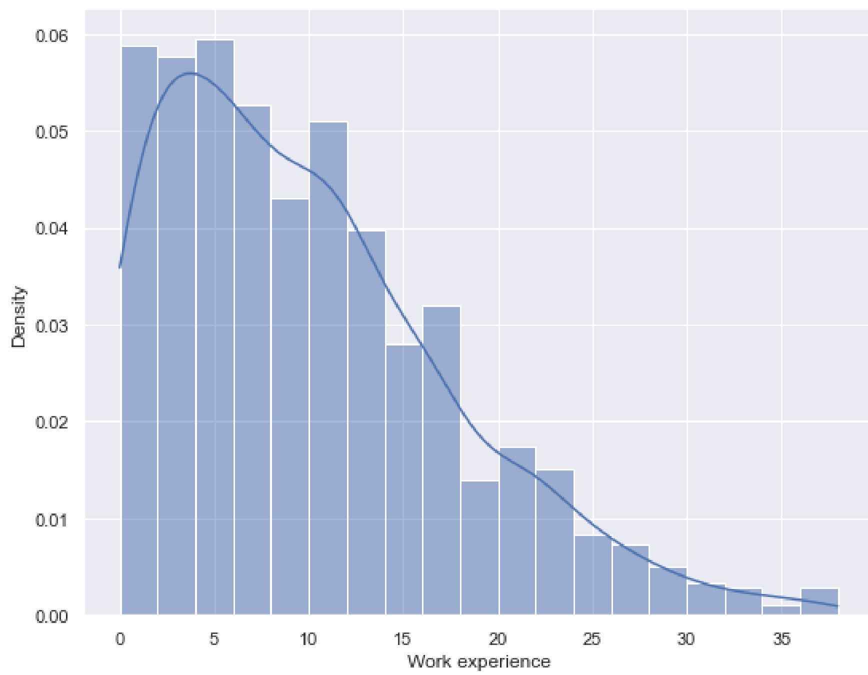

Figure 4 - Distribution of the studied objects by work experience

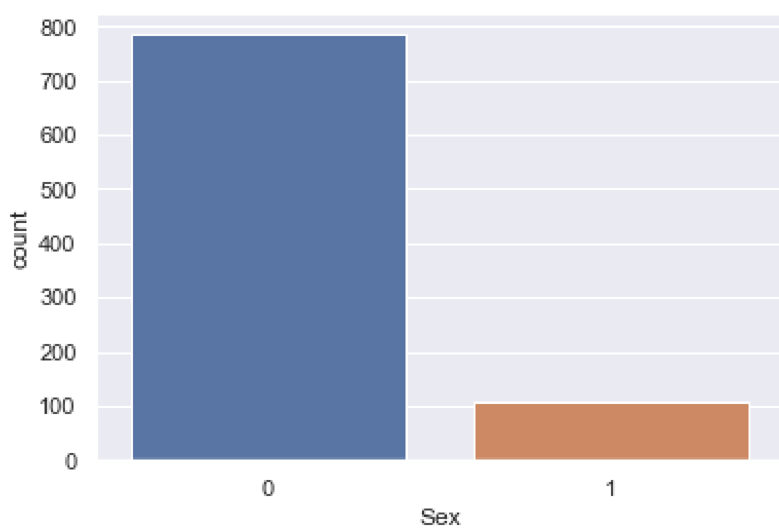

Figure 5 - Gender distribution of the studied

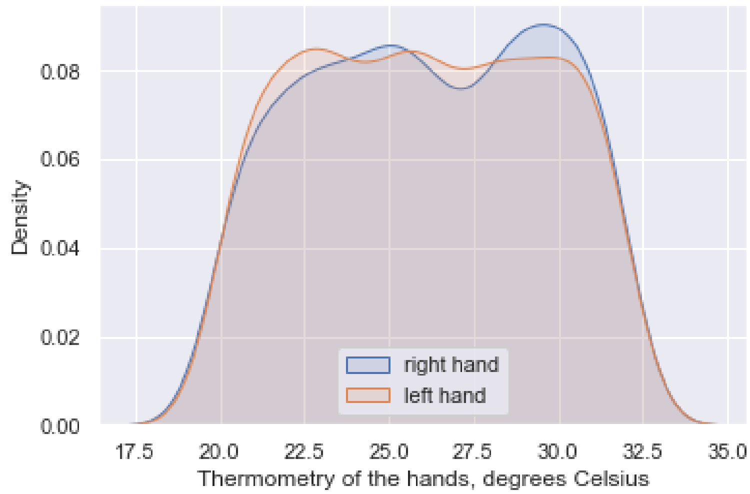

Figure 6 - Presented thermometry of the hands (distribution)

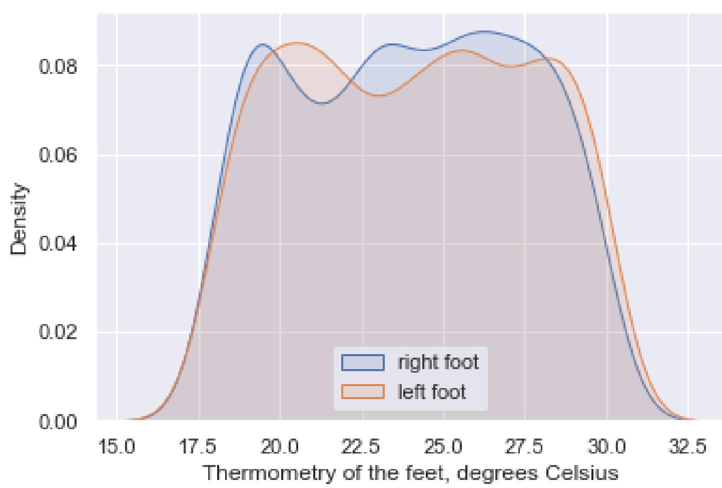

Figure 7 - Presented stop thermometry (distribution)

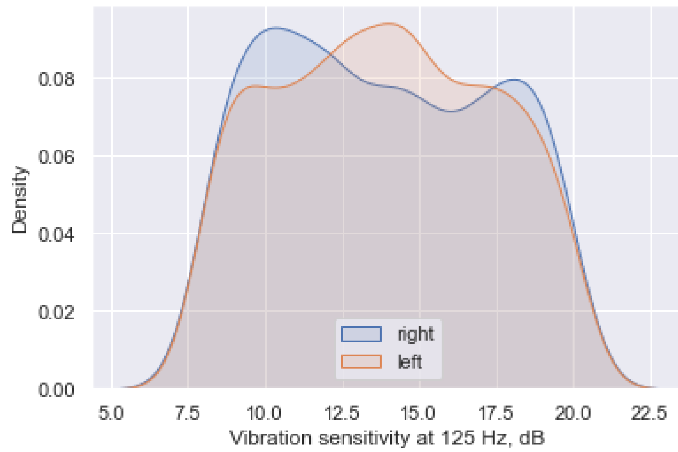

Figure 8 - Vibration sensitivity at 125 Hz (distribution)

When considering these graphs and others for other distinguished features, we can get an idea of the average patient we will examine. This stage of the work is very important, because here we can see the features and anomalies in the data.

The type of the problem to be solved and the target variables

Target variables represent different risk groups for occupational diseases. In relation to training samples, target variables are determined by expert evaluation, or based on data from long-term (the period of the forecast foundation) prospective occupational pathology medical and medico-social observations comparable to a given period of anticipation of the risk forecast.

Each of the target variables represents different risk groups that take a value from 1 to 5. Thus, each patient is assigned according to one of the 5 risk groups of PD. This task can be considered as a classical classification problem, however, the comparability of risk groups among themselves will not be taken into account, although in reality it is known that the higher the group, the worse, that is, the risk is higher. This feature and information are used in the development of models.

With the standard approach to solving and dividing the data into 5 groups, the possibility of comparing the level of risk among patients of the same group and comparing groups among themselves is lost. Therefore, they change the type of problem to be solved from a classification problem to a regression problem, given that the target variables of different risk groups are comparable to each other. This allows for gradation between employees who have the same risk group, arranging them in ascending or descending order of existing risks.

One of the approaches to solving this problem may be to replace the type of problem being solved with a classification problem with a regression problem. At the same time, we will take into account that the target variables of different classes are comparable to each other. In addition, it will be possible to arrange gradation between objects of the same class, building them in the form of increasing or decreasing existing risks.

Figures 9-11 show the types of target variables and proportions of risk groups that are in our training sample in relation to professional sensorineural hearing loss and vibration disease. The provided sample is not balanced, which must be taken into account when splitting the data into training and test samples and when analyzing the results obtained and indicative metrics.

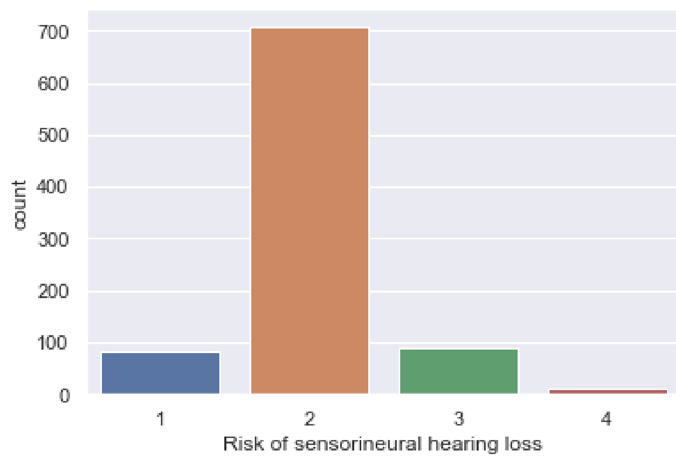

Figure 9 - Distribution of the results of the assessment of the risk of sensorineural hearing loss obtained by expert means

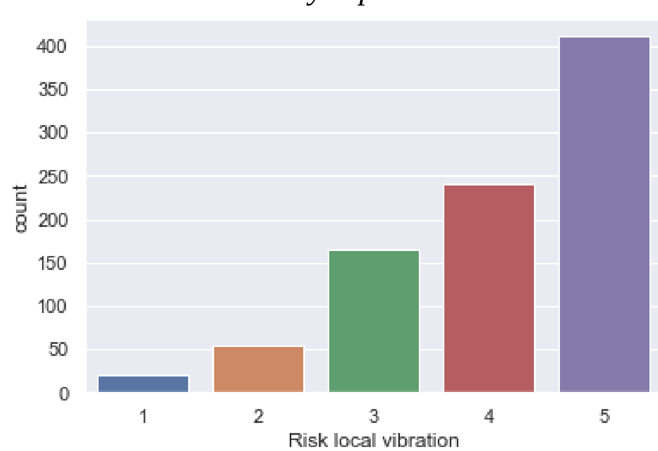

Figure 10 - Distribution of risks of vibration disease (local vibration) obtained by expert means

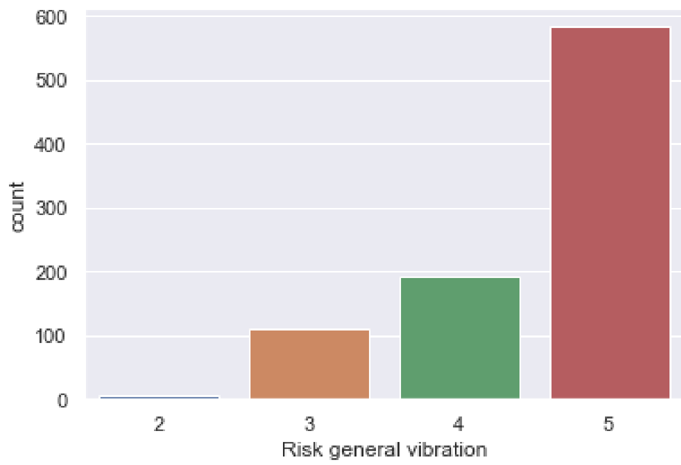

Figure 11 - Distribution of vibration disease risks (general vibration) obtained by expert means

In all three tasks, we see that the presented samples are extremely unbalanced. This circumstance should be taken into account when splitting the data into training and test samples. This fact should be taken into account when analyzing the results and indicative metrics.

Trained models and their results

Below are various models with their brief descriptions and the results they showed on the tasks being solved. In each case, the available data is divided into training and test samples (datasets) in a ratio of 4:1, training takes place on a training dataset, verification and the presented results are checked on a deferred (test) sample. In each case, various metrics are considered that characterize the correctness of the predictions of the obtained models.

Linear regression model

Coefficients are selected for each feature in such a way that the target variable is expressed through a general formula:

$$y^{(k)} = w_0 + \sum_{i=1}^{num\_features} w_i x_i^{(k)}$$

The advantages of this model are that the resulting formula is simple and clear. By its appearance and coefficients, it is clear how each feature affects the result. The disadvantage of this approach is that it is quite simplified, for optimal results and analysis of coefficients in formulas, preprocessing of features is needed.

Determining the risk of sensorineural hearing loss

For the first target feature, the model showed the following metrics on a deferred sample (Table 3):

Model metrics type	*Metrics values*
Accuracy score	0.7877
Mean absolute error (mae)	0.3227
Mean squared error (mse)	0.1784

Table 3. Metrics of sensorineural hearing loss linear regression model

Here and further, respectively, the standard concepts and abbreviations accuracy score are used - accuracy score, that is, the proportion of correct answers of the algorithm; MAE - the average absolute error, MSE - the standard error. It can be seen that the model is being built, and its accuracy is reasonably correlated with the size of the training sample already on the basic technologies of model construction.

For example, below are top 5 features with the highest values of weights (Table 4):

top5 "+" features	*values of weights*
The year of the current employment	0.2468
Work experience	0.2359
Hearing threshold at 4000 Hz,(dB)	0.1998
The hearing threshold at the frequency of 4000 Hz(dB)	0.1245
The hearing threshold at the frequency of 2000 Hz(dB)	0.0952

Table 4. Top five features with the highest values of weights

Similarly, we can give an example of 5 features with the smallest (negative) weights:

top5 "-" features	*values of weights*
The hearing threshold at the frequency of 1000 Hz	-0.0948
27.6.	-0.0524
Hearing threshold at 250 Hz, (dB)	-0.0444
Lower limb right (no changes-a, there are changes-b)	-0.0410
Untreated ear diseases: chronic purulent otitis	-0.0368

Table 5. Top five features with the smallest values of weights

These weights allow the researchers to see the contribution to the result of each feature describing the object. In this case, the features are pre-normalized (the average is subtracted from the features and divided by the amount of variance), so it is correct to compare such coefficients.

Determination of the risk of vibration disease (local vibration)

For the second target value, the linear regression model showed the following metrics:

Model metrics type	*Metrics values*
Accuracy score	0.4692
Mean absolute error (mae)	0.6513
Mean squared error (mse)	0.6827

Table 6. Metrics of vibration disease linear regression model (local vibration)

Top 5 features with the highest values of weights:

top5 "+" features	*values of weights*
3.4.1.	0.1901
The year of the current employment	0.1200
1.2.2.	0.1071
Paint brushes (Normal color-a, Marble-cyanotic–b)	0.1033
Work experience	0.0974

Table 7. Top five features with the highest values of weights

Similarly, example of 5 features with the smallest negative weights:

top5 "-" features	*values of weights*
Thermometry of the left hand, degrees Celsius	-0.5786
Thermometry of the right hand, degrees Celsius	-0.5786
1.3.5.	-0.1540
Pain in the hands / and forearms (yes-a, no-b)	-0.1028
3.4.2.	-0.1009

Table 8. Top five features with the smallest values of weights

These weights allow the researchers to see the contribution to the result of each feature describing the object. In this case, the features are pre-normalized (the average is subtracted from the features and divided by the variance value), so it is correct to compare such coefficients.

Determination of the risk of vibration disease (general vibration)

For the third target value, similar metrics and results are provided. The linear regression model showed the following metrics:

Model metrics type	*Metrics values*
Accuracy score	0.5474
Mean absolute error (mae)	0.5423
Mean squared error (mse)	0.4683

Table 9. Metrics of vibration disease linear regression model (general vibration)

Top 5 features with the highest values of weights:

top5 "+" features	*values of weights*
1.2.37.	0.0646
Hearing threshold at 250 Hz, (dB)	0.0519
The upper limb to the left (there are no changes – a, there are changes – b)	0.0515
Perforation Mt-2	0.0510
2.	0.0500

Table 10. Top five features with the highest values of weights

Similarly, example of 5 features with the smallest negative weights:

top5 "-" features	*values of weights*
The year of the current employment	-0.3297
Work experience	-0.3182
Thermometry of the left hand, degrees Celsius	-0.2801
Thermometry of the right foot, degrees Celsius	-0.2189
Thermometry of the left foot, degrees Celsius	-0.2165

Table 11. Top five features with the smallest values of weights

Weights allow the researchers to see the contribution to the result of each feature describing the object. In this case, the features are pre-normalized (the average is subtracted from the features and divided by the amount of variance), so it is correct to compare such coefficients.

Based on the first results, we can say that the first task is solved better than the next two.

The accuracy of solving the first problem is satisfactory, the quality of solutions to the 2nd and 3rd problems is low. Perhaps this model is too simple for such a set of features and does not take into account the patterns inherent in the features.

The decision tree model

The decision tree model is an algorithm that provides an answer by making decisions at various levels, at each of which it checks the object for a certain condition on the selected attribute. Schematically, this algorithm can be represented as a binary tree. To get an answer, the researchers need to go down from its root to one of the leaves. The big advantage of this algorithm is its intuitive clarity.

Determining the risk of sensorineural hearing loss

For the first target feature, the model showed the following metrics on a deferred sample:

Model metrics type	*Metrics values*
Accuracy score	0.8994
Mean absolute error (mae)	0.1184
Mean squared error (mse)	0.0884

Table 12. Metrics of sensorineural hearing loss decision tree model

A part of the calculated decision tree is shown in Figure 12.

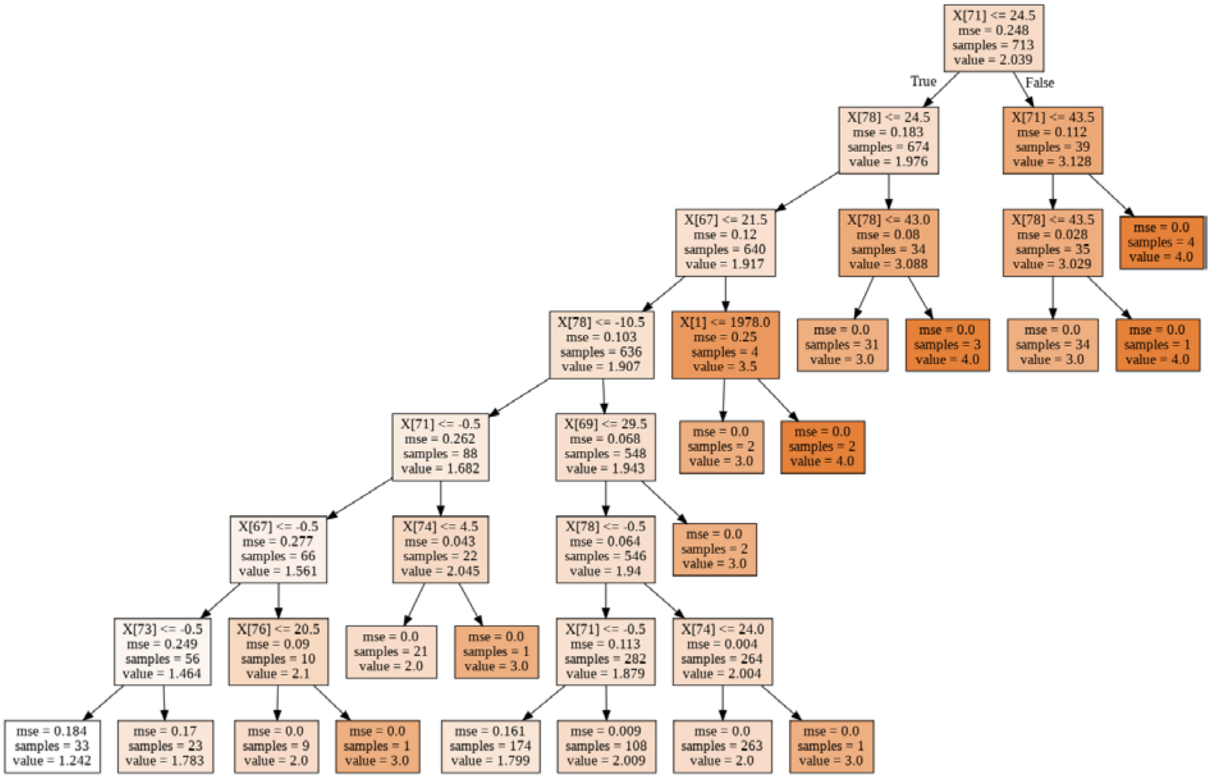

Figure 12 - Visual representation of the decision tree - a model that gives the most accurate results for the task of predicting the risk of sensorineural hearing loss

In this model, the maximum depth of tree construction is limited. The accuracy indicated above and other metrics are thereby determined by at least a small part of the original features. Let's list the features on the basis of which the greatest number of decisions were made when the algorithm was running:

feature	*importance*
The hearing threshold at the frequency of 4000 Hz(dB)	0.4105
Hearing threshold at 4000 Hz,(dB)	0.3856
The hearing threshold at the frequency of 2000 Hz(dB)	0.0700
Hearing threshold at a frequency of 500 Hz, (dB)	0.0422
The hearing threshold at the frequency of 500 Hz (dB)	0.0375
The hearing threshold at the frequency of 250 Hz (dB)	0.0275
Lower limb right (no changes - a, there are changes - b)	0.0069
The hearing threshold at the frequency of 6000 Hz(dB)	0.0069
The mucous membrane of the posterior pharyngeal state	0.0066
Hearing threshold at a frequency of 1000 Hz, (dB)	0.0062

Table 13. A list of the features on the basis of which the greatest number of decisions were made for estimation of the risk of sensorineural hearing loss.

We see that the column names are repeated, most likely each of the repeated features refers to the left and right sides.

Determination of the risk of vibration disease (local vibration)

For the second target value, the decision tree model showed the following metrics:

Model metrics type	*Metrics values*
Accuracy score	0.8491
Mean absolute error (mae)	0.1799
Mean squared error (mse)	0.1885

Table 14. Metrics of decision tree model of vibration disease (local vibration)

The completed decision tree is shown in Figure 13.

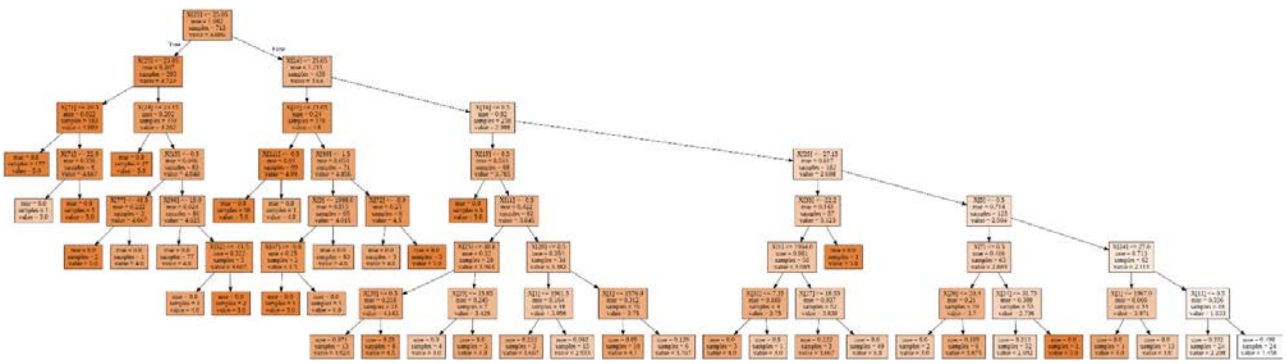

Figure 13 - Visual representation of the decision tree - a model that gives the most accurate results for the task of predicting the risk of vibration disease (local vibration)

Top of the features by which this decision tree works:

feature	*importance*
Thermometry of the right hand, degrees Celsius	0.4629
Thermometry of the left hand, degrees Celsius	0.3472
Hyperhidrosis of the hands	0.0769
Pain in the hands / and forearms (yes - a, no - b)	0.0257
Attacks of whiteness / blueness of the fingers	0.0163
Hypalgesia and/or hypesthesia of the feet	0.0155
Numbness and/or paresthesia of the hands	0.0107
Age	0.0095

Symptom of white spot and/or Bogolepovs test (negative - a, positivs - b)	0.0056
The hearing threshold at the frequency of 2000 Hz(dB)	0.0050
Vibration sensitivity at 250 Hz on the left hand	0.0049
The hearing threshold at the frequency of 250 Hz (dB)	0.0045
Thermometry of the right foot, degrees Celsius	0.0026
The left lower limb (no changes - a, there are changes - b)	0.0022
Numbness and/or paresthesia of the feet	0.0020
Paint brushes (Normal color - a, Marble-cyanotic - b)	0.0017
Hearing threshold at 250 Hz, (dB)	0.0016
1.1.	0.0014
The hearing threshold at the frequency of 1000 Hz(dB)	0.0013
Vibration sensitivity at 63 Hz on the left hand	0.0010
The hearing threshold at the frequency of 4000 Hz(dB)	0.0009
Work experience	0.0007

Table 15. A list of the features on the basis of which the greatest number of decisions were made for estimation of the risk of vibration disease (local vibration).

Again, we see that the defining parameters for decision-making in this case are a limited set of initial features.

Determination of the risk of vibration disease (general vibration)

For the third target value, similar metrics and results are provided. The decision tree model showed the following metrics:

Model metrics type	*Metrics values*
Accuracy score	0.9218
Mean absolute error (mae)	0.1058
Mean squared error (mse)	0.0837

Table 16. Metrics of vibration disease decision tree model (general vibration)

The general view of the model (the resulting tree) with the features and conditions of transition to the lower levels is shown in Figure 14.

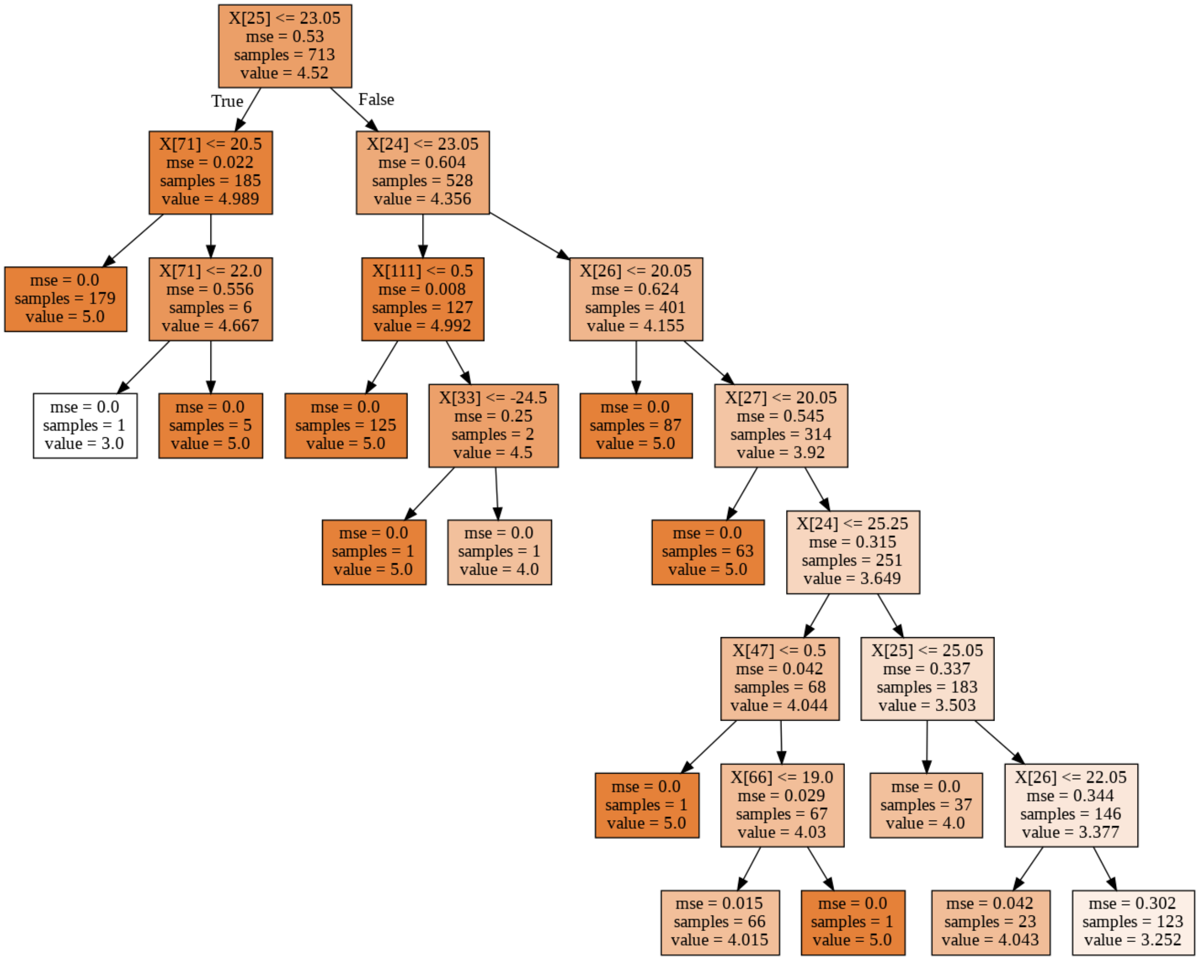

Figure 14 - Visual representation of the decision tree - a model that gives the most accurate results for the task of predicting the risk of vibration disease (general vibration)

The features that the algorithm focuses on when descending to the leaves of the tree and making the resulting decision:

feature	*importance*
Thermometry of the left foot, degrees Celsius	0.2726
Thermometry of the right foot, degrees Celsius	0.2702
Thermometry of the right hand, degrees Celsius	0.2426
Thermometry of the left hand, degrees Celsius	0.1959
Vibration sensitivity at 250 Hz on the left	0.0098
Hearing threshold at 250 Hz, (dB)	0.0028
The mucous membrane of the posterior pharyngeal state	0.0027
The hearing threshold at the frequency of 4000 Hz(dB)	0.0019
1.1.	0.0015

Table 17. A list of the features on the basis of which the greatest number of decisions were made for estimation of the risk of vibration disease (general vibration).

We see that this simple but effective algorithm has shown significantly better results than the linear regression model, while the interpretability of the results and the logic of decision-making are just as well understood.

Random Forest Model

This method is based on the construction of a set of algorithms for the decision tree. At the same time, the displayed quality should be better than the quality of individual algorithms. In addition, this algorithm allows you to evaluate the quality of the available features, their importance in decision-making.

Determining the risk of sensorineural hearing loss

For the first target feature, the model showed the following metrics on a deferred sample:

Model metrics type	*Metrics values*
Accuracy score	0.9218
Mean absolute error (mae)	0.1310
Mean squared error (mse)	0.0612

Table 18. Metrics of sensorineural hearing loss random forest model

The random forest model allows you to evaluate how important the features that describe the objects are. The following are the most important features for the first task.

feature	*importance*	*Loss reduction sum*
The hearing threshold at 4000 Hz (dB)	0.3338	0.3338
The hearing threshold at of 4000 Hz (dB)	0.3291	0.6629
The hearing threshold at of 500 Hz (dB)	0.0406	0.7035
The hearing threshold at of 500 Hz (dB)	0.0385	0.7421
The hearing threshold at of 2000 Hz (dB)	0.0338	0.7758
The hearing threshold at of 2000 Hz (dB)	0.0181	0.7940
Untreated ear diseases: chronic purulent otitis	0.0157	0.8096
The hearing threshold at of 250 Hz (dB)	0.0110	0.8206
The hearing threshold at of 6000 Hz, (dB)	0.0102	0.8308
Thermometry of the left foot, degrees Celsius	0.0093	0.8402

Somatic diseases: diseases of the circulatory system	0.0085	0.8486
Vibration sensitivity at of 32 Hz on the left, dB	0.0082	0.8569
The hearing threshold at of 3000 Hz (dB)	0.0071	0.8639
Vibration sensitivity at 125 Hz on the left, dB	0.0071	0.8710
Vibration sensitivity at 63 Hz on the left, dB	0.0064	0.8774

Table 19. The most important features for the sensorineural hearing loss random forest model

Let's look at how the total share of the relative loss decrease grows. This value can be interpreted as what part of the decisions we can make using only these first features. In other words, what proportion of the information needed to solve this problem and contained in the data, we use.

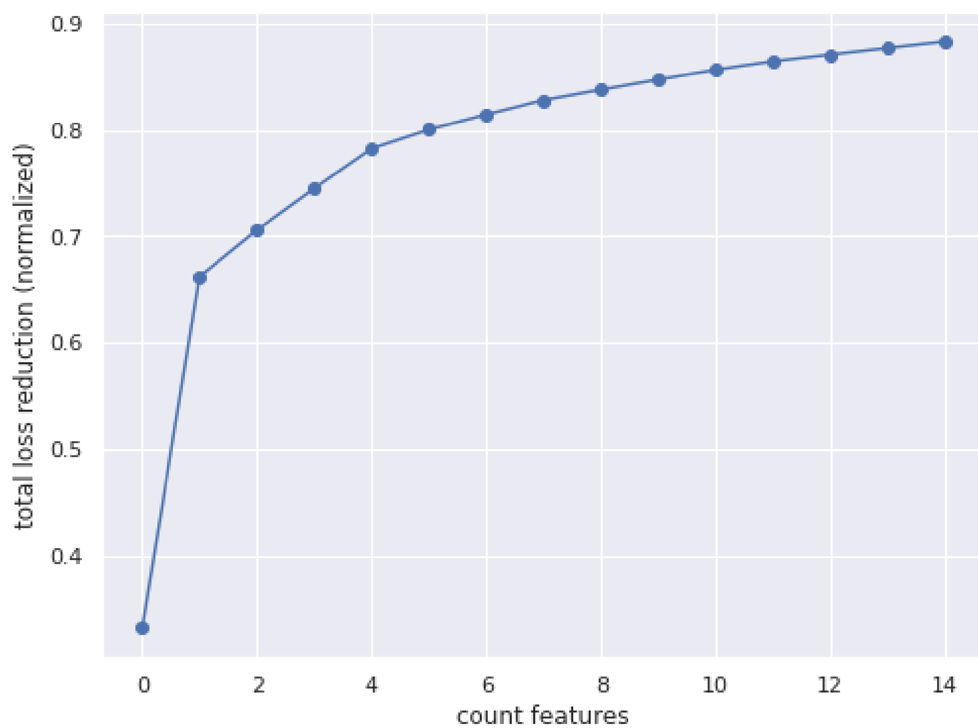

Figure 15 - Dependence of the normalized accuracy of the prediction of the risk class of the model on the number of features taken into account in the model (sensorineural hearing loss)

These data allow us to highlight information that is primarily worth paying attention to when making decisions. We can identify a set of features that ensures maximum awareness of the decision-making system for this type of tasks.

Determination of the risk of vibration disease (local vibration)

For the second target feature, the model showed the following metrics on a deferred sample:

Model metrics type	*Metrics values*
Accuracy score	0.9218
Mean absolute error (mae)	0.1460
Mean squared error (mse)	0.0977

Table 20. Metrics of vibration disease random forest model (local vibration)

Below are the top 15 most important features for decision-making when determining the second target feature:

feature	*importance*	*Loss reduction sum*
Thermometry of the right hand, degrees Celsius	0.4022	0.4022
Thermometry of the left hand, degrees Celsius	0.3715	0.7737
Numbness and/or paresthesia of the hands	0.0373	0.8110
Hyperhidrosis of the hands	0.0286	0.8397
Paint brushes (Normal color - a, Marble-cyanotic - b)	0.0162	0.8559
Pain in the hands / and forearms (yes - a, no - b)	0.0159	0.8717
Attacks of whiteness / blueness of the fingers	0.0157	0.8874
Age	0.0128	0.9002
Symptom of white spot and/or Bogolepovs test	0.0105	0.9107
Hypalgesia and/or hypesthesia of the feet	0.0076	0.9183
Vibration sensitivity at 125 Hz on the left, dB	0.0043	0.9226
The hearing threshold at the frequency of 4000 Hz (dB)	0.0036	0.9262
Vibration sensitivity at 250 Hz on the left, dB	0.0032	0.9294
Vibration sensitivity at 250 Hz on the right, dB	0.0032	0.9326
Work experience	0.0030	0.9357

Table 21. The most important features for vibration disease random forest model (local vibration)

How does the total share of information about the patient in the issue of solving this type of tasks grow from the first important features:

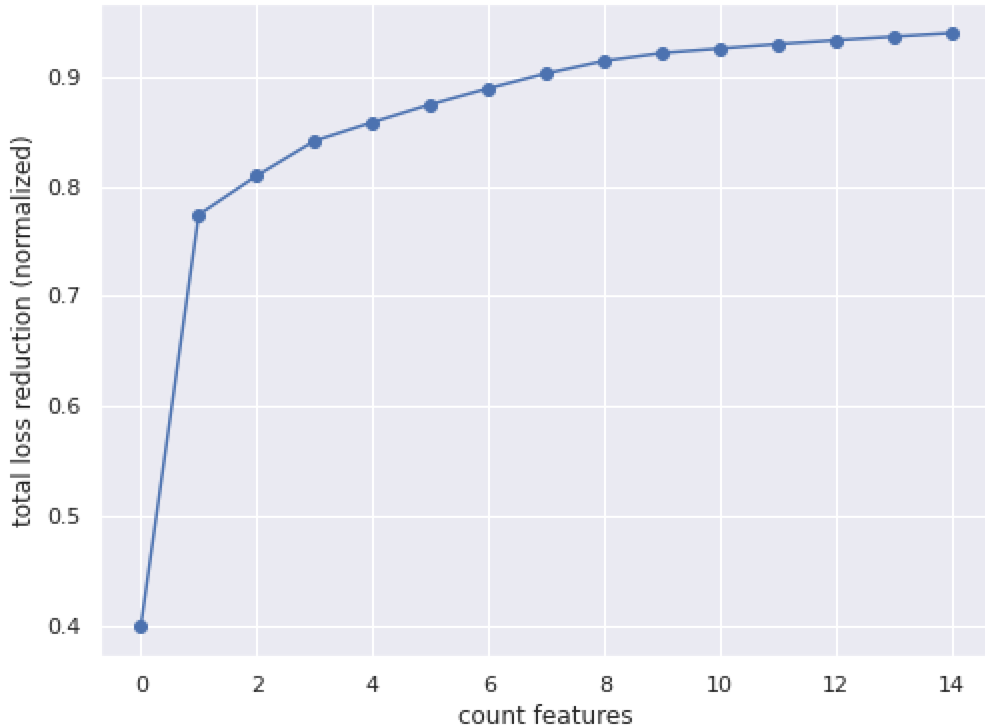

Figure 16 - Dependence of the normalized accuracy of the prediction of the vibration disease risk class by the model on the number of features taken into account in the model (local vibration)

First of all, it is worth paying attention to these features when making decisions on determining the risk of local vibration.

Determination of the risk of vibration disease (general vibration)

For the third target feature, the model showed the following metrics on a deferred sample:

Model metrics type	*Metrics values*
Accuracy score	0.9596
Mean absolute error (mae)	0.0738
Mean squared error (mse)	0.0363

Table 22. Metrics of vibration disease random forest model (local vibration)

Below are the top 15 most important features for decision-making when determining the third target feature.

feature	*importance*	*Loss reduction sum*
Thermometry of the left foot, degrees Celsius	0.2479	0.2479
Thermometry of the right foot, degrees Celsius	0.2465	0.4944
Thermometry of the right hand, degrees Celsius	0.2280	0.7225
Thermometry of the left hand, degrees Celsius	0.1923	0.9147
Age	0.0109	0.9256
The hearing threshold at the frequency of 4000 Hz (dB)	0.0062	0.9318
Vibration sensitivity at a frequency of 32 Hz	0.0046	0.9364
The hearing threshold at a frequency of 6000 Hz (dB)	0.0039	0.9403
Vibration sensitivity at 63 Hz on the right, dB	0.0033	0.9436
Tinnitus (constant) (yes - a, no - b)	0.0033	0.9468
Hypalgesia and/or hypesthesia of the feet	0.0029	0.9497
Vibration sensitivity at 250 Hz on the left, dB	0.0025	0.9523
Vibration sensitivity at 250 Hz on the right dB	0.0025	0.9548
The hearing threshold at the frequency of 250 Hz (dB)	0.0023	0.9571
Work experience	0.0023	0.9594

Table 23. The most important features for vibration disease random forest model (general vibration)

The growth of the total "importance" of features with an increase in their number (we take the top of the most important features):

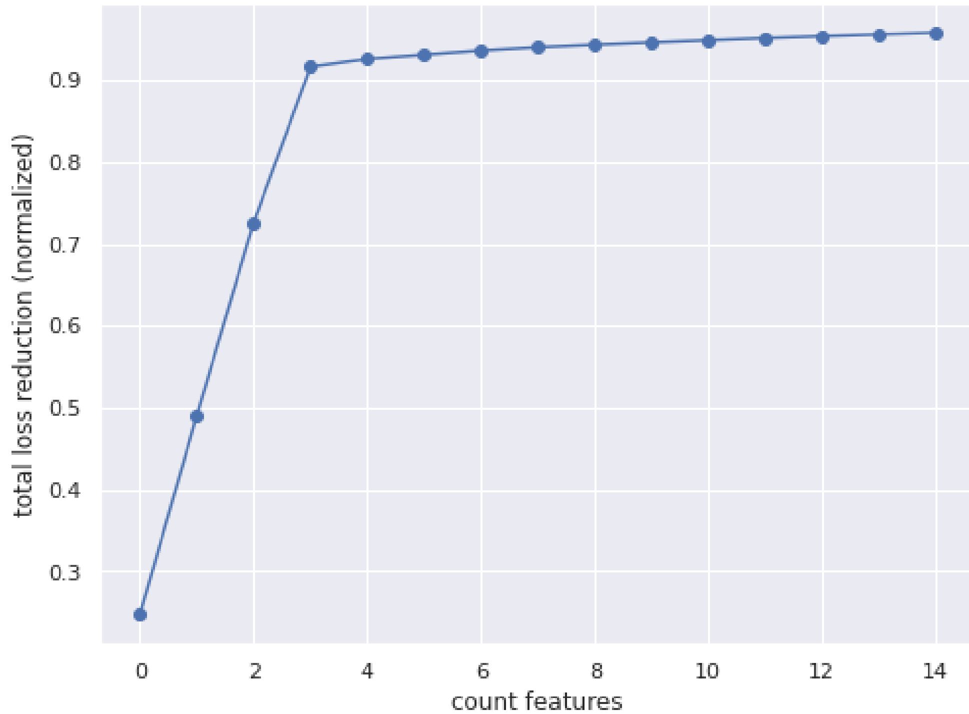

Figure 17 - Dependence of the normalized accuracy of the prediction of the vibration disease risk class by the model on the number of features taken into account in the model (total vibration)

These features should first of all be paid attention to when making decisions to determine the risk of general vibration.

We see that this algorithm has allowed us to further improve the results obtained.

Predictive risk model

For the correct formulation of the problem of risk forecasting and for a meaningful study of forecasts, data is needed in which a time series for each patient would be presented. In other words, such a cross-section of parameters that we have at the moment must be repeatedly reproduced in time. When predicting the result at a future moment in time, to a certain extent, it relies on information about the current moment in time (or on information in past moments of time). At this stage, we do not have this reference point on which the created model should have been based. This is necessary not only for prediction, but also for training the model itself.

At the current time, this disadvantage can be compensated by the fact that we can change certain features that will naturally change with the passage of time, and look at the result of the model. At the same time, we are forced to assume that all other features, clearly independent of time, will remain unchanged. In reality, other features that fully describe the object can and will change. So, a person's working conditions may change, certain symptoms and observations may appear. All of the above certainly depends on the length of service, working conditions and changes over time. To take into account such subtle effects, it is necessary to assess how these features change over time. For such models, we come back to the question of the data provided, because to train such algorithms, time series with variable parameters are needed. This conclusion confirms the fact that, judging by the results of the models (random forest and decision tree), the main contribution to the result is made by specific medical parameters, which only indirectly depend on time. However, as a first approximation, we can consider how the risk changes in the linear regression model.

For example, consider the problem of determining the risk of sensorineural hearing loss when solving it with a linear regression model. Consider a person in a certain risk group. Next, we will change his features of seniority and age, which will grow over time. And let's estimate how many years the risk will increase to the next degree.

In the model we trained, weights for features of age and seniority have positive values. This corresponds to our expectations that the longer a person works and the more experience he has, the greater the risk of getting an occupational disease.

We will change the features of age and seniority and monitor how the results that our model produces change (at the same time, we will not forget that other features can also change over time, but in this approach we neglect this). For such a change, we get:

Year	*Risk estimation*
0	2.132
5	2.306
10	2.480
15	2.653
20	2.827
25	3.001

Table 24. Risk of sensorineural hearing loss forecasted by a linear regression model

We see that over 25 years of work, with the current features, the risk will increase to 3.

It is worth noting once again that most likely, during this time, other features that describe employees should change. In order to correctly predict the risk, we must take into account such changes.

As an alternative, we can offer a statistical regression assessment of changes in features, which can be used to calculate time-varying parameters for their further transmission to the model.

Conclusions and model selection

The study presents the results of using various mathematical models to determine the risks of neural network hearing loss, local and global vibration. The results of the work show that the problem is solved, the accuracy of the results obtained can be called good. Tables 1-3 below show comparative indicators of all the presented models.

Table 1 - Risk of sensorineural hearing loss

Accuracy\ Model	*Linear Regression*	*Decision Tree*	*Random Forest*
Accuracy (the proportion of exact solutions of the model)	0.788	0.894	0.933
MAE (average absolute error)	0.325	0.154	0.127
MSE (Standard error)	0.183	0.099	0.060

Table 25. Metrics of sensorineural hearing loss risk models

Accuracy\ Model	*Linear Regression*	*Decision Tree*	*Random Forest*
Accuracy (the proportion of exact solutions of the model)	0.464	0.838	0.927
MAE (average absolute error)	0.651	0.186	0.148
MSE (Standard error)	0.682	0.183	0.094

Table 26. Metrics of vibration disease risk models (local vibration)

Accuracy\ Model	*Linear Regression*	*Decision Tree*	*Random Forest*
Accuracy (the proportion of	0.542	0.927	0.966

exact solutions of the model)			
MAE (average absolute error)	0.543	0.095	0.077
MSE (Standard error)	0.466	0.061	0.036

Table 27. Metrics of vibration disease risk models (general vibration)

We see that the random forest model performed best in all three tasks. The advantages of this model include the fact that minimal preprocessing of features is required. The decision tree model also showed good results.

Nevertheless, the results of this study can be improved if more time is devoted to analyzing existing features, generating new features, configuring hyperparameters of the algorithms used, considering new metrics and the possibilities of using the results obtained.

The encouraging results of this work do not allow us to speak with confidence about the possible application of these models for complex risk forecasting over time. Now we can only give an estimate for these parameters. The paper indicates ways to solve this issue: first of all, we need appropriate data that would show us how the features of objects change over time. But even in this form, the models shown can demonstrate the optimal ways to work with staff and their healthcare.

An example of an explicitly interpreted DSS fragment based on an automated generated and optimized model

Here is an example of a constructed decision tree up to a depth of 2 (Figure 18). In this case, the upper-level branching criteria can be considered as decisive rules and expert conditions that can be used when making decisions:

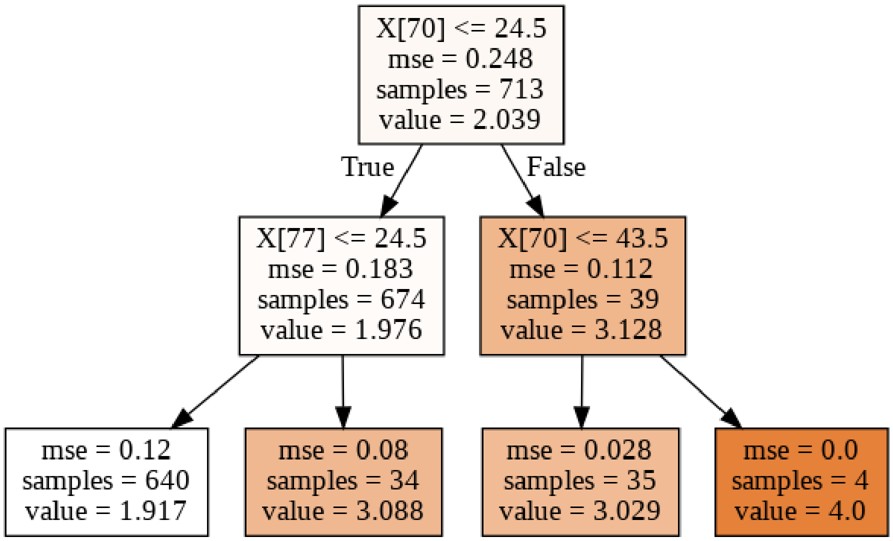

Figure 18 - Fragment of the decision tree model (example)

In this image, the conditions X[70] ≤ 24.5 are given in the first block. This suggests that we should look at the 70th feature of the description of the object – in this case, it is "The threshold of hearing at a frequency of 4000 Hz (dB)". Accordingly, for further decision-making, we descend either to the left or to the right branch, depending on the result of our comparison. At the next level, we check the 77th attribute in the left branch, and the 70th again in the right branch, but at a new value.

As a result, we finally descend to a certain answer that determines the risk in question.

In this case, we can stop at the level we need (or at the available one) and already make the necessary assessment on it.

Justification of risk prediction error estimation

To estimate the error, we will use the results obtained as a result of the inference of trained models on deferred data.

The first approach to estimating the risk prediction error

In the first approach, for each resulting risk level, we estimate the resulting average absolute deviation from the true values of the predicted class. As a result, after memorizing the received error level, we will be able to predict the interval in which the true value of risk lies with a high degree of probability.

This approach can be described formally. Let us define $\hat{y}_j$ - predicted risk value for the object j, y_j - the true value of risk. Then for the risk r=i the prediction error will be determined as follows:

$$err_i = \frac{\sum_{j|\hat{y}_j == i} \left| y_j - \hat{y}_j \right|}{\sum_j I\left[\hat{y}_j == i\right]}.$$

After making calculations, we can give the error of the obtained estimate, using also other statistics of the obtained deviation modules.

Note that this estimate will be the better, the larger the deferred sample we carry out these calculations.

The second approach to estimating the error of the risk forecast

The second approach complements the first approximation in a certain way. In it, we can estimate the resulting error using subsamples of features of objects. This information will be especially useful in the case of limited information about objects.

In this approximation, it is necessary to repeat the logic of error calculations, using only the corresponding subspaces of features.

In this case, the error interval will naturally narrow with an increase in the number of features for the object under study. Using an estimate for the importance of features of objects (in any of the presented models), we can identify several sets of features that are responsible for, for example, a basic, extended and detailed description of the object, and then make an estimate for the resulting errors for each set of features. With this logic of combining features, the extended set will include the basic one, and the detailed one will include both of them.

In general, we can distinguish N such groups of features [g_1, g_2, …,g_N], then for each of these groups for risk level i we will have an error vector [e_1, e_2, …,e_N], each component of the vector will be expressed similarly to the first approach:

$$e_j^i = \frac{\sum_{k|\hat{y}_k(x_j)==i} \left| y_k - \hat{y}_k(x_j) \right|}{\sum_k I\left[\hat{y}_k == i\right]}.$$

Here $\hat{y}_k(x_j)$ - prediction of a model for an object in the description of which there are only features from the group g_j.

It is also worth noting that these estimates require an even larger dataset for accurate estimates of the errors obtained.

The presented study presents the results obtained using the first approach.

Conclusions

In this study there were represented some data-based health risks estimation and forecasting models for workplaces with high levels of noise and vibrations. Features extraction and models quality metrics comparison were described. Similar approaches could be applied for medical and lifestyle personalized data analysis in various areas and tasks.

Acknowledgements

This work has been performed with partial support from National Technological Initiative (M.M.Deminov, Health Heuristics Project)

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics approval

The original Study was performed with the approval of the Ethics Committee of the Institute of Occupational Medicine (Moscow)

References

1. Dobie RA (2008) The burdens of age-related and occupational noise-induced hearing loss in the United States. Ear Hear 29(4):565-77. DOI: 10.1097/AUD.0b013e31817349ec
2. Ekman L, Lindholm E, Brogren E, Dahlin LB (2021) Normative values of the vibration perception thresholds at finger pulps and metatarsal heads in healthy adults. PLoS One 16(4):e0249461. DOI: 10.1371/journal.pone.0249461
3. GOST R ISO 1999-2017. National standard of the Russian Federation. Acoustics. Estimation of noise-inducted hearing loss. Group T34 (in Russian). https://docs.cntd.ru/document/1200157242. Accessed 1 August 2022
4. ISO 1999:2013. Acoustics — Estimation of noise-induced hearing loss. ICS : 13.140 Noise with respect to human beings
5. ISO 2631-5:2018. Mechanical vibration and shock — Evaluation of human exposure to whole-body vibration — Part 5: Method for evaluation of vibration containing multiple shocks
6. Krut’ko VN, Deminov MM, Briko NI, Mitrokhin OV, Chichua DT (2021) Problems of health and quality of life management: intelligent digital platform “Health Heuristics”.

National Health Care (Russia). 2(2):55-63 (in Russian). https://doi.org/10.47093/2713-069X.2021.2.2.55-63. Accessed 1 August 2022

7. Mahbub MH, Hase R, Yamaguchi N, Hiroshige K, Harada N, Bhuiyan ANH, Tanabe T (2020) Acute Effects of Whole-Body Vibration on Peripheral Blood Flow, Vibrotactile Perception and Balance in Older Adults. Int J Environ Res Public Health 17(3): 1069. DOI: 10.3390/ijerph17031069
8. Ntlhakana L, Nelson G, Khoza-Shangase K (2020) Estimating miners at risk for occupational noise-induced hearing loss: A review of data from a South African platinum mine. S. Afr J Commun Disord 67(2):e1-e8. DOI: 10.4102/sajcd.v67i2.677
9. Roberts B, Seixas NS, Mukherjee B, Neitzel RL (2018) Evaluating the Risk of Noise-Induced Hearing Loss Using Different Noise Measurement Criteria. Ann Work Expo Health 62(3):295-306. DOI: 10.1093/annweh/wxy001
10. Sliwinska-Kowalska M (2020) New trends in the prevention of occupational noise-induced hearing loss. Int J Occup Med Environ Health 33(6):841-848. DOI: 10.13075/ijomeh.1896.01600
11. WHO (2020) Deafness and hearing loss. WHO Fact sheets: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed 1 August 2022