Neural networks and regression analysis in the diagnosis of breast cancer

Victoria Dubovskaya¹, Alexander Losev¹
Volgograd State University, 100, Prospect Universitetsky, Volgograd, Russia
E-mail: {dubovskajav, alexander.losev}@volsu.ru

Abstract. This work is devoted to the study of the dependence of the temperature fields of the mammary glands on external conditions and the parameters of the anamnesis, and preliminary examination of patients. As a result, it was possible to significantly improve the space of thermometric diagnostic signs intended for the intelligent system. The initial set of highly informative diagnostic thermometric signs was earlier obtained by A.G. Losev and V.V. Levshinsky. To take into account the influence of external factors on the temperature during the formation of the feature space, regression models were proposed. They were built by the method of neural network modeling. These models have sufficient performance and low error value, which allows them to be used in practice. The use of neural networks made it possible to scale the database of thermometric data obtained using a combined and EMC-sensor. As a consequence, it became possible to analyze the influence of the previously revealed heterogeneity of data in the context of age and diameter of the mammary glands on the effectiveness of highly informative diagnostic signs.

1. Introduction
Active use of artificial intelligence is a distinctive feature of modern medical diagnostics. Mathematical modeling of physical processes in biological tissues also have important applied value. Advisory systems are able to account for subtle classifications and complex signs relationships. Neural networks have a huge potential in this direction. The use of networks in working with medical data allows to increase the accuracy and reliability of detecting diseases [1-3].

Breast cancer is one of the most common and dangerous oncological diseases. At the moment, there is a tendency to increase the proportion of young women in relation to the entire array of breast cancer patients. The result of treatment is greatly influenced by the timely detection and differential diagnosis of breast formations. If a tumor less than 5 mm is detected, in most cases a complete cure is possible. However, doctors diagnose stage III-IV of the disease in approximately 30% of patients with the primary examination. Survival rates vary depending on the level of medical development. Untimely treatment of patients and the lack of effective methods for diagnosing early stages of breast cancer cause a high frequency of detection of advanced forms.

Therefore, early differential diagnosis of breast diseases is a complex and urgent problem of modern medicine. Traditional methods of examination (pulption, mammography, cytology etc.), have not solved the problem of early detection of breast cancer. In recent years, the method of microwave radiothermometry has been actively developing. This method allows to early diagnosis by the non-invasive determination of the temperature inside the patient's organs [4-8]. The advantage of RTM is an effective and safe breast examination for women of all age groups. Safety plays a key role, as benign
breast formations are detected in every 4th woman under the age of 30 years. In patients over 40 years of age, pathological conditions of the mammary glands are diagnosed in 60% of cases.

The microwave radiometry method uses passive radar techniques. Passive radar in the decimeter and centimeter wavelength ranges allows you to observe changes in the thermal fields of internal organs. Microwave radiometry is an effective and affordable tool for conducting mass preventive examinations, as it is a very sensitive indicator of emerging deviations in the functioning of human organs. The measurements are carried out using the antenna-applicators. They are set on the surface of the patient's body. The measuring system registers electromagnetic radiation proportional to the radio brightness temperature associated with the physical temperature of tissues and the degree of absorption of electromagnetic waves in them. The principle of operation of the method is based on the use of natural electromagnetic radiation by arbitrary objects whose temperature is different from absolute zero. Such bodies emit almost all wavelengths with intensities obeying Planck’s law.

Different antenna applicators have their own characteristics [9]. One of the most important problem of RTM diagnostics is the problem of noise immunity, since radiation of human tissues is very weak and amounts to 10^{-14} W. Electronic devices around us often radiate much stronger. Recent advances in science have made it possible to create noise-immune antennas (EMC-sensor) in the 10-centimeter diapason. They can work without special shielding. The doctor does not know in advance where the temperature source (heat anomaly) is, closer to the skin or in depth. In the first case, the skin temperature will be more informative, in the second - the internal one. It should be noted that it is difficult to carry out sequential measurements of self-radiation by two sensors at the same points, which leads to additional errors. Therefore, to expand this approach, a combined sensor was created, in which the IR and RTM sensors are integrated simultaneously.

Since its inception, the main area of application of the RTM method is mammology. A number of studies have noted that the use of microwave radiothermometry in combination with artificial intelligence methods shows high efficiency in early diagnosis of diseases (see, for example [1, 10]).

In the process of applying the method of microwave radiothermometry, a significant influence on the temperature fields of factors such as ambient temperature, breast diameter, patient's age, etc. was revealed. In this study, it was decided to take these factors into account when forming the feature space using regression models built on the basis of the neural networks.

2. Research materials and methods

The RTS-01-RES device is used to diagnose the mammary glands. Internal and skin temperatures are measured at 44 specific points in the mammary glands.

Pic. 1 – The measurement scheme of the breast

The sample of temperature data can be represented as:
where \( t_i^j \), \( i = 0, \ldots, 9 \) – temperature of the i-th point of the right breast of the j-th patient (RTM-range);
\( t_i^j \), \( i = 10, \ldots, 19 \) – temperature of [i-10]-th point of the right breast of the j-th patient (IR-range);
\( t_i^j \), \( i = 20, \ldots, 29 \) – temperature [i-20]-th point of the left breast of the j-th patient (RTM-range);
\( t_i^j \), \( i = 30, \ldots, 39 \) – temperature [i-30]-th point of the left breast of the j-th patient (IR-range);
\( T_{10} \), \( T_{40} \) = T1, \( T_{41} \) = T2 – temperatures of the reference points of the j-th patient (RTM-range);
\( T_{42} \) = T1, \( T_{43} \) = T2 – temperatures of the reference points of the j-th patient (IR-range).

As a result of the studies carried out by specialists, a number of highly effective thermometric signs of the disease were identified, for example: the difference in temperature of the nipples, increased thermoasymmetry between the same points of the right and left mammary glands, increased temperature spread between individual points of the affected mammary gland, the ratio of skin and depth temperatures, and others [1, 4, 10].

Information about the mammary glands of patients is contained in the thermometric database, which was created on the basis of surveys conducted in a number of Russian cancer centers. In addition to the above data, the database also contains information about the patient's age, ambient temperature, breast diameter, cycle day, height and weight. Therefore, it was concluded that these factors are significant, since they are used by a specialist in the diagnostic process.

The dataset consists of 229 patients of which 145 are classified as healthy, and 83 classified as potentially cancerous (EMC-sensor) and 10428 patients of which 9509 are classified as healthy, and 919 classified as potentially cancerous (combined sensor).

3. Results of investigate
In previous works, the hypothesis of homogeneity of samples of thermometric data by the diameter of the breast, the place of examination, the age of the patient, and the type of sensor was tested [11]. The nonparametric Mann-Whitney test and the Kruskal-Wallis test were used to determine the homogeneity of the samples. The statistical programs Statistica was chosen to automate the process of calculating criteria for features.

The following results were obtained:
1. The hypothesis about the homogeneity of the samples of different sensors was rejected. Large heterogeneity is characteristic of the healthy control class.
2. Pairwise comparison of samples of age groups showed that the data for 40-50 and >50 years are homogeneous. Therefore, it is possible to combine these groups into one during research. Samples of groups of adjacent ages are also homogeneous.
3. The hypothesis about the dependence of the detection of the disease on the location was rejected.
4. In the course of testing the hypothesis of homogeneity for samples with different diameters of the mammary glands, absolute homogeneity was revealed for patients with breast cancer.

Based on the results obtained, it was decided to divide the data into two groups by age and diameter of the breast. Accordingly, neural networks were built for each group separately.

The Statistica program was also used to build regression models. External factors such as ambient temperature, cycle day, age, diameter of the patient's breast and mass index acted as independent parameters (EMC-sensor). Only the first 4 factors were used for the combined sensor data, since there was no information on weight and height. At the first stage of the study, models were considered in which the temperature of the mammary gland at a particular point acts as the dependent variable. To
extend this approach, regression models were built not only for the temperatures of the mammary glands, but also for the modeling functions. These functions represent a mathematical formalization of the known heuristics and describe the nuances of temperature behavior in the breast tissue of patients from various control classes.

The multilayer perceptron was chosen as the architecture of neural networks, since it is the most suitable for solving nonlinear regression problems [12]. The number of hidden processing nodes that actually determine the prediction equation is selected by Statistica. The minimum number of hidden neurons is 3, the maximum is 11. The function of activation of hidden and output neurons is determined in a similar way. The following functions are presented: identity, logistic, hyperbolic, exponential, and sinusoidal. Regularization of the weights was used to minimize the effect of overfitting.

All statistics are generated independently for training and test samples. The purpose of the search algorithm for a suitable neural network is to iterate over a number of neural network configurations and select the best one in terms of the minimum error at the network output, which is a function of errors averaged over the entire set, and maximum performance. The performance indicator characterizes the quality of network forecasting, i.e. the degree to which the output matches the target.

Some of the results obtained are shown below. For example, the temperature in specific points, the maximum temperatures, the average temperature of the breast.

Table 1. Regression models constructed by neural network modeling tools (samples by breast diameter)

| Signs | Diameter | Network | Performance | Errors |
|-------|----------|---------|-------------|--------|
|       |          |         | Training    | Test   | Training | Test   |
| t₀    | General  | MLP 5-9-1 | 0.38        | 0.32   | 0.01     | 0.01   |
|       | <21      | MLP 5-10-1 | 0.39        | 0.43   | 0.02     | 0.01   |
|       | 22-26    | MLP 5-7-1 | 0.53        | 0.53   | 0.01     | 0.01   |
|       | >27      | MLP 5-9-1 | 0.47        | 0.58   | 0.02     | 0.07   |
| t₁    | General  | MLP 5-9-1 | 0.23        | 0.50   | 0.01     | 0.01   |
|       | <21      | MLP 5-9-1 | 0.42        | 0.51   | 0.01     | 0.01   |
|       | 22-26    | MLP 5-7-1 | 0.43        | 0.89   | 0.01     | 0.07   |
|       | 22-26    | MLP 5-7-1 | 0.43        | 0.89   | 0.01     | 0.07   |
| t₈    | General  | MLP 5-7-1 | 0.34        | 0.41   | 0.01     | 0.01   |
|       | <21      | MLP 5-5-1 | 0.42        | 0.44   | 0.01     | 0.02   |
|       | 22-26    | MLP 5-6-1 | 0.42        | 0.89   | 0.01     | 0.04   |
|       | 22-26    | MLP 5-7-1 | 0.62        | 0.51   | 0.01     | 0.01   |
| ∑ tᵢ  | General  | MLP 5-7-1 | 0.29        | 0.50   | 0.01     | 0.02   |
|       | <21      | MLP 5-10-1 | 0.39        | 0.43   | 0.02     | 0.01   |
|       | 22-26    | MLP 5-6-1 | 0.49        | 0.56   | 0.01     | 0.01   |
|       | >27      | MLP 5-9-1 | 0.38        | 0.36   | 0.2      | 0.01   |
| max tᵢ| General  | MLP 5-7-1 | 0.29        | 0.5    | 0.01     | 0.02   |
|       | <21      | MLP 5-5-1 | 0.4        | 0.33   | 0.01     | 0.01   |
|       | 22-26    | MLP 5-10-1 | 0.54       | 0.47   | 0.05     | 0.01   |
Based on the data shown in table 1, it can be concluded that dividing the data into separate samples relative to the breast diameter increases network performance and reduces error. In relation to temperatures at specific points, the effect is expressed more significantly.

Table 2. Regression models built by neural network modeling tools (samples by age)

| Signs                   | Age   | Network | Performance | Errors |
|-------------------------|-------|---------|-------------|--------|
|                         |       |         | Training    | Test   |
|                         |       |         | Training    | Test   |
| General                 |       | MLP 5-9-1 | 0.38    | 0.32  |
|                         | <40   | MLP 5-5-1 | 0.46    | 0.9   |
|                         | >40   | MLP 5-7-1 | 0.31    | 0.86  |
| General                 | <40   | MLP 5-10-1 | 0.43   | 0.63  |
|                         | >40   | MLP 5-10-1 | 0.55   | 0.61  |
| General                 | <40   | MLP 5-6-1 | 0.44    | 0.88  |
|                         | >40   | MLP 5-9-1 | 0.44    | 0.68  |
| General                 | <40   | MLP 5-10-1 | 0.39    | 0.43  |
|                         | >40   | MLP 5-6-1 | 0.22    | 0.49  |
| General                 | <40   | MLP 5-10-1 | 0.29    | 0.5   |
|                         | >40   | MLP 5-5-1 | 0.29    | 0.8   |

Similar to the previous result, in most cases dividing the data into age groups improved the model's performance. It may be due to the heterogeneity of data related to structural changes that occur in the mammary glands of patients after 40 years.

Neural networks can be used to scale a database of thermometric data. It is assumed to bring the temperature data to the optimal conditions. A comparison was made of the effectiveness of the initial signs in the database (set 1) and signs obtained using regression (set 2). Sensitivity and specificity are used to evaluate the effectiveness of elements of the signs space and diagnostic methods. In the case of unbalanced samples, the efficiency criterion is a measure defined as the geometric mean of specificity and sensitivity [10].

The following algorithm was implemented.
1. Finding the optimal values for the input parameters of multiple regression function.
2. Calculation of the regression function value for parameters corresponding to a specific patient from the survey database.
3. Calculation of the increment — the difference between the value of the modelling function and the value of the regression function.
4. Adding the value of the regression function for the initial values of the parameters and the increment.

5. Calculation and comparison of the effectiveness of the signs.

Table 3. Effectiveness of signs (EMC-sensor)

| Signs   | Interval (Set 1) | Interval (Set 2) | Effectiveness (Set 1) | Effectiveness (Set 2) |
|---------|------------------|------------------|-----------------------|-----------------------|
| $t_0$   | $\infty$, 34.15  | $\infty$, 35.90  | 0.53                  | 0.63                  |
| $t_1$   | $\infty$, 33.90  | $\infty$, 35.00  | 0.38                  | 0.59                  |
| $t_s$   | $\infty$, 34.02  | $\infty$, 36.00  | 0.39                  | 0.63                  |
| $\sum_{i=1}^{n} t_i$ | $\infty$, 33.93  | $\infty$, 36.00  | 0.53                  | 0.63                  |
| $\max_{i=0.9} t_i$    | $\infty$, 34.45  | $\infty$, 36.02  | 0.44                  | 0.63                  |

Table 4. Effectiveness of signs (combined sensor)

| Signs   | Interval (Set 1) | Interval (Set 2) | Effectiveness (Set 1) | Effectiveness (Set 2) |
|---------|------------------|------------------|-----------------------|-----------------------|
| $t_0$   | $\infty$, 34.45  | $\infty$, 34.57  | 0.45                  | 0.89                  |
| $t_1$   | $\infty$, 34.99  | $\infty$, 34.20  | 0.50                  | 0.70                  |
| $t_s$   | $\infty$, 34.73  | $\infty$, 34.60  | 0.44                  | 0.55                  |
| $\sum_{i=1}^{n} t_i$ | $\infty$, 33.92  | $\infty$, 33.24  | 0.62                  | 0.75                  |
| $\max_{i=0.9} t_i$    | $\infty$, 35.80  | $\infty$, 35.25  | 0.51                  | 0.62                  |

Based on the data obtained, it can be concluded that the application of this approach at temperatures is more rational. The calculated ones can serve as a basis for creating other more informative features. A similar result is observed for both sensors, but in the case of a combined one, it is expressed more significantly.

4. Conclusion

Thus, the results of the study confirm the need to take into account data heterogeneities in the context of age and diameter of the mammary glands, especially when creating an advisory system and when forming a training sample.

The accuracy provided by neural networks allows them to be used in expert systems. The neural network modeling tools made it possible to identify the influence of all external factors on the temperature of the mammary glands. The existing signs space was refined to take into account external parameters.

Scaling of the thermometric database is possible using neural networks. As a result, a higher efficiency of signs is achieved and the accuracy and reliability of diagnostic results is increased.
The result of using regression models is significantly noticeable in relation to the combined sensor. Since this sensor is more modern and frequently used, it is advisable to conduct further research in relation to it.

Acknowledgments
This work was supported by the RFBR, project no. 19-01-00358 "Mathematical models of radiation fields and analysis of microwave radiothermometry data in the early diagnosis of breast cancer".

References
[1] Levshinskii V, Galazis C, Ovchinnikov L, Vesnin S, Losev A and Goryanin I 2020 Communications in Computer and Information Science 1211 265
[2] Yasnitsky L N, Dumler A A, Poleshuk A N, Bogdanov C V and Cherepanov F M 2015 Biology and Medicine (Aligarh) 2(7) BM-095-15
[3] Arajo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polnia A and Campilho A 2017 PLOS ONE 12(6) 1
[4] Vesnin S, Turnbull A K, Dixon J M, Goryanin I 2017 Journal of Molecular Imaging & Dynamics 7(2) 136
[5] Goryanin I, Karbainov S, Shevelev O, Tarakanov A, Redpath K, Vesnin S, and Ivanov Y 2020 Drug Discovery Today 25(4) 757
[6] Toutouzas K, Synetos A, Nikolaou C, Stathogiannis K, Tsiamis E and Stefanadis C 2012 Cardiovascular diagnosis and therapy 2(4) 290
[7] Rodrigues D B, Stauffer P R, Pereira P J S and Maccarini P F 2018 Electromagnetic Technologies for Brain Diseases Diagnostics, Monitoring and Therapy 87
[8] Polyakov M V, Khoperskov A V and Zamechnic T V 2017 Lecture Notes in Computer Science 10594 128
[9] Vesnin S G, Sedankin M K, Gudkov A G, Leushin V Y, Sidorov I A, Porokhov I O et al 2020 Biomedical Engineering 54 79
[10] Losev A G and Levshinskii V V 2017 Mathematical Physics and Computer Simulation 20(5) 49
[11] Mazepa E A and Dubovskaya V I 2019 Modern science and innovation 4 39
[12] Goodfellow I, Bengio Y and Courville A 2016 Deep Learning: The MIT Press 19 800