

Article

Method for Detecting Pathology of Internal Organs Using Bioelectrography

Yulia Shichkina ^{1,*} , Roza Fatkueva ¹ , Alexander Sychev ¹ and Anatoliy Kazak ^{2,*} ¹ Department of Computer Science and Engineering, Saint Petersburg Electrotechnical University LETI, 197022 St. Petersburg, Russia² Humanitarian Pedagogical Academy, V.I. Vernadsky Crimean Federal University, 295007 Simferopol, Russia

* Correspondence: strange.y@mail.ru (Y.S.); kazak_a@mail.ru (A.K.)

Abstract: This article considers the possibility of using the bioelectrography method to identify the pathology of internal organs. It is shown that with the currently existing methods, there is no possibility of the automatic detection of diseases or abnormalities in the functioning of a particular organ, or of the definition of combined pathology. It has been revealed that the use of various classifiers makes it possible to expand the field of pathology and choose the most optimal method for determining a particular disease. Based on this, a method for detecting the pathology of internal organs is developed, as well as a software package that allows the detection of diseases of the internal organs based on the bioelectrography results. Machine-learning models such as logistic regression, decision tree, random forest, xgboost, KNN, SVM and HyperTab are used for this purpose. HyperTab, logistic regression and xgboost turn out to be the best among them for this task, achieving a performance according to the f1-score metric in the order of 60–70%. The use of the developed method will, in practice, allow us to switch to combining various machine-learning models for the identification of certain diseases, as well as for the identification of combined pathology, which will help solve the problem of detecting pathology during screening studies and lead to a reduction in the burden on the staff of medical institutions.

Keywords: bioelectrography method; detecting pathology of internal organs; machine-learning model



Citation: Shichkina, Y.; Fatkueva, R.; Sychev, A.; Kazak, A. Method for Detecting Pathology of Internal Organs Using Bioelectrography. *Diagnostics* **2024**, *14*, 991. <https://doi.org/10.3390/diagnostics14100991>

Received: 18 March 2024

Revised: 1 May 2024

Accepted: 7 May 2024

Published: 9 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The increase in the growth of computing power of modern computer systems accompanied by the decrease in their element base leads to the integration of knowledge, both in related subject areas and in the formation of multidisciplinary research, for example, in medical practice. However, this approach has a contradiction. On the one hand, this leads to an increase in the number of detected diseases. On the other hand, the processes of automating diagnosis require reducing the number of medical personnel involved in routing the patient. The requirements for highly specialized specialists who need to adapt to a rapidly changing hardware and software environment are also increasing. Against this background, there is a need to develop screening programs that allow you to quickly identify a particular pathology.

The purpose of the study described in this article was to increase the effectiveness of detecting the pathology of internal organs through data analysis using machine-learning methods. As a result, a method for detecting the pathology of internal organs using bioelectrography was developed. The method analyzes which datasets are needed to assess violations, identifies the patterns and correlations existing in them on the basis of which pathology is searched, and also identifies errors and data distortions that may affect the results of the processing. A distinctive feature of the developed method is the possibility of synthesizing new records, which allows you to solve the problem of data imbalance. This made it possible to develop a software package for the practical implementation of the

above method and to form an ensemble of machine-learning models that identify pathology according to the description of GRV-grams.

The novelty of the proposed method consists of:

- the study of various machine-learning models for use in the detection of certain diseases using gas discharge imaging;
- the development of a method for detecting the pathology of internal organs using bioelectrography data, characterized by the use of machine-learning methods to identify the pathology of internal organs, which allows the use of a combination of different classifiers to determine combined pathology;
- the development of a model that allows you to obtain a dataset based on a sample of diseases, as well as generate training, validation and test kits in semi-automatic mode.

Using the developed method in practice will allow us to move to combining various machine-learning models to identify certain diseases, as well as to identify combined pathology, which will help solve the problem of detecting pathology during screening studies and reduce the workload of employees of medical institutions.

The remaining portions of this paper are organized as follows. Section 2 presents the results of the analysis of related work. Section 3 discusses the method of detecting the pathology of internal organs using bioelectrography, as well as the software package used in this work. Section 4 presents the experimental results from a dataset with a sample of 170 patients. Section 5 presents a discussion of the results obtained.

2. Related Work

Such methods include studies that allow, on the basis of the X-ray images and machine-learning methods used, us to identify the pathology of the respiratory organs [1–3]. The methods based on the classification of untreated lung sounds and the detection of pneumonia and chronic obstructive pulmonary diseases can be attributed [4–6]. A separate area is the work on the detection of oncological diseases [7–9]. The use of machine-learning methods also makes it possible to identify the pathology of the genitourinary system [1]. Image segmentation based on machine-learning models and neural networks makes it possible to identify and classify diseases of the gastrointestinal tract [10–13]. In terms of identification studies, possible heart diseases based on clinical data in different patients were considered in [14], but their application in practice relies on the complexity of the data and the presence of correlations between them. The listed methods have a high resource intensity in terms of the applied software and hardware. Summarizing the results [1–14], it is advisable to note that with almost all the methods, there is no possibility of automating the detection of pathology or deviations in the functioning of a particular organ (Table 1). In the table, a «+» sign indicates the presence of this property from the column header in the method from the first column in this source, «−» its absence.

Table 1. Comparison of methods for determining human organ pathologies.

Method	The Possibility of Detecting a Combined Pathology	Assessment of Correlation with Other Pathologies	Resource Intensity of the Method	Automation of Pathology Detection
Machine-learning methods for detecting respiratory pathology	+ [1]/− [2,3]	+ [1]/− [2,3]	+ [1–3]	− [1–3]
Methods based on the classification of pulmonary noises	− [4–6]	− [4–6]	+ [4–6]	− [4–6]
Detection of oncological diseases	+ [8,9]/− [7]	+ [9]/− [7,8]	+ [7–9]	− [7–9]
Machine-learning methods for detecting pathology of the genitourinary system	+ [1]	+ [1]	+ [1]	− [1]
Diseases of the gastrointestinal tract	− [10–13]	+ [11,12]/− [10,13]	+ [10–13]	− [10–13]
Detection of cardiovascular diseases	− [15]	+ [15]	+ [15]	+ [15]

Existing methods are usually limited to considering the pathology of only one of the body's systems. As a rule, the methods for detecting combined pathology are not considered in studies [1–24].

One of the methods with low resource consumption and the ability to quickly identify diseases is the bioelectrography method based on gas discharge imaging (GDV) [15–24].

To identify relevant works, we, as well as the authors of [21], analyzed publications in the Google Scholar, ResearchGate and Elibrary databases using the following keywords: visualization of gas discharge (GDV), Kirlian effect, bioelectrography. The study revealed that during the period from 2000 to 2005, 9 works were published on the research topic; from 2005 to 2010, 48 works; from 2010 to 2015, 72 works; from 2015 to 2020, 51 works; and from 2020 to the present, 18 works. Thus, the peak of publications occurred in the period from 2010 to 2015 (72 works). From the works presented for analysis over the past decade, one can single out a study [17] aimed at establishing normative bioelectrography data for the healthy population of India in order to increase the accuracy of its measurement and interpretation, since according to the authors [17], the data for this population group differ from Europeans. In [18], causal relationships have been established between the parameters of gas discharge imaging and the main neuroendocrine adaptation factors, indicating the informativeness of this method. Using other methods, the authors of [19] reached a similar conclusion, in which it was shown that the method of gas-discharge imaging using a Bio-Well device makes it possible to determine the effect of diagnostic ultrasound on human homeostasis, which opens up prospects for using a Bio-Well device to assess the effects of various medical technologies, both diagnostic and therapeutic, on the human body.

In [20–23], a classification of the practical application of bioelectrography methods in medicine is provided: analysis of the vegetative status of the body and individual functional systems; monitoring of the body's reactions to therapy; assessment of the likelihood of organ pathology. However, all the studies have a number of disadvantages, in particular:

- the difficulty in ensuring the completeness of the volume and quality of the sample of the studied data due to both obtaining the sample itself and decoding medical diagnoses and data;
- the lack of methods for detecting the correlation of organ dysfunction in combined pathology;
- the data transmitted from the bioelectrography device is stored in an unordered form and requires further processing;
- the marking of bioelectrography zones does not always correctly reflect possible pathologies of internal organs.

These problems require additional research to identify a particular pathology. The use of various classifiers allows you to increase the space of diseases and choose the most optimal method that identifies a particular pathology. To identify violations of internal organs in this work, a software package is proposed that allows you to obtain a dataset based on a sample of diseases, as well as to generate training, validation and test kits in automatic mode.

3. Materials and Methods

Before starting to process the data, it is necessary to conduct an analysis to determine which of the sets are necessary for the assessment of pathology, to derive patterns that allow you to determine the correlation of their characteristics, and to form an algorithm of actions to achieve the goal. This approach allows you to identify errors and data distortions that may affect the results of the processing.

To accomplish this, two types of data are generated, the type of data of the characteristics of the disease received from the patient and the type of data of the thresholds, the excess of which characterizes the presence of the disease. Next, to form a dataset, data are received from patients and their feature space is normalized with the removal of irrelevant features and outliers. This allows you to create 3 samples with data: training, validation

and test. In the process of model learning, the set of applied models is refined. This makes it possible to leave only informative models from the formed complex, as well as to clarify the vectors of the data type responsible for detecting pathology and the vector identifying a particular pathology. Let us look at this method in more detail:

Step 1. The formation of a feature space of two types. The first type forms the characteristic space of bioelectrography characteristics, while the second type of data forms the characteristic space of pathology of internal organs. Comparing both types with each other in the following steps makes it possible to identify the characteristics of bioelectrography with a particular type of disease. Let us take a closer look at the formation of 2 types of datasets:

Type 1: data suggesting pathology of internal organs. For example, type 1 sets may represent data that can be easily obtained, but it is not possible to draw a conclusion based on them. The feature space of data characteristics of type 1 are denoted as:

$$B_s = \{b_1, b_2, \dots, b_i, \dots, b_s\} \quad (1)$$

where b_i , $i = \overline{1, s}$ —are the characteristics of the feature space. For example, for bioelectrography, the element of the set b_1 may correspond to the number of pixels of the image of the GRV-gram of any zone of a certain finger, and b_2 to the radius of the inner circle of the GRV-gram zone, etc.

For a more visual representation, it is advisable to provide an example of a snapshot of a gram (Figure 1).

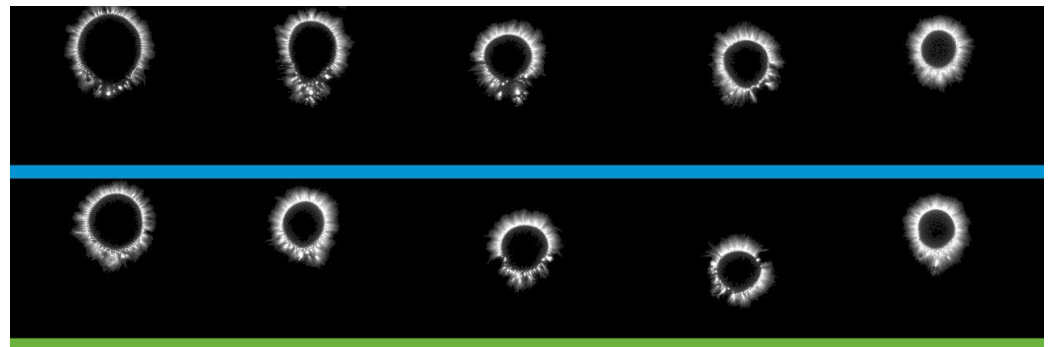


Figure 1. An example of a GRV-gram for a patient with kidney pathology.

The GRV-grams obtained were processed using computer vision algorithms. This made it possible to form the B_s vector of the GRV-gram feature space. The vector contains about a thousand elements. The main ones include the following:

- Ellipse dimensions—the dimensions of the inscribed ellipse in X and Y in pixels;
- Radius of the inner circle—the radius of the inscribed circle in pixels;
- Area—the number of pixels of the sector image;
- Area (K)—the ratio of the glow area of the finger to the glow area of the calibration cylinder of the sector;
- Normalized area—the ratio of the area of the glow to the area of the inner oval of the sector;
- Intensity—the average intensity of all the pixels in the sector;
- Inner area—the total number of pixels in the inner oval of the sector;
- Internal noise—the number of noise pixels in the inner oval of the sector;
- Internal noise (%)—the ratio of the internal noise of the internal area of the sector as a percentage;
- Energy—the glow energy in joules ($\times 10^{-2}$) of the sector;
- Energy (K)—the energy adjusted for the angular size of the sector;
- Shape coefficient (CF)—the ratio of the square of the length of the outer contour to the area of the image;

- Entropy coefficient (KE)—the ratio of the length of the outer contour to the length of the inner sector;
- Inner contour length—the length of the inner contour in pixels of the sector;
- Inner contour radius—the radius of the inner contour in pixels of the sector;
- Length of the outer contour—the length of the outer contour in pixels of the sector;
- Outer contour radius—the radius of the outer contour in pixels of the sector.

If the feature space is too large, it is possible to submit not all of them at once, but some of them can be used, based on algorithms for reducing the feature space.

Type 2: data of the feature space, which can be used to uniquely identify pathology, with the possibility of using them for machine-learning models. The second type of data may contain elements of many patterns of pathology description. Such patterns will be further designated as follows:

$$C = \{c_1, c_2, \dots, c_j, \dots, c_r\} \quad (2)$$

where c_j , $j = \overline{1, R}$ —are the characteristics of the feature space identifying the pathology.

To perform this, the target signs of the pathology for each patient are taken from the conclusions of ultrasound diagnostics specialists (an example of the protocol is shown in Figure 2).

For example, the element c_1 may indicate that the patient has hepatomegaly, and c_2 , diffuse changes in the structure of the thyroid gland.

In a particular case, it can be said that datasets from bioelectrography belong to type 1, and data from ultrasound belong to type 2. It should be clarified that both types should be collected based on data from the same patients, in quantity N .

Step 2. Identification of the patient's characteristic space. This step is necessary to compare the individual characteristics of the patient obtained from bioelectrography with the target characteristics of his pathology. Then, for each patient, within the types selected in step 1, we leave only those that characterize his individual characteristics:

2.1. Type 1 data obtained from a special device is decomposed into elements of the feature space B_s for each patient. That is, each patient will be characterized by their own characteristic space B_s^i , where $i = \overline{1, N}$ is the patient's number. The order of the signs is the same in all the patients, but the values are different. For example, a type 1 data feature space for a patient $i = 1 \dots N$ may look like this:

$$B_s^i = \{b_1^i, b_2^i, \dots, b_i^i, \dots, b_s^i\} \quad (3)$$

2.2. For each patient, their own set of data feature spaces of type 2 is compiled:

$$C^i = \{c_2^i, c_5^i, c_8^i, \dots, c_k^i\} \quad (4)$$

To clarify, the number i for the type 1 data and type 2 data should indicate the same patient. If the patient has no pathology, then the element of the set is skipped, for example, for a patient with $C^i = \{c_2, c_5, c_8, \dots, c_k\}$, $k \leq r$, signs of hepatomegaly pathology, the element is c_2 missing from the set.

Step 3. Normalization of the data feature space for type 1. Since the data obtained for type 1 are heterogeneous, it becomes necessary to bring them to a single scale and range of values. To accomplish this, we introduce the operator F_{norm} :

$$F_{\text{norm}}: B_s^i \rightarrow B_{sN}^i \quad (5)$$

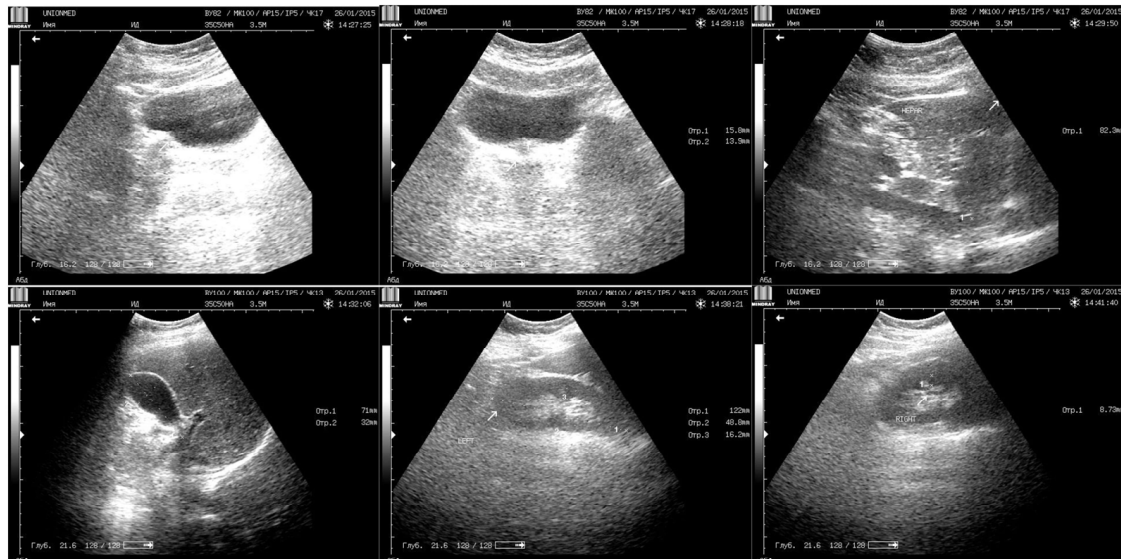
where B_{sN}^i is a set of normalized feature space for the i -th patient.

Normalization is necessary to convert the data into a single form, improve the convergence of the optimization algorithm and level the weights of features when training models in the future. The range of values during normalization and the normalization method itself depend on the task.

Protocol ULTRASOUND

FULL NAME: [REDACTED] Age: [REDACTED]

Liver: enlarged, vertical size of the right side 160 mm, thickness of the left side 82 mm. The structure is diffusely heterogeneous, echogenicity is average. The edge is not rounded, intrahepatic channels are not dilated, portal vein 10 mm, splenic vein 7 mm, choledochus 4.5 mm.



Gallbladder: 73x33 mm, pear-shaped, walls irregularly thickened, contours smooth clear. Anechogenic heterogeneous content with sediment in the lumen.

Pancreas: 24x18x10 mm, not enlarged, contours smooth, indistinct, structure homogeneous. Echogenicity is insignificantly higher than the echogenicity of the liver. Virsung's duct is not visible.

Spleen: 118x65 mm, not enlarged, the structure is not changed. There is no free fluid in the abdominal cavity. No enlarged lymph nodes are visualized.

Right kidney: 107x46 mm, contours smooth, mobility preserved. Parenchyma up to 17 mm, hypoechogenic. The calyx-lochanous complex is not dilated.

Left kidney: 116x50 mm, contours smooth, mobility preserved. Parenchyma up to 18 mm, hypoechogenic. The calyx-lochanous complex is not dilated.

Sinuses are unevenly compacted due to small single hyperechogenic inclusions up to 3-4 mm.

In vertical position the kidneys are at the level of the 12th rib.

Urinary bladder: The shape is rounded. Filling is average. The wall is irregularly thickened in the triangle area up to 13 mm, hypoechogenic. Internal contours are even. The content is homogeneous, anechogenic. Stones are not revealed.

CONCLUSION: Echographic picture of diffuse changes in the liver structure. Gallbladder wall thickening, violation of bile rheologic properties. Indirect ultrasound signs of mineral metabolism disorders. Ultrasound signs of inflammatory changes in the bladder wall.

RECOMMENDED: Consultation of gastroenterologist, urologist.

Thyroid gland

It's normally located. Insignificantly enlarged for age.



Figure 2. Cont.

Size, mm	Right lobe	Left lobe	Gland volume according to the Brunn method: total $V = 10.7 \text{ cm}^3$. Age level of the upper boundary of the volume – 13 cm^3 . The contours of the gland are smooth, indistinct. The gland tissue echostructure is diffusely heterogeneous. Regional thyroid lymph nodes along the carotid arteries, one on each side, up to 5 mm in diameter, with unchanged structure.
	17	17	
	14	15	
	45	45.8	
Volume, ml	5.1	5.6	
Thyroid isthmus	2.4		
Total volume of the gland, ml		10.7	
CONCLUSION: Echographic picture of increased volume, changes in the structure of the thyroid gland. Nodular mass of the right lobe.			
RECOMMENDED: TTG, T3sv, T4sv, ATPO. Ultrasound control in 3 – 6 months. Endocrinologist consultation. MRI of the thyroid area.			

Figure 2. An example of an ultrasound protocol of the internal organs (liver, kidneys, spleen, adrenal glands, pancreas, etc., in Russian).

Step 4. Reducing the dimension of the feature space to reduce the number of features used to describe objects, while preserving the most important properties of the data. To improve the training of models in the future and reduce the complexity of the data at this step, depending on the task, the dimension of the type 1 feature data space is reduced. Let us introduce the operator F_{red} :

$$F_{\text{red}}: B_{\text{SN}} \rightarrow B \quad (6)$$

Let us denote the elements B^i for the i -th patient as follows:

$$B^i = \{b_1^i, b_2^i, b_3^i, \dots, b_n^i\}, i = \overline{1, N} \quad (7)$$

This step is necessary to simplify the data structure. Each gram is described by a large number of features, many of which are probably insignificant. Such preprocessing will simplify the process of learning classifiers.

After applying this operator, the number of elements in the set of attributes of type 1 data will decrease. Depending on the task, the operator F_{red} will be different. It is important to take into account that the operator F_{red} decreases the same signs for each patient. The operator can also change the basis of the feature space while reducing the feature space. Let us denote the size of the set B after applying the operator F_{red} with a symbol n .

Step 5. The introduction of a set P^i describing the pathology of the patient number i . This step is necessary to specify and unify the structure of each sample element. The values for the presence or absence of pathology are set for each sample element. To achieve this, we denote the elements P^i as follows:

$$P^i = \{p_1^i, p_2^i, p_3^i, \dots, p_z^i, \dots, p_m^i\} \quad (8)$$

where $z = \overline{1, M}$.

For each element of the set C^i obtained in step 2.1, we compare the set P^i using the operator F_p . The order of pathologies should be the same P^i for everyone:

$$F_p: C \rightarrow P \quad (9)$$

Let us denote the size of the set P after applying the operator F_p with the symbol m .

In a particular case, 1 is assigned to each element of the set P^i when a pathology is detected, and 0 in its absence. For example, $P^i = \{0, 1, 0, 1\}$, means: absence of heart pathology, presence of pathology of the left kidney, absence of liver pathology, presence of pathology of the spleen. Separately, we can note the case in which the value of the elements in the set can also be a symbol, meaning that there is no data on pathology (for example, if there are no liver data for a patient number k , then the third element p_3^k of the set P^k will be equal to this symbol. In the future, such a symbol will be designated as follows: “-”).

The order of the elements in the set, their value and number, as well as the function itself, depend on the task and the selected machine-learning (ML) models.

Step 6. Declaring a set of tuples to identify the pathology. To accomplish this, we form the matrix M_B^j , in which a row is responsible for each feature space of the set of characteristics obtained by bioelectrography in step 2.1, and the matrix M_P^j , in which the column values are assigned the presence or absence of pathology obtained in step 5. This allows you to proceed to the formation of datasets. Let us look at this step in more detail:

6.1. Defining a tuple (M_B^j, M_P^j) that consists of two elements:

(1) Matrices M_B^j of dimension (N, n) . The row i of the matrix represents the elements p_j^i , where $P^i, i = 1 \dots N, j = 1 \dots m$.

6.2. Formation of a set O consisting of tuples defined in step 6.1:

$$O = \left\{ (M_B^1, M_P^1), (M_B^2, M_P^2), (M_B^3, M_P^3), \dots, (M_B^m, M_P^m) \right\} \quad (10)$$

At the moment, all the matrices M_B^j are the same. That is, the following is performed: $\forall j = 1 \dots m, \forall k = 1 \dots m M_B^j = M_B^k$. An example of matrices M_B^j and M_P^j :

$$M_B^j = \begin{pmatrix} b_1^1 & \dots & b_n^1 \\ \vdots & \ddots & \vdots \\ b_1^N & \dots & b_n^N \end{pmatrix}, \quad M_P^j = \begin{pmatrix} p_j^1 \\ \vdots \\ p_j^N \end{pmatrix} \quad (11)$$

6.3. Formation of a set of operations on the matrices. For example, the operation $M[i]$ will mean taking the i -th row from the matrix $M_B^j[i] = (b_1^i, b_2^i, b_3^i, \dots, b_n^i)$. The entry $\text{size}(M_P^j)$ means an action that will output the number of rows in the matrix M_B^j .

Step 7. Delete irrelevant data and outliers. The initial data do not contain information about all the pathologies for each patient. For example, one group of patients has information about thyroid pathology, while the other does not. This step is necessary to remove irrelevant pathology data from patients who do not have information about the pathology in question. Each element p_j^i where $i = 1 \dots N, j = 1 \dots m$ can also have a symbol corresponding to the fact that there are no data ("–"). In this case, such data must be deleted. That is:

$$\forall i = 1 \dots N, \forall j = 1 \dots m \text{ if } p_j^i = \text{"–"} \text{ then delete } M_B^j[i], M_P^j[i]$$

It is also possible to delete explicit outliers or erroneous data.

Step 8. Refinement of the characteristic data space of type 1, taking into account data of type 2.

The matrix M_P^j corresponds to a set of pathologies of the internal organs. At this step, if necessary, you need to reduce the character space of the lines M_B^j but take into account M_P^j . Let us introduce a set of operators: F_{rep}

$$F_{\text{rep}} = \left\{ F_{\text{rep}}^1, F_{\text{rep}}^2, F_{\text{rep}}^3, \dots, F_{\text{rep}}^m \right\} \quad (12)$$

Each operator is responsible for its own tuple of the set O :

$$F_{\text{rep}}^j: (M_B^j[i], M_P^j[i]) \rightarrow M_B^j[i], i = 1 \dots N, j = 1 \dots m \quad (13)$$

After applying the operator F_{rep}^j , the number of elements in the row $M_B^j[i]$ will decrease. It should be noted F_{rep}^j that it reduces the same elements for everyone $i = 1 \dots N$.

For example, M_P^3 may be "responsible" for liver pathology. Based on this, it is necessary to reduce the feature space of the matrix rows M_B^3 , taking into account the fact that these data will reveal liver pathology.

Step 9. Forming data samples. At this step, for further training of the ML models, it is necessary to form 3 samples with data: training, validation and test. To accomplish this, for each M_p^j and M_p^j , $j = 1 \dots m$, three disjoint lists of numbers from 1 to $\text{size}(M_p^j)$ are formed in such a way that the number of elements in the list when they are combined is equal $\text{size}(M_p^j)$. These 3 sets will be denoted by:

- set Z_{train}^j —set indexes of the training sample, the number of elements is approximately 50–80% of $\text{size}(M_p^j)$;
- set Z_{val}^j —set indexes of the validation sample, the number of elements is approximately 10–25% of $\text{size}(M_p^j)$;
- set Z_{test}^j —set indexes of the test sample, the number of elements is approximately 10–25% of $\text{size}(M_p^j)$.

The number of elements in the set varies depending on the number N .

Step 10. Creating a set of synthetic data. If there are little data (equivalent to the fact that Z_{train}^j indicates there are few elements), then the process of synthesizing new records and augmentation is carried out. It is important to note that only the matrix strings $\{M_B^j[i] \mid i \in Z_{\text{train}}^j\}$ can be synthesized and augmented. The choice of synthesis and augmentation method depends on the field of work, data, and ML models.

To achieve this, we introduce an operator F_{syn} that will form a new tuple along the rows of the matrix:

$$F_{\text{syn}}: \{(M_B^j[i], M_P^j[i]) \mid i \in Z_{\text{train}}^j\} \rightarrow (M_{\text{Bsyn}}^j, M_{\text{Psyn}}^j) \quad (14)$$

We assign a new index equal to $N + 1$ to such a tuple and add it to the set Z_{train}^j , and then we increment it N . Then, we add a row M_{Bsyn}^j to the matrix M_B^j , and a row M_{Psyn}^j to the matrix M_P^j .

The synthesis of new data can also help to cope with the problem of imbalance. There are two ways to solve the problem here: to make the major and minor classes equal, or to make the minor class major.

Step 11. Training of ML models based on the data obtained in steps 9–10. At this step, a general list of ML models with different parameters is compiled. Each of these models will be further trained and validated. At this step, you need to create many models of ML:

$$ML = \{v_1, v_2, v_3, \dots, v_L\} \quad (15)$$

where v_L is the machine-learning model. The choice of the model depends on the subject area and the data. At this stage, you also need to decide on the type of metrics (precision, recall, f1-score or another). After composing the set ML , it is advisable for each element of the set O (10) to train each model from (15) on $\{(M_B^j[i], M_P^j[i]) \mid i \in Z_{\text{train}}^j\}$, selecting hyperparameters on $\{(M_B^j[i], M_P^j[i]) \mid i \in Z_{\text{test}}^j\}$, with accuracy checking on $\{(M_B^j[i], M_P^j[i]) \mid i \in Z_{\text{test}}^j\}$. Based on these results, it is advisable to choose the best model for each element of the set O based on the selected metric.

Step 12. Refinement of the set of applied models. Here, from the entire set of models obtained in the previous step, those that are the most informative for each type of pathology are selected. This allows you to form an ensemble of models with the pathology necessary for analysis. After choosing the best models, they must be combined into an ensemble V :

$$V = \{v_1, v_2, v_3, \dots, v_m\}$$

where each ML , v_i model, $j = 1 \dots m$ is responsible for a certain pathology.

Now, when receiving new data of type 1, they can be transferred to the ensemble V , which in turn forms a vector P :

$$P = \{p_1^{\text{pred}}, p_2^{\text{pred}}, p_3^{\text{pred}}, \dots, p_m^{\text{pred}}\} \quad (16)$$

where each element p_j^{pred} , $j = 1 \dots m$, of the vector P describes the presence of pathology. For example, if the vector p_3^{pred} is 1, then this may mean, for example, liver pathology.

4. Software Package for Detecting Pathology of Internal Organs

To identify the pathology of the internal organs, a software package was formed (Figure 3), which includes a module with the implementation of gradient boosting: “xgboost”; libraries for ML “scikit-learn”; “pandas” and “NumPy”; “imblearn”; and “pyplot”. The following models were chosen to compose the ensemble: decision tree; KNN; logistic regression; random forest.

Logistic regression is a machine-learning method used to predict the probability of an observed object belonging to a certain class. The goal of logistic regression is to find the optimal hyperplane that separates objects of two classes as accurately as possible based on the values of their features. This hyperplane can be used to predict the probability of an object belonging to one of the classes.

To train the logistic regression model, the maximum likelihood method is used, which allows you to find such values of the model parameters at which the probability of the observed data will be the maximum.

An important advantage of logistic regression is the ability to interpret the results. The coefficients of the model, which determine the contribution of each feature to the prediction of a class, can be used to determine which features are most important for classifying objects.

A **decision tree** is a machine-learning method that is a tree in which each node represents a decision when choosing a specific feature to divide data into smaller groups.

The process of building a tree begins with the root node, which contains all the available data. The tree is then divided into two or more branches based on the value of the selected attribute to divide the data into smaller groups. Each of these new nodes can also be divided into smaller groups using a different attribute, and so on until the specified stopping criteria are reached. The result of the tree is rules describing how the data should be divided into groups in order to obtain the best prediction result for the target variable.

The decision tree has the property of a simple interpretation of the received rules.

Random forest is a machine-learning algorithm used for tasks such as classification or regression. It is an ensemble of decision trees, that is, it consists of several trees, each of which solves a classification or regression problem for the input data.

When classifying data, a random forest uses majority voting—each tree “votes” for a certain class, and the final result is chosen by majority vote.

Gradient boosting is a machine-learning method used to solve classification and regression problems. It is based on the idea of combining several weak learning models, such as decision trees, to create a stronger model.

Unlike a random forest, where each tree is created independently of the others, in gradient boosting, each new tree is created by taking into account the errors of previous trees. The gradient-boosting algorithm builds a new decision tree at each iteration, which takes into account the errors of the previous model.

One of the most popular and effective implementations of the gradient-boosting algorithm is XGBoost.

The **support vector machine** (SVM) method is a machine-learning method that is used for both classification and regression tasks. It is based on the construction of an optimal hyperplane that maximally separates the data of two classes in a multidimensional feature space. That is, the separating hyperplane is located at the maximum distance from the classes in the feature space.

In addition, a SVM can work with non-linearly separable data by using cores that project data into a higher-dimensional space.

HyperTab is a classifier for small datasets based on a hypernet. The smaller the dataset, the greater the advantage of HyperTab over other algorithms. The principle of operation of HyperTab is the generation of target networks by a hypernetwork, thus obtaining an ensemble of networks. The learning parameters are the hypernet parameters. To form an element of the ensemble, an augmentation mechanism is used for a subset of points in the feature space, which allows you to actually expand the original dataset. Augmentation in HyperTab terminology refers to the imposition of a mask on the features of an element.

The individual software modules included in the complex allow the following:

- Download module—download from the program where the GRV-grams and their description for each patient are stored;
- A data parser with a description of the GRV-grams—convert data describing the GRV-grams into a feature space (2);
- A script replacing the full name of patients with an ID—assign a unique ID to each patient and organize a repository with records about it according to (3).
- A search script and a conclusion parser—search throughout the array of documents with the conclusions of ultrasound diagnostics specialists are those that are necessary for training ML models according to (4);
- Automatic filling script—add columns of pathologies for each organ to the table with patient IDs and their conclusions. These columns have a Boolean value: “yes” or “no”. “Yes” means that there is a pathology (8).
- The unifying script is to combine tables, forming all the data into one view, in which the columns contain: patient IDs, descriptions of their GRV-grams and pathologies (11). Machine-learning models can work with such a table in the future.

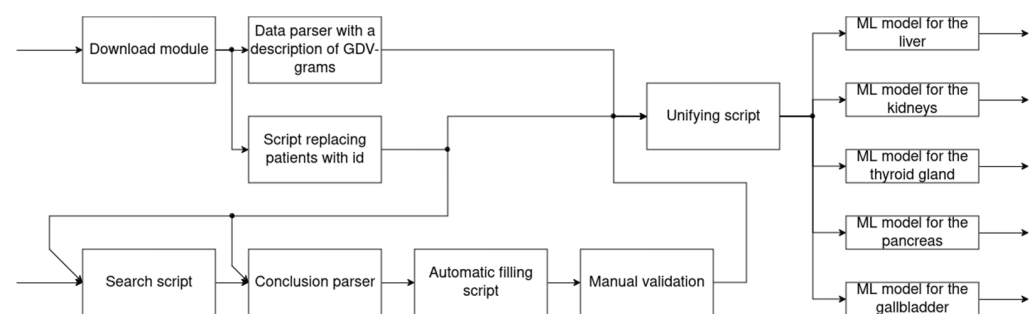


Figure 3. Software for detecting the pathology of internal organs.

Figure 3 shows the result of the obtained vector (16), formed at step 12, defining the entire set of pathologies, where each pathology has its own place in the vector. Furthermore, the characteristic responsible for a certain pathology is extracted from this vector, and a decision is made on the presence or absence of it.

The chosen scripts and modules of the software package provide automation and standardization of the data processing and analysis process, which simplifies the work of specialists and increases the efficiency of their activities. It also allows you to significantly reduce the time spent on data preparation and analysis, which is especially important in conditions of intensive medical practice.

5. Experimental Results and Analysis

The table obtained at the last step of the combining script (Figure 3) allows you to proceed to the selection of machine-learning models. To train machine-learning models, a dataset was compiled based on a sample of 170 patients' diseases. The data were provided by the medical center, and the sample included 130 women and 40 men. The distribution of the pathologies of the internal organs is shown in Figure 4.

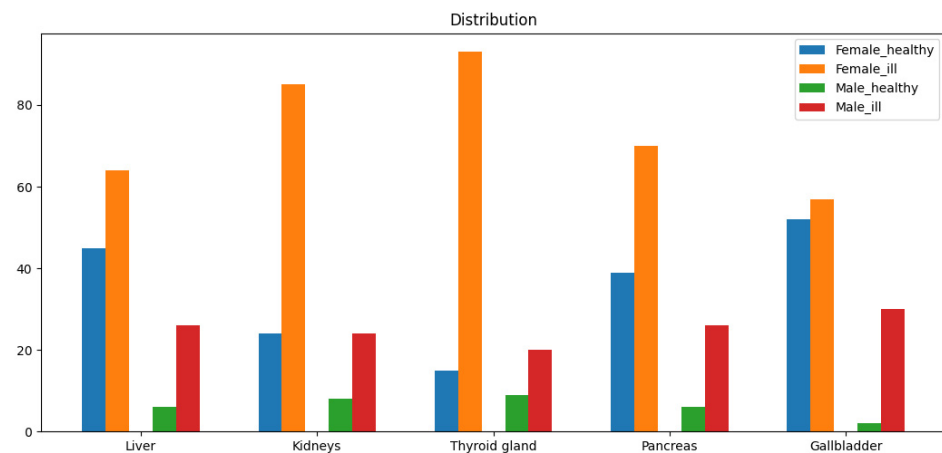


Figure 4. Histogram of the distribution of the pathologies of the internal organs.

Several conclusions can be drawn from the histogram. The data are highly unbalanced. There are more patients with pathology than healthy ones, and significantly so for the following organs: liver, kidneys, thyroid gland, pancreas and gallbladder (Table 2 shows a numerical description of the distribution of patients). Such an imbalance must be taken into account when training ensemble models. The data describing the spleen is critically unbalanced. Pathology is not observed in a large majority of patients (which is true). With such a strong imbalance, it is impossible to train the model, and therefore, the spleen data will not be taken into account in the future.

Table 2. Distribution of diseases (imbalance).

	Liver	Kidneys	Thyroid Gland	Pancreas	The Gallbladder
Healthy	51	32	24	45	54
Pathology	90	109	113	96	87
Total relevant data	141	141	137	141	141
There is no data	29	29	33	29	29

Since each object in the feature space is characterized by over 1000 features, it is obvious that there is a need to reduce the feature space. In general, data modification can be divided into three directions: leave the dataset unchanged; reduce the feature space; synthesize new data. We would like to note right away that with regard to the validation and test datasets, there can be no question of synthesizing new objects in the feature space. Therefore, the third direction was applied exclusively to the training dataset. The dataset was processed using StandardScaler from scikit-learn 1.4.2 (<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>, accessed on 10 February 2024). To reduce the feature space, the methods of factor analysis and principal component analysis were used. When synthesizing new data, the following strategies were used: alignment of the major and minor classes; change of the minor class to the major one. The second strategy is used in some cases, as presented in [20]. Weights were also used to combat the unbalance when training models and selecting metrics. The distribution of data in the training, validation and test datasets is shown in Table 3.

From the table, you can see that the imbalance of classes in the datasets for the gallbladder and liver is minimal compared to the rest. This means that it will be easier for machine-learning models to learn from them. It should be noted that with the thyroid, the situation is the opposite—there is an imbalance of classes in the dataset, which is why the machine-learning model can very quickly choose a strategy to always predict the pathology class.

Table 3. Distribution of survey results by dataset.

Dataset	Pathology	Liver	Kidneys	Thyroid Gland	Pancreas	The Gallbladder
Training (60%)	Healthy	31	19	14	27	32
	Pathology	54	65	68	58	52
Validation (15%)	Healthy	8	5	4	7	8
	Pathology	14	16	17	14	13
Test (25%)	Healthy	12	8	6	11	14
	Pathology	22	28	28	24	22

The resulting dataset made it possible to train machine-learning models to identify the pathology of internal organs, particularly the pathology of the gastrointestinal tract, kidneys and thyroid gland and spleen. The following classifiers of machine-learning methods were selected for the study: decision tree; KNN; logistic regression; random forest; SVC; XGBoost; HyperTab. The learning outcomes are presented in Table 4.

At the training stage, the dataset was split randomly. Also, during the splitting, preprocessing was performed, which included alignment, normalization, and reduction of the feature space. This made it possible to create several different datasets. Cross-validation performed with the parameter $k = 5$ was used, but it showed approximately the same results as its absence. This made it possible to decide to leave the division into three subsamples, in which the training and validation were changed, and the test was fixed. The training was conducted in several rounds. It was not possible to identify a universal model for all types of pathology.

Different models predict good results for each individual pathology. In the results were selected models that identified each pathology, the data on which are summarized in Table 4.

Table 4. Learning outcomes of an ensemble of machine-learning models.

Model	Metrics					Confusion Matrix
	Organ	Precision	Recall	F1-Score	Type of Feature	
HyperTab	Liver	0.53	0.67	0.59	Pathology is not revealed	[[8 4] [7 15]]
		0.79	0.68	0.73	Pathology revealed	
		0.66	0.67	0.66	Average	
		0.70	0.68	0.68	Weighted	
Logistic Regression	Kidneys	0.67	0.50	0.57	Pathology is not revealed	[[4 4] [2 26]]
		0.87	0.93	0.90	Pathology revealed	
		0.77	0.71	0.73	Average	
		0.82	0.83	0.82	Weighted	
HyperTab	Thyroid gland	0.75	0.50	0.60	Pathology is not revealed	[[3 3] [1 27]]
		0.90	0.96	0.93	Pathology revealed	
		0.82	0.73	0.77	Average	
		0.87	0.88	0.87	Weighted	
XGB Classifier	Gallbladder	0.64	0.64	0.64	Pathology is not revealed	[[9 5] [5 17]]
		0.77	0.77	0.77	Pathology revealed	
		0.71	0.71	0.71	Average	
		0.72	0.72	0.72	Weighted	
Decision Tree	Pancreas	0.47	0.82	0.60	Pathology is not revealed	[[9 2] [10 14]]
		0.88	0.58	0.70	Pathology revealed	
		0.67	0.70	0.65	Average	
		0.75	0.66	0.67	Weighted	

From Table 3, we can conclude that the presented models have good values for both precision and recall. The best results are shown by models where the class imbalance is the smallest. In the thyroid gland, where the class imbalance is high, the recall metric is unsatisfactory.

The KNN classifier and SVM were used, but they were excluded from further analysis because they did not show the best results in detecting pathology. Together with the fact that the training of these models is performed over a sufficiently long time interval, it did not make sense to use them in the future. The GDA and Naïve classifiers were also excluded from consideration under the assumption that the features that describe the objects of the dataset are related.

The effectiveness of the training was assessed using the constructed histograms of the distribution of diseases and metrics: precision, recall, f1-score. In the process of predicting the model, an error matrix (confusion matrix) was compiled, an example of which is shown in Figure 5, where TP is the number of correctly predicted elements of class 0; TN is the number of correctly predicted elements of class 1; and FP is the number of elements that the model assigned to class 0, but this element actually belongs to Class 1. FN is the number of elements that the model assigned to Class 1, but this element actually belongs to class 0.

	Class 0	Class 1	
True Class	TP	FN	Class 0
	FP	TN	Class 1
	Predicted Class		

Figure 5. Confusion matrix.

The error matrix expands if there are more than two classes. The effectiveness of the training can be assessed using metrics: precision, recall, f1-score. The metrics were calculated using the following formulas:

$$\text{precision} = \frac{TP}{TP + FP},$$

$$\text{recall} = \frac{TP}{TP + FN},$$

$$f1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The effectiveness of the training was assessed using the constructed histograms of the distribution of diseases and metrics: precision, recall, f1-score. The average value of the metrics (the third row of values in Table 4) was considered as an arithmetic mean and weighted as half of the sum of the metrics for individual classes multiplied by the coefficients depending on the unbalance of classes.

6. Discussion

Analyzing the data obtained, we can say that the percentage of errors is lower in those models where the distribution of the number of patients with pathology and healthy patients approaches normal.

The use of the HyperTab classifier has shown the possibility of its application for small tabular datasets based on a hypernet. To form an element of the HyperTab ensemble, the

augmentation mechanism of a subset of the feature space points was used. Augmentation in HyperTab terminology refers to the imposition of a mask on the attributes of an element. This model has shown good results, which suggests the possibility of using neural networks where the input data will be a GRV-gram image.

However, it must be borne in mind that learning and using machine learning and neural network methods require large computational resources, and they also have weak interpretability. In this regard, this method can only be used if there are enough datasets: both the GRV-gram and the target attribute for each object in the sample. This requirement applies to every pathology, since data preprocessing and machine-learning models are used differently from case to case. For example, the analysis of the data in Table 2 showed the possibility of using this method with restrictions on the ratio of the number of patients with pathology of the internal organs and patients without such pathology. It can be seen from the table that there is an imbalance in the data that is due to the fact that there are too many patients with pathologies in the sample in relation to healthy ones. Models that were trained on those datasets where there is an “imbalance” (kidneys and thyroid gland) poorly identify a class of patients who do not have this pathology. Despite this, the application of the method has shown the possibility of its use and extension to other classes of tasks related to image processing.

The limitations of the method include non-specific limitations that are characteristic of all machine-learning methods, such as the following:

1. Necessity of large computational resources, as machine-learning methods require a significant amount of computational resources for training and data processing;
2. Poor interpretability of results when a new pathology arises that has not been previously described or is not available in the dataset.
3. The method also has specific limitations:
4. The requirement of a sufficient number of datasets of both GRV-grams and the target trait for each subject in the sample. This requirement applies for each pathology because different data preprocessing and machine-learning models are used for different types of ensembles.
5. The requirement for the balance of patients with pathologies and healthy patients in the dataset. Models that have been trained on those datasets where there is an “imbalance” are poor at identifying the class of patients who do not have the pathology.

The use of the developed method will, in practice, allow us to switch to combining various machine-learning models for the identification of certain diseases, as well as the for identification of combined pathology.

Further areas of research should include the formation of multiple zones of patients' GRV-grams, with the solution of the problem of their clustering for the possibility of determining a particular pathology or their combinations based on machine-learning models. This will allow us to move on to solving the problem of detecting pathology during screening studies, which will lead to an increase in the detection of diseases and a reduction in the burden on the staff of medical institutions.

Author Contributions: Conceptualization, R.F.; methodology, Y.S.; software, A.S.; validation, A.S. and A.K.; formal analysis, Y.S.; writing—original draft preparation R.F.; writing—review and editing, Y.S. and R.F.; visualization, A.K.; supervision, R.F.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded and supported by the Ministry of Science and Higher Education of the Russian Federation by the Agreement No. 075-15-2022-291, dated 15 April 2022, on the provision of a grant in the form of subsidies from the federal budget for the implementation of state support for the establishment and development of the world-class scientific center “Pavlov center Integrative physiology for medicine, high-tech healthcare, and stress-resilience technologies”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all the subjects involved in the study.

Data Availability Statement: There are no associated data to be shared related to this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu, B.; Moeckel, G. Application of digital pathology and machine learning in the liver, kidney and lung diseases. *J. Pathol. Inform.* **2023**, *14*, 100184. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Albahli, S.; Ahmad Hassan Yar, G.N. AI-driven deep convolutional neural networks for chest X-ray pathology identification. *J. X-Ray Sci. Technol.* **2022**, *30*, 365–376. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Arora, M.; Davis, C.M.; Gowda, N.R.; Foster, D.G.; Mondal, A.; Coopersmith, C.M.; Kamaleswaran, R. Uncertainty-Aware Convolutional Neural Network for Identifying Bilateral Opacities on Chest X-rays: A Tool to Aid Diagnosis of Acute Respiratory Distress Syndrome. *Bioengineering* **2023**, *10*, 946. [\[CrossRef\]](#)
4. Heitmann, J.; Glangetas, A.; Doenz, J. DeepBreath—Automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries. *npj Digit. Med.* **2023**, *6*, 104. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Gelman, A.; Furman, E.G.; Kalinina, N.M.; Malinin, S.V.; Furman, G.B.; Sheludko, V.S.; Sokolovsky, V.L. Computer-Aided Detection of Respiratory Sounds in Bronchial Asthma Patients Based on Machine Learning Method. *Sovrem. Tehnol. Med.* **2022**, *14*, 45. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Demirci, B.A.; Koçyiğit, Y.; Kızıllırmak, D.; Havlucu, Y. Adventitious and Normal Respiratory Sound Analysis with Machine Learning Methods. *Celal Bayar Univ. J. Sci.* **2021**, *18*, 169–180. [\[CrossRef\]](#)
7. Jiang, H.; Sun, W.-J.; Guo, H.-F.; Zeng, J.-Y.; Xue, X.; Li, S. A review of intelligent diagnosis methods of imaging gland cancer based on machine learning. *Virtual Real. Intell. Hardw.* **2023**, *5*, 293–316. [\[CrossRef\]](#)
8. Chhillar, I.; Singh, A. An Insight into Machine Learning Techniques for Cancer Detection. *J. Inst. Eng. India Ser. B* **2023**, *104*, 963–985. [\[CrossRef\]](#)
9. Ramaha, N.T.A.; Mahmood, R.M.; Hameed, A.A.; Fitriyani, N.L.; Alfian, G.; Syafrudin, M. Brain Pathology Classification of MR Images Using Machine Learning Techniques. *Computers* **2023**, *12*, 167. [\[CrossRef\]](#)
10. Aliyi, S.; Dese, K.; Raj, H. Detection of gastrointestinal tract disorders using deep learning methods from colonoscopy images and videos. *Sci. Afr.* **2023**, *20*, e01628. [\[CrossRef\]](#)
11. Obayya, M.; Al-Wesabi, F.N.; Maashi, M.; Mohamed, A.; Hamza, M.A.; Drar, S.; Yaseen, I.; Alsaied, M.I. Modified Salp Swarm Algorithm with Deep Learning Based Gastrointestinal Tract Disease Classification on Endoscopic Images. *IEEE Access* **2023**, *11*, 25959–25967. [\[CrossRef\]](#)
12. Vania, M.; Tama, B.A.; Maulahela, H.; Lim, S. Recent Advances in Applying Machine Learning and Deep Learning to Detect Upper Gastrointestinal Tract Lesions. *IEEE Access* **2023**, *11*, 66544–66567. [\[CrossRef\]](#)
13. Roy, J.; Abdel-Dayem, A. Segmentation of the gastrointestinal tract MRI using deep learning. *Int. J. Artif. Intell. Appl. (IJAAIA)* **2023**, *14*, 41–56. [\[CrossRef\]](#)
14. Radzhabov, A.G. Decision Making Support System for the Diagnostics of the Cardiovascular System Pathologies by the X-ray Images of the Chest. *Dokl. BGUIR* **2023**, *21*, 98–103. (In Russian) [\[CrossRef\]](#)
15. Korotkov, K.G. *Principles of GDV Bioelectrography Analysis*; Renome: Saint Petersburg, Russia, 2007; 286p.
16. Novickij, I.O.; Deschenya, P.A. Method of bioelectrography at the present stage. *Achiev. Sci. Educ.* **2018**, *2*, 30–32.
17. Kushwah, K.K.; Srinivasan, T.M.; Nagendra, H.R.; Ilavarasu, J.V. Development of normative data of electro photonic imaging technique for healthy population in India: A normative study. *Int. J. Yoga* **2016**, *9*, 49–56. [\[CrossRef\]](#) [\[PubMed\]](#) [\[PubMed Central\]](#)
18. Babelyuk, V.Y.; Gozhenko, A.I.; Dubkova, G.I.; Korolyshyn, T.A.; Kikhtan, V.V.; Babelyuk, N.V.; Zukow, W.; Popovych, I.L. Causal relationships between the parameters of gas discharge visualization and principal neuroendocrine factors of adaptation. *J. Phys. Educ. Sport* **2017**, *17*, 624–637.
19. Krashenyuk, A.; Korotkov, K.; Kuryleva, N. Study of the influence of diagnostic ultrasound on the human aqua-system with bio-well device. *Hum Aqua Syst.* **2017**, *9*, 5–15.
20. Yakovleva, E.G. Diagnostic capabilities of the GDV-bioelectrography method. *Bull. New Med. Technol.* **2013**, *41*.
21. Pushkin, A.S.; Kamshilin, S.A.; Drugova, E.D. Fundamentals of gas discharge imaging—A promising toxicological method. *Chem. Org. Technol.* **2019**, *1*, 66–80. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Solovevskaya, N.L.; Yanovskaya, E.E.; Yusubov, R.R.; Belisheva, N.K. Assessment of psychophysiological effects of geocosmic agents using the gas discharge imaging method. *Proc. Kola Sci. Cent. Russ. Acad. Sci.* **2019**, *10*, 125–137. [\[CrossRef\]](#)
23. Shichkina, Y.A.; Fatkueva, R.R.; Sychev, A.Y.; Prasad, M.S.; Verma, N.S. Assessment of the Applicability of Machine Learning Methods for the Detection of Pathology of Internal Human Organs based on the Results of Bioelectrography. In Proceedings of the 2023 IV International Conference on Neural Networks and Neurotechnologies (NeuroNT), Saint Petersburg, Russia, 16 June 2023; pp. 29–32. [\[CrossRef\]](#)
24. Bista, S.; Jasti, N.; Bhargav, H.; Sinha, S.; Gupta, S.; Ramarao, P.; Chaturvedi, S.K.; Gangadhar, B.N. Applications of Gas Discharge Visualization Imaging in Health and Disease: A Systematic Review. *Altern. Ther. Health Med.* **2023**, *29*, AT6764. [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.