New Formalism for Statistical Similarity Based XAI

Dmitriy Klyushin

Taras Shevchenko National University of Kyiv, Ukraine, 03680, Kyiv, Akademika Glushkova Avenue, 4D

Abstract

The paper describes a non-parametric approach to the Similarity Based XAI, based on Shepard's universal law of generalization, which states that "the probability of a response to one stimulus being generalized to another is a function of a "distance" between the two stimuli in a psychological space". In the theory of pattern recognition, the analogue of this law is the compactness postulate of Averyanov and Braverman, which states that objects of one class in a feature space, as a rule, are located closer to each other than to objects of other classes. It is common to consider objects to be similar if they are close in a feature space. Meanwhile, features of objects in real life often are random values. Such objects are described not by a feature vector, but by a random sample or several samples of features and the compactness hypothesis should be replaced by a statistical homogeneity hypothesis. Objects are considered homogeneous if their features obey the same distributions, and their similarity measures form vectors in a similarity space. The chapter describes a non-parametric measure of homogeneity and provides an illustration of their use in medical applications, in particular for the diagnosis of breast cancer in the framework of the Similarity-Based XAI. We formulate new statistical postulates of machine learning and propose to consider a machine learning algorithm explainable and interpretable if it satisfies these postulates. The proposed concepts are illustrated by an application of XAI principles for the breast cancer diagnosis.

Keywords [1]

Explainable artificial intelligence, similarity-based XAI, non-parametric statistics, machine learning postulates.

1. Introduction

The concept of explainable artificial intelligence has become the subject of intensive research, since the mechanical application of artificial intelligence does not meet the natural requirements that apply to human-machine systems. In automatic systems, one subsystem unconditionally accepts the results of the work of another subsystem, since this feature is inherent in their designs. Medical applications are of a completely different nature, since they can be schematically described as a patient–physician–artificial intelligence–doctor–patient interaction. As you can see, the role of a person in this scheme is many times greater than the role of artificial intelligence. Instead of a mechanical interaction that implies blind trust and automatism, a person dictates to artificial intelligence the properties that it should possess: trust, causality, portability, and informativeness [1]. According to [1], interpretable artificial intelligence must inspire confidence in the algorithms of its work, identify cause-and-effect relationships between initial data, intermediate results, and the final conclusion, allow application to new data, and also allow extracting new information from the constructed model. In addition to the criteria listed above, XAI classification is also used based on the purpose of its use. This classification is based on four criteria: validity of inferences, control of inferences, improvement of algorithms, and acquisition of new knowledge [2]. Currently, various artificial intelligence models are widely used both in medicine in general [3–5], and in oncology in

Information Technology and Implementation (IT&I-2023), November 20 – 21, 2023, Kyiv, Ukraine
EMAIL: dmytroklyushin@gmail.com (D. Klyushin)
ORCID: 0000-0003-4554-1049 (D. Klyushin)

particular [6, 7]. Among them, it is worth highlighting the models of explainable artificial intelligence used in the analysis of medical images and diagnostics [8–10]. The works listed above give a complete overview of modern applications of artificial and explainable artificial intelligence in medicine. At the same time, such a variety is not yet described by a unified mathematical theory that allows formalizing the process of interpreting explanations [11].

Particular attention is drawn to the cognitive aspects of the explainability of deep learning and machine learning algorithms. First, we note that in many works (see, for example, [12]) machine learning is unreasonably associated exclusively with artificial neural networks. Perhaps this is due to the fact that in the last 10 years, artificial neural networks have become mainstream, while other algorithms have fallen into the shadows. However, machine learning is much broader than learning neural networks. Secondly, all the criteria proposed for interpreting the process and conclusions that machine learning leads to in artificial neural networks boil down to attempts to answer the question: "What is in the black box?" There is a lot of cognitive dissonance at this point. If you already decided to use the black box from the very beginning, then why do you need to know what is in it if it gives the right answers? Who needs it? Does a pilotless taxi passenger need to know how his mechanics work? Obviously not. At the same time, a technician who maintains such a vehicle must know its construction, because otherwise he will not be able to set it up. However, when machine learning algorithms are built according to the black box principle, the situation is more complicated. Here, the technician gives the artificial intelligence the possibility of self-tuning, which is expressed in the independent choice of features of objects. The fact that the researcher does not build the feature space himself, but places the responsibility for it on the algorithm, leads to unpleasant consequences. First, there is an exponential growth in the number of features, because the algorithm is limited only by the computing power of the computer on which it runs, and only the author of the algorithm can set the task of optimizing and selecting only the most significant features. Secondly, the semantics of the features themselves becomes completely uninterpretable from the point of view of a person.

In [12], an attempt was made to reveal the meaning of the concepts of explainability and interpretability. Having analyzed a large amount of literature, the authors reduced the concept of explainability to answers to three questions: 1) explainability of the data; 2) explainability of the results; 3) explainability of the algorithm.

The answer to the first question is reduced to explaining what kind of data is used for training and why. In particular, it requires the formulation of a working hypothesis, which reflects the choice of data. For example, in medical applications in the field of oncology, the input of algorithms is information collected from patients with various types of cancer, as well as from healthy people. It is logical to assume that this information has a different value for diagnosis. Who and how will select the necessary information to ensure the greatest accuracy of the diagnosis? The classical approach implies the participation in this process of an expert in the subject area, who has a priori knowledge about the significance of certain data. This expert can, for example, use prior knowledge that in the body of a person with cancer, biochemical reactions occur that affect the distribution of chromatin in the nuclei of buccal epithelial cells. In this case, the expert will offer the algorithm photographs of cell nuclei taken from sick and healthy people. At the same time, a second, no less important question arises: how objective are these data? For example, if a photograph contains some kind of systematic artifact (label), then the classification apriori will be compromised. Therefore, a data validation mechanism is needed that guarantees their unbiased and random selection from the entire data set. By the way, this raises, for example, the question of the criteria for the randomness of the available data.

The second question follows from the first. If the algorithm, acting on the basis of a black box, received input data, it can independently select features that have neither biological nor physical meaning, nor any other meaning except statistical. In other words, the results to which it will lead will be correlated rather than causal. This does not affect the recognition accuracy in any way, because the algorithm strives to classify objects as accurately as possible, but the recognition process itself in this case can only be described from a statistical point of view. This makes it difficult to acquire new knowledge. Yes, we have correctly divided the two sets, but what does it give from a meaningful point of view. What features did not the human eye see in these photographs, and what is their biological or physical meaning? The modern theory of machine learning does not provide meaningful answers to these questions. As a rule, the features are some function or a combination of several functions that were chosen by the algorithm without any knowledge of the content of the photo.

The third question concerns the design of the model. In the context of artificial neural networks, it comes down to architecture analysis and layer manipulation. A variety of architectures and methods of working with them leave the impression of blind juggling. If this or that architecture has led to a successful result, it seems superfluous to explain why. At the same time, such a network can be like a house of cards that will crumble as soon as new data enters it. The ability of such algorithms to generalize is very problematic, and their stability is questionable. An interesting classification of classification methods in terms of their explainability is given in [13]. According to the authors, linear, and logistic regression, decision trees, the nearest neighbor method, rule-based algorithms, generalized additive models and Bayesian models do not require explanation, that is, by definition, they belong to explainable artificial intelligence. At the same time, the authors consider the random forest method, the support vector machine, as well as multilayered, convolutional, and recurrent neural networks, to be partially interpretable with the help of special methods.

It seems to us that such a classification of methods in the field of artificial intelligence applications in medicine is not entirely accurate. If you think about it, it becomes obvious that this classification evaluates explainability from the point of view of the researcher, that is, it makes sense only for the developer of the algorithm. At the same time, in the patient–doctor–AI–doctor–patient scheme, there are two more people who do not belong to the category of developers, but users of algorithms. A doctor wants to understand and trust an algorithm, and a patient wants to trust and understand the doctor. Therefore, explainability must be considered from their point of view too. There are both similarities and differences in these approaches.

From the point of view of data explainability, the doctor, like the developer, wants and is obliged to understand the meaning and quality of the input data. Here, their interests coincide. In addition to the increased quality of data preparation by a competent doctor who will correctly set up the mechanics with an understanding of the specifics of input data preparation (for example, set the correct illumination level, choose the right reagent, and set the required microscope resolution), a priori understanding of the input data will allow you to get feedback in time, if there is a drift of concept, due to which the algorithm may be compromised.

The explainability of the results also affects the level of confidence of the doctor and the patient in the results obtained. A doctor verifies the conclusions made by the algorithm, one way or another, using additional methods (for example, by directing the patient to be examined by other methods). If the algorithm constantly demonstrates high accuracy, then the doctor and the patient can make the right clinical decisions, rightly believing that this technique has been tested and is reliable.

At the same time, neither a doctor nor a patient should be interested in how many layers are used in the neural network architecture. This information is beyond his competence and has nothing to do with explainability from the point of view of either the doctor or the patient. Therefore, the explainability of the algorithm for the doctor and the patient is of no value.

Let’s move on to interpretability, the second aspect of XAI [12]. According to [12], interpretability is synonymous with the intelligibility of the model (algorithm and data) to the observer. In our patient–doctor–AI–doctor–patient interaction scheme, there is an implicit presence of a third person, namely, a developer of the algorithm. Consider the roles of stakeholders in accordance with the paradigm proposed in [1, 12, 13]. It is necessary to analyze how understandable a model should be every stakeholder. It is quite natural that the model should be absolutely understandable to its developer; otherwise, he will not be able to guarantee its high quality. Should it be understandable to the doctor? To answer this question, it is necessary to answer what we mean by a model in this paradigm. From our point of view, a model is a combination of an algorithm and data. The algorithm is tuned to a certain type of input data and produces a certain type of output data. This means that algorithms and data are inseparable. In this case, it must be recognized that the doctor must correctly interpret the input data, but is not obliged to delve into the structure of the algorithm. The interests of the doctor and the developer partially overlap at this point. The structure of the model is also not important for the patient (as is the construction of any medical equipment). For him, the accuracy with which the model works is important. In this case, the interpretability of a model from the patient’s point of view may increase the confidence in it. If we return to the explanability criteria listed at the beginning of the chapter, then we can agree that the interpretability of results, but not the iterpretability of a model, is important for a patient.

Closely related to the concept of interpretability is the concept of transparency, which boils down to three properties: imitability, decomposability, and algorithmic learning [1]. According to the authors of the work [14], the transparency of the model means the a priori ability of a person to understand the operation of the mechanism underlying the model. In [13], decision trees are categorized as completely transparent because the inference rules are formulated in human-readable language, and the nearest neighbor method, rule-based algorithms, generalized additive models, and Bayesian models are categorized as categories of methods, for the understanding of which mathematical knowledge is necessary. At the same time, the authors consider the random forest method, the support vector machine, as well as multilayer, convolutional, and recurrent neural networks to be completely opaque. The imitability of a model means that it is simple and easy to reproduce it on new input data in order to obtain the expected results. We have already implicitly used the possibility of decomposition above, when we divided the model into data and algorithms for their processing. Algorithmic means the comprehensibility of the learning process of the algorithm. In algorithms based on intuitive concepts (for example, the concept of proximity, the concept of linear separability, and so on), the learning algorithmic is quite high. This can be said, for example, about linear and logistic regression, kNN and SVM methods. Of course, the learning process of a neural network is understandable only to high-level specialists, and for an ignorant observer (doctors and patients), it inevitably looks like manipulation with a black box.

A detailed and extensive analysis carried out in the works [12, 13] leads to the conclusion that, at present, a formal and rigorously justified mathematical apparatus has not yet been developed that allows evaluating the explainability of machine learning algorithms. For each specific algorithm, a certain subjective assessment is given, depending on the author's point of view.

Note also remarkable papers of Rudin [15, 16]. These publications reflect an alternative and well-documented point of view on the interpretability of machine learning. Rudin makes a clear distinction between interpretable machine learning and explainable artificial intelligence. According to Rudin, interpretive machine learning is not part of the theory of explainable artificial intelligence, since the goal of this theory is to explain the black box model by approximating it with simpler and more understandable models, and the goal of interpretable machine learning is to choose an initially interpretable model that provides high accuracy [16]. According to Rudin, one should not explain the operation of the black box, but build transparent, interpretable models. This is especially important for AI medical applications, where the risk of error comes at a high cost.

However, without being able to open the black box and interfere with its operation, we can try to explore how well it meets the criteria of explainability and interpretability. This is what our work is devoted to, in which we show, what criteria can be used to evaluate the interpretability of a black box, if one had to apply it, and how to build a transparent Rudin-interpretable model for diagnosing breast cancer that also meets these criteria. Thus, we propose a compromise between these two alternatives.

As we will show below, the construction of a formal XAI theory based on a rigorous mathematical approach leads to the conclusion that the explanation of the work and conclusions of artificial intelligence is impossible without the application of machine learning postulates. The goal of this chapter is to propose a new XAI formalism based on alternative statistical postulates of machine learning.

2. Formalism

Confidence, causality, portability, and informativeness are essential properties of any machine learning method. To prove this thesis, consider the general scheme of any machine learning method.

Let X be a set of objects, Y be a set of class lables, and $f\colon X \to Y$ be a function attaining values at elements of a training sample drawn from X. The goal of the algorithm learning it to extent $f\colon X \to Y$ on all the set X to construct a solving function $g\colon X \to Y$ mapping all objects to their lables and mimimizing a risk function (for example, an error).

Since the very definition of a machine learning problem includes minimizing the error and specifying the function on the entire set X, accuracy and portability (ability to generalize) are guaranteed by default. At the same time, the concepts of causality and informativeness are of an informal nature, which is difficult to describe with the help of a mathematical apparatus. However, difficult does not mean impossible.

One of the most meaningful mathematical formalisms of interpretable artificial intelligence was proposed in [11]. This formalism is based on the Bayesian approach and uses the concept of explanee's inference and Sheppard's universal law, which states that the probability of generalizing a response from one stimulus to another is a function of the similarity between two stimuli in psychological space. Introducing a metric space instead of a psychological space of proximity and interpreting the probability of generalizing a response to a new stimulus as the probability of recognizing a new object similar to the learning object, the authors propose a measure of the explainability of machine learning results in the form of a likelihood function. The validity of the proposed approach has been demonstrated by an experiment.

It should be noted that this formalism has a significant drawback, explained by the Bayesian nature of the entire method. It requires knowledge of the a priori probability of recognizable classes, which must be given by experts. The authors also propose to calculate the Sloman similarity function as a cosine measure between the vectors generated by the silency-map and the experts participating in the experiment. It is obvious that all this makes the methodology local and subjective. Its correctness depends on the competence of specific experts.

We propose to use a more rigorous and objective approach, based not on the subjective assessments of experts, but on objectively determined values of a measure of the object's homogeneity with other objects and its statistical depth. In short, we propose to consider the results of the work of artificial intelligence as interpretable from the point of view of a doctor if they satisfy the postulate of statistical compactness. On the other hand, we suggest that these results be considered interpretable from the patient's point of view if they allow one to determine the patient's individual risk. The accuracy of any machine learning algorithm in medical diagnostics is treated according to Fisher. Assume that the sensitivity of an algorithm is 95%. This means that out of 100 randomly selected patients, with repeated repetition of this choice, the diagnosis is correctly determined on average in 95 patients and 5 patients receive a false diagnosis. If the algorithm has a specificity of 95%, this means that out of 100 randomly selected healthy people, if this choice is repeated many times, 5% will receive the diagnosis "sick". These ratings do not answer the natural questions that every patient has: "What is the probability that I am really sick? Which group do I fall into: 95 correct diagnoses or 5 false ones?" This question can be answered using two concepts: statistical homogeneity and statistical depth. On the one hand, by comparing the patient's data with the patient's etalon, one can assess the degree of similarity between them, on the other hand, by evaluating the statistical depth of the data one can assess the typicality for a given diagnosis.

3. Statistical similarity

Machine learning is predicated on two premises, both of which we have previously mentioned implicitly above: The first postulate is that an object should be represented as a feature vector in a feature vector space, and the second is the Averyanov and Braverman postulate on compactness [17]. The first postulate represents machine learning experts' innate inclination to apply the tools of algebra, geometry, and optimization techniques. This postulate essentially enables us to reduce the machine learning problem to an optimization problem, that is, to a problem of minimization or maximizing of a certain function under particular constraints. The second premise is as logical but less clear. He contends that feature vectors of things in the same class are situated closer to one another in a feature space than are feature vectors of objects in a different class. The requirement that these sets of vectors in a feature space be divided by a reasonably straightforward function is frequently included with this postulate. Fisher's linear discriminant methods, the support vector method, and the closest neighbor approach are three prominent examples of techniques developed on the foundation of these postulates.

Despite the obvious success of the methods listed above, it cannot be denied that the postulates of vector space and compactness in the form formulated above are not applicable to all problems. In many medical and biological studies, the patient does not correspond to a vector of features, that is, an ordered set of numbers characterizing his various properties, but a random sample, that is, an unordered set of random results of measurements of certain parameters (for example, nuclear area, optical density of nuclei and the like). Such samples arise, for example, when analyzing probes taken from a patient. These probes typically contain several tens of cells, so the patient does not correspond to a single point in the vector space, but rather to a cloud of points that are not in any way arranged in any particular order. It is true that this process can be made simpler by computing the average sample values and using the accepted postulates, but it is also clear that in this scenario, a significant portion of the information regarding the distribution of the measured parameters is lost. Imagine you were given no further information save the fact that the median of a conventional Gaussian distribution, even an empirical one, is zero.

Alternative statistical hypotheses that we put forth include: 1) that things can be represented by sample parameter values, 2) that parameters of objects within the same class have similar distributions, and that parameters of objects within different classes have different distributions. With the use of this strategy, we can reduce the challenge of determining how similar two things are to verifying whether two or more samples are homogeneous. The approach we suggest using to determine how similar two things are is shown below. It has statistical universality, which means that it performs equally well for samples with different means and the same standard deviations as it does for samples with the same means but different standard deviations, unlike traditional methods like Kolmogorov-Smirnov statistics and Mann-Whitney-Wilcoxon statistics.

Take into account a sample of size n made up of continuous random variables coming from a swappable distribution. The likelihood that a random value from the same distribution is larger than i-th and less than j-th order statistics of the sample equals (j-i)/(n+1), according to the Hill's assumption [18, 19]. As can be seen, the only variables that affect this probability are the sample size and the order number of the order statistics. This information allows us to examine the homogeneity of the two samples hypothesis. To do this, we order the first sample in ascending order to acquire its order statistics, and then we count the relative frequency of the occurrence of an element from the second sample having an order statistic that is more than the i-th and less than the j-th order statistics of the first samples. With these relative frequencies, we can create a confidence interval (for instance, the Wilson interval) for the binomial proportion in the generalized Bernoulli scheme under consideration. The coverage of $(j–i)/(n+1)$ by this confidence interval is then tested. Thus, we can calculate the so-called p-statistics by measuring the relative frequency of this event. Finally, using a predetermined significance threshold, we create a confidence interval for the p-statistics. If a confidence interval built on these samples does not encompass $1–\alpha$, we reject the null hypothesis regarding homogeneity [20].

4. Statistical depth

For a detailed review of the various concepts of statistical depth, see [21]. Their goal is to order multidimensional random variables.

Consider some distribution F in R^n. A depth function is a function $D_F(x)$ that orders points x from distribution F monotonically decreasing from center to outward. A depth of $x \in F$ is a value attained by $D_F(x)$ at x [22]. A center of a distribution may be the median, centroid, geometric center of distribution, etc.). A depth function must have the following properties [22]: 1) a depth function is independent from the used coordinate system and affine transformations (affine invariant); 2) a depth function attains the maximum at a center of distribution (maximum at most deep point); 3) a depth function monotonically decreases from the most deep point to the least deep points (monotonicity); 4) when a distance from a point x from a distribution center tends to infinity $D_F(x)$, it must tend to zero (limit property).

When we have no information about a distribution F but have a sample containing *n* points from F , we shall use notation $D_n(x)$. Let us consider some examples of a depth function.

Tukey depth [23]. First, introduce some necessary concepts. A center of a sample is such a point that every hyperplane passing through it divides the sample into two almost equal subsets. When this point is an element of the sample, it is the median of the sample. The Tukey depth of a sample element x is the minimum number of sample elements lying on one side of random hyperplane passing through it.

Convex hulls peeling [24]. The convex hull of a set of points is the minimum polygon containing the given points. Convex hull peeling is a procedure of consequent finding and removing enclosed convex hulls. Vertices of the same convex hull have the same statistical depth.

Oja depth [25]. The Oja depth of a sample element x is the average simplex volume based on d random sample points and x.

The simplex depth [26]. The simplex depth of a sample element x is a number of simplexes based on a random sample of points and containing x.

Zonoid depth [27]. The zonoid depth of a sample element x is the number $d(x\,|\,x_1,...,x_n)=\sup\{\alpha : y \in D_\alpha(x_1,...,x_n)\}$, where

$$D_\alpha(x_1,...,x_n)=\left\{\sum_{i=1}^{n}\lambda_i x_i : \sum_{i=1}^{n}\lambda_i = 1,\ 0 \le \lambda_i, \forall i : \alpha\lambda_i \le \frac{1}{n}\right\}.$$

Mahalanobis depth [22]. The Mahalanobis distance is a generalization of Mahalanobis distance and is defined by the formula $MHD_F(x)=(1+\mathrm{d}^2(x,E(F)))^{-1}$, where $d^2(x,y)=(x-y)^T\Sigma_F^{-1}(x-y)$, $E(F)$ is the distribution expectation, amd Σ_F is the covariance matrix.

Elliptical statistical depth [28]. Elliptical statistical depth is a function that maps points of sample to increasing ranks using the confidence Petunin ellipsoids [29]. These ellipsoids are concentric and cover a sample. Thus, we have a sequence of ellipsoids $E_1 \subset E_2 \subset ... \subset E_n$. Every sample point lies on a surface of only one ellipsoid, and the probability that a random point from F lies in E_n is $\frac{n-1}{n+1}$. Thus, the elliptical statistical depth is a monotonous function that attains a maximum at the deepest point and decrease from the center to outward.

Depth-ordered regions [30] is a set of points where the statistical depth is greater or equal to a given value $D_\alpha(F)=\{x \in R^d : D_F(x) \ge \alpha\}$, where $D_F(x)$ is a statistical depth of the point x obeing F. Depth-ordered regions are affine equivariant, nested, monotonical, compact, and subadditive. Obviously, the Petunin ellipsoids are depth-ordered regions.

Consider is a set of random points $X=\{x_1,...,x_n\}$, $x_i \in \square^d$. Let us divide the description of the algorithm for construction of Petunin's ellipsoids into two cases: two-dimensional $d = 2$ and multidimensional $d > 2$.

Petunin's ellipses. Find a convex hull of $X=\{(x_1,y_1),...,(x_n,y_n)\}$ and a diameter of this convex hull with ends (x_k,y_k) and (x_l,y_l). Connect these point by a segment L. Find points (x_r,y_r) and (x_q,y_q) that are most distant from L. Find segments L_1 and L_2 passing through (x_r,y_r) and (x_q,y_q) parallel to L. Find segments L_3 and L_4 passing through (x_k,y_k) and (x_l,y_l) orthogonal to L and. Segments L_1, L_2, L_3 and L_4 are sides of a rectangle Π. Let us denote by a a short side and by b a long side.

Translate, rotate and shrink Π with a coefficient $\alpha=\frac{a}{b}$ to obtain a square Π' with a center (x_0',y_0'). The random points $X=\{(x_1,y_1),...,(x_n,y_n)\}$ are mapped to points (x_1',y_1'), (x_2',y_2'), ..., $(x_n',y_n') \in \Pi'$. Find distances $r_1,r_2,...,r_n$ between (x_0',y_0') and (x_1',y_1'), (x_2',y_2'), ..., $(x_n',y_n') \in \Pi'$. Find $R=\max(r_1,r_2,...,r_n)$. Consider a circle C with the center (x_0',y_0') and radius R containing (x_1',y_1'), (x_2',y_2'), ..., (x_n',y_n'). Perform inverse transformations of C. As a result, we obtain an ellipse E containing points $X=\{(x_1,y_1),...,(x_n,y_n)\}$.

Petunin's ellipsoids. Find a convex hull of $X=\{x_1,...,x_n\}$, $x_i \in \square^d$. Find a diameter of the convex hull with ends (x_k,y_k) and (x_l,y_l). Rotate and translate the diameter aligning it along to Ox_1'. Project (x_1',y_1'), (x_2',y_2'), ..., (x_n',y_n') to the orthogonal complement of Ox_1'. Find a convex hull of projections, rotate and translate it up to a two-dimensional rectangle Π. Construct a minimum volume axisaligned parallelogram in *d*-dimensional space containing the projections of input points. Shrink this parallelogram to hypercube. Find its center x_0 and distances $r_1, r_2, ..., r_n$ from x_0 to $x'_1, ..., x'_n$. Find $R = \max(r_1, r_2, ..., r_n)$. Consider a hypersphere with the center x_0 and radius R. Perform inverse transformations. As a result, we obtain the Petunin's ellipsoid covering $X = \{x_1, ..., x_n\}$.

The Petunin's ellipsoids have remarkable properties. They uniquely arrange random multivariate points according their statistical depth due to the fact that the surface of Petunin's ellipsoid contains only one point from the initial set. Moreover, the probability that a Petunin's ellipsoid covers random points obeying the same distribution is equal to $\frac{n-1}{n+1}$. Therefore, we can find most and least probable points of a sample.

The most difficult problem of the construction of the Petunin's ellipses and ellipsoid in a space of high dimension is the finding of two most distant points (ends of the convex hull diameter). at the plane this is not a difficult problem because the computational complexity of the optimal algorithms in this case is *O*(*n* ln *n*). When в *d* > 2, the computational complexity of the construction of a convex hull in $\square^{d}$ is $O\left(n^{\left\lfloor \frac{n}{2} \right\rfloor}\right)$. However, using effective and fast procedures [2], [5] we can solve this problem.

5. Universality

We provide empirical results of numerical experiments with samples of various sizes from Gaussian distribution with various parameters for estimation the sensitivity and specificity of the tests. We generated 100 pairs of samples that each contained 10, 20, 30, 40, and 100 random numbers with the same mean and different standard deviations, as well as different means and the same standard deviation. Then, we used both tests to empirically evaluate their specificity and sensitivity. The frequency of the occurrence when the p-value for various distributions is less than 0.05 is the Wilcoxon signed-rank test's sensitivity. The frequency of the occurrence where the upper confidence bound for the p-statistics is greater than 0.95 for identical distributions is the Klyushin-Petunin test's specificity. The frequency of the occurrence when the p-value is less than 0.05 for identical distributions is the specificity of the Wilcoxon signed-rank test. Table 1 presents the outcomes.

Table 1
Sensitivity and specificity of the Klyushin–Petunin and the Wilcoxon signed-rank tests

	Sensitivity				Specificity	
	N(0,1) vc N(0,5, 1)		N(0,1) vc N(0, 2)		N(0,1) vc N(0, 1)	
Size of sample	Klyushin-Petunin	Wilcoxon	Klyushin-Petunin	Wilcoxon	Klyushin-Petunin	Wilcoxon
10	0.00	0.10	0.03	0.12	0.98	0.00
20	0.53	0.42	0.30	0.20	0.91	0.00
30	0.75	0.57	0.67	0.17	0.95	0.31
40	0.94	0.62	0.80	0.15	0.90	0.00
100	1.00	0.91	1.00	0.14	0.96	0.00

As we see, when the samples are drawn from the distributions N(0,1) and N(0.5,1) the sensitivity of the Klyushin–Petunin test and the Wilcoxon signed-rank tests depend on the sample size (see Table 1). Both tests fail for the small sample (*n*=10) and have the high sensitivity when sample size is more that 40, but the sensitivity of the Klyushin–Petunin tests exceeds the sensitivity of the Wilcoxon signed-rank test.

The Klyushin-Petunin test is successful when the sample size is more than 40 when the compared samples are selected from the distributions N(0,1) and N(0,2). The Wilcoxon signed-rank test, meanwhile, is always invalid. This impact may be explained by the fact that the Klyushin-Petunin test validates the overall hypothesis of equivalence of the distribution functions whereas the Wilcoxon signed-rank test validates the claim concerning equality of means.

For estimation of specificity of the Klyushin–Petunin test and the Wilcoxon signed-rank test we again used 30 pairs of samples containing 10, 20, 30, 40 and 100 random numbers from the normal distribution N(0,1) (see Table 3). In contrasts to the previous experiments, now the Klyushin–Petunin test demonstrates stable results in every range of the samples sizes and its specificity is comparable with the specificity of the Wilcoxon signed-rank test in every range of the sample sizes.

Thus, we may conclude that the Klyushin–Petunin test has high sensitivity when $n \geq 40$ and high specificity for all samples sizes and all possible variants of means and standard deviations. The Wilcoxon signed-rank test is valid for testing the hypothesis about distribution location shift but is invalid when different distributions have the same location and different standard deviations (hypothesis of distribution scale). Therefore, the results cited in Table 1 provide the technical validity of proposed test. It works better than the Wilcoxon signed rank test in all the hypothetical situations.

6. Conclusion

Evaluation of the criteria for explainable artificial intelligence according to existing methods is subjective. We propose an objective approach and a universal criterion of explainability for XAI. Artificial intelligence is explainable if the results of its application satisfy two statistical postulates: 1) objects can be represented by sample values of their parameters; 2) parameters of objects belonging to the same class have the same distributions, and the parameters of objects belonging to different classes have different distributions. In addition, XAI must allow for an estimate of the statistical depth of the results (for example, the individual risk in the healthycare context). As a mathematical tool for this formalism, we propose to use the Klyushin–Petunin test. It is universal due to high sensitivity and specificity when size of a sample is more than 40 (this is rational query that may be easily satisfied). Moreover, it robustly tests both hypotheses on location shift and scale shift of distribution.

The improvement of statistical techniques for evaluating the homogeneity of samples and their statistical depth, as well as the application of the suggested idea to other algorithms, are likely to be linked to the future development of the proposed methodology.

7. References

[1] Z. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, ACM Queue (2018) 16(3): 31–57, doi:10.1145/3236386.3241340.

[2] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access (2018) 6: 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[3] R,-K. Sheu, M. Pardeshi, A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System, Sensors (2022) 22, 8068. doi:10.3390/ s22208068

[4] Y. Zhang, Y. Weng, J. Lund, Applications of Explainable Artificial Intelligence in Diagnosis and Surgery, Diagnostics (2022) 12(2): 237. doi:10.3390/diagnostics12020237.

[5] J. Amann et al. To explain or not to explain? — Artificial intelligence explainability in clinical decision support systems, PLOS Digital Health (2022) 1(2), e0000016, doi:10.1371/journal.pdig.0000016

[6] W. Bi et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: A Cancer Journal for Clinicians (2019) 69(2): 127–157. doi:10.3322/caac.21552.

[7] Z. Chen et al. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine, Cancer Communications (2021) 41(11): 1100–1115. doi:10.1002/cac2.12215.

[8] K. Borys et al. Explainable AI in medical imaging: An overview for clinical practitioners — Saliency-based XAI approaches, European Journal of Radiology (2023) 162: 110787. doi:10.1016/j.ejrad.2023.110787.

[9] A. Chaddad, J. Peng, J. Xu, A. Bouridane, Survey of Explainable AI Techniques in Healthcare. Sensors (2023) 23(2): 634. doi:10.3390/s23020634.

[10] S. Nazir, D. Dickson, M. Akram, Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks, Computers in Biology and Medicine, (2023) 156:106668. doi:10.1016/j.compbiomed.2023.106668.

[11] S. Yang, T. Folke, P. Shafto, A psychological theory of explainability. In: Proceedings of the 39th International Conference on Machine Learning (2022), Baltimore, Maryland, USA, PMLR 162.

[12] P. Love et al. Explainable Artificial Intelligence (XAI): Precepts, Methods, and Opportunities for Research in Construction, arXiv:2211.06579v2 (2022). doi:10.48550/arXiv.2211.06579.

[13] A. B. Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI, Information Fusion (2022) 58: 82–115, doi:10.1016/j.inffus.2019.12.012

[14] K. Sokol, P. Flach, Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches, arXiv:1912.05100v (2019). doi:10.1145/3351095.3372870.

[15] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence (2019) 1: 206–215. doi:10.1038/s42256-019-0048-x.

[16] C. Rudin et al. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistical Surveys (2022), 16: 1–85, 2022. doi:10.1214/21-SS133.

[17] V. Mottl, O. Seredin, O. Krasotkina, Compactness Hypothesis, Potential Functions, and Rectifying Linear Space in Machine Learning. In: International Conference Commemorating the 40th Anniversary of Emmanuil Braverman's Decease, Boston, MA, USA, April 28-30, 2017, Invited Talks (2017). doi:10.1007/978-3-319-99492-5_3.

[18] B. Hill, Posterior distribution of percentiles: Bayes' theorem for sampling from a population. Journal of American Statistical Association (1968) 63: 677–691.

[19] B. Hill. De Finetti's theorem, induction, and A(n) or Bayesian nonparametric predictive inference (with discussion). In: D. V. Lindley, J. M. Bernardo, M. H. DeGroot, & A. F. M. Smith (Eds.), Bayesian statistics (1988, Vol. 3, pp. 211–241). Oxford: Oxford University Press.

[20] D. Klyushin, Yu. Petunin, A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples. Ukrainian Mathematical Journal (2003) 55(2): 181–198.

[21] K. Mosler, P. Mozharovskyi, Choosing among notions of multivariate depth statistics. Statistical Science (2022) 37(3): 348–368. doi:10.1214/21-sts827.

[22] Y. Zuo, R. Serfling, General notions of statistical depth function, Annals of Statistics (2000) 28: 461–482. doi:10.1214/aos/1016218226.

[23] J. Tukey, Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematician, Montreal, Canada, 1975, pp. 523–531.

[24] V. Barnett, The ordering of multivariate data, Journal of the Royal Statistical Society, Series A (General) (1976), 139 (3): 318–355.

[25] H. Oja, Descriptive statistics for multivariate distributions, Statistics and Probability Letters (1983) 1: 327–332. doi:10.1016/0167-7152(83)90054-8.

[26] R. J. Liu, On a notion of data depth based on random simplices, Annals of Statistics (1990) 18: 405–414.

[27] G. Koshevoy, K. Mosler, Zonoid trimming for multivariate distributions. Annals of Statistics (1997) 25: 1998–2017. doi:10.1214/aos/1069362382.

[28] S. Lyashko, D. Klyushin, V. Alexeyenko, Mulrivariate ranking using elliptical peeling. Cybernetic and Systems Analysis (2013) 49(4): 511–516. doi:10.1007/s10559-013-9536-x

[29] Yu. Petunin, B. Rublev, Pattern recognition using quadratic discriminant functions. Numerical and Applied Mathematics (1996) 80: 89–104.

[30] I. Cascos, Depth function as based of a number of observation of a random vector. Working Paper 07-29, Statistic and Econometric Series (2007) 2: 1–28.