Statistical concept of interpretable artificial intelligence

Dmytriy Klyushin[1]

[1] *Taras Shevchenko National University of Kyiv, Volodymyrska Str., 60, 01601, Kyiv, Ukraine*

Abstract

The paper describes the concept of interpretable artificial intelligence and machine learning based on the assessment of statistical homogeneity of classification objects described by features that are random variables. Objects are considered homogeneous if random values of their features have identical distributions. Mathematical theories of machine learning use two postulates: 1) the feature space is a vector space, i.e. the classified object can be represented as a vector of numbers (the vector space postulate), and 2) objects belonging to the same class form a compact set with a relatively simple boundary and the distance between them is less than to objects belonging to another class (the compactness postulate). However, in many practically essential situations, for example, in biomedical research, objects are associated not with feature vectors, but with samples of measured random variables. Therefore, we must suppose alternatives for the vector space and compactness postulates in such cases. The paper describes components of the suggested theory (measure of homogeneity, prediction set, and statistical depth).. Based on the proposed statistical postulates of machine learning, machine learning algorithms would be classified as interpretable if they comply with them.

Keywords

interpretability, explainability, statistical homogeneity

1. Introduction

Computer scientists are intensively researching the concepts of explainable and interpretable artificial intelligence. Although the black box model provides high classification accuracy, it no longer fully satisfies researchers or users regarding its interpretability [1]. The issues of trust in machine-human systems come to the fore. This fact is especially evident in medical applications with extremely high error costs. The problem of trust in the conclusions of artificial intelligence is closely related to understanding the logical mechanism. Four requirements are imposed on explainable and interpretable artificial intelligence: it should inspire trust, demonstrate logical functioning, have the property of generalization, and be able to discover new data [2]. In other words, interpretable artificial intelligence should not raise doubts about the correctness of its algorithms, should be able to identify logical cause-and-effect relationships between the original data and the final result, generalize them to new data, and generate new knowledge. Interestingly, in recent works, authors have begun to consider explainable and interpretable artificial intelligence as different entities, although previously they were considered interchangeable concepts. Interpretability is treated as the development and application of an understandable model, and explainability now means understanding the relationship between input data and the result. A typical example of an interpretable model is a decision tree, in which each step of logical inference is understandable, and, for example, a convolutional neural network is an example of an

Information Technology and Implementation (IT&I-2025), November 20-21, 2025, Kyiv, Ukraine

* Corresponding author.

dmytroklyushin@knu.ua (D. Klyushin)

0000-0003-4554-1049 (D. Klyushin)

explainable but not interpretable model, since we know that the neural network minimizes the loss function, but do not know what features it generates and includes in the model.

The concepts of explainability and interpretability have a pronounced psychological and subjective nature [3]. For example, a mathematician considers a linear regression model interpretable because he knows how it is structured and how it works. Still, a physician does not know this information and, as a result, this model will remain a black box for him. This fact creates some difficulties in perceiving the results of such systems on the part of physicians, who, for natural reasons, mistrust the conclusions of such models. To eliminate such mistrust, proposing a concept of explainable and interpretable artificial intelligence that would be a priori understandable to both specialists and non-specialists in computer science is necessary. Such a concept should appeal not to mathematical competence, but to common sense, which all thinking beings have. We believe the statistical idea of machine learning in artificial intelligence satisfies this requirement best.

Artificial intelligence models have found wide application in medicine [4–7]. Explainable artificial intelligence models that analyze medical images have proven particularly useful [8–10]. However, these models' utility results from a compromise between explainability and interpretability since physicians do not understand how the model is constructed and limit themselves to explaining the input and output data (explainability). In contrast, interpretability remains the prerogative of mathematicians [11].

The content of the explainability concept is disclosed in the work [12], in which the authors reduced it to three points: 1) explainability of input data; 2) explainability of output data; 3) explainability of the algorithm. Based on the above, we must recognize that this structure requires clarification, since the third point appeals to the user's competence. In the work [13], the authors classified the known models by the degree of their explainability, arguing that linear regression, logistic regression, decision trees, kNN method, rule-based inference algorithms, generalized additive models, and Bayesian models are self-evident. The authors consider the random forest method, SVM, and various neural networks less explainable. In this case, there is a typical aberration of the professional point of view. After all, the complexity of explaining each of these models depends on the degree of professionalism of the mathematician, and for a non-mathematician, all of them are equally incomprehensible. In our view, this deficiency can be addressed by shifting the emphasis from explainability to interpretability, which is sometimes called model transparency.

In this direction, it is worth highlighting the works of Cynthia Rudin [15, 16], devoted to studying the interpretability of machine learning. Rudin considers explainability and interpretability to be different properties of machine learning and suggests not to explain the work of black boxes, but to develop transparent, interpretable models. This approach is correct, but Rudin also does not go beyond traditional models, ignoring the subjectivity of interpretability assessments if they depend on the degree of competence of a specialist. This fact is especially evident in medical applications. For example, how can a mathematician explain the input data if he has no idea what condensed chromatin is in Feulgen-stained buccal epithelial nuclei? In turn, a doctor cannot say anything about a nonparametric criterion for assessing the homogeneity of samples containing measurements of the level of condensed chromatin in healthy people and patients with breast cancer. They do not have a common point of view. We propose such a point of view as the concept of results typicality expressed by the elliptical statistical depth. The explanation of typicality does not require mathematical knowledge, but it is based on common sense: the greater the statistical depth of a result, the more typical it is. For example, the more statistical depth of a patient's features, the more likely the patient is sick. This does not mean that she is more seriously ill. It means a higher probability (but not the subjective confidence of the doctor) that the patient is sick.

An excellent analysis of the psychological foundations of explainability and interpretability was given by David Broniatowski [3]. Based on the analysis of the literature on experimental

psychology in the field of interpretation of numerical data, the author proves that the concepts of interpretability and explainability reflect different requirements for machine learning algorithms. From the author's point of view, interpretability is the ability to understand to what extent the output of a machine learning model corresponds to its intended purpose, as well as to the goals and preferences of its users. In turn, explainability means the ability to accurately understand the mechanism of obtaining the result in order to improve the algorithm. The author analyzes the psychological aspects of decision-making, in particular, distinguishing between users who prefer to make decisions based on detailed explanations and users who want to receive meaningful interpretations of the model's output. This aspect is clearly manifested in diagnostic systems used by doctors and patients. The doctor must be confident in the diagnosis, since he is legally responsible for it, so explainability and interpretability are equally important for him, and it is important for the patient to know the level of reliability of the diagnosis in order to make a decision on further treatment, so interpretability is more important for him than explainability.

It is obvious that machine learning systems should have both explainability and interpretability. The only question is in what proportions. Currently, more attention is paid to explainability, and relatively little attention has been paid to the interpretability of machine learning models. As research in the field of experimental psychology of numerical stimuli shows, people understand the concept of interpretability by spacing and connecting the output of a model with its inference engine by spacing. According to Broniatowski, it is necessary to study to what extent this issue can be automated, since this problem is still poorly understood. Summarizing his analysis, Broniatowski argues that interpretable models should take into account the context of the user's knowledge and present the results in a simple form, justifying their reliability.

According to Rudin, many machine learning models are too complex for humans to understand. For this reason, their explanation usually boils down to a theoretical description of the inference mechanism rather than a description of its actual implementation. Focusing on the explainability of machine learning models and ignoring issues related to their interpretability hinders the widespread use of machine learning models. Cynthia Rudin makes several points about interpretability: 1) accurate models do not have to be complex; 2) explanations of machine learning methods often do not match the computations of the original model; 3) explanations are often meaningless or unclear; 4) unexplainable systems should not be used in high-risk situations; 5) unexplainable systems complicate the human decision-making process.

As noted above, reliable explanations are essential for machine learning models involving high-risk decisions (particularly in medical applications). It is natural to use algorithms that reveal the decision-making mechanism for such models. At the same time, the commercial interests of corporations put the protection of decision-making mechanisms from copying to the forefront, preventing their explanation and interpretation. There is even a separate line of research devoted to finding a compromise between explainability and interpretability, on the one hand, and preserving commercial secrets, on the other [17].

Of course, many statistical methods are already widely used in machine learning (logistic regression, Bayesian methods, and many others), but each has limitations in its explainability and interpretability. For example, logistic regression allows you to find the probability of a particular event only if the probability distribution of this event obeys the Bernoulli distribution with a specific parameter. Bayesian classification methods are relative; they allow you to compare estimates of the probability of an event, but do not estimate these probabilities themselves. As a result, their explainability and interpretability are pretty weak.

We propose 1) new postulates of statistical machine learning; 2) a new method for assessing the homogeneity of objects; 3) a new method for assessing the typicality of an object based on its statistical depth in the prediction set; 4) a new method for ranking random points in a multidimensional space; 5) a new concept of interpretability of machine learning algorithms; 6) dimensionality reduction method. Consider them step by step.

2. Homogeneity measure

Machine learning is based on two key principles, which we have previously alluded to: first, objects should be represented as feature vectors within a feature vector space; second, feature vectors representing items within the same class are closer together in the feature space than those from different classes. The first principle reflects the tendency of machine learning practitioners to utilize algebra, geometry, and optimization techniques. This allows us to frame machine learning challenges as optimization problems, explicitly focusing on minimizing or maximizing a function under certain constraints. The second principle suggests a relatively simple function can separate these vector sets. Techniques such as Fisher's linear discriminant, support vector machines, and the nearest neighbor approach are notable examples built on these foundations.

However, despite the success of these methods, it is essential to recognize that the vector space and compactness postulates do not apply universally. In many medical and biological contexts, a patient is not represented by a single feature vector—an ordered set of numerical characteristics—but rather by a random sample, an unordered collection of measurements (e.g., nuclear area, optical density). For instance, when analyzing samples from a patient, which may consist of dozens of cells, the patient is represented as a cloud of points rather than a single point in vector space. While averaging these sample values can simplify the process and allow for the application of the established postulates, it also results in a loss of significant information regarding the distribution of the measured parameters.

We propose alternative statistical hypotheses: 1) sample parameter values can represent objects, and 2) parameters of objects within the same class exhibit similar distributions, while those from different classes show distinct distributions. This approach allows us to tackle the challenge of assessing similarity between objects by verifying whether two or more samples are homogeneous. The method we suggest for determining similarity is outlined below. It possesses statistical universality, meaning it performs consistently across samples with varying means and identical standard deviations, as well as samples with the same means but differing standard deviations—unlike traditional methods such as the Kolmogorov-Smirnov and Mann-Whitney-Wilcoxon tests.

Consider a sample of size n composed of continuous random variables drawn from a exchangeable distribution. According to Hill's assumption, the probability that a random value from the same distribution falls between the i-th and j-th order statistics of the sample is given by $(j\text{-}i)/(n+1)$ [18, 19]. Notably, the only factors influencing this probability are the sample size and the order numbers of the statistics.

This insight enables us to test the hypothesis of homogeneity between two samples. To do this, we first arrange the elements of the first sample in ascending order to obtain its order statistics. Next, we calculate the relative frequency of occurrences from the second sample that fall between the i-th and j-th order statistics of the first sample.

Using these relative frequencies, we can construct a confidence interval (for example, the Wilson interval) for the binomial proportion in the generalized Bernoulli framework we are examining. We then assess whether this confidence interval covers the value $(j–i)/(n+1)$.

To quantify this, we compute the so-called p-statistics by measuring the relative frequency of the event in question. Finally, we establish a confidence interval for the p-statistics based on a predetermined significance level. If this confidence interval does not include $1 - \alpha$, were α is the given confidence level, we reject the null hypothesis of homogeneity [20].

3. Statistical depth

For a comprehensive overview of the various concepts related to statistical depth, refer to [21]. The primary objective of these concepts is to establish an ordering of multidimensional random variables..

Consider a distribution *D*. A depth function *d* is defined to order points from this distribution in a manner that monotonically decreases from the center outward. The depth of a point *x* is represented as *d*(x) [22]. The center of a distribution can be defined in various ways, such as the median, centroid, or geometric center. A depth function must satisfy the following properties [22]:

1. **Affine Invariance**: The depth function should be independent of the coordinate system used and should remain unchanged under affine transformations.
2. **Maximum at the Center**: The depth function attains its maximum value at the center of the distribution, which is the point of greatest depth.
3. **Monotonicity**: The depth function must decrease monotonically from the deepest point to the least deep points.
4. **Limit Property**: As the distance from a point *x* to the center of the distribution approaches infinity, the depth must approach zero.

When we lack specific information about the distribution *D* but have a sample containing *n* points from it, we denote this sample as *S*. Below are examples of different depth functions.

1. **Tukey Depth** [23]: To understand Tukey depth, we first need to define a center of a sample as a point such that every hyperplane passing through it divides the sample into two nearly equal subsets. When this point is part of the sample, it corresponds to the sample's median. The Tukey depth of a sample element *x* is defined as the minimum number of sample elements that lie on one side of a random hyperplane passing through *x*.
2. **Convex Hulls Peeling** [24]: The convex hull of a set of points is the smallest polygon that encompasses all the given points. Convex hull peeling is a method that involves sequentially identifying and removing enclosed convex hulls. All vertices of the same convex hull share the same statistical depth.
3. **Oja Depth** [25]: The Oja depth of a sample element *x* is calculated as the average volume of the simplex formed by *d* random sample points and the point *x*.
4. **Simplex Depth** [26]: The simplex depth of a sample element *x* is defined as the number of simplexes formed by a random sample of points that include *x*.
5. **Zonoid depth** [27]. The zonoid depth of a sample element x is the number $d(x \mid x_1,...,x_n) = \sup\{\alpha : y \in D_\alpha(x_1,...,x_n)\}$, where
$$D_\alpha(x_1,...,x_n) = \left\{ \sum_{i=1}^{n} \lambda_i x_i : \sum_{i=1}^{n} \lambda_i = 1,\ 0 \le \lambda_i, \forall i : \alpha\lambda_i \le \frac{1}{n} \right\}$$
6. **Mahalanobis depth** [22]. The Mahalanobis distance is a generalization of the Mahalanobis distance. It is defined by the formula $MHD_F(x) = (1 + d^2(x, E(F)))^{-1}$, where $d^2(x,y) = (x-y)^T \Sigma_F^{-1}(x-y)$, $E(F)$ is the distribution expectation, and Σ_F is the covariance matrix/
7. **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.
8. **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.

4. Petunin ellipsoids

Consider is a set of random points $X = \{x_1,...,x_n\}$, $x_i \in \mathbb{R}^d$. For simplicity and easy visualization, we shall descride the case $d = 2$ (Petunin 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 $X = \{(x_1, y_1),...,(x_n, y_n)\}$.

We can generalize this algorithm to construct a Petunin ellipsoids. Construct a convex hull of $X = \{x_1,...,x_n\}$, $x_i \in \mathbb{R}^d$. Find ends of a diameter of the convex hull (x_k, y_k) and (x_l, y_l). Align the diameter along to Ox'_1. Project points (x'_1, y'_1), (x'_2, y'_2), ..., (x'_n, y'_n) to the orthogonal complement of Ox'_1. Construct a convex hull of projections, rotate and translate it up to a two-dimensional rectangle Π. Construct an axis-aligned parallelogram of minimum volume in d-dimensional space containing the projections of input points. Shrink this parallelogram to hypercube. Find center x_0 of the hypercube and the distances $r_1, r_2,..., r_n$ from x_0 to $x'_1,...,x'_n$. Compute $R = \max(r_1, r_2,..., r_n)$. Construct a hypersphere with the center x_0 and radius R. Make the inverse transformations. The result of the operations is the Petunin ellipsoid covering $X = \{x_1,...,x_n\}$.

The Petunin's ellipsoids allow uniquely arranging random multivariate points according their statistical depth because the surface of Petunin's ellipsoid contains only one point from the initial set. In addition, the probability that a Petunin's ellipsoid covers random points obeying the same distribution is equal to $\frac{n-1}{n+1}$.

We also note an essential property of Petunin ellipsoids: their concentricity. This property allows for automatic and unambiguous ranking of multidimensional points. Ellipses in this case are chosen for convenient visualization.

Therefore, we can find most and least probable points of a sample. The deepest point has the highest statistical depth.

5. Similarity space and Petunin ellipses

Duin and Pekalska [31–34] and others in their works proposed the concept of relational discriminant analysis. They suggested replacing the feature vector of an object with an estimate of its proximity to some training set using a metrics or a measure of proximity between random

samples. This approach is well suited to solving problems often encountered in biomedical research. Let's say a researcher studies the parameters of a set of cells. In this case, it gets samples of real numbers, not an ordered vector. In such cases, the metric is not applicable and the only useful tool is the uniformity measure.

The homogeneity measure described above can be used to solve dimensionality reduction and feature selection. To do this, we calculate the measure of homogeneity between the samples from G_1 and G_2 for two features, i-th and j-th, and consider the matrices of features of the k-th object from G1 and the l-th object from G2, where is the number of features, and is the number of measured values of each feature.

$$U_k = \begin{pmatrix} u_{11}^{(k)} & u_{12}^{(k)} & \dots & u_{1n}^{(k)} \\ u_{21}^{(k)} & u_{22}^{(k)} & \dots & u_{2n}^{(k)} \\ \dots & \dots & \ddots & \dots \\ u_{m1}^{(k)} & u_{m2}^{(k)} & \dots & u_{mm}^{(k)} \end{pmatrix}, \; V_l = \begin{pmatrix} v_{11}^{(l)} & v_{12}^{(l)} & \dots & v_{1n}^{(l)} \\ v_{21}^{(l)} & v_{22}^{(l)} & \dots & v_{2n}^{(l)} \\ \dots & \dots & \ddots & \dots \\ v_{m1}^{(l)} & v_{m2}^{(l)} & \dots & v_{mn}^{(l)} \end{pmatrix}.$$

Denote the ith columns corresponding to ith feature from u_k and v_l as $U_i^{(k)} = \left(u_{1i}^{(k)}, u_{2i}^{(k)}, ..., u_{mi}^{(k)}\right)^T$ and $V_i^{(l)} = \left(v_{1i}^{(l)}, v_{2i}^{(l)}, ..., v_{mi}^{(l)}\right)^T$. Then, compute p-statistics for samples $U_i^{(k)}$ and $V_i^{(l)}$ and find the vector of p-statistics for u_k and v_l with respect to every feature:

$$\mu_{kl}^{(1)} = \rho\left(U_1^{(k)}, V_1^{(l)}\right), \; \mu_{kl}^{(2)} = \rho\left(U_2^{(k)}, V_2^{(l)}\right), \; ..., \; \mu_{kl}^{(n)} = \rho\left(U_N^{(k)}, V_N^{(l)}\right).$$

Then, compute the average p-statistics.

$$\nu_k^{(1)} = \frac{1}{N}\sum_{t=1}^{N} \mu_{kt}^{(1)}, \; \nu_k^{(2)} = \frac{1}{N}\sum_{t=1}^{N} \mu_{kt}^{(2)}, \; ..., \; \nu_k^{(n)} = \frac{1}{N}\sum_{t=1}^{N} \mu_{kt}^{(n)}.$$

for U_k and an object from G_2 with respect to ith feature. This scheme allows estimating the proximity of U_k to other object from G_1.

Pairing p-statistics we form a proximity vector space corresponding to ith and jth features: $\left(\nu_t^{(i)}, \nu_t^{(j)}\right)$ and $\left(\bar{\nu}_s^{(i)}, \bar{\nu}_s^{(j)}\right)$, $i, j = 1, 2, ..., m; t, s = 1, 2, ..., n.$ Thus, we have two sets of points consisting of average interclass homogeneity measure and average intraclass homogeneity measure in the proximity space but not feature space. This allow using any method of classification developed for metric spaces but in the proximity measure of less dimensions. The average intraclass homogeneity measure allows estimating intrinsic diversity of objects in the population, and the average intraclass homogeneity measure allows estimating the feature significance.

6. Uncertainty and Petunin ellipses

When using Petunin ellipses to classify an object, uncertainty may arise: the point corresponding to the object may not fall into any ellipses or into their intersection. In turn, the intersection may also be such that one ellipse completely covers the other. In this case, you can use the remarkable property of Petunin ellipses, namely, their concentricity. Since at the penultimate stage of constructing the Petunin ellipse, we obtain concentric circles containing only one point, we can automatically rank the points by statistical depth, simply by calculating the circle number relative to the center of gravity of the points. Next, we alternately include the point under study in one or another set of training samples and find its statistical depth in each of them. By comparing these statistical depths, we assign the point to the set with greater statistical depth.

Knowing the number of the ellipse on which the test point lies, we can even estimate the probability with which it belongs to the class. We can decide with a given significance level by constructing a confidence interval for this probability. This fact allows for a significant increase in classification sensitivity by eliminating uncertainty

7. Conclusion

Based on the homogeneity of random variables, the paper proposes consider that objects belong to the same class if the random values of their features have the same distribution. The paper describes the constituent parts of the proposed theory (homogeneity measure, prediction set as Petunin ellipse, statistical depth based on Petunin ellipsoids, and decreasing of dimensionality using the similarity space). Based on the statistical postulates of machine learning, it is shown that the proposed machine learning algorithms can be classified as interpretable and explainable. We proposed and justified new postulates of statistical machine learning; a new method for assessing the homogeneity of objects; a new method for assessing the typicality of an object based on its statistical depth in the prediction set; a new method for ranking random points in a multidimensional space; a new concept of interpretability of machine learning algorithms, and dimensionality reduction method.

8. Declaration on Generative AI

The author have not employed any Generative AI tools.

References

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

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

[3] D. Broniatowski, Psychological Foundations of Explainability and Interpretability in Artificial Intelligence, NISTIR (2021) 8367. doi: 10.6028/NIST.IR.8367.

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

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

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

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

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

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

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

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

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

[13] 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.
[14] A. Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI, Information Fusion 58 (2022) 82–115, doi:10.1016/j.inffus.2019.12.012.
[15] K. Sokol, P. Flach, Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches, arXiv:1912.05100v. (2019) doi:10.1145/3351095.3372870.
[16] C. Rudin et al., Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistical Surveys 16 (2022) 1–85. doi:10.1214/21-SS133.
[17] C. Zhong, P. Chen, C. Rudin, Models That Are Interpretable But Not Transparent. arXiv:2502.19502 (2025). doi: 10.48550/arXiv.2502.19502.
[18] B. Hill, Posterior distribution of percentiles: Bayes' theorem for sampling from a population. Journal of American Statistical Association 63 (1968) 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 55(2) (2003) 181–198.
[21] K. Mosler, P. Mozharovskyi, Choosing among notions of multivariate depth statistics. Statistical Science 37(3) (2022) 348–368. doi:10.1214/21-sts827.
[22] Y. Zuo, R. Serfling, General notions of statistical depth function, Annals of Statistics 28 (2000) 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) 139 (3) (1976) 318–355.
[25] H. Oja, Descriptive statistics for multivariate distributions, Statistics and Probability *Letters* 1 (1983) 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 18: 405–414 (1990).
[27] G. Koshevoy, K. Mosler, Zonoid trimming for multivariate distributions. Annals of Statistics 25 (1997) 1998–2017. doi:10.1214/aos/1069362382.
[28] S. Lyashko, D. Klyushin, V. Alexeyenko, Mulrivariate ranking using elliptical peeling. *Cybernetic and Systems Analysis* 49(4): 511–516. doi:10.1007/s10559-013-9536-x (2013)
[29] Yu. Petunin, B. Rublev. Pattern recognition using quadratic discriminant functions. Numerical and Applied Mathematics 80 (1996) 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 2 (2007) 1–28.
[31] R.P.W. Duin, D. de Ridder, D.N.J. Tax, Experiments with a featureless approach to pattern recognition, Pattern Recognit Lett 18 (1997) 1159–1166. doi: 10.1016/S0167-8655(97)00138-4.
[32] R.P.W. Duin, E. Pekalska, D. de Ridder, Relational discriminant analysis, Pattern Recognition Letters 20 (1999) 1175–1181. doi: 10.1016/S0167-8655(99)00085-9 .
[33] E. Pekalska, R.P.W. Duin, On combining dissimilarity representations, in: J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, LNCS, vol. 2096, Springer-Verlag, 2001, pp. 359–368. doi: 10.1007/3-540-48219-9_36.
[34] E. Pekalska, R.P.W. Duin, The Dissimilarity Representation for Pattern Recognition, Foundations and Applications, World Scientific, Singapore, 2005.