Computing, Telecommunication and Control, 2026, Vol. 19, No. 3, Pp. 29–45.
Информатика, телекоммуникации и управление. 2026. Том 19, № 3. С. 29–45.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19303
UDC 004.852

# KERNEL REGRESSION WITH TREE-INDUCED METRICS

*M.I. Goriunov* , *A.V. Konstantinov* ✉ , *F.A. Novikov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ andrue.konst@gmail.com

**Abstract.** This paper presents a family of kernel models based on metrics derived from the structure of decision trees and random forests. Unlike the standard random-forest proximity, which only checks whether two observations fall into the same leaf, the proposed metrics also use information about decision paths, the depth of their divergence, and the depth of their lowest common ancestor. A weighted variant with trainable depth-dependent parameters is also introduced. The tree-based metrics are aggregated over a forest and used to construct exponential kernels. They can also be combined with the Euclidean distance in the original feature space. We prove that the resulting kernels are positive semidefinite and can therefore be used in standard kernel methods, including Nadaraya–Watson regression and kernel ridge regression. The proposed models are evaluated on classification datasets and in a controlled source-to-target regression experiment with a limited number of labeled target observations. The results show that information about the internal structure of decision trees can be useful beyond simple leaf coincidence. Several proposed variants improve upon conventional geometric kernels and the standard random-forest proximity kernel, while the best-performing metric depends on the prediction method and the dataset.

**Keywords:** kernel regression, decision trees, tree-based metrics, random forests, Nadaraya–Watson estimator, kernel ridge regression

**Acknowledgements:** The research was financially supported by the Russian Science Foundation, project “New methods and models of multi-variant learning in medical diagnostic tasks” (Agreement No. 25-21-00103; available online: https://rscf.ru/project/25-21-00103/).

**Citation:** Goriunov M.I., Konstantinov A.V., Novikov F.A. Kernel Regression with Tree-Induced Metrics. Computing, Telecommunications and Control, 2026, Vol. 19, No. 3, Pp. 29–45. DOI: 10.18721/JCSTCS.19303

© Goriunov M.I., Konstantinov A.V., Novikov F.A., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19303
УДК 004.852

# ЯДЕРНАЯ РЕГРЕССИЯ С ИНДУЦИРОВАННЫМИ ДЕРЕВЬЯМИ МЕТРИКАМИ

*М.Ю. Горюнов* , *А.В. Константинов* ✉ , *Ф.А. Новиков*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ andrue.konst@gmail.com

**Аннотация.** В работе предложено семейство ядерных моделей, основанных на метриках, формируемых по структуре деревьев решений и случайных лесов. В отличие от стандартной меры близости случайного леса, которая учитывает только попадание двух объектов в один лист, предложенные метрики также используют информацию о путях в дереве, глубине их расхождения и глубине наименьшего общего предка. Кроме того, рассматривается взвешенная метрика с обучаемыми параметрами, зависящими от глубины дерева. Метрики отдельных деревьев агрегируются по лесу и используются для построения экспоненциальных ядер. Они также могут сочетаться с евклидовым расстоянием в исходном пространстве признаков. Для полученных ядер доказана положительная полуопределенность, что позволяет применять их в стандартных ядерных методах, включая регрессию Надарая–Уотсона и ядерную гребневую регрессию. Предложенные модели исследованы на задачах классификации и в контролируемом эксперименте по переносу структуры из исходной задачи в целевую задачу регрессии с малым числом размеченных наблюдений. Результаты показывают, что внутренняя структура деревьев может содержать полезную информацию, не ограниченную совпадением листьев. Несколько предложенных вариантов превосходят традиционные геометрические ядра и стандартное ядро близости случайного леса, при этом выбор наиболее подходящей метрики зависит от метода прогнозирования и набора данных.

**Ключевые слова:** ядерная регрессия, деревья решений, метрики на основе деревьев, случайные леса, ядерные методы, оценка Надарая–Уотсона, ядерная гребневая регрессия

**Финансирование:** Исследование выполнено при финансовой поддержке Российского научного фонда в рамках реализации проекта «Новые методы и модели многовариантного обучения в задачах медицинской диагностики» (Соглашение №25-21-00103; https://rscf.ru/project/25-21-00103/).

**Для цитирования:** Goriunov M.I., Konstantinov A.V., Novikov F.A. Kernel Regression with Tree-Induced Metrics // Computing, Telecommunications and Control. 2026. Т. 19, № 3. С. 29–45. DOI: 10.18721/JCSTCS.19303

## Introduction

The random forest algorithm [1, 2] is one of the most widely used ensemble methods in machine learning. Its practical appeal stems from several complementary properties: averaging over many decision trees reduces overfitting, the resulting model is robust to noise and outliers, and feature-importance estimates provide a useful degree of interpretability. Extremely randomized trees [3] further develop this paradigm by choosing split thresholds at random. This additional randomization reduces the computational cost of training, can improve robustness on noisy data, and often yields strong generalization performance in both classification and regression.

In addition to producing predictions, a forest defines an adaptive partition of the feature space. This observation has motivated the use of random-forest proximities as kernels: two observations are considered similar if they frequently fall into the same terminal leaf. Such proximities are attractive

© Горюнов М.Ю., Константинов А.В., Новиков Ф.А., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

because they inherit the nonlinear and data-dependent geometry of the trained ensemble, but the usual shared-leaf score is a coarse summary of tree structure. It treats all leaf mismatches equally, even though two leaves may diverge just below a deep internal node or immediately below the root. It also ignores whether leaves occupy related subtrees and whether early or late path divergences should be penalized differently.

This limitation is especially relevant when a tree ensemble is used as a representation rather than only as a predictor. In kernel classification and regression, the Gram matrix determines the geometry available to the downstream estimator. It is also relevant in source-to-target structure reuse: a forest trained on one response may encode a useful hierarchical organization of the input space, and this organization can be tested as a prior similarity for a related response with fewer labels. In this paper, we use such a setting as a controlled diagnostic experiment rather than as a full transfer-learning benchmark. A proximity based solely on terminal-node equality may therefore underuse the hierarchy that made the forest informative in the first place.

This paper addresses the lack of a compact framework with explicit theoretical guarantees for converting different aspects of tree hierarchy into positive semidefinite kernels. We focus on distances and similarities that are still simple enough to compute from a fitted tree, but richer than the standard leaf-coincidence score. The proposed constructions use path divergence, lowest common ancestors, path-indicator embeddings, inverse-depth penalties, and learnable depth weights. These structural quantities are then aggregated over a forest and combined with Euclidean geometry in the original feature space.

The contributions of the paper are threefold. First, we formulate several single-tree structural distances on the quotient space induced by terminal leaves and show how they extend to forests. Second, we prove positive semidefiniteness of the corresponding exponential and mixed kernels using conditional negative definiteness and Schoenberg's theorem. Third, we evaluate the resulting models in supervised classification and in a controlled source-to-target regression experiment that tests reuse of source-trained tree structure for a related target response with fewer labeled examples. The source code of the proposed distances is available at https://github.com/MaxGoryunov/tree-distance-kernel.

***Related work***

Beyond their direct use as predictive models, random forests can also induce kernels and proximity functions. In [4], the kernel connection function, or KeRF, is introduced, defined as the fraction of trees in which two observations fall into the same leaf, and, in the population limit, as the probability of this event. This view makes it possible to interpret random forests as adaptive kernel estimators and to incorporate forest-induced similarities into kernel methods. The related adaptive-nearest-neighbor interpretation of random forests was developed in [5] with forest predictions expressed through data-dependent weights determined by terminal cells. Foundational analyses establish consistency for averaging classifiers [6] and characterize the behavior of a random-forest model under sparse signal structure [7].

Several related constructions have been proposed. In [8], centered and uniform KeRF estimators are analyzed, improved convergence rates are derived, and the reproducing-kernel Hilbert space associated with the centered construction is investigated. In a complementary unsupervised setting [9] random-forest leaf proximities are used as a data-adaptive similarity for clustering and visualization.

Random-forest-induced similarities have also been used beyond direct supervised prediction. In [10], an unsupervised random-forest procedure for mixed data that retains the usual advantages of forest models while identifying clusters is developed. For class-probability estimation [11], aggregation rules and terminal-node sizes can materially affect calibration. Related forest weighting constructions have been used to estimate full conditional distributions and conditional quantiles [12].

The geometry represented by forest proximities has been examined in [13]. The authors argue that traditional proximity measures, including Breiman's original definition, may fail to preserve the

geometry learned by the forest, which can introduce bias and limit their usefulness as kernels. Their RF-GAP proximity is designed to preserve both geometry and predictive accuracy, making it better suited for kernel-based learning. Other work has considered distances between trees themselves: the Weighted Splitting Variable (WSV) distance [14] measures the difference between trees through weighted variable usage and can be used to identify representative trees, thereby improving interpretability of complex ensembles.

Random forests have also been used for metric learning. The Random Forest Distance (RFD) [15] learns an implicitly position-dependent distance function through a pairwise mapping $\phi(x_i, x_j)$ that encodes both the relative position of two points and their absolute location in feature space. This enables a single metric to adapt to different regions of the input space, combining the locality of metric-learning methods with the efficiency of global models. Distances can also be computed in the presence of missing data [16]: instead of sending an observation with a missing split value along a single path, the method constructs a multi-path subtree $P_T^{MP}(x)$ that contains all possible paths obtained by propagating the observation to both children. This approach has been applied to RatioRF and adapted to other RF distances, including Shi, Zhu, and Ting, and remains robust even with substantial missingness.

Additional lines of work exploit distances either between classes or between observations. The RSCT decision-tree algorithm [17] augments impurity criteria with distances between classes, thereby favoring splits that separate distant classes; both Euclidean and cosine distances are considered. Experiments show improved classification accuracy relative to a classical Gini-based decision tree, both for a single tree and for a random-forest ensemble. RF distances are also known to exhibit robustness in high-dimensional settings [18]: nearest-neighbor classification with RF distances such as Shi and RatioRF can continue to improve as the number of features grows, because tree partitions tend to amplify distinctions between observations and counteract distance concentration. Related approximation results [19] show that observations that are close in the true metric are likely to remain close under RF distances computed from ensembles of extremely randomized trees; the probability of poor approximation decreases exponentially with the number of trees, with RatioRF giving tighter bounds than Shi. Instead, our framework treats tree structure through conditionally negative definite distances, establishes positive semidefiniteness guarantees under forest aggregation, and introduces learnable depth-dependent penalties compatible with kernel prediction.

## Materials and methods

**Decision Trees and Paths.** Throughout the paper, $\mathrm{I}[A]$ denotes the indicator of an event or logical statement $A$: it equals 1 when $A$ is true and 0 otherwise.

A decision tree $T$ is represented as a finite rooted binary tree with vertex set $V$, root $v_0$, internal-node set $J$, and leaf set $L$. Each internal node $v$ from $J$ is equipped with a split rule, typically of the form $x_j < t$, which sends a $r$-dimensional input $x$ to either the left or the right child. Each leaf $l$ from $L$ stores a local prediction, such as a class distribution in classification or an empirical response mean in regression.

For every input $x$, the split rules determine a unique root-to-leaf path

$$p_T(x) = \left(v_0, v_1, \ldots, v_{\tau_T(x)}\right),$$

where the last component is the terminal node reached by $x$, and $\tau_T(x)$ is the depth of that leaf. The depth of the root is 0, and the depth of any other node is the number of edges on the path from the root to that node.

Throughout the single-tree constructions below, the tree $T$ is assumed to be fixed. To avoid cumbersome notation, the dependence of structural distances and auxiliary quantities on $T$ is suppressed

whenever no ambiguity can arise. The corresponding tree indices are restored in forest-level definitions and aggregation formulae.

**Leaf-Induced Quotient Sets.** Each leaf $l$ from $L$ of a tree $T$ defines an equivalence class

$$[\ell] = \{x \in \mathbb{R}^r : x \text{ falls into leaf } l \text{ of tree } T\}.$$

The corresponding quotient set is

$$X_T := \mathbb{R}^r / \sim_T = \{[l] : l \in L\}.$$

When the tree is fixed, we write $X$ for $X_T$. All single-tree structural distances below are defined on these classes. This makes the value of a distance independent of the particular representative chosen from a leaf: two observations in the same terminal node are indistinguishable for the structural component associated with that tree.

**Random Forests.** A random forest is an ensemble of decision trees

$$F = \{T_1, T_2, \ldots, T_M\}.$$

The trees are trained using randomized procedures, such as bootstrap resampling of observations and random selection of candidate split variables. Extremely randomized trees additionally randomize split thresholds. Given a new input, the forest aggregates the predictions of its individual trees, for example by majority vote in classification or by averaging in regression.

In this paper, the forest is also treated as a learned partitioning mechanism. Each tree assigns an input to a leaf, while the whole ensemble assigns it to a vector of leaves. The method developed below uses these assignments and the paths leading to them to construct structural distances and kernels.

**Nadaraya–Watson Kernel Regression.** Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a supervised training sample, and let $K(x, x_i) \geq 0$ be a kernel similarity between a query point $x$ and a training point $x_i$. The Nadaraya–Watson estimator predicts a weighted average of the observed responses:

$$\hat{y}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n K(x, x_i) y_i}{\sum_{i=1}^n K(x, x_i)}.$$

For classification, the same formula can be applied to one-hot encoded class labels, yielding estimated class probabilities. Thus, kernel regression provides a direct way to turn a similarity function into either a regression model or a probabilistic classifier.

**Kernel Ridge Regression (KRR).** KRR is a regularized kernel method in a reproducing kernel Hilbert space. Given a kernel $K$, let $K_{\text{train}}$ be the Gram matrix with entries

$$(K_{\text{train}})_{ij} = K(x_i, x_j),$$

and let

$$k_x = (K(x, x_1), \ldots, K(x, x_n))^\top.$$

For the regularized least-squares objective with penalty parameter $\lambda > 0$, the representer theorem gives coefficients

$$c = \left(K_{\text{train}} + \lambda I_n\right)^{-1} y.$$

The prediction at a new point $x$ is therefore

$$\hat{y}_{\text{KRR}}(x) = k_x^{\top} c = k_x^{\top} \left(K_{\text{train}} + \lambda I_n\right)^{-1} y.$$

For multi-class classification, the same expression can be applied to one-hot encoded labels or to one-vs-rest target vectors.

**Positive Semidefiniteness Criterion.** We use the standard notion of conditional negative definiteness. A symmetric function $d$ with $d(x, x) = 0$ is conditionally negative definite (CND) if, for all finite sets $x_1, \ldots, x_n$ and real coefficients $a_1, \ldots, a_n$ with $\sum_i a_i = 0$,

$$\sum_{i,j=1}^{n} a_i a_j d\left(x_i, x_j\right) \le 0.$$

By Schoenberg's theorem [20], $d$ is CND if and only if $\exp(-\gamma d)$ is positive semidefinite (PSD) for every $\gamma > 0$.

***Problem statement***

Let $\mathcal{X} \subseteq \mathbb{R}^r$ be the input space. In classification, the response space is $\mathcal{Y} = \{1, \ldots, C\}$; in regression, $\mathcal{Y} = \mathbb{R}$. Given a supervised sample

$$\mathcal{D}_{\text{train}} = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}, \quad \left(x_i, y_i\right) \in \mathcal{X} \times \mathcal{Y},$$

the standard objective is to construct a prediction rule $\hat{g} : \mathcal{X} \to \mathcal{Y}$ that generalizes to unseen observations. In this paper, $\hat{g}$ is obtained from a kernel estimator, so the main modeling problem is to construct a valid kernel

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}.$$

The central object of interest is a tree-induced structural function. Let a fitted forest $F$ assign each observation $x$ to a collection of terminal leaves. We seek a function $f_F(x, y)$ that measures how far two observations are with respect to this learned forest structure, and then define a kernel of the form

$$K\left(x, y\right) = \exp\left(-D\left(x, y\right)\right),$$

where $D$ may combine $f_F$ with the Euclidean distance in $\mathbb{R}^r$. The mathematical requirement is that the resulting Gram matrix $\left(K\left(x_i, x_j\right)\right)_{i,j=1}^{n}$ be positive semidefinite for every finite sample.

The motivation for this formulation is that standard random-forest proximities usually measure only how often two observations reach the same terminal leaf. This information is useful but coarse: it ignores where two paths separate, how deep their common ancestor is, and how leaves are arranged inside the tree. The problem is therefore to build tree-induced kernels that use richer structural information while remaining mathematically valid and computationally practical for kernel-based classification and regression.

We also consider a controlled source-to-target structure-reuse setting. A source task provides a comparatively large labeled dataset $\mathcal{D}_{\text{scr}}$, from which a tree ensemble and the associated structural similarity can be learned. A related target task provides a much smaller labeled dataset $\mathcal{D}_{\text{tgt}}$. The goal is not to learn a general domain-adaptation model, but to test whether the source-induced tree structure

remains informative when used as a kernel or metric for a related target response with limited labeled data.

This formulation leads to four requirements. First, the resulting distances must be well defined on the leaf quotient spaces induced by individual trees. Second, forest-level aggregation must produce positive semidefinite kernels suitable for kernel methods. Third, the construction should allow a controlled combination of tree-induced structure and Euclidean geometry in the original feature space. Fourth, the resulting kernels should be usable in supervised classification, supervised regression, and controlled source-to-target structure-reuse scenarios.

***Method***

The method turns a trained tree ensemble into a kernel model in three steps. First, for each tree $T$, observations are projected to the quotient space $X_T$ determined by terminal leaves. A single-tree structural distance or dissimilarity $f$ is then computed on this quotient space. Second, the single-tree quantities are averaged over the forest. Third, the resulting forest-level structural distance is inserted into an exponential kernel, optionally mixed with the Euclidean distance in the original feature space.

Thus, each choice of $f$ defines not only a structural interpretation of tree proximity, but also a complete kernel learning model through the induced Gram matrix. The constructions below differ in how much of the tree hierarchy they use: terminal-leaf agreement, depth of path divergence, explicit Euclidean embeddings of leaves, inverse-depth penalties, or learnable depth-dependent weights. For an ensemble $F$, the single-tree quantities are aggregated by averaging:

$$f_F(x,y) = \frac{1}{M}\sum_{m=1}^{M} f_{T_m}(x,y).$$

The associated exponential forest kernel is

$$K_\gamma(x,y) = \exp\left(-\gamma f_F(x,y)\right) = \prod_{m=1}^{M} \exp\left(-\frac{\gamma}{M} f_{T_m}(x,y)\right),$$

which is positive semidefinite whenever the corresponding single-tree exponential kernels are positive semidefinite. The forest induces an equivalence relation in which two observations are equivalent if they are assigned to the same terminal leaf in every tree.

The proposed approach combines two sources of information about object proximity. The structural component captures similarity with respect to data-dependent tree partitions and can represent nonlinear local structure that is not apparent under purely Euclidean distances. The geometric component preserves sensitivity to direct distances between feature vectors. Once a structural forest distance $f_F$ has been obtained, we use the mixed kernel

$$K_{\text{mix}}(x,y) = \exp\left(-\left[\alpha\frac{\Delta_{\text{E}}(x,y)}{h} + (1-\alpha)\frac{f_F(x,y)}{\sigma_F}\right]\right),$$

where $\Delta_{\text{E}}(x,y) = \|x-y\|_2^2$, $\alpha \in [0,1]$ is the mixing coefficient, and $\sigma_F > 0$, $h > 0$ are scaling parameters for the structural and Euclidean components, respectively. As $\alpha$ tends to 0, the model emphasizes forest structure; as $\alpha$ tends to 1, it approaches a Euclidean kernel model.

The resulting kernel can be used directly in the Nadaraya–Watson estimator or in KRR. In both cases, each choice of the structural component $f$ induces a different Gram matrix and therefore a different learning model.

We introduce four structural choices for $f$: a shared-leaf distance based on $1-s(x, y)$, a path-based distance $\rho$, an inverse-depth metric $r$ based on the lowest common ancestor, and a learnable weighted metric $w$ with depth-dependent weights. The metrics are introduced to capture progressively richer information contained in a fitted tree: terminal-node agreement, the depth at which two decision paths separate, explicit finite-dimensional representations of leaves, and learnable depth-dependent penalties.

**Path-Based Structural Kernel Model.** Let $d \geq 1$ be the maximum depth of a nontrivial tree $T$. For an object $x$, write its path as

$$p_T(x) = \left(v_0, v_1(x), \ldots, v_d(x)\right),$$

where paths ending before depth $d$ are padded with leaf-specific dummy states. Thus, two observations in the same leaf have identical padded paths, while observations in different leaves differ immediately after their lowest common ancestor. For two objects $x$, $y$, define the first depth at which their padded paths differ by

$$\mathrm{diff}_T(x, y) = \min\left\{i \in \{1, \ldots, d\} : v_i(x) \neq v_i(y)\right\},$$

when such an index exists. The depth of the lowest common ancestor is

$$l(x, y) = \begin{cases} \mathrm{diff}_T(x, y) - 1 & p_T(x) \neq p_T(y) \\ d, & p_T(x) = p_T(y) \end{cases}.$$

This quantity is the basic structural statistic used by the path-based, inverse-depth, and weighted-depth metrics. It equals $d$ for observations assigned to the same leaf under the padded-path convention, which makes the normalized path distance vanish for leaf coincidences.

For a single tree, this model treats two observations as structurally close when their paths diverge only at a large depth, even if they do not end in the same terminal node. The path-based structural distance is the complement:

$$\rho(x, y) = 1 - \frac{l(x, y)}{d}.$$

Hence $\rho(x, y)$ is zero for objects in the same leaf and increases when their paths separate closer to the root.

**Proposition 1** (Path-Based Structural Kernel Model). *For the path-based model defined by* $\rho(x, y) = 1 - l(x, y)/d$, *the function* $\rho$ *is CND. Hence the forest-aggregated path-based kernel is PSD.*

Proof. Let $\mathcal{V}^*$ be the set of non-root states in the padded tree representation, including the leaf-specific padding states, and define the path-indicator feature map

$$\psi_T(x)_v = \mathrm{I}\left[v \in p_T(x)\right], \quad v \in \mathcal{V}^*.$$

Then $\left\langle \psi_T(x), \psi_T(y) \right\rangle$ equals the number of shared non-root states on the two padded paths, which is $l_T(x, y)$. Hence

$$k(x, y) = l(x, y)/d = \left\langle d^{-1/2}\psi_T(x),\ d^{-1/2}\psi_T(y) \right\rangle$$

is a PSD kernel with unit diagonal. Therefore, ρ is CND. Since CND functions are closed under non-negative linear combinations, the forest average is also CND. The squared Euclidean distance is CND as well, so the mixed exponent is CND. The PSD property of the corresponding forest and mixed kernels follows by Schoenberg's theorem.

**Leaf-Coincidence Kernel Model.** For a single tree, this model uses only terminal-node coincidence and ignores the depth at which two paths diverge. Let $\lambda_T(x) \in \mathrm{L}$ denote the leaf reached by $x$. The single-tree similarity and distance are

$$s_T(x,y) = \mathrm{I}\left[\lambda_T(x) = \lambda_T(y)\right], \quad d_T^{\text{leaf}}(x,y) = 1 - s_T(x,y).$$

The corresponding forest-level similarity is

$$s(x,y) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{I}\left[p_{T_m}(x) = p_{T_m}(y)\right],$$

the fraction of trees in which $x$ and $y$ lie in the same leaf. The associated exponential kernel is

$$K(x,y) = \exp\left(-t\left(1 - s(x,y)\right)\right), \quad t > 0.$$

This is a distance-based version of the classical random-forest proximity kernel, with $t$ controlling how strongly leaf mismatches reduce similarity. The same distance can also be used in the mixed kernel to combine leaf coincidence with Euclidean geometry.

**Proposition 2** (Leaf-Coincidence Kernel Model). *The forest similarity s of the leaf-coincidence model is PSD. Consequently*, $1 - s$ is *CND and* $\exp(-t[1 - s])$ is PSD for every $t > 0$.

*Proof.* Let $\lambda_m(x)$ denote the leaf reached by $x$ in tree $T_m$, and let $\eta_m(x)$ be a $|L_m|$-dimensional unit vector with $\lambda_m(x)$-th component equal to one. Then

$$\left\langle \eta_m(x), \eta_m(y) \right\rangle = \mathbb{I}\left[\lambda_m(x) = \lambda_m(y)\right].$$

With the concatenated feature map

$$\Phi(x) = \frac{1}{\sqrt{M}}\left(\eta_1(x)^{\top}, \ldots, \eta_M(x)^{\top}\right)^{\top},$$

we obtain

$$s(x,y) = \left\langle \Phi(x), \Phi(y) \right\rangle.$$

Thus, $s$ is PSD and has constant diagonal $s(x, x) = 1$. Therefore $1 - s$ is CND, and the exponential kernel follows from Schoenberg's theorem.

**LCA-Depth Kernel Model.** Unlike the path-based distance ρ, which depends linearly on the normalized LCA depth, this model uses an inverse-depth penalty. For $x, y \in X_T$ define

$$r(x,y) = \begin{cases} 0 & x = y \\ \dfrac{1}{1 + l(x,y)} & x \neq y \end{cases}.$$

Using the metric $r$ produces a kernel model that assigns a strong penalty to pairs separating near the root and a smaller penalty to pairs whose lowest common ancestor is deep in the tree. This metric is useful when early splits are expected to encode coarse, high-level distinctions between observations.

**Proposition 3** (LCA-Depth Kernel Model $r$). *For the LCA-depth model, $r$ is an ultrametric on the leaf quotient space. In particular, $r$ is CND, and therefore* $\exp(-\gamma r)$ *and the corresponding forest-aggregated mixed kernel are PSD.*

*Proof.* Let $x, y, z \in X_T$, and set

$$a = l(x,y), \quad b = l(y,z), \quad c = l(x,z).$$

If two of the three classes coincide, the ultrametric inequality is immediate. Otherwise, the common ancestors of $(x, y)$ and $(y, z)$ imply that $x$ and $z$ share an ancestor at depth at least $\min\{a, b\}$. Hence $c \geq \min\{a, b\}$. Since $u \mapsto (1 + u)^{-1}$ is decreasing,

$$r(x,z) = \frac{1}{1+c} \leq \max\left\{\frac{1}{1+a}, \frac{1}{1+b}\right\} = \max\left\{r(x,y), r(y,z)\right\}.$$

Thus, $r$ is an ultrametric. Every finite ultrametric is CND: it can be written as a nonnegative linear combination of partition semimetrics induced by its level sets, and each partition semimetric is CND. Forest averaging and the Euclidean component preserve CND, and Schoenberg's theorem completes the proof.

**Weighted-Depth Kernel Model.** Consider a tree $T$ of depth $d$. Introduce parameters $g_j \in \mathbb{R}$, $j = 0, \ldots, d - 1$, and define positive increments

$$s_j = \exp(g_j),$$

which determine a decreasing sequence of distances

$$z_i = \sum_{j=1}^{d-1} \exp(g_j) = \sum_{j=1}^{d-1} s_j, \quad i \in \overline{0,d}, \quad z_d = 0.$$

By construction, $z_i > z_i + 1$ for $i < d$, so the distance decreases as the shared part of the path becomes longer.

The tree distance is then defined as

$$w(x,y) = \begin{cases} 0 & \text{if the leaves containing } x \text{ and } y \text{ coincide} \\ \sum\limits_{j=l(x,y)}^{d-1} \exp(g_j) & \text{otherwise} \end{cases}.$$

Equivalently, the second case can be written as

$$w(x,y) = z_0 - \sum_{j=0}^{l(x,y)-1} \exp(g_j),$$

which is the difference between the maximum path length and the length accumulated along the common part of the two paths.

The exponential parametrization $s_j = \exp(g_j)$ ensures positive, differentiable depth weights, while the use of absolute depth distinguishes subtrees with similar internal structure but different positions in the tree.

**Proposition 4** (Weighted-Depth Kernel Model *w*). *For the weighted-depth model, with positive increments* $s_j = \exp(g_j)$, *the distance w is CND. Hence* $\exp(-\gamma w)$ *and the corresponding forest-aggregated mixed kernel are PSD.*

*Proof.* Use the padded tree representation from the path-divergence definition. For each non-root padded state *v* at depth $u \in \{1, ..., d\}$, set $a_v = s_{u-1}$ and define the weighted path feature map

$$\chi(x)_v = \sqrt{a_v}\,\mathbb{I}\left[v \in p_T(x)\right].$$

The inner product

$$k(x, y) = \left\langle \chi(x), \chi(y) \right\rangle$$

equals the total weight accumulated along the common part of the two padded paths. Its diagonal is constant and equal to $z_0 = \sum_{j=0}^{d-1} s_j$. If two observations lie in the same leaf, then $k(x, y) = z_0$; otherwise $k(x, y) = \sum_{j=0}^{l(x,y)-1} s_j$. Therefore $z_0 - k(x, y)$ is CND and coincides with $w(x, y)$. Forest averaging and the Euclidean component preserve CND, and Schoenberg's theorem implies the stated PSD guarantees.

## Numerical experiments

The proposed tree-induced kernel models were evaluated in two experimental settings. First, we study standard supervised learning on classification tasks, using real benchmark datasets. Second, we study a controlled source-to-target regression setting, where the tree structure and the induced metric are learned from a source response with more labeled data and then used as a structural similarity for prediction on a related target response with fewer labeled examples. The experiments compare several variants of the mixed structural-geometric kernel with classical kernel baselines, random forests, and nearest-neighbor methods.

**Models and baselines.** Across all experiments, we compare the proposed tree-induced kernel models with standard kernel methods and non-kernel baselines. The first group comprises variants of the proposed models, with the mixing parameter α tuned over the set {0.0, 0.25, 0.50, 0.75}. Here, zero corresponds to a purely structural kernel without a Euclidean component.

The learnable variant uses the weighted-depth distance *w*. In the classification setting, the depth-weight parameters, mixing parameter α and the bandwidths σ and *h* of the structural and Euclidean components are optimized jointly during training by gradient descent. In the source-to-target regression setting, only the bandwidth and depth-weight parameters are optimized by gradient descent, while α is selected from the set {0.0, 0.25, 0.50, 0.75} as the value yielding the lowest training loss. Therefore, we refer to this learnable weighted-depth model as Grad. The trainable parameters of the Grad model are optimized for 300 epochs using AdamW with a learning rate of $5\times10^{-3}$ and weight decay 0.01. The learning rate is adjusted according to a cosine annealing schedule implemented with CosineAnnealingLR. Since tree depths are different for the forest trees, for each tree we train a separate depth-weight parameter vector during joint optimization. The optimization objective is the mean squared error for regression tasks and the cross-entropy loss for classification tasks. To reduce the trivial self-similarity effect, each training observation is excluded from the Nadaraya–Watson sum when its own prediction is computed. This yields a leave-one-out training objective for all trainable

parameters. In the classification setting, the parameter α is optimized by gradient descent, and it is parameterized as the sigmoid of an unconstrained trainable variable to restrict values of α to the range [0,1]. The bandwidth parameters σ and $h$ are parameterized as exponentials of unconstrained trainable variables in both learning settings.

We additionally consider three Nadaraya–Watson kernel predictors based on an RBF kernel (RBF), a Laplace kernel (LAPL), and a mixture of Euclidean and Laplace distances in an exponential kernel (R+L), where the Laplace kernel employs $l_1$-norm.

Another baseline is a Nadaraya–Watson predictor based on the random forest kernel (KeRF), defined as the fraction of trees in which two observations fall into the same terminal leaf. We also evaluate an extremely randomized random forest (ET) consisting of 100 extremely randomized trees with a tuned maximum depth on a grid from 1 to 15. The same ensemble of trees is used to construct the structural component of the proposed kernel models. Finally, we include a k-nearest-neighbor predictor (KNN), for which the integer-valued number of neighbors $K$ is tuned on a grid from 1 to 10.

Unless stated otherwise above for the Grad model, hyperparameters are selected using grid search by the target metric AUROC for classification and $R^2$ for regression. Each trial is evaluated using 2-fold cross-validation performed exclusively within the corresponding training partition. For every inner cross-validation split, preprocessing and the tree ensemble were fitted using only the corresponding inner-training subset. Thus, the test folds used for the final evaluation do not influence hyperparameter selection. For all kernel-based methods we tune bandwidth parameters in range $[0.01 \cdot d_{\max}, 2 \cdot d_{\max}]$, where $d_{\max}$ is the maximum value over distance matrix, for which the bandwidth is tuned, 20 logarithmically scaled samples are used.

In the supervised classification setting, we evaluate the models using four times repeated stratified holdout with train size 20%. In each of four repetitions, the data were partitioned into five stratified folds. Each fold was used once as the 20% outer training set, while the union of the remaining four folds was used as the 80% outer test set. This produced 20 outer train-test splits. Hyperparameters were selected by 2-fold cross-validation performed only within each outer training set. Categorical features are one-hot encoded and continuous features are standardized. Missing values are filled by median. All preprocessing transformations are fitted exclusively on each training fold and then applied to the corresponding test fold. In the supervised learning setting, the training and test observations are drawn from the same distribution. These experiments therefore examine whether tree-induced structural kernels improve predictive performance in the absence of transfer between tasks or domains. Classification performance is measured using the area under the receiver operating characteristic curve (AUROC). For the multiclass Dermatology and Website Phishing datasets, we use micro-averaged one-vs-rest AUROC. The experiments use classification datasets obtained from the UCI Machine Learning Repository [21]. Their main characteristics are summarized in Table 1.

Table 2 reports mean AUROC values over 20 outer test folds (5 folds × 4 repetitions), multiplied by 100. The two best values in each row are highlighted. Grad achieves the best result in most cases, while simpler non-learnable variants also often exceed classical baselines, including the KeRF built from the same forest.

For compactness, standard deviations are shown separately in Table 3. These values characterize the variability of the results across repeated data splits.

We use a controlled source-to-target regression setting to test whether the hierarchical partition learned by a forest for a source response can be reused as a structural representation for a related target response. Because the source-target relationship is known by construction, the experiment isolates structure reuse from uncontrolled changes in domains, covariates, or label semantics.

The controlled setting contains a source task A and a target task B. The complete data-generation, splitting, training, tuning, and evaluation procedure was repeated in 12 independent experiments. Each generated dataset contains 210 observations and is divided as follows:

Table 1

**Characteristics of the classification datasets**

| Title | Samples | Features | Classes |
|---|---|---|---|
| Wisconsin Breast Cancer (WBC) | 569 | 30 | 2 |
| Diabetic Retinopathy (DR) | 1151 | 19 | 2 |
| EEG Eyes (EEG) | 14980 | 14 | 2 |
| Haberman (Hab) | 306 | 3 | 2 |
| Ionosphere (Ion) | 351 | 34 | 2 |
| Tic-Tac-Toe (TTT) | 958 | 9 | 2 |
| Website Phishing (WPh) | 1353 | 9 | 3 |
| Dermatology (Drm) | 366 | 34 | 6 |

Table 2

**AUROC mean, %**

| | *s* | ρ | *r* | Grad | RBF | LAPL | R+L | KeRF | ET | KNN |
|---|---|---|---|---|---|---|---|---|---|---|
| WBC | 99.29 | 99.27 | 99.23 | **99.34** | 97.49 | 98.79 | 97.91 | 99.19 | 98.10 | 88.60 |
| DR | 78.94 | 78.45 | 80.28 | **81.93** | 74.21 | 75.90 | 74.81 | 81.92 | 74.18 | 61.33 |
| EEG | 93.87 | 94.17 | **94.52** | 92.14 | 74.37 | 92.64 | 87.23 | 88.69 | 94.08 | 92.50 |
| Hab | 65.13 | 64.49 | 64.96 | **65.92** | 64.97 | 65.13 | 64.93 | 65.77 | 64.23 | 55.88 |
| Ion | 97.13 | 96.93 | 96.67 | 96.53 | 89.58 | 95.14 | 95.10 | 96.52 | **97.38** | 77.27 |
| TTT | **95.60** | 94.13 | 93.53 | 94.53 | 94.42 | 94.42 | 94.53 | 92.97 | 94.61 | 88.78 |
| WPh | 96.48 | 97.01 | 97.47 | 97.60 | 96.37 | 96.72 | 96.71 | 97.13 | 96.83 | 82.41 |
| Drm | 99.82 | 99.81 | 99.74 | 99.82 | 76.84 | 97.78 | 97.84 | 99.81 | 99.54 | 82.30 |

Table 3

**AUROC standard deviation, %**

| | *s* | ρ | *r* | Grad | RBF | LAPL | R+L | KeRF | ET | KNN |
|---|---|---|---|---|---|---|---|---|---|---|
| WBC | 0.76 | 0.83 | 0.74 | 0.79 | 0.66 | 0.68 | 0.81 | 0.67 | 0.73 | 0.96 |
| DR | 2.12 | 1.95 | 1.82 | 2.01 | 2.95 | 1.80 | 2.18 | 1.58 | 2.54 | 2.82 |
| EEG | 1.31 | 0.97 | 1.17 | 1.06 | 2.97 | 0.98 | 1.43 | 0.98 | 1.03 | 1.30 |
| Hab | 0.60 | 0.97 | 0.29 | 0.51 | 0.81 | 0.91 | 0.78 | 0.33 | 0.80 | 0.73 |
| Ion | 0.69 | 0.91 | 0.84 | 0.83 | 1.11 | 0.72 | 0.94 | 0.76 | 0.66 | 2.00 |
| TTT | 0.69 | 1.06 | 1.03 | 0.98 | 1.01 | 1.28 | 1.39 | 1.09 | 1.43 | 1.00 |
| WPh | 0.66 | 0.70 | 0.73 | 0.83 | 0.86 | 0.74 | 0.83 | 0.91 | 0.89 | 1.70 |
| Drm | 0.92 | 0.70 | 0.76 | 0.69 | 1.76 | 0.81 | 0.93 | 0.75 | 0.70 | 1.52 |

1. $D_A$ contains 105 observations (50%) and is used only to train the source forest and construct the structural metrics.

2. $D_B$ contains 35 observations (16.7%) and provides the limited target-task labels used for adaptation. The source trees remain fixed. Baseline models are trained directly on the same DB subset.

3. $D_T$ contains 70 observations (33.3%) and is used only for final evaluation. All target-model hyperparameters are selected without access to DT, and performance is reported using $R^2$.

For classification, preprocessing and the tree ensemble were refitted within each inner-training fold. In the source-to-target experiment, the source forest was fitted on $D_A$ and kept fixed, while target-model hyperparameters were selected by 2-fold cross-validation on $D_B$. This protocol prevents target-test information from influencing either the source partition or model selection. It therefore evaluates whether a representation learned from task A remains useful when only a small labeled sample from task B is available.

The selected synthetic dataset is a multidimensional Archimedean spiral. A parameter $t$ is randomly sampled from the uniform distribution over the interval from 0 to $6\pi$, and the four informative coordinates follow the smooth embedding $x = (t \sin t, t \cos t, t \sin 2t, t \cos 2t)$. Four additional Gaussian noise coordinates are appended, giving eight input features in total. The scale of each noise coordinate is matched to the corresponding informative coordinate using $\sigma_i = \text{median}(|X_i - \text{median}(X_i)|)$. The additional coordinates act as nuisance variables for the source task and make ambient Euclidean neighborhoods less reliable, while the informative observations remain ordered along one continuous curved manifold.

For the Linear A source task, the response is $y_A = 3t + 2$. The Simple B target response is $y_B = 0.5y_A + 0.3\|x\|_2 + \varepsilon$, where $\varepsilon$ is independent Gaussian noise with standard deviation 0.1. The target therefore retains a substantial source-task component but also depends directly on feature-space geometry. This construction is useful for the present purpose because it does not predetermine that either a purely structural distance or a purely Euclidean distance must be optimal.

The mixing coefficient $\alpha$ controls the Euclidean contribution: $\alpha = 0$ gives a purely tree-structural model, whereas $\alpha > 0$ combines the transferred structure with direct feature-space geometry. For the Grad model, $\alpha$ is selected from the set {0.0, 0.25, 0.50, 0.75} using the leave-one-out training loss. For KRR, $\lambda$ is selected from 15 logarithmically spaced values between $10^{-10}$ and $10^4$. All proposed models and baselines are tuned under the same data-access restrictions.

Table 4 reports means and standard deviations over all 12 independent experiments, no best-seed or best-split selection is used. All competing methods are presented depending on the type of kernel regression (KRR or NW) when applicable. Grad is evaluated only with NW because its trainable metric parameters are optimized through the NW leave-one-out objective.

Table 4

**$R^2$ (mean ± standard deviation) for the source-to-target structure-reuse task**

| Method | KRR | NW |
|---|---|---|
| *s* | 0.45 ± 0.07 | 0.85 ± 0.04 |
| ρ | 0.80 ± 0.05 | 0.88 ± 0.03 |
| *r* | 0.89 ± 0.04 | 0.86 ± 0.04 |
| Grad (*w*) | – | 0.90 ± 0.02 |
| RBF | –0.06 ± 0.13 | 0.08 ± 0.10 |
| LAPL | 0.43 ± 0.07 | 0.16 ± 0.08 |
| R+L | 0.35 ± 0.09 | 0.15 ± 0.09 |
| KeRF | –0.13 ± 0.14 | 0.84 ± 0.03 |
| ET | 0.70 ± 0.05 | |
| KNN | 0.03 ± 0.11 | |

The strongest baseline is KeRF with NW, which reaches $R^2$ 0.84. Several proposed configurations improve on this reference: *s* with NW reaches 0.85, ρ with NW reaches 0.88, *r* with KRR reaches 0.89, and Grad reaches 0.90. The absolute improvement of Grad over the strongest baseline is 0.06. Expressed in terms of unexplained variance, $1 - R^2$ decreases from 0.16 to 0.10, a reduction of 37.5%. This is a practically meaningful gain in a setting where only 35 target-labeled observations are available for adaptation.

The results also reveal a useful distinction between fixed and learned structural representations. The fixed LCA-Depth metric *r* is strongest with KRR and $\alpha = 0$, indicating that the source forest alone captures most of the transferable ordering. The table does not show uniform dominance of every proposed configuration, which is important for interpreting the result. For example, *s* combined with KRR reaches only 0.45, whereas *s* with NW reaches 0.85; conversely, *r* is stronger with KRR than with NW. The benefit therefore depends on compatibility between the structural metric and the prediction rule. Grad provides the highest mean score, but the fixed *r* configuration is close at 0.89. The evidence consequently supports a family of tree-induced kernels and an adaptive hybrid model, not a claim that one metric must dominate on every task.

The real-data classification experiments show a more heterogeneous but still favorable pattern. The proposed structural kernels improve over classical geometric baselines on several datasets and often outperform the standard leaf-coincidence kernel. This suggests that terminal-leaf equality is not the only useful signal contained in a trained forest: the depth and organization of the paths also contribute to predictive performance. At the same time, the results are not uniformly dominated by a single structural metric. This is expected, since different datasets may favor different levels of locality: leaf coincidence is very restrictive, path-based metrics introduce graded similarity, and the weighted-depth model can adapt the contribution of shallow and deep splits.

The source-to-target regression experiment supports the central motivation for using tree-induced structure as a reusable representation. A forest trained only on task A remains informative for task B under limited target supervision, and the strongest results are obtained when the transferred organization is retained while a controlled Euclidean contribution is allowed.

The scope of this conclusion is deliberately limited. The experiment uses a controlled synthetic relationship, 210 observations, and a known source-target construction; it does not reproduce the full variety of real transfer-learning shifts. The reported values are averages over 12 complete independent repetitions. Finally, positive semidefiniteness guarantees that the proposed functions are valid kernels, not that every forest partition will be useful. Performance depends on whether the source forest captures structure that is genuinely relevant to the target task.

## Conclusion

This paper introduced a family of tree-induced kernel models that combine structural information from random forests with Euclidean geometry in the original feature space. The construction is based on several tree-defined distances and similarities: shared-leaf coincidence, path divergence, an inverse-depth LCA metric, and a learnable weighted-depth distance. For these constructions, we established positive semidefiniteness of the associated exponential and mixed kernels under forest aggregation.

The framework turns each structural metric into a distinct kernel model that can be used with Nadaraya–Watson prediction or KRR. Empirical results support its usefulness in supervised classification on real datasets and in a controlled source-to-target regression experiment that tests reuse of source-trained tree structure for a related target response. In these settings, structural information beyond leaf equality often improves over the standard random-forest proximity kernel and classical geometric baselines.

Future work will apply the proposed framework to real transfer-learning and domain-adaptation problems in tabular-data settings. A possible theoretical extension is to use the learned tree metrics to handle auxiliary feature vectors in multiple-instance learning. Another interesting direction is to study the sensitivity of the models to tree depth and bandwidth selection, and further develop adaptive depth-dependent weighting schemes. Such extensions would make it possible to calibrate the model to particular datasets while preserving its connection with positive semidefinite kernel theory.

## REFERENCES


1. **Breiman L.** Random Forests. *Machine Learning*, 2001, Vol. 45, Pp. 5–32. DOI: 10.1023/A:1010933404324

2. **Biau G., Scornet E.** A random forest guided tour. *TEST*, 2016, Vol. 25, Pp. 197–227. DOI: 10.1007/s11749-016-0481-7

3. **Geurts P., Ernst D., Wehenkel L.** Extremely randomized trees. *Machine Learning*, 2006, Vol. 63, Pp. 3–42. DOI: 10.1007/s10994-006-6226-1

4. **Scornet E.** Random Forests and Kernel Methods. *IEEE Transactions on Information Theory*, 2016, Vol. 62, No. 3, Pp. 1485–1500. DOI: 10.1109/TIT.2016.2514489

5. **Lin Y., Jeon Y.** Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association*, 2006, Vol. 101, No. 474, Pp. 578–590. DOI: 10.1198/016214505000001230

6. **Biau G., Devroye L., Lugosi G.** Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 2008, Vol. 9, Pp. 2015–2033.

7. **Biau G.** Analysis of a random forests model. *Journal of Machine Learning Research*, 2012, Vol. 13, Pp. 1063–1095.

8. **Iakovidis I., Arcozzi N.** Improved convergence rates for some kernel random forest algorithms. *Mathematics in Engineering*, 2024, Vol. 6, No. 2, Pp. 305–338. DOI: 10.3934/mine.2024013

9. **Shi T., Horvath S.** Unsupervised Learning with Random Forest Predictors. *Journal of Computational and Graphical Statistics*, 2006, Vol. 15, No. 1, Pp. 118–138. DOI: 10.1198/106186006X94072

10. **Mantero A., Ishwaran H.** Unsupervised random forests. *Statistical Analysis and Data Mining*, 2021, Vol. 14, No. 2, Pp. 144–167. DOI: 10.1002/sam.11498

11. **Sage A.J., Genschel U., Nettleton D.** Tree aggregation for random forest class probability estimation. *Statistical Analysis and Data Mining*, 2020, Vol. 13, No. 2, Pp. 134–150. DOI: 10.1002/sam.11446

12. **Meinshausen N.** Quantile Regression Forests. *Journal of Machine Learning Research*, 2006, Vol. 7, No. 35, Pp. 983–999.

13. **Rhodes J.S., Cutler A., Moon K.R.** Geometry- and Accuracy-Preserving Random Forest Proximities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, Vol. 45, No. 9, Pp. 10947–10959. DOI: 10.1109/TPAMI.2023.3263774

14. **Laabs B.-H., Westenberger A., König I.R.** Identification of representative trees in random forests based on a new tree-based distance measure. *Advances in Data Analysis and Classification*, 2024, Vol. 18, Pp. 363–380. DOI: 10.1007/s11634-023-00537-7

15. **Xiong C., Johnson D., Xu R., Corso J.J.** Random forests for metric learning with implicit pairwise position dependence. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, Pp. 958–966. DOI: 10.1145/2339530.2339680

16. **Bicego M., Cicalese F.** Computing Random Forest-distances in the presence of missing data. *ACM Transactions on Knowledge Discovery from Data*, 2024, Vol. 18, No. 7, Art. no. 180. DOI: 10.1145/3656345

17. **Lee S., Lee C., Mun K.G., Kim D.** Decision Tree Algorithm Considering Distances Between Classes. *IEEE Access*, 2022, Vol. 10, Pp. 69750–69756. DOI: 10.1109/ACCESS.2022.3187172

18. **Bicego M., Cicalese F.** An Interesting Property of Random Forest Distances with Respect to the Curse of Dimensionality. In: *Structural, Syntactic, and Statistical Pattern Recognition* (eds. A. Torsello, L. Rossi, L. Cosmo, G. Minello), 2025, Vol. 15444, Pp. 188–198. DOI: 10.1007/978-3-031-80507-3_19

19. **Bicego M., Cicalese F.** On the Good Behaviour of Extremely Randomized Trees in Random Forest-Distance Computation. *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference* (*ECML PKDD*), 2023, Vol. IV, Pp. 645–660. DOI: 10.1007/978-3-031-43421-1_38

20. **Schoenberg I.J.** Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 1938, Vol. 44, Pp. 522–536. DOI: 10.1090/S0002-9947-1938-1501980-0

21. **Kelly M., Longjohn R., Nottingham K.** The UCI Machine Learning Repository [online]. Available: https://archive.ics.uci.edu (Accessed 30.08.2026).

## INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

**Maksim I. Goriunov**
**Горюнов Максим Юрьевич**
E-mail: goryunov_myu@spbstu.ru
ORCID: https://orcid.org/0009-0005-6368-0001

**Andrei V. Konstantinov**
**Константинов Андрей Владимирович**
E-mail: andrue.konst@gmail.com
ORCID: https://orcid.org/0000-0002-1542-6480

**Fedor A. Novikov**
**Новиков Федор Александрович**
E-mail: fedornovikov51@gmail.com

*Submitted: 27.06.2026; Approved: 07.09.2026; Accepted: 17.09.2026.*
*Поступила: 27.06.2026; Одобрена: 07.09.2026; Принята: 17.09.2026.*