Pattern Recognition 177 (2026) 113337

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# InsNet: Deep indefinite spectral kernel network

Yanfang Xue [a,b], Hui Xue [a,b,*], Shipeng Zhu [a,b]

[a] School of Computer Science and Engineering, Southeast University, Nanjing, 210096, China
[b] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing, 211189, China

ARTICLE INFO

*Keywords:*
Deep kernels
Indefinite kernels
Complex-valued representations

ABSTRACT

Deep learning dominated by neural networks (NN) shares profound connections with kernel methods, offering rich opportunities for mutual advancement between these two paradigms. Consequently, researchers tend to combine deep learning with kernel methods, leading to deep kernel learning techniques. In particular, the integration of kernel methods offers a principled approach to mitigate the "black-box" nature of deep models by imposing structural inductive biases on deep models, thereby enhancing their interpretability. Conversely, the integration of hierarchical NN enables kernel methods to adopt hierarchical parameterizations, significantly enhancing their expressive power and broadening their range of applications. Despite these advances, existing deep kernel learning approaches remain constrained by their reliance on positive definite kernels, limiting their expressiveness in capturing complex structures and relationships within the data. To address this issue, we propose deep indefinite spectral kernel networks (**InsNet**), a novel framework that generalizes the conventional Hilbert space formulation by relaxing the positive definiteness constraint, thereby enabling more flexible and expressive modeling of intricate data dependencies. Concretely, an indefinite spectral kernel mapping is first estimated based on the decomposition of a signed measure, comprising both positive and negative definite components with complex-valued representation. The devised mapping is then stacked to construct InsNet, facilitated by a novel initialization scheme. Beyond the architectural innovation, we provide a rigorous theoretical analysis of InsNet, examining its structural properties and generalization bounds. Extensive experiments on synthetic and real-world data demonstrate InsNet's superior capability, underscoring its practical advantages.

## 1. Introduction

Deep learning [1–3] and kernel methods [4–6] have achieved remarkable success by leveraging their unique advantages. However, they each face fundamental limitations. Deep learning models, despite their powerful empirical expressiveness, are commonly regarded as black-box models, lacking process-level interpretability. Conversely, kernel methods are grounded in rigorous mathematical foundations, offering clear structural interpretability, but are inherently shallow and thus struggle to handle complex tasks. Interestingly, these two paradigms exhibit profound connections [7], offering mutual advancement opportunities. Consequently, researchers tend to combine deep learning with kernel methods, leading to deep kernel learning techniques. On the one hand, the integration of kernel methods offers a principled approach to mitigate the "black-box" nature of deep models, thereby enhancing their interpretability. This design imposes structural inductive biases on deep models, enabling them to encode prior assumptions on smoothness, reciprocal patterns, and non-Euclidean relationships, which are difficult to control explicitly in standard deep neural networks (DNNs). Specifically, deep spectral kernels admit a layer-wise spectral interpretation, where each layer can be understood as imposing specific spectral constraints on the learned representation. On the other hand, the introduction of DNNs enables kernel methods to adopt hierarchical parameterizations, significantly enhancing their expressive power and broadening their range of applications.

For deep kernel learning, researchers have devoted extensive efforts to researching and have achieved significant advancements. A pioneering work, Arc-cosine kernels [8], was proposed by introducing recursive kernel maps, where input transformations are iteratively applied $L$ times to emulate NN architectures. Building upon this foundation, researchers have developed increasingly sophisticated deep kernel methods capable of modeling complex data structures and their intricate interactions. For example, Zhang et al. [9] proposed a stacked kernel network with stationary positive definite kernels to capture nonlinear patterns behind data. Xue et al. [10] extended this scheme to non-stationary positive definite kernels for capturing the long-range dependence of data in amore

* Corresponding author.
*E-mail addresses:* hzxyanfang@163.com (Y. Xue), hxue@seu.edu.cn (H. Xue), shipengzhu@seu.edu.cn (S. Zhu).

https://doi.org/10.1016/j.patcog.2026.113337
Received 27 April 2025; Received in revised form 16 February 2026; Accepted 17 February 2026
Available online 20 February 2026

**Table 1**
Commonly presented notations and their definitions.

| Notation | Definition | Notation | Definition |
|---|---|---|---|
| $\mathbb{R}$ | real number space | $\mathbb{C}$ | complex number space |
| $\mathbb{R}^n$ | $n$-dimensional Euclidean space | $\mathbb{C}^n$ | $n$-dimensional complex number space |
| $\mathbb{R}^{m\times n}$ | the space of $m\times n$ real-valued matrix | $\mathbb{C}^{m\times n}$ | the space of $m\times n$ complex-valued matrix |
| $\mathcal{K}$ | Kreĭn space | $\mathcal{F}$ | function space |
| $N$ | sample number | $d$ | sample dimension |
| $\langle\cdot,\cdot\rangle$ | inner product | $k(\cdot,\cdot)$ | kernel function |
| $\mathbb{E}$ | expectation | $\|\cdot\|_p$ | $L_p$ norm |
| $\mathcal{H}$ | Hilbert space | $\mathcal{H}_k$ | RKHS |
| $(\cdot)^\top$ | transposition | $\|\cdot\|_F$ | $F$-norm |

compact way. Furthermore, Xue et al. [11] generalized it to a complex-valued representation to enhance the representational power. These developments demonstrate the promising potential of deep kernel methods in combining the theoretical rigor of kernel methods with the expressiveness of deep learning.

Nevertheless, existing deep spectral kernel methods typically rely on positive definite kernels, a choice inherited from classical kernel constructions based on Bochner's theorem. This constraint restricts the admissible spectral representations to non-negative measures, limiting their ability to mine the complex hierarchical structures and reciprocal relationships within the data. The former denotes interactions between data [12], such as the mutual promotion and suppression in biomedicine [13]. The latter refers to the multi-scale organizational structure [14], such as the ranging local-to-global and low-to-high level information in the sequential data [15] and image data [14]. Remarkably, existing studies have demonstrated that indefinite kernels provide a rich representational capability for modeling these characteristics in a reproducing kernel Kreĭn space. However, Bocher's theorem no longer holds for indefinite kernels, which generally involve both positive and negative components. The fundamental differences between positive definite and indefinite kernels hinder the straightforward transfer of probabilistic sampling and stacking. Therefore, more efforts are required to develop a new framework that can break the positive definiteness constraints in kernel networks.

In this paper, we propose deep Indefinite spectral kernel Network (InsNet), a novel framework that generalizes the deep spectral kernel networks constructed by stacking explicit spectral kernel mappings derived from the Fourier transform of positive definite kernels to indefinite settings. This scheme extends the standard Hilbert space by relaxing the positive definiteness constraint, enabling more expressive modeling of complex hierarchical structures and relationships within the data. Specifically, we first derive an indefinite spectral kernel mapping based on the Bochner theorem [16] and signed measures [17], yielding complex-valued representations. Here, the real and imaginary parts correspond to positive definite and negative definite components, respectively. Then, we stack the indefinite spectral kernel mappings, adhering to the rules of complex number operations, to construct InsNet. Note that a novel initialization scheme is also introduced for the weight matrix, where the real and imaginary parts are initialized with block diagonal and anti-diagonal matrices, respectively. This initialization scheme retains the statistical characteristics of indefinite kernels. Beyond the architectural innovation, we provide a rigorous theoretical analysis of InsNet, examining its structural properties and generalization bounds. Furthermore, we evaluate InsNet through comprehensive experiments on both synthetic and real-world data. Empirical results demonstrate consistent improvements over state-of-the-art deep spectral kernel methods across all evaluation metrics.

The remainder of this paper is organized as follows. In Section 2, we introduce the notation, preliminary knowledge, and related works. We provide the details of the proposed InsNet in Section 3. In Section 4, we theoretically analyze our InsNet in terms of components and generalization. In Section 5, we conduct a set of experiments on synthetic and real-world datasets, indicating the practical advantages of InsNet. Finally, we simply conclude this paper in Section 6.

## 2. Preliminary and related works

### 2.1. Preliminary

This section introduces the necessary preliminary knowledge to better illustrate the proposed InsNet. Throughout this paper, matrices, vectors, and scalars are denoted by bold capital letters (*e.g.,* $\boldsymbol{X}$), bold lower-case letters (*e.g.,* $\boldsymbol{x}$) and lower-case letters (*e.g.,* $x$), respectively. A complex number $\boldsymbol{z}\in\mathbb{C}^n$ is represented as $\boldsymbol{z}=\boldsymbol{u}+i\boldsymbol{v}$ with a real part $\boldsymbol{u}$ and an imaginary part $\boldsymbol{v}$. $\boldsymbol{z}^*=\boldsymbol{u}-i\boldsymbol{v}$ denotes the complex conjugate of $\boldsymbol{z}$. For any two complex numbers $\boldsymbol{z}_1=\boldsymbol{u}_1+i\boldsymbol{v}_1, \boldsymbol{z}_2=\boldsymbol{u}_2+i\boldsymbol{v}_2\in\mathbb{C}^n$, $\boldsymbol{z}_1+\boldsymbol{z}_2=(\boldsymbol{u}_1+\boldsymbol{u}_2)+i(\boldsymbol{v}_1+\boldsymbol{v}_2)$, $\boldsymbol{z}_1^\top\boldsymbol{z}_2=(\boldsymbol{u}_1^\top\boldsymbol{u}_2-\boldsymbol{v}_1^\top\boldsymbol{v}_2)+i(\boldsymbol{u}_1^\top\boldsymbol{v}_2+\boldsymbol{v}_1^\top\boldsymbol{u}_2)$. A set of commonly presented notations is summarized in Table 1.

Kernel methods are well known to suffer from scalability issues due to high memory and computational costs. To tackle this problem, the random Fourier feature scheme has been developed to approximate the kernel function using explicit kernel mapping by Rahimi & Recht [18].

**Theorem 1** (Bochner's Theorem [16])**.** *A continuous and stationary kernel $k(\boldsymbol{x},\boldsymbol{x}')=k(\boldsymbol{\tau}), \boldsymbol{\tau}=\boldsymbol{x}-\boldsymbol{x}'$ on $\mathbb{R}^d$ is positive definite if and only if it can be formulated as: $k(\boldsymbol{\tau})=\int_{\mathbb{R}^d}e^{i\boldsymbol{\omega}^\top\boldsymbol{\tau}}d\mu(\boldsymbol{\omega})$, where $\mu$ is a bounded non-negative measure on $\mathbb{R}^d$.*

Based on Bocher's theorem, if $\mu$ is absolutely continuous with respect to the Lebesgue measure, *i.e.,* $d\mu(\boldsymbol{\omega})=s(\boldsymbol{\omega})d\boldsymbol{\omega}$, there exists a one-to-one correspondence between the kernel $k(\boldsymbol{\tau})$ and the spectral density $s(\boldsymbol{\omega})$. such that:

$$k(\boldsymbol{\tau})=\int_{\mathbb{R}^d}s(\boldsymbol{\omega})e^{i\boldsymbol{\omega}^\top\boldsymbol{\tau}}d\boldsymbol{\omega},\quad s(\boldsymbol{\omega})=\int_{\mathbb{R}^d}k(\boldsymbol{\tau})e^{-i\boldsymbol{\omega}^\top\boldsymbol{\tau}}d(\boldsymbol{\tau}). \tag{1}$$

Theorem 1 establishes a bijective correspondence between a stationary kernel $k(\boldsymbol{\tau})$ and its spectral density $s(\omega)$, which is associated with a probability density $p(\omega)$. As a result, the stationary kernel in Theorem 1 can be rewritten as follows:

$$k(\boldsymbol{\tau})=\mathbb{E}_{\boldsymbol{\omega}\sim p(\cdot)}\Big[\exp(i\boldsymbol{\omega}^\top\boldsymbol{x})\exp(i\boldsymbol{\omega}^\top\boldsymbol{x}')^*\Big]. \tag{2}$$

Since the kernel is a real-value function, removing the imaginary part and using Monte Carlo sampling, the stationary kernel can be approximated as:

$$k(\boldsymbol{\tau})\approx\frac{1}{M}\sum_{m=1}^{M}\cos(\boldsymbol{\omega}_m^\top\boldsymbol{x})\cos(\boldsymbol{\omega}_m^\top\boldsymbol{x}')=\langle\Phi(\boldsymbol{x}),\Phi(\boldsymbol{x}')\rangle, \tag{3}$$

where $\Phi(\boldsymbol{x})=\frac{1}{\sqrt{M}}\Big[\cos(\boldsymbol{\omega}_1^\top\boldsymbol{x}),\ldots,\cos(\boldsymbol{\omega}_M^\top\boldsymbol{x})\Big]^\top$ is called random Fourier features. $\{\boldsymbol{\omega}_m\}_{m=1}^M$ are the frequencies and are sampled from $p(\boldsymbol{\omega})$ independently.

**Definition 1** (Positive, Negative, and Indefinite Kernel [19])**.** Let $X$ be a nonempty set. A function $k: X\times X\to\mathbb{R}$ is a positive definite kernel if and only if

$$\sum_{i,j=1}^{N}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\geq 0 \tag{4}$$

for all $N \in \mathbb{N}$, $\{\boldsymbol{x}_i\}_{i=1}^N \subset X$ and $\{\alpha_i\}_{i=1}^N \subset \mathbb{R}$. We call the function $k$ is a negative definite kernel if and only if it is Hermitian and

$$\sum_{i,j=1}^{N} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq 0 \tag{5}$$

for all $N \in \mathbb{N}$, $\{\boldsymbol{x}_i\}_{i=1}^N \subset X$ and $\{\alpha_i\}_{i=1}^N \subset \mathbb{R}$ with $\sum_{i=1}^N \alpha_i = 0$. Otherwise, it is called an indefinite kernel.

**Definition 2** (Kreĭn Space [20])**.** An inner product space $(\mathcal{K}, \langle\cdot,\cdot\rangle_{\mathcal{K}})$ is a Kreĭn space if there exist two Hilbert space $\mathcal{H}_+$, $\mathcal{H}_-$ spanning $\mathcal{K}$ such that:

- All $f \in \mathcal{K}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$.
- For any $f, g \in \mathcal{K}$, $\langle f, g\rangle_{\mathcal{K}} = \langle f_+, g_+\rangle_{\mathcal{H}_+} - \langle f_-, g_-\rangle_{\mathcal{H}_-}$.

### 2.2. Related works

#### 2.2.1. Indefinite kernel learning

Indefinite kernels, as a generalization of positive definite kernels, have gained research interest due to their flexibility and promising performance in exploring non-Euclidean metrics [21,22]. For example, Münch et al. [23] proposed a complex-valued embedding framework for generic proximity data, in which pairwise relations, including indefinite and asymmetric similarities, are represented through inner products in a complex Hilbert space. Learning indefinite kernels in RKKS is often computationally expensive and challenging to apply directly to large-scale learning tasks. To enhance the scalability of indefinite kernels, researchers have developed various approaches. For instance, Gisbrecht et al. [24] proposed an integrative combination of Nyström approximation, potential double centering, and eigenvalue correction to obtain valid kernel matrices at linear costs in the number of samples. Subsequently, by leveraging Nyström approximation, researchers tend to seek a low-rank representation to approximate indefinite kernels in a data-dependent way [25–28]. The core idea of these methods is to approximate the original kernel matrix using a small-sized matrix, which is calculated by a subset of training samples or eigenvalues.

In addition, the random Fourier feature (RFF) is also a widely used method to approximate an indefinite kernel in a data-independent way. Pennington et al. [29] first proposed to approximate stationary indefinite kernels with the Gaussian mixture model and introduced spherical random Fourier features, delivering a compact approximation to polynomial kernels for data on the unit sphere. Based on [29], Liu et al. [30] proposed a double-infinite Gaussian mixture model in random Fourier feature by placing the Dirichlet process prior, which takes full advantage of high flexibility on the number of components and has the capability of approximating indefinite kernels on a wide scale. Additionally, Liu et al. [31] transformed the positive decomposition of indefinite kernels to measure decomposition, developing the RFF-based algorithm for the indefinite kernels. Furthermore, Luo et al. [32] proposed the generalized orthogonal random Fourier features, an unbiased estimation with lower variance.

#### 2.2.2. Deep kernel learning

Deep kernel learning techniques are propelling the field toward a more principled integration of NNs and kernel methods, effectively addressing the inherent limitations of traditional kernel approaches, particularly their architectural and scalability challenges. One strategy integrates a deep module, such as DBN [33], deep GP (DGP) [34], and DNNs [35], as the front-end of a kernel to form a synergy model. This framework was subsequently generalized to a structured kernel interpolation framework [36], which integrates the inducing point method with structure-exploiting techniques to derive a sparse approximation of the original kernel. Wilson et al. [37] reformulated the framework using stochastic variational inference, introducing stochastic variational deep kernel learning (SV-DKL). Furthermore, Matias et al. [38] introduced amortized variational deep kernel learning (AVDKL). This approach involves amortized inducing points and a parameter-sharing scheme. Loría & Bhadra [39] introduced a deep kernel posterior learning framework using infinite-variance prior weights, deriving an $\alpha$-stable infinite-width limit with conditionally Gaussian structure, enabling recursive random kernels and effective posterior inference while preserving representation learning in deep Bayesian neural networks. D'Amore [40] proposed a model-level decomposition approach for deep kernel learning that addresses the curse of dimensionality by splitting both the model's operators and network into decomposed parts to improve computational efficiency and scalability.

An alternative strategy constructs deep kernels through the stacking of kernels, implicitly encoding hierarchical feature interactions. In this strategy, spectral kernels, constructed from the inverse Fourier transform, are typically applied to modeling deep kernels. For example, Tian et al. [41] proposed a copula-nested approach based on Yaglom's theorem, which introduces copula networks into the design of the spectral density. Fang et al. [42] developed an innovative generative network framework for stationary kernels, where the sampling distribution is implicitly learned via an NN. In addition, Tonin et al., [43] proposed deep kernel principal component analysis, which found that a negative regularization is helpful for deep feature learning. More deep kernel methods can be found in [44–47].

## 3. Deep indefinite spectral kernel network

This section first presents the overall architecture of the proposed InsNet. Then, each module within the architecture is explicitly provided in the subsequent section.

### 3.1. Overall architecture

The construction of InsNet begins with estimating indefinite spectral kernel mapping through the Bochner theorem and the decomposition of a signed measure. Then, the indefinite kernel is hierarchically incorporated into the neural network via stacking the estimated indefinite spectral kernel mappings with a novel initialization scheme. The stack operation is illustrated in Fig. 1.

Concretely, we notate the estimated indefinite spectral kernel mapping as $\Phi(\cdot)$, which satisfies $k(\boldsymbol{x}, \boldsymbol{x}') \approx \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}')\rangle_{\mathcal{K}}$ for an indefinite kernel. $\mathcal{K}$ is a Kreĭn space. The stack operation (or feedforward) is denoted as $\Psi(\cdot)$, defined as $\Psi^l(\cdot) = \sigma(\boldsymbol{W}^\top \Psi^{l-1}(\cdot))$. $\sigma$ is the activation function. $\boldsymbol{W}$ is the weight matrix, generated by the proposed initialization scheme. Based on the above framework, InsNet with $l$ layers is formulated as follows:

$$\text{InsNet}(\boldsymbol{x}) = \Psi^{l-1}(\cdots \Psi^1(\Phi(\boldsymbol{x}))). \tag{6}$$

The corresponding deep indefinite spectral kernel (DiSK) is defined by:

$$\begin{aligned} k^l(\boldsymbol{x}, \boldsymbol{x}') &= \langle \Psi^{l-1}(\cdots \Psi^1(\Phi(\boldsymbol{x}))), \Psi^{l-1}(\cdots \Psi^1(\Phi(\boldsymbol{x}'))) \rangle \\ &= \langle \text{InsNet}(\boldsymbol{x}), \text{InsNet}(\boldsymbol{x}') \rangle. \end{aligned} \tag{7}$$

### 3.2. Indefinite spectral kernel mapping

As established in Section 2.1, the mapping estimation of a positive definite kernel predominantly depends on sampling from a probability density function $p(\cdot)$, which fundamentally constitutes a non-negative measure under the framework of Bochner's theorem. However, the derivation strategy employed for Eq. (3) is inapplicable to the indefinite case. To bridge this methodological gap between positive definite and indefinite kernels, following the paradigm of [31], the signed measure and its Jordan decomposition are introduced to estimate the indefinite spectral kernel mapping.

**Definition 3** (Signed Measure [17])**.** Let $(X, \mathcal{A})$ be a measurable space, where $X$ is a set and $\mathcal{A}$ is a $\sigma$-algebra of subsets of $X$. A signed measure is a function $\mu : \mathcal{A} \to \mathbb{R}$ that satisfies:

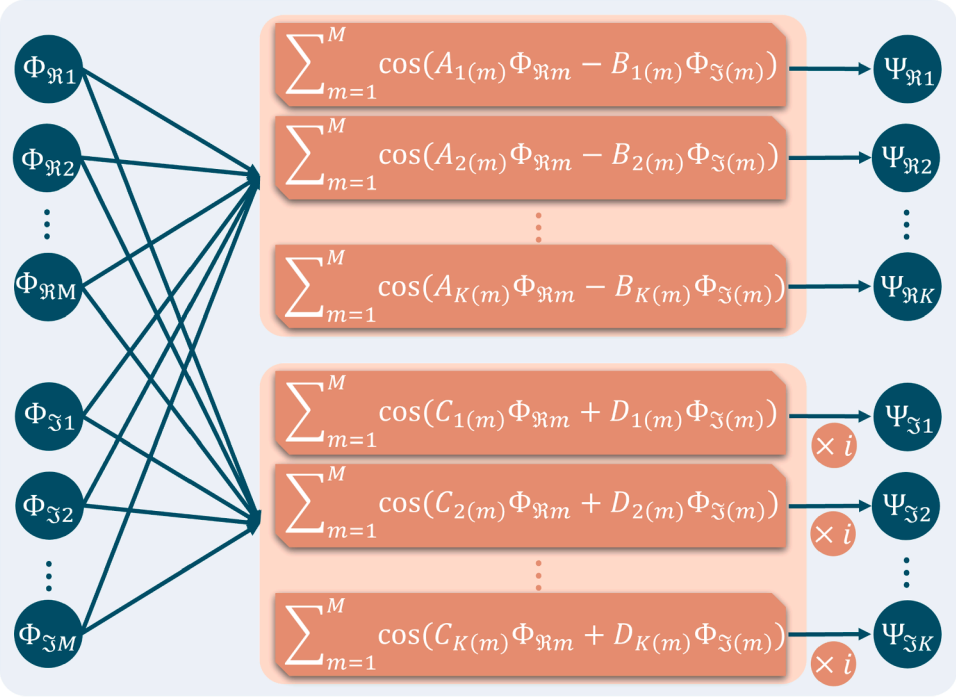


**Fig. 1.** The framework of InsNet for the stack operation.

- $\mu(A) \in \mathbb{R}$ for every $A \in \mathcal{A}$.
- $\delta$-additivity. For any countable collection of disjoint sets $\{A_i\}_{i=1}^{\infty} \subset \mathcal{A}$, $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

**Definition 4** (Jordan Decomposition [17])**.** Given a signed measure $\mu$ on a measurable space $(X, \mathcal{A})$, its Jordan decomposition expresses as:

$$\mu = \mu_+ - \mu_-, \tag{8}$$

where $\mu_+$ is a positive measure, called the positive part of $\mu$. $\mu_-$ is a positive measure, called the negative part of $\mu$.

Based on the Fourier-Stieltjes representation theorem, a stationary indefinite kernel can be defined as:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^\top(\boldsymbol{x}-\boldsymbol{x}')} d\mu(\boldsymbol{\omega}), \tag{9}$$

where $\mu$ denotes a finite signed measure. According to Definition 4, the indefinite kernel can be further formulated by:

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{x}') &= \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_+^\top(\boldsymbol{x}-\boldsymbol{x}')} d\mu_+(\boldsymbol{\omega}_+) - \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_-^\top(\boldsymbol{x}-\boldsymbol{x}')} d\mu_-(\boldsymbol{\omega}_-) \\ &= \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_+^\top(\boldsymbol{x}-\boldsymbol{x}')} s_+(\boldsymbol{\omega}_+) d\boldsymbol{\omega}_+ - \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_-^\top(\boldsymbol{x}-\boldsymbol{x}')} s_-(\boldsymbol{\omega}_-) d\boldsymbol{\omega}_-, \end{aligned} \tag{10}$$

where $\boldsymbol{\omega}_+$ and $\boldsymbol{\omega}_-$ denote integration variables over the same frequency domain $\mathbb{R}^d$, associated with the positive measures $\mu_+$ and $\mu_-$, respectively. $\int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_+^\top(\boldsymbol{x}-\boldsymbol{x}')} s_+(\boldsymbol{\omega}_+) d\boldsymbol{\omega}_+$ and $\int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_-^\top(\boldsymbol{x}-\boldsymbol{x}')} s_-(\boldsymbol{\omega}_-) d\boldsymbol{\omega}_-$ represent two positive definite kernels, corresponding to the spctral densities $s_+(\boldsymbol{\omega}_+)$ and $s_-(\boldsymbol{\omega}_-)$, respectively.

Eq. (10) provides a measure-based representation of the indefinite kernel decomposition. By further simplifying this equation and using Monte Carlo sampling for each positive definite component, the indefinite kernel can be approximated as follows:

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{x}') &= \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_+^\top(\boldsymbol{x}-\boldsymbol{x}')} s_+(\boldsymbol{\omega}_+) d\boldsymbol{\omega}_+ - \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}_-^\top(\boldsymbol{x}-\boldsymbol{x}')} s_-(\boldsymbol{\omega}_-) d\boldsymbol{\omega}_- \\ &\approx \frac{1}{M} \sum_{m=1}^{M} \left[ \cos(\boldsymbol{\omega}_{+,m}^\top \boldsymbol{x}) \cos(\boldsymbol{\omega}_{+,m}^\top \boldsymbol{x}') - \cos(\boldsymbol{\omega}_{-,m}^\top \boldsymbol{x}) \cos(\boldsymbol{\omega}_{-,m}^\top \boldsymbol{x}') \right] \\ &= \frac{1}{M} \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ i \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix}^\top \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}') \\ i \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}') \end{bmatrix} \\ &= \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle, \end{aligned} \tag{11}$$

where $\boldsymbol{\Omega}_+ = [\boldsymbol{\omega}_{+,1}, \ldots, \boldsymbol{\omega}_{+,M}]$ and $\boldsymbol{\Omega}_- = [\boldsymbol{\omega}_{-,1}, \ldots, \boldsymbol{\omega}_{-,M}]$ are the frequency matrices, sampling from $s_+(\boldsymbol{\omega}_+)$ and $s_-(\boldsymbol{\omega}_-)$, respectively. $M$ is the number of samples. The approximation bound, established in [31], provides a preliminary theoretical foundation for the indefinite spectral kernel mapping estimation.

As a result, the indefinite spectral kernel mapping with the complex-valued representation is defined as follows:

$$\Phi(\boldsymbol{x}) = \frac{1}{\sqrt{M}} \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ i \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{M}} \cos(\boldsymbol{\omega}_{+,1}^\top \boldsymbol{x}) \\ \vdots \\ \frac{1}{\sqrt{M}} \cos(\boldsymbol{\omega}_{+,M}^\top \boldsymbol{x}) \\ i \frac{1}{\sqrt{M}} \cos(\boldsymbol{\omega}_{-,1}^\top \boldsymbol{x}) \\ \vdots \\ i \frac{1}{\sqrt{M}} \cos(\boldsymbol{\omega}_{-,M}^\top \boldsymbol{x}) \end{bmatrix}. \tag{12}$$

### 3.3. InsNet

Rewriting Eq. (12) as follows:

$$\boldsymbol{h} = \Phi(\boldsymbol{x}) = \frac{1}{\sqrt{M}} \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ \boldsymbol{0} \end{bmatrix} + i \frac{1}{\sqrt{M}} \begin{bmatrix} \boldsymbol{0} \\ \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix}, \tag{13}$$

where $\boldsymbol{0} \in \mathbb{R}^M$, and all the entries are 0. After that, we stack it to construct InsNet. To ensure the sub-network from the first layer to the $l$th layer ($l \geq 2$) can be rigorously interpreted as an indefinite spectral kernel, the complex-valued weight matrix with a block diagonal and an

anti-diagonal matrix is defined as:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} + i \begin{bmatrix} \boldsymbol{0} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix}, \tag{14}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are real-valued matrices. $\boldsymbol{0}$ denotes an all-zero matrix. Based on Eqs. (13) and (14), the stack operation $\Psi(\cdot) = \sigma(\boldsymbol{W}^\top(\cdot))$ ($\sigma$ is the activation function) can be formulated as follows:

$$\begin{aligned}
&\Psi(\boldsymbol{h}) = \sigma(\boldsymbol{W}^\top \boldsymbol{h}) \\
&= \sigma\left[\left(\begin{bmatrix} \boldsymbol{A}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}^\top \end{bmatrix} + i \begin{bmatrix} \boldsymbol{0} & \boldsymbol{B}^\top \\ \boldsymbol{C}^\top & \boldsymbol{0} \end{bmatrix}\right) * \left(\begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ \boldsymbol{0} \end{bmatrix} + i \begin{bmatrix} \boldsymbol{0} \\ \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix}\right)\right] \\
&= \begin{bmatrix} \sigma\left[\boldsymbol{A}^\top \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) - \boldsymbol{B}^\top \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x})\right] \\ i\sigma\left[\boldsymbol{C}^\top \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) + \boldsymbol{D}^\top \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x})\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma\left[\begin{bmatrix} \boldsymbol{A}^\top & -\boldsymbol{B}^\top \end{bmatrix} \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix}\right] \\ i\sigma\left[\begin{bmatrix} \boldsymbol{C}^\top & \boldsymbol{D}^\top \end{bmatrix} \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix}\right] \end{bmatrix}.
\end{aligned} \tag{15}$$

By defining $\sigma(\cdot)$ as the cosine function, we obtain

$$\Psi(\boldsymbol{h}) = \begin{bmatrix} \cos(\boldsymbol{W}_+^\top \hat{\boldsymbol{h}}) \\ \boldsymbol{0} \end{bmatrix} + i \begin{bmatrix} \boldsymbol{0} \\ \cos(\boldsymbol{W}_-^\top \hat{\boldsymbol{h}}) \end{bmatrix}, \tag{16}$$

where $\hat{\boldsymbol{h}} = \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix}$ is the real-valued representation of indefinite spectral kernel mapping. $\boldsymbol{W}_+^\top = \begin{bmatrix} \boldsymbol{A}^\top & -\boldsymbol{B}^\top \end{bmatrix}$, $\boldsymbol{W}_-^\top = \begin{bmatrix} \boldsymbol{C}^\top & \boldsymbol{D}^\top \end{bmatrix}$ are the weight matrices, corresponding to two different positive measures of the indefinite kernel. The output $\begin{bmatrix} \cos(\boldsymbol{W}_+^\top \hat{\boldsymbol{h}}) \\ \boldsymbol{0} \end{bmatrix} + i \begin{bmatrix} \boldsymbol{0} \\ \cos(\boldsymbol{W}_-^\top \hat{\boldsymbol{h}}) \end{bmatrix}$ is considered as an indefinite spectral kernel mapping and input to the next layer.

As a result, InsNet can be constructed as follows:

$$\text{InsNet}(\boldsymbol{x}) = \Psi^{l-1}(\cdots \Psi^1(\Phi(\boldsymbol{x}))), \tag{17}$$

and the corresponding DiSK with $l$ layers is defined as:

$$\begin{aligned}
k^l(\boldsymbol{x}, \boldsymbol{x}') &= \langle \Psi^{l-1}(\cdots \Psi^1(\Phi(\boldsymbol{x}))), \Psi^{l-1}(\cdots \Psi^1(\Phi(\boldsymbol{x}'))) \rangle \\
&= \langle \text{InsNet}(\boldsymbol{x}), \text{InsNet}(\boldsymbol{x}') \rangle.
\end{aligned} \tag{18}$$

## 4. Analysis of InsNet

In this section, we rigorously analyze the proposed InsNet, including the components, approximation, and generalization error bound.

### 4.1. Positive-definite and negative-definite parts

Based on Definition 2, any $f$ in a Kreĭn space $\mathcal{K}$ admits a decomposition into $f = f_+ + f_-$, where $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$. The inner product in $\mathcal{K}$ is defined as $\langle f, g \rangle_{\mathcal{K}} = \langle f_+, g_+ \rangle_{\mathcal{H}_+} - \langle f_-, g_- \rangle_{\mathcal{H}_-}$ for any $f, g \in \mathcal{K}$. If $\mathcal{H}_+$ and $\mathcal{H}_-$ are two RKHSs, $\mathcal{K}$ is an RKKS that is associated with an indefinite kernel $k$ and admits a Kolmogorov decomposition into two positive definite kernels *i.e.,* $k = k_+ - k_-$.

According to Eqs. (16) and (18), DiSK can be defined and further decomposed by:

$$\begin{aligned}
k^l(\boldsymbol{x}, \boldsymbol{x}') &= \begin{bmatrix} \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}^{l-1}) \\ i \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1}) \end{bmatrix}^\top \begin{bmatrix} \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}'^{l-1}) \\ i \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}'^{l-1}) \end{bmatrix} \\
&= k_+^l(\boldsymbol{x}, \boldsymbol{x}') - k_-^l(\boldsymbol{x}, \boldsymbol{x}'),
\end{aligned} \tag{19}$$

where $\boldsymbol{W}_+^l$ and $\boldsymbol{W}_-^l$ are weight matrices of the $l^{th}$ layer. $\hat{\boldsymbol{h}}^{l-1}$ and $\hat{\boldsymbol{h}}'^{l-1}$ are the real-valued representations of $\Psi^{l-2}(\cdots \Psi^1(\Phi(\boldsymbol{x})))$ and $\Psi^{l-2}(\cdots \Psi^1(\Phi(\boldsymbol{x}')))$, respectively, and

$$\begin{aligned}
k_+^l(\boldsymbol{x}, \boldsymbol{x}') &= \langle \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}^{l-1}), \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}'^{l-1}) \rangle, \\
k_-^l(\boldsymbol{x}, \boldsymbol{x}') &= \langle \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1}), \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}'^{l-1}) \rangle,
\end{aligned} \tag{20}$$

are two stationary positive definite kernels with the RFF representation. Therefore, we deem the DiSK associated with an RKKS, which can be decomposed into two RKHSs, induced by two positive definite kernels.

We denote the RKKS, induced by the kernel $k^l(\boldsymbol{x}, \boldsymbol{x}')$, as $\mathcal{K}^l$. The corresponding two RKHS are denoted as $\mathcal{H}_+^l$ and $\mathcal{H}_-^l$. For $f \in \mathcal{K}^l$, it can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{H}_+^l$, $f_- \in \mathcal{H}_-^l$, and $f_+(\boldsymbol{x}) = \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}^{l-1})$, $f_-(\boldsymbol{x}) = \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1})$. Furthermore, by defining $f(\boldsymbol{x}) = \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}^{l-1}) + \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1})$, $f(\boldsymbol{x}') = \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}'^{l-1}) + \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}'^{l-1})$, the inner product can be written as:

$$\begin{aligned}
\langle f(\boldsymbol{x}), f(\boldsymbol{x}') \rangle &= \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}^{l-1}) \cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}'^{l-1}) - \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1}) \cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}'^{l-1}) \\
&= \langle f_+(\boldsymbol{x}), f_+(\boldsymbol{x}') \rangle - \langle f_-(\boldsymbol{x}), f_-(\boldsymbol{x}') \rangle,
\end{aligned} \tag{21}$$

preserving the property of the inner product in Definition 2.

Reconsidering this inner product and rewriting it as:

$$\langle f(\boldsymbol{x}), f(\boldsymbol{x}') \rangle = \langle f_+(\boldsymbol{x}), f_+(\boldsymbol{x}') \rangle + (-\langle f_-(\boldsymbol{x}), f_-(\boldsymbol{x}') \rangle), \tag{22}$$

where $\langle f_+(\boldsymbol{x}), f_+(\boldsymbol{x}') \rangle$ is associated with a positive definite kernel and corresponds to the feature mappings $\cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}^{l-1})$ and $\cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}'^{l-1})$. Since for any $\boldsymbol{x} \neq \boldsymbol{0}$, $-\langle f_-(\boldsymbol{x}), f_-(\boldsymbol{x}) \rangle = -\cos^2(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1}) \leqslant 0$, $-\langle f_-(\boldsymbol{x}), f_-(\boldsymbol{x}') \rangle$ is connected with an negative definite kernel.

Returning to Eq. (21), it shows that $-\langle f_-(\boldsymbol{x}), f_-(\boldsymbol{x}') \rangle$ corresponds to the feature mappings $\cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1})$ and $\cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}'^{l-1})$. Consequently, $\cos(\boldsymbol{W}_+^{l\top} \hat{\boldsymbol{h}}^{l-1})$ and $\cos(\boldsymbol{W}_-^{l\top} \hat{\boldsymbol{h}}^{l-1})$ can be, respectively, seen as positive definite and negative definite parts of InsNet and hierarchically capture the complex patterns and structures inherent in the data. This advancement is verified in the experiment section.

### 4.2. Theoretical results

This section investigates the generalization ability of InsNet. By analyzing its generalization error bound, we demonstrate that InsNet achieves improved generalization guarantees compared with positive definite stationary spectral kernel-based models.

**Definition 5** (Empirical Rademacher Complexity). Set $\mathcal{F}$ to be a class of uniformly bounded functions. The empirical Rademacher complexity of the class of functions $\mathcal{F}$, $\hat{\mathcal{R}}_N(\mathcal{F})$, is defined as:

$$\hat{\mathcal{R}}_N(\mathcal{F}) = \mathbb{E}_\xi \sup_{f \in \mathcal{F}} | \frac{1}{N} \sum_{i=1}^N \xi_i f(\boldsymbol{x}_i) |, \tag{23}$$

where $\{\boldsymbol{x}_i\}_{i=1}^N \in \mathcal{X}$ are i.i.d samples. $\{\xi_i\}_{i=1}^N$ are random variables.

**Theorem 2.** *Assume the loss function $\ell$ is L-Lipstchitz in $\mathbb{R}^d$ and $\mathcal{F}$ is a hypothesis space. With probability at least $1 - \delta$, the following risk bound holds*

$$\epsilon_{\mathcal{F}} - \hat{\epsilon}_{\mathcal{F}} \leq 4\sqrt{2} L \hat{\mathcal{R}}(\mathcal{F}) + \mathcal{O}(\sqrt{\frac{log 1/\delta}{N}}), \tag{24}$$

*where $\epsilon_{\mathcal{F}}$ denotes the expected risk and $\hat{\epsilon}_{\mathcal{F}}$ denotes the empirical risk.*

By Eq. (12), the RKKS, induced by InsNet, is defined as follows:

$$\begin{aligned}
\mathcal{K} &:= \{\Phi(\cdot) | \boldsymbol{\omega}_{+,m}, \boldsymbol{\omega}_{-,m} \in \mathbb{R}^d\}, \\
\Phi(\boldsymbol{x}) &= \begin{bmatrix} \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}) \\ i \cos(\boldsymbol{\Omega}_-^\top \boldsymbol{x}) \end{bmatrix}.
\end{aligned} \tag{25}$$

As discussed in Section 4.1, the indefinite spectral kernel can be decomposed into two positive definite stationary kernels. Considering the part $k_+$ of the indefinite spectral kernel, it leads to a positive definite stationary spectral kernel. The corresponding plain spectral kernel mapping $\Phi_{spd}(\boldsymbol{x})$ and RKHS $\mathcal{H}_{spd}$ are defined as follows:

$$\begin{aligned}
\Phi_{spd}(\boldsymbol{x}) &= \cos(\boldsymbol{\Omega}_+^\top \boldsymbol{x}), \\
\mathcal{H}_{spd} &:= \{\Phi_{spd}(\cdot) | \boldsymbol{\omega}_{+,m} \in \mathbb{R}^d\}.
\end{aligned} \tag{26}$$

As a result, we have the following theorem considering the empirical Rademacher complexities of these hypothesis spaces $\mathcal{K}$ and $\mathcal{H}_{spd}$.

**Theorem 3.** *Following the notation and considering a normalized training set* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. *Suppose* $\mathcal{X}$ *is compact and* $||\boldsymbol{x}||_2 = 1$ *for any* $\boldsymbol{x}$ *in* $\mathcal{X}$. *The empirical Rademacher complexity of different models is bounded by*

$$\begin{aligned}\hat{\mathcal{R}}_N(\mathcal{K}) &\leq ||\xi||_2||\boldsymbol{w}||_2|\exp(-2||\boldsymbol{\omega}_+||^2) - \exp(-2||\boldsymbol{\omega}_-||^2)| \\ \hat{\mathcal{R}}_N(\mathcal{H}_{spd}) &\leq ||\xi||_2||\boldsymbol{w}||_2\exp(-2||\boldsymbol{\omega}_+||^2).\end{aligned} \tag{27}$$

**Proof.** Based on the definition of empirical Rademacher complexity, we have

$$\begin{aligned}\hat{\mathcal{R}}_N(\mathcal{F}) =& \mathbb{E}_\xi \sup_{f\in\mathcal{F}} | \frac{1}{N}\sum_{i=1}^N \xi_i f(\boldsymbol{x}_i) | \\ =& \mathbb{E}_\xi \sup_{f\in\mathcal{F}} | \frac{1}{N}\sum_{i=1}^N \xi_i \boldsymbol{w}^\top \Phi(\boldsymbol{x}_i) | \\ \leq& \frac{1}{N}\mathbb{E}_\xi \sup_{f\in\mathcal{F}} | \boldsymbol{\xi}^\top(\boldsymbol{w}^\top\boldsymbol{\Phi}) | \\ \leq& \frac{1}{N}||\boldsymbol{\xi}^\top|| \cdot ||\boldsymbol{w}^\top|| \cdot ||\boldsymbol{\Phi}||_F,\end{aligned} \tag{28}$$

where $\boldsymbol{\xi} = [\xi_1; \xi_2; \cdots; \xi_N], \boldsymbol{\Phi} = [\Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2), \ldots \Phi(\boldsymbol{x}_N)]$.

According to Eq. (28), we can observe that the empirical Rademacher complexity is closely related to $||\boldsymbol{\Phi}||_F$. Next, we calculate $||\boldsymbol{\Phi}||_F$ and have

$$\begin{aligned}||\boldsymbol{\Phi}||_F =& (tr(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}))^{\frac{1}{2}} \\ =& (\sum_{i=1}^N \phi(\boldsymbol{x}_i)^\top\phi(\boldsymbol{x}_i))^{\frac{1}{2}} \\ =& (\sum_{i=1}^N k(\boldsymbol{x}_i, \boldsymbol{x}_i))^{\frac{1}{2}}.\end{aligned} \tag{29}$$

We further calculate $\sum_{i=1}^N k(\boldsymbol{x}_i, \boldsymbol{x}_i)$ under two cases, our proposed method and the stationary positive definite case. For our proposed method, we have

$$\begin{aligned}\sum_{i=1}^N k(\boldsymbol{x}_i, \boldsymbol{x}_i) &= \sum_{j=1}^N \left[\mathbb{E}_{\boldsymbol{\omega}_+\sim P_+}[\cos(\boldsymbol{\omega}_+^\top\boldsymbol{x}_j)]^2 - \mathbb{E}_{\boldsymbol{\omega}_-\sim P_-}[\cos(\boldsymbol{\omega}_-^\top\boldsymbol{x}_j)]^2\right] \\ &= \sum_{j=1}^N \frac{1}{2}\left[(1+\exp(-2\|\boldsymbol{\omega}_+\|^2\|\boldsymbol{x}_j\|^2) - (1+\exp(-2\|\boldsymbol{\omega}_-\|^2\|\boldsymbol{x}_j\|^2)\right] \\ &\leq \frac{N}{2}|\exp(-2\|\boldsymbol{\omega}_+\|^2) - \exp(-2\|\boldsymbol{\omega}_-\|^2)|.\end{aligned} \tag{30}$$

For the stationary positive definite case, it can be seen as the special case involving $\boldsymbol{\omega}_+$. Thereby, we have

$$\begin{aligned}\sum_{i=1}^N k_{spd}(\boldsymbol{x}_i, \boldsymbol{x}_i) &= \sum_{j=1}^N \mathbb{E}_{\boldsymbol{\omega}_+\sim P_+}[\cos(\boldsymbol{\omega}_+^\top\boldsymbol{x}_j)]^2 \\ &= \sum_{j=1}^N \frac{1}{2}(1+\exp(-2\|\boldsymbol{\omega}_+\|^2\|\boldsymbol{x}_j\|^2) \\ &= \frac{N}{2}\exp(-2\|\boldsymbol{\omega}_+\|^2).\end{aligned} \tag{31}$$

As a result, we have

$$\begin{aligned}\hat{\mathcal{R}}_N(\mathcal{K}) &\leq \|\xi\|_2\|\boldsymbol{w}\|_2|\exp(-2\|\boldsymbol{\omega}_+\|^2) - \exp(-2\|\boldsymbol{\omega}_-\|^2)| \\ \hat{\mathcal{R}}_N(\mathcal{H}_{spd}) &\leq \|\xi\|_2\|\boldsymbol{w}\|_2\exp(-2\|\boldsymbol{\omega}_+\|^2).\end{aligned} \tag{32}$$

Since $|\exp(-2\|\boldsymbol{\omega}_+\|^2) - \exp(-2\|\boldsymbol{\omega}_-\|^2)| \leq \exp(-2\|\boldsymbol{\omega}_+\|^2)$, guarantee $\hat{\mathcal{R}}_N(\mathcal{K}) \leq \hat{\mathcal{R}}_N(\mathcal{H}_{spd})$. □

By Theorems 2 and 3, it can guarantee that our InsNet has a better generalization ability than the model that is induced by the stationary positive definite kernel.

**Table 2**
The detailed information of the involved dataset. Specifically, the input size denotes the number of time points or features for the time-series classification task.

| Dataset | Type | Input size | Train.Data | Test.Data | Class |
|---|---|---|---|---|---|
| FordA | Sensor | 500 | 3601 | 1320 | 2 |
| FordB | Sensor | 500 | 3636 | 810 | 2 |
| Wine | Spectro | 234 | 57 | 54 | 2 |
| ECG200 | ECG | 96 | 100 | 100 | 2 |
| ECG5000 | ECG | 140 | 500 | 4500 | 5 |
| Herring | Image | 512 | 64 | 64 | 2 |
| Ham | Spectro | 431 | 109 | 105 | 2 |
| Proximal | Image | 80 | 400 | 139 | 6 |

**Table 3**
The average results of time series classification with 20 repetitions. The best results are highlighted in **bold** and the suboptimal results are highlighted in underline.

| Dataset | InsNet | SRFF | DSKN | ASKL | CosNet | CokeNet |
|---|---|---|---|---|---|---|
| ECG200 | **91.10**$_{\pm 1.25}$ | 87.95$_{\pm 2.19}$ | 88.85$_{\pm 3.25}$ | 88.90$_{\pm 0.62}$ | 90.60$_{\pm 1.05}$ | 90.00$_{\pm 1.65}$ |
| ECG5000 | **94.07**$_{\pm 0.01}$ | 93.25$_{\pm 0.43}$ | 92.46$_{\pm 0.17}$ | 92.75$_{\pm 0.01}$ | 93.68$_{\pm 0.04}$ | – |
| FordA | **81.36**$_{\pm 0.00}$ | 80.34$_{\pm 0.76}$ | 80.55$_{\pm 0.46}$ | 72.66$_{\pm 3.45}$ | 80.33$_{\pm 0.00}$ | 72.29$_{\pm 0.00}$ |
| FordB | **71.27**$_{\pm 1.09}$ | 69.01$_{\pm 1.44}$ | 69.60$_{\pm 1.09}$ | 64.63$_{\pm 8.03}$ | 70.41$_{\pm 1.06}$ | – |
| Ham | **74.33**$_{\pm 1.62}$ | 72.24$_{\pm 12.06}$ | 70.81$_{\pm 9.67}$ | 68.24$_{\pm 0.69}$ | 71.29$_{\pm 2.21}$ | 73.57$_{\pm 3.22}$ |
| Herring | **65.55**$_{\pm 11.94}$ | 57.58$_{\pm 4.18}$ | 57.73$_{\pm 9.89}$ | 60.39$_{\pm 3.15}$ | 65.00$_{\pm 7.30}$ | 62.34$_{\pm 10.77}$ |
| Wine | **87.13**$_{\pm 9.92}$ | 77.87$_{\pm 0.00}$ | 75.65$_{\pm 0.00}$ | 68.43$_{\pm 145.65}$ | 83.98$_{\pm 5.51}$ | 78.52$_{\pm 22.16}$ |
| Proximal | **83.98**$_{\pm 0.66}$ | 78.98$_{\pm 4.11}$ | 77.88$_{\pm 6.22}$ | 82.49$_{\pm 9.39}$ | 82.88$_{\pm 7.84}$ | 82.80$_{\pm 1.70}$ |

## 5. Experiment

As discussed previously, InsNet extends the standard Hilbert space formulation to Kreĭn space via relaxing the positive definiteness constraint, thereby enhancing its capacity to capture complex structures within the data. Moreover, the associated DiSK provides an effective means to measure the relationships between data. To assess the efficacy of the proposed InsNet and DiSK, we perform extensive experiments on multiple public datasets and synthetic data. All implementations are based on PyTorch [48] and executed on a workstation with an NVIDIA RTX 3090 GPU, an AMD R7-5700X 3.40GHz 8-core CPU, and 32 GB of memory.

### 5.1. Can InsNet capture the complex structures?

To verify the capability of InsNet in capturing intricate structures, we conduct experiments on two tasks involving data with intricate architectures in the dimensions of time and space: time series classification and image segmentation.

#### 5.1.1. Time series classification

**Dataset**: 8 sub-datasets with default training and testing data splitting from the **UCI Archive** [49] dataset are involved in this experiment, and their statistics are shown in Table 2.

**Compared methods**: We compare InsNet with mainstream deep spectral kernel methods: **DSKN** [10]: Deep Spectral Kernel Network, which embedded non-stationary spectral kernel into deep architectures; **SRFF** [9]: Stacked Kernel Network, which stacks random Fourier features with stationary kernels; **ASKL** [44]: Automated Spectral Kernel Learning, which incorporates the process of finding suitable kernels and model training in a learning framework; **CosNet** [11]: A Generalized Spectral Kernel Network, which generalizes spectral kernel mapping in real number domain to complex number domain; **CokeNet** [41]: Copula-Nested Spectral Kernel Network, which introduces copula networks into the design of the spectral density.

**Results**: The results of time series classification in Table 3 reveal the following insights: (1) Our proposed InsNet consistently outperforms

mainstream deep spectral kernel methods. This is attributed to its hierarchical indefinite spectral kernel mappings, which enhance representational ability, coupled with the inherent flexibility of Kreĭn space compared to conventional positive definite kernels in Hilbert space. (2) CosNet achieves generally subpar performance, likely due to its non-stationarity and the complex-valued representation that inherently exists in the time-sequential data. (3) SRFF exhibits limited performance, which is caused by its stationarity that relies solely on the distance between data, ignoring the long-range dependence of the data. These findings collectively underscore InsNet's efficacy in extracting the complex patterns of time series data.

#### 5.1.2. Image segmentation

As discussed in Section 4.1, InsNet consists of positive and negative definite components, enabling the learning of complementary feature representations. To empirically verify this, we extend InsNet to convolutional neural networks (CNNs), proposing the indefinite spectral convolutional network (InsCoNet), and evaluate its performance on a segmentation task using two public datasets. Like InsNet, the construction of InsCoNet includes two steps: estimation of indefinite spectral convolutional mapping and stacking of convolutional mappings.

Specifically, the indefinite spectral convolutional mapping is defined as follows:

$$\Phi_{con}(\boldsymbol{x}) = \frac{1}{\sqrt{M}}\begin{bmatrix}\cos(\boldsymbol{\Omega}_+^\top * \boldsymbol{x}) \\ i\cos(\boldsymbol{\Omega}_-^\top * \boldsymbol{x})\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{M}}\cos(\boldsymbol{\omega}_{+,1}^\top * \boldsymbol{x}) \\ \vdots \\ \frac{1}{\sqrt{M}}\cos(\boldsymbol{\omega}_{+,M}^\top * \boldsymbol{x}) \\ i\frac{1}{\sqrt{M}}\cos(\boldsymbol{\omega}_{-,1}^\top * \boldsymbol{x}) \\ \vdots \\ i\frac{1}{\sqrt{M}}\cos(\boldsymbol{\omega}_{-,M}^\top * \boldsymbol{x})\end{bmatrix} \tag{33}$$

where $\boldsymbol{\Omega}_+$ and $\boldsymbol{\Omega}_-$ are filters, $M$ is the number of filters.

The stack operation with the matrix form is defined as follows:

$$\begin{aligned}\Psi_{con}(\boldsymbol{h}) &= \sigma(\boldsymbol{W}^\top \boldsymbol{h}) \\ &= \sigma\left[\left(\begin{bmatrix}\boldsymbol{A}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}^\top\end{bmatrix} + i\begin{bmatrix}\boldsymbol{0} & \boldsymbol{B}^\top \\ \boldsymbol{C}^\top & \boldsymbol{0}\end{bmatrix}\right)\right. \\ &\left.* \left(\begin{bmatrix}\cos(\boldsymbol{\Omega}_+^\top * \boldsymbol{x}) \\ \boldsymbol{0}\end{bmatrix} + i\begin{bmatrix}\boldsymbol{0} \\ \cos(\boldsymbol{\Omega}_-^\top * \boldsymbol{x})\end{bmatrix}\right)\right] \\ &= \begin{bmatrix}\sigma\left[\boldsymbol{A}^\top * \cos(\boldsymbol{\Omega}_+^\top * \boldsymbol{x}) - \boldsymbol{B}^\top * \cos(\boldsymbol{\Omega}_-^\top * \boldsymbol{x})\right] \\ i\sigma\left[\boldsymbol{C}^\top * \cos(\boldsymbol{\Omega}_+^\top * \boldsymbol{x}) + \boldsymbol{D}^\top * \cos(\boldsymbol{\Omega}_-^\top * \boldsymbol{x})\right]\end{bmatrix}\end{aligned} \tag{34}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are the filters. Moreover, InsCNet with $l$ layers is defined as:

$$InsCNet(\boldsymbol{x}) = \Psi_{con}^{l-1}(\ldots\Psi^1(\Phi_{con}^1(\boldsymbol{x}))). \tag{35}$$

**Dataset**: CrackForest dataset, an annotated database of road cracks consisting of 156 images, 118 of which have the corresponding ground truth. We select 97 images with binary masks from 118 images, of which 77 are randomly selected for training and 20 for testing. The segmentation subset of **PASCAL VOC** dataset, including 2913 images with an 80-20 split ratio allocated for model training and testing, respectively.

**Experimental Setting**: The model in this experiment consists of an encoder $F_{en}$ and a decoder $F_{de}$. Specifically, the encoder is an extension of our InsNet in a convolutional form with $L$ layers. It encodes the image into two parts, corresponding to positive definite and negative definite components discussed in Section 4.1. The operation is defined as $F_{en} : \mathbb{R}^{C_0\times H_0\times W_0} \to \mathbb{R}^{C_L\times H_L\times W_L}$, where $C_0$ is the number of channels ($C_0 = 3$ for the RGB image), $H_0$ and $W_0$ denote the height and width of the input image, respectively. $C_L \times H_L \times W_L$ is the size of output features. The first $\frac{C_L}{2}$ channels correspond to the positive definite features and the remaining $\frac{C_L}{2}$ channels refer to the negative definite features. The decoder is the deconvolutional network with $L$ layers, decoding the positive definite and negative definite features, respectively. The operation is defined as $F_{de} : \mathbb{R}^{\frac{C_L}{2}\times H_L\times W_L} \to \mathbb{R}^{1\times H_0\times W_0}$. In this experiment, we set $L = 3$, and ReLU is selected as the activation in the decoder. The model is trained using the Adam [50] algorithm with cross-entropy loss. The learning rate is equal to 0.001.

**Table 4**
Segmentation results. The best results are highlighted in **bold**. (↑) indicates the larger, the better.

| Dataset | Metrics | InsNet | DSKN | CosNet | 1-InsNet |
|---|---|---|---|---|---|
| CrackForest | Dice (↑) | **0.5113** | 0.5021 | 0.4869 | 0.2464 |
| | IoU (↑) | **0.3660** | 0.3546 | 0.3422 | 0.1492 |
| PASCAL VOC | Dice (↑) | **0.5550** | 0.5391 | 0.5357 | 0.3537 |
| | IoU (↑) | **0.4116** | 0.3955 | 0.3937 | 0.2304 |

**Results**: To quantitatively assess the effectiveness of our InsCoNet, we compare InsCoNet with the convolutional variant of three baselines, including DSKN, CosNet, and one-layer InsNet (1-InsNet), under two evaluation metrics, Dice Coefficient (Dice) and Intersection over Union (IoU). Here, DSKN contains two non-linear maps in each layer, and CosNet is composed of real and imaginary parts. Different parts within DSKN or CosNet are applied to encode the image into two parts, respectively. One-layer InsNet, which includes negative definite components, is a 'shallow' model. The results are reported in Table 4. We can observe that (1) InsNet outperforms DSKN and CosNet, which are induced by positive kernels. This is attributed to the negative definite components. (2) InsNet is significantly superior to 1-InsNet, highlighting the necessity of developing deep indefinite kernels.

**Visualization**: The segmentation results, visualized in Fig. 2, demonstrate the following information: (1) The predictions achieve remarkable alignment with the actual crack locations, indicating the effectiveness of the proposed InsNet. (2) InsNet shows outstanding performance in capturing sharp turns or intricate details of the crack pattern, as highlighted by the red rectangle in Fig. 2(e). This is because of the introduction of negative definite components, which enables us to supervise the information from different hierarchies. (3) InsNet, in its convolution form, not only segments the crack structures but also detects the additional texture, *i.e.,* the cracks at the bottom in Fig. 2(c) and the cracks at the top in Fig. 2(d). This demonstrates the robustness of our approach in separating the abnormal structure from the background.

### 5.2. Can DiSK capture reciprocal relationships?

In practice, obtaining ground-truth correlations to directly validate the ability of a model to capture complex relationships remains challenging. In this part, we evaluate DiSK through the simulation experiment and further the real-world task (*i.e.,* the functional brain network estimation for brain disorder classification).

#### 5.2.1. Simulation experiment

Our simulation experiment begins by generating an indefinite covariance matrix (*i.e.,* correlation matrix) $\boldsymbol{\Sigma} \in \mathbb{R}^{5\times5}$ governing the relationships among five random variables $v_1, v_2, \ldots, v_5$. For each variable, we draw 200 independent samples from a distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, resulting in a data matrix $\boldsymbol{V} \in \mathbb{R}^{5\times200}$. We then employ DiSK to estimate the underlying correlation structure, denoted as $\hat{\boldsymbol{\Sigma}}$. In addition, we perform eigenvalue decomposition on both the ground-truth $\boldsymbol{\Sigma}$ and the estimated $\hat{\boldsymbol{\Sigma}}$. All results are visualized in Fig. 3.

Results in Fig. 3 show that $\Sigma$ and $\hat{\Sigma}$ are highly similar, indicating that our InsNet can mine the inherent patterns of the data and explore their complex relationships. In Fig. 3(c), we visualize the eigenvalues of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$, where eig_Sigma (blue dots) and eig_InsNet (red dots) denote the eigenvalues of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$, respectively. This indicates that the proposed approach can effectively capture eigenvalues, whether positive, negative, or near zero, which makes it capable of modeling the various modes of data.

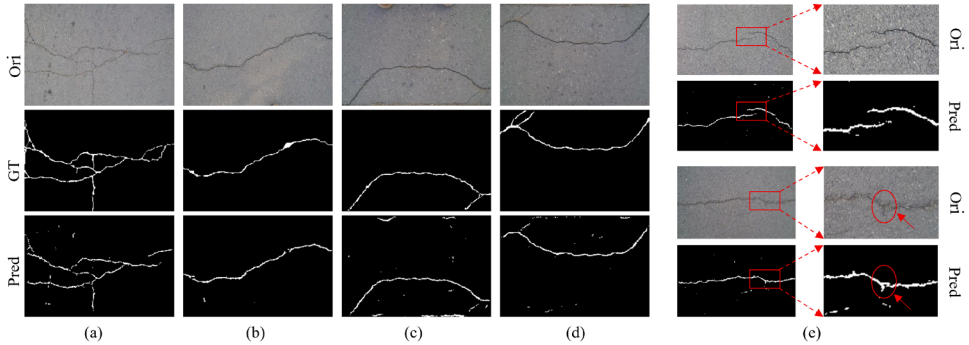


**Fig. 2.** The segmentation results on the CrackForest dataset. *Ori* denotes the original image. *GT* denotes the ground truth of the binary mask, and *Pred* denotes the prediction using the convolution form of InsNet.

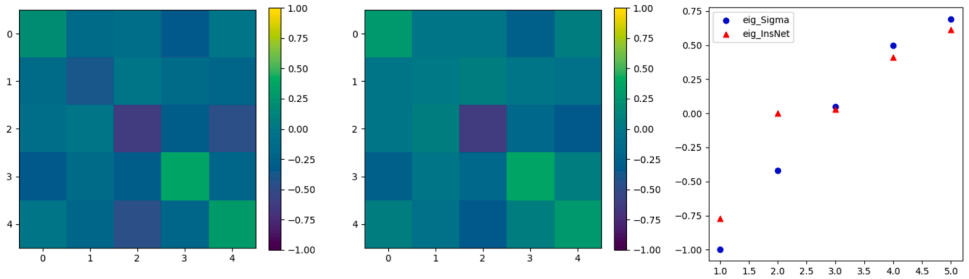


**Fig. 3.** Results on synthetic data. (a) The ground-truth covariance matrix $\boldsymbol{\Sigma}$. (b) The estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$. (c) The eigenvalues of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$, where eig_Sigma and eig_InsNet denote the eigenvalues of $\boldsymbol{\Sigma}$ (blue dots) and $\hat{\boldsymbol{\Sigma}}$ (red dots), respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

#### 5.2.2. Brain network estimation for MCI detection

Brain networks, estimated by modelling the functional interactions arising from cross-regional temporal dependencies of signals among distinct brain regions, provide a valuable framework for exploring our brain and mining biomarkers for the progressive neurodegenerative disorders, such as Alzheimer's disease (AD). In this section, we employ InsNet to estimate the functional brain network and subsequently perform a detection task to distinguish between the early stage of AD (Mild Cognitive Impairment (MCI)) and Normal Control (NC).

This experiment consists of three key steps: (1) Signal matrix extraction. For each subject, we extract a signal matrix $\boldsymbol{X} \in \mathbb{R}^{N\times B}$ from the resting-state functional magnetic resonance imaging (rs-fMRI) data, where $N$ is the number of brain regions, and $B$ is the length of the brain signals. (2) Functional connectivity estimation. The proposed DiSK is utilized to compute the correlation matrix $\boldsymbol{S} \in \mathbb{R}^{N\times N}$ for each subject, capturing pairwise functional interactions between brain regions. (3) Detection. We perform MCI vs. NC detection based on the derived correlation.

**Dataset**: Alzheimer's Disease Neuroimaging Initiative (ADNI)[1] dataset is utilized in this experiment, comprising 165 MCI and 154 NC. The rs-fMRI data of all subjects were acquired by a Philips 3.0T scanner with the following imaging parameters: flip angle = 80, TR/TE = 3000ms/300ms, voxel thickness = 3.3mm, and image matrix = $64 \times 64$. The scanning lasted 7 min, which generated 140 volumes for each subject. The information on these subjects is reported in Table 5.

**Compared methods**: **GRFF** [42]: Generative Random Fourier Features, a one-stage kernel learning approach that models some latent distribution of the kernel via a generative network based on the random Fourier features; **CosNet** [11]: A Generalized Spectral Kernel Network, which generalizes spectral kernel mapping in the real number domain to the complex number domain; **CokeNet** [41]: Copula-Nested Spectral Kernel Network, which introduces copula networks into the design of the spectral density.

**Table 5**
The information of subjects.

| Database | Category | Subjects | Age | Brain regions | Volumes |
|---|---|---|---|---|---|
| ADNI | MCI (+1) | 165 | 72.03±7.71 | 116 | 140 |
| | NC (-1) | 154 | 75.36±6.16 | 116 | 140 |

**Table 6**
Detection results. The best results are highlighted in **bold**. (↑) indicates the larger, the better.

| Metric | DiSK | CosNet | CokeNet | GRFF |
|---|---|---|---|---|
| ACC (↑) | **79.69%** | 68.75% | 64.06% | 67.19% |
| Precision (↑) | **80.46%** | 68.65% | 63.04% | 66.18% |
| Recall (↑) | **79.12%** | 68.43% | 64.73% | 68.27% |
| F1 (↑) | **79.28%** | 68.47% | 62.52% | 65.77% |

**Results in MCI detection**: As shown in Tables 6, the proposed InsNet achieves the following results: ACC = 79.69%, Precision = 80.46%, Recall = 79.12%, and F1 = 79.28%, outperforming all compared methods across all evaluation metrics. Specifically, InsNet achieves 15.91% accuracy increment (68.75% → 79.69%), 17.2% precision increment (68.65% → 80.46%), 15.62% recall increment (68.43% → 79.12%), and 15.79% F1 score increment (68.47% → 79.28%) compared to the suboptimal model (CosNet). The quantitative results reveal the following insights: (1) Our DiSK demonstrates exceptional performance. This advantage can be attributed to the fact that, compared with the positive definite kernel-based approaches that mainly capture correlations, DiSK is capable of representing reciprocal relationships, including both mutual promotion and suppression between brain regions. These findings suggest that indefinite kernels offer a more expressive representation for

[1] https://adni.loni.usc.edu/

**Table 7**
The detailed experimental settings.

| Task | Dataset | Methods | lr | Batch size | Initialization (spectral sampling) | Network architecture |
|---|---|---|---|---|---|---|
| Time series classification | All | InsNet | 0.01 | train samples | Gaussian | $d \times 1024 \times 256 \times 128 \times Class$ |
| | | SRFF | 0.01 | train samples | Gaussian | $d \times 512 \times 128 \times 64 \times Class$ |
| | | DSKN | 0.01 | train samples | Gaussian | $d \times 1024 \times 256 \times 64 \times Class$ |
| | | ASKL | 0.01 | train samples | Gaussian | – |
| | | CosNet | 0.01 | train samples | Gaussian | $d \times 512 \times 128 \times 64 \times Class$ |
| | | CokeNet | 0.01 | train samples | Gaussian | $d \times 1024 \times 256 \times 64 \times Class$ |
| Image segmentation | CrackForest | InsNet | 0.01 | 4 | Gaussian | $in\_channel \times 256 \times 512 \times 1024$ |
| | | DSKN | 0.01 | 4 | Gaussian | $in\_channel \times 256 \times 512 \times 1024$ |
| | | CosNet | 0.01 | 4 | Gaussian | $in\_channel \times 256 \times 512 \times 1024$ |
| | | Single-layer InsNet | 0.01 | 4 | Gaussian | $in\_channel \times 256$ |
| | PASCAL VOC | InsNet | 0.001 | 16 | Gaussian | $in\_channel \times 256 \times 512 \times 1024$ |
| | | DSKN | 0.001 | 16 | Gaussian | $in\_channel \times 256 \times 512 \times 1024$ |
| | | CosNet | 0.001 | 16 | Gaussian | $in\_channel \times 256 \times 512 \times 1024$ |
| | | Single-layer InsNet | 0.001 | 16 | Gaussian | $in\_channel \times 256$ |
| MCI detection | ADNI | InsNet | 0.0001 | 32 | Gaussian | $137 \times 128 \times 256 \times 256$ |
| | | CosNet | 0.0001 | 32 | Gaussian | $137 \times 128 \times 256 \times 256$ |
| | | CokeNet | 0.001 | 32 | Gaussian | $137 \times 256$ |
| | | GRFF | 0.001 | 32 | Gaussian | – |

complex brain functional interactions. (2) The performance improvement is more remarkable in terms of Recall and F1 score. This emphasizes that our method identifies subjects with MCI more accurately. Meanwhile, it balances false positives and false negatives more effectively. All these observations further demonstrate that our InsNet offers a more robust solution for the MCI detection task, underscoring its practical advantages.

### 5.3. Discussion and limitation

In this section, we provide a comprehensive discussion of the proposed InsNet, focusing on its computational complexity, hyperparameter sensitivity, component interpretability, and scalability.

**Computational Complexity**: The proposed method has the same computational complexity as the baseline methods when applied to the same task. In the classification task, the proposed InsNet is employed as a feature extractor. The computational complexity is $\mathcal{O}(NM)$. When the method is used as a kernel function to model reciprocal relationships between data points, it incurs a computational complexity of $\mathcal{O}(N^2)$. Compared with baselines, the proposed InsNet involves two components, the positive definite part and the negative definite part. This design increases the model size in terms of parameters, thereby requiring more memory. To further evaluate the efficiency of models, we compare the proposed method with baselines under three key metrics, including performance, training time, and memory footprint, on the **FordA** dataset. The result, shown in Fig. 4, demonstrates that our InsNet achieves superior classification accuracy while reducing training time, thereby indicating its computational efficiency.

**Hyperparameter Sensitivity**: All the experimental settings are summarized in Table 7. It can be observed that the proposed model involves several key hyperparameters, such as the number of features $M$, the number of layers $l$, and the choice of distributions for spectral sampling. Within the proposed framework, the number of features and layers can be interpreted as the width and depth of neural networks, respectively. Empirical evidence suggests that these two parameters have a noticeable impact on model performance. Specifically, experiments on the image segmentation task also indicate that the deeper architecture (InsNet) achieves better performance than the shallower counterpart (single-layer InsNet). By contrast, the choice of distribution appears to have a limited influence on model performance, as it is only used to initialize the spectral sampling procedure. During training, the spectrum is further optimized, which reduces the impact of the initial distribution on the final performance.

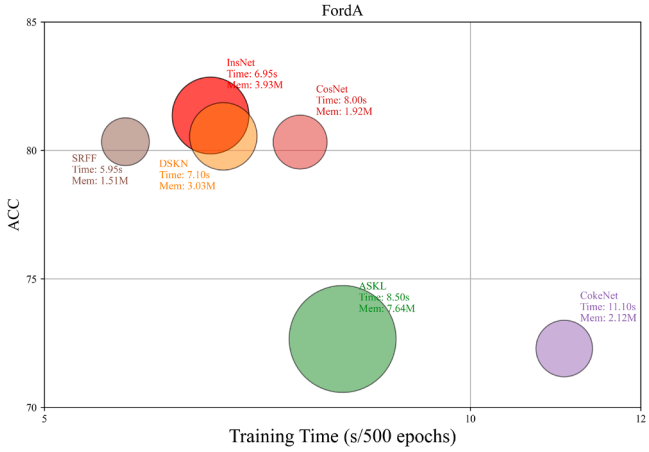


**Fig. 4.** Model efficiency comparison on FordA.

**Interpretability of Components**: The proposed model incorporates both positive definite and negative definite components. Both theoretical analysis and experimental results demonstrate that the cooperation between these components enhances the overall performance of the model. Nevertheless, the individual contributions of each component in real-world applications are not yet well understood and constitute an interesting direction for future research.

**Scalability**: The scalability of the proposed method primarily depends on the type of task. If the method is employed as a feature extractor, its neural-network-like architecture enables it to scale to large datasets. In contrast, when the method is used as a kernel function to model interactions between data points, it incurs a computational complexity of $\mathcal{O}(N^2)$, which restricts its applicability to large-scale tasks.

## 6. Conclusion

In this paper, we propose InsNet, a novel deep kernel method. This method extends the standard Hilbert space by breaking the positive definiteness constraint, enabling the model to effectively capture complex reciprocal connections and hierarchical structures within the data. The theoretical analysis further characterizes the structural properties of InsNet and establishes its approximation and generalization guarantees, offering solid theoretical support for the proposed design. Extensive experiments on both synthetic and real-world datasets consistently demonstrate that our proposed InsNet can effectively capture

complex correlations and structures inherent in the data, leading to significant performance improvements over state-of-the-art relevant deep kernel methods. Overall, by tightly integrating the architecture of InsNet with rigorous theoretical analysis and comprehensive experimental validation, this paper presents a coherent deep kernel learning framework, highlighting the practical potential and theoretical advantages of incorporating indefinite kernels into deep kernel networks.

## CRediT authorship contribution statement

**Yanfang Xue:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis; **Hui Xue:** Funding acquisition; **Shipeng Zhu:** Writing – review & editing.

## Data availability

All the data are public.

## Declaration of competing interest

All authors declare no competing interests.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62476056 and T24B2005), the Fundamental Research Funds for the Central Universities (2242025K30024). Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

## References

[1] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, (2012) arXiv:1207.0580.
[2] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, ICLR, 2013.
[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classific ation with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, NeurIPS, 25, 2012.
[4] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[5] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning, MIT Press, 2006.
[6] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[7] R.M. Neal, Priors for infinite networks, Bayesian Learning for Neural Networks, Lecture Notes in Statistics, 118, Springer, New York, (1996) 29–53.
[8] Y. Cho, L. Saul, Kernel methods for deep learning, Adv. Neural Inf. Process. Syst. (NeurIPS) 22 (2009) 342–350.
[9] S. Zhang, J. Li, P. Xie, Y. Zhang, M. Shao, H. Zhou, M. Yan, Stacked kernel network, (2017) arXiv:1711.09219.
[10] H. Xue, Z.-F. Wu, W.-X. Sun, Deep spectral kernel learning, in: International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 4019–4025.
[11] Y. Xue, P. Fang, J. Tian, S. Zhu, H. Xue, CosNet: a generalized spectral kernel network, in: Advances in Neural Information Processing Systems (NeurIPS), 2023.
[12] A.A. Freitas, Understanding the crucial role of attribute interaction in data mining, Artif. Intell. Rev. 16 (2001) 177–199.
[13] J. Bijsterbosch, S.M. Smith, C. Beckmann, An Introduction to Resting State fMRI Functional Connectivity, Oxford University Press, 2017.
[14] S. Liwicki, S. Zafeiriou, G. Tzimiropoulos, M. Pantic, Efficient online subspace learning with an indefinite kernel for visual tracking and recognition, IEEE Trans. Neural Netw. Learn. Syst. 23 (10) (2012) 1624–1636.
[15] Z. Zeng, J. Sun, Z. Han, W. Hong, SAR automatic target recognition method based on multi-stream complex-valued networks, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–18.
[16] S. Bochner, et al., Lectures on Fourier Integrals, 42, Princeton University Press, 1959.
[17] J.-F. Le Gall, Measure Theory, Probability, and Stochastic Processes, Springer, 2022.
[18] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in: Advances in Neural Information Processing Systems (NeurIPS), 2007, pp. 1177–1184.
[19] C. Berg, J.P.R. Christensen, P. Ressel, Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions, 100, Springer, 1984.
[20] C.S. Ong, X. Mary, S. Canu, A.J. Smola, Learning with non-positive kernels, in: International Conference on Machine Learning (ICML), 2004, p. 81.
[21] F. Schleif, P. Tiňo, Indefinite proximity learning: a review, Neural Comput. 27 (10) (2015) 2039–2096.
[22] F. Schleif, P. Tiňo, Indefinite core vector machine, Pattern Recognit. 71 (2017) 187–195.
[23] M. Münch, M. Straat, M. Biehl, F. Schleif, Complex-valued embeddings of generic proximity data, in: Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshops, S+SSPR, 12644 of Lecture Notes in Computer Science, (2021), pp. 14–23.
[24] A. Gisbrecht, F. Schleif, Metric and non-metric proximity transformations at linear costs, Neurocomputing 167 (2015) 643–657.
[25] S. Mehrkanoon, X. Huang, J.A.K. Suykens, Indefinite kernel spectral learning, Pattern Recognit. 78 (2018) 144–153.
[26] F.-M. Schleif, A. Gisbrecht, P. Tino, Probabilistic classifiers with low rank indefinite kernels, (2016) arXiv:1604.02264.
[27] D. Oglic, T. Gärtner, Scalable learning in reproducing kernel Krein spaces, in: International Conference on Machine Learning, PMLR, 2019, pp. 4912–4921.
[28] S. Heilig, M. Münch, F.-M. Schleif, Memory efficient kernel approximation for non-stationary and indefinite kernels, in: 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 1–8.
[29] J. Pennington, F.X.X. Yu, S. Kumar, Spherical random features for polynomial kernels, Adv. Neural Inf. Process. Syst. (NeurIPS) 28 (2015) 1846–1854.
[30] F. Liu, X. Huang, L. Shi, J. Yang, J.A.K. Suykens, A double-variational Bayesian framework in random Fourier features for indefinite kernels, IEEE Trans. Neural Netw. Learn. Syst. 31 (8) (2019) 2965–2979.
[31] F. Liu, X. Huang, Y. Chen, J. Suykens, Fast learning in reproducing kernel krein spaces via signed measures, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 388–396.
[32] Q. Luo, K. Fang, J. Yang, X. Huang, Towards unbiased random features with lower variance for stationary indefinite kernels, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.
[33] R. Salakhutdinov, G.E. Hinton, Using deep belief nets to learn covariance kernels for gaussian processes, in: Advances in Neural Information Processing Systems (NeurIPS), 2007, pp. 1249–1256.
[34] A.C. Damianou, N.D. Lawrence, Deep Gaussian processes, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS), 31, 2013, pp. 207–215.
[35] A.G. Wilson, Z. Hu, R. Salakhutdinov, E.P. Xing, Deep kernel learning, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 51, 2016, pp. 370–378.
[36] A.G. Wilson, H. Nickisch, Kernel interpolation for scalable structured Gaussian processes (KISS-GP), in: Proceedings of the International Conference on Machine Learning (ICML), 37, 2015, pp. 1775–1784.
[37] A.G. Wilson, Z. Hu, R.R. Salakhutdinov, E.P. Xing, Stochastic variational deep kernel learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 2586–2594.
[38] A.L.S. Matias, C.L.C. Mattos, J.P.P. Gomes, D. Mesquita, Amortized variational deep kernel learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2024.
[39] J. Loría, A. Bhadra, Deep kernel posterior learning under infinite variance prior weights, in: The Thirteenth International Conference on Learning Representations (ICLR), 2025.
[40] L. D'Amore, Resource-efficient model for deep kernel learning, Comput. Inf. 44 (1) (2025) 1–25.
[41] J. Tian, H. Xue, Y. Xue, P. Fang, Copula-nested spectral kernel network, in: Forty-first International Conference on Machine Learning (ICML), 2024.
[42] K. Fang, F. Liu, X. Huang, J. Yang, End-to-end kernel learning via generative random Fourier features, Pattern Recognit. 134 (2023) 109057.
[43] F. Tonin, Q. Tao, P. Patrinos, J.A.K. Suykens, Deep kernel principal component analysis for multi-level feature learning, Neural Netw. 170 (2024) 578–595.
[44] J. Li, Y. Liu, W. Wang, Automated spectral kernel learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 4618–4625.
[45] J. Li, Y. Liu, W. Wang, Convolutional spectral kernel learning with generalization guarantees, Artif. Intell. 313 (2022) 103803.
[46] E. Milsom, B. Anson, L. Aitchison, Convolutional deep kernel machines, in: The Twelfth International Conference on Learning Representations (ICLR), 2024.
[47] D. Gao, X. Kang, W. Huang, C. Zheng, Data-driven random feature selection for deep kernel learning with kernel alignment, in: Advanced Intelligent Computing Technology and Applications International Conference (ICIC), 15859 of Lecture Notes in Computer Science, 2025, pp. 359–369.
[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: an imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. (NeurIPS) 32 (2019).
[49] H.A. Dau, A. Bagnall, K. Kamgar, C.-C.M. Yeh, Y. Zhu, S. Gharghabi, C.A. Ratanamahatana, E. Keogh, The UCR time series archive, IEEE/CAA J. Autom. Sin. 6 (6) (2019) 1293–1305.
[50] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015.