EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
2026, Vol. 19, Issue 1, Article Number 5825
ISSN 1307-5543 – ejpam.com
Published by New York Business Global

Integrating Hierarchical Structures in Medical Data Classification:A Kernel-Based Method

Seydou Nourou Sylla

[1] *Alioune Diop University, Senegal*

Abstract. Medical diagnosis systems frequently rely on structured information collected during physician–patient interviews. These data naturally follow a hierarchical organization, where general questions are followed by more specific sub-questions. Such a structure should be explicitly incorporated into similarity measures used in classification algorithms, as it reflects dependencies between symptoms and contributes essential diagnostic information.

In this work, we introduce a kernel that simultaneously accounts for (i) the hierarchical structure linking main questions to their subordinate items and (ii) interactions among sub-variables. The kernel is integrated into the pgpDA classification framework, allowing the method to embed prior knowledge on how variables are organized and how symptoms interact. The proposed kernel is designed for binary data arranged in two-level tree structures and supports interaction modeling of any given order.

Experiments conducted on simulated data and a real verbal autopsy dataset from Senegal demonstrate consistent improvements over classical kernels, and a deep-learning benchmark confirms that the structured kernel retains strong predictive power even in modern architectures. The methodology may be extended to mixed data types or adapted to graph-structured symptom networks.

2020 Mathematics Subject Classifications: AMS classification code

Key Words and Phrases: Kernel methods; hierarchical data; pgpDA; verbal autopsy; medical diagnosis; deep learning

1. Introduction

Diagnostic procedures in medicine often rely on structured interviews in which physicians ask a sequence of general and more specific questions to assess symptoms and contextual factors. This process naturally generates hierarchically organized binary data, where a main symptom is queried first and its associated sub-questions are collected only if the primary symptom is present, as commonly observed in verbal autopsy investigations that rely on multi-level questionnaires to assign probable causes of death [1–3].

Despite this hierarchical structure, many statistical and machine learning approaches treat binary predictors as independent variables, thereby ignoring the dependencies induced

DOI: https://doi.org/10.29020/nybg.ejpam.v19i1.5825

Email address: seydounourou.sylla@uadb.edu.sn (S. N. Sylla)

by the data acquisition process. As a result, similarity measures between individuals may fail to reflect clinically meaningful relationships among symptoms, limiting their effectiveness for diagnostic and classification tasks. Incorporating hierarchical relationships into similarity measures can substantially improve the analysis of medical data by aligning the representation of variables with clinical reasoning and underlying disease mechanisms. Such hierarchical representations are also central beyond medicine, notably in psychology, socio-behavioral sciences, and education, where multi-level constructs guide both data collection and interpretation.

In machine learning, several methods have been proposed to handle structured or heterogeneous data, particularly through multiple-kernel learning [4–9] and hierarchical kernel formulations [10–13]. While these approaches combine kernels from multiple sources or encode structural relationships among variables, few explicitly integrate both the hierarchical dependence linking main variables to their sub-variables and the interaction effects among sub-variables, especially in the context of high-dimensional binary data commonly encountered in medical surveys and verbal autopsy studies [14–16].

To address this limitation, we propose a kernel specifically designed for binary predictors organized in a two-level hierarchical structure. The proposed kernel simultaneously captures similarities between main variables, similarities between their associated sub-variables, and interaction effects among sub-variables up to a chosen order, thereby modeling both the hierarchical occurrence of symptoms and their potential combined effects, which are often critical in diagnostic reasoning.

The kernel is integrated into the pgpDA classification framework introduced by Sylla et al. [17], which models each class as a parsimonious Gaussian process in a reproducing kernel Hilbert space. By embedding structured prior knowledge directly into the kernel, the pgpDA framework benefits from a more informative and clinically meaningful representation of the predictors. The proposed approach is evaluated using both simulated data and a large verbal autopsy dataset collected in rural Senegal [3, 18–21]. This dataset involves numerous symptoms and multiple causes of death, making classification particularly challenging. Experimental results demonstrate that the hierarchical kernel consistently improves classification accuracy compared with standard kernels, with performance increasing as higher-order interactions are incorporated. Comparisons with deep-learning benchmarks further confirm that the proposed kernel provides highly informative representations for modern neural architectures.

The remainder of the paper is organized as follows. Section 2 reviews multiple and hierarchical kernel methods. Section 3 presents the pgpDA framework. Section 4 introduces the proposed hierarchical interaction kernel for binary data. Section 5 reports the experimental results, and Section 6 concludes the paper with perspectives for future work.

2. Multiple and Hierarchical Kernels

Kernel methods provide a flexible framework for learning from complex data by implicitly mapping observations into high-dimensional feature spaces where linear techniques become powerful and expressive [22–25]. In many applications, data are heterogeneous or naturally structured, which motivates the development of kernel formulations capable of integrating multiple sources of information, hierarchical dependencies, or interactions among variables.

2.1. Multiple Kernel Learning

Multiple Kernel Learning (MKL) extends classical kernel methods by allowing several kernels to be combined, each capturing complementary aspects of the data. Instead of relying on a single similarity measure, MKL constructs a weighted combination of base kernels

$$\kappa(x,x') = \sum_{m=1}^{M} d_m \, \kappa_m(x,x'), \qquad d_m \geq 0, \quad \sum_{m=1}^{M} d_m = 1, \tag{1}$$

thus enabling the method to adaptively select or emphasize specific sources of information.

Early foundational work includes the semidefinite programming approach of Lanckriet et al. [4], the conic duality formulation of Bach et al. [5], and the efficiency improvements proposed by Rakotomamonjy et al. [26]. Comprehensive algorithmic developments and theoretical perspectives can be found in Gönen and Alpaydın [6], which remains a standard reference in MKL research.

MKL has proved particularly useful for integrating heterogeneous data such as genomic measurements [7], multimodal images [27], and remote sensing signals [28, 29]. More recent works also investigate MKL for unsupervised learning and data fusion, for example Mariette and Villa-Vialaneix [8] and Dai and Shao [9], highlighting its relevance when data originate from different measurement processes or modalities.

2.2. Hierarchical and Structured Kernels

In many real-world problems, predictors exhibit inherent structure—temporal, spatial, syntactic, semantic, or hierarchical. Standard kernels cannot adequately capture such relationships, which has motivated the development of *structured kernels* and *hierarchical kernels*.

Notable early contributions include convolution kernels for sequences, trees, and graphs developed by Haussler [30], Collins and Duffy [31], Kashima and Koyanagi [32], and Lodhi et al. [33]. These kernels exploit the structure of discrete objects by decomposing them into subcomponents and summing contributions from shared substructures. Surveys of kernels for structured data can be found in Gärtner [34].

Hierarchical kernels, in particular, encode multi-level representations of data. Hierarchical Gaussian kernels have been investigated by Steinwart et al. [11], while deep kernel compositions have recently been formalized by Huang, Lu, and Zhang [12, 13]. These works demonstrate that hierarchical compositions of kernels can capture complex interactions, improve function approximation, and provide interpretable, structure-preserving similarity measures.

Hierarchical and deep formulations have also been applied in various domains:

- text classification through hierarchical deep learning architectures [35];

- medical image classification, where anatomical structures exhibit natural hierarchical organization [36];
- remote sensing and multi-source image analysis, using kernels defined over hierarchical image representations [37];
- graph-structured data, where hierarchical graph kernels exploit multi-scale structural information [38, 39].

These approaches highlight the importance of designing kernels that reflect the multi-level or nested nature of the data.

2.3. Hierarchical Kernels for Binary and Heterogeneous Data

While hierarchical kernels have been extensively developed for structured objects such as images, graphs, or sequences, fewer contributions focus on *binary data organized in multi-level structures*. In many medical or epidemiological studies—particularly verbal autopsies—variables follow conditioned or hierarchical relationships, yet most classification methods treat them as flat binary vectors.

Our work differs from existing hierarchical kernels in that it focuses specifically on:

(i) two-level hierarchical binary variables (main items and sub-items),

(ii) interactions among sub-variables, which are often diagnostically meaningful,

(iii) integration within the pgpDA framework [17], enabling parsimonious modeling of class-specific latent structures in a kernel space.

This fills a methodological gap between classical MKL approaches, which treat kernels independently, and hierarchical kernels developed for structured sequences or images, which do not naturally apply to binary questionnaire data.

3. Binary Classification Using a Kernel Function

In this section, we recall the classification framework underlying the pgpDA method, initially introduced for binary predictors in [17]. The pgpDA classifier belongs to the family of parsimonious Gaussian process models [40, 41], where each class is assumed to lie in a low-dimensional subspace of a reproducing kernel Hilbert space (RKHS). This parsimonious representation avoids the instability that typically arises from inverting large kernel matrices [42, 43].

Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ denote a training sample, where each $x_i \in \{0,1\}^p$ is a binary vector and $y_i \in \{1, \ldots, K\}$ indicates the class label. For class C_k, we define

$$n_k = \sum_{i=1}^{n} \mathbf{1}_{\{y_i = k\}}, \qquad C_k = \{\, x_i : y_i = k \,\}.$$

Let $\kappa : \{0,1\}^p \times \{0,1\}^p \to \mathbb{R}$ be a symmetric non-negative kernel function. Following standard kernel methods [22–24], the centered kernel associated with class C_k is defined as

$$\rho_k(x,x') = \kappa(x,x') - \frac{1}{n_k}\sum_{x_\ell \in C_k} \kappa(x_\ell, x') - \frac{1}{n_k}\sum_{x_\ell \in C_k} \kappa(x, x_\ell) + \frac{1}{n_k^2}\sum_{x_\ell, x_{\ell'} \in C_k} \kappa(x_\ell, x_{\ell'}). \quad (2)$$

For each class C_k, we construct the $n_k \times n_k$ matrix

$$M_k = \frac{1}{n_k}\left[\rho_k(x_\ell, x_{\ell'})\right]_{\ell,\ell'}.$$

Let $\lambda_{k1} \geq \cdots \geq \lambda_{kn_k}$ denote its eigenvalues and $\beta_{k1}, \ldots, \beta_{kn_k}$ the associated normalized eigenvectors.

The key assumption of pgpDA is that the data of class C_k lie in a subspace of low intrinsic dimension $d_k < n_k$ within the RKHS induced by κ. This assumption, also used in kernel mixture models [41, 43], ensures numerical stability and yields a compact representation of each class. The intrinsic dimension d_k is estimated via the classical Cattell scree test [44].

Let $d_{\max} = \max_{k=1,\ldots,K} d_k$. The pgpDA decision rule assigns a new observation x to the class minimizing the discriminant score

$$\begin{aligned} D_k(x) = {} & \frac{1}{n_k}\sum_{j=1}^{d_k} \frac{1}{\lambda_{kj}} \left(\sum_{x_\ell \in C_k} \beta_{kj,\ell}\, \rho_k(x, x_\ell) \right)^2 + \frac{1}{\lambda}\rho_k(x,x) \\ & + \sum_{j=1}^{d_k} \log \lambda_{kj} + (d_{\max} - d_k)\log\lambda - 2\log(n_k), \end{aligned} \quad (3)$$

where λ is a shrinkage parameter given by

$$\lambda = \frac{\sum_{k=1}^K n_k \left(\mathrm{trace}(M_k) - \sum_{j=1}^{d_k} \lambda_{kj}\right)}{\sum_{k=1}^K n_k (r_k - d_k)}, \quad (4)$$

and r_k denotes the dimension of class C_k in the feature space. For nonlinear kernels, $r_k = n_k$.

The discriminant rule (3) provides a nonlinear extension of parsimonious Gaussian mixture discriminant analysis [17, 40], with the advantage of avoiding large matrix inversions while leveraging the expressivity of kernel embeddings. The performance of pgpDA depends crucially on the choice of kernel κ, which motivates the introduction of a new hierarchical kernel in Section 4.

4. Hierarchical Kernel for Binary Observations

Many medical questionnaires, including verbal autopsy instruments, are designed with an inherent hierarchical structure: a *main question* is asked first, and a set of *sub-questions* is asked only if the main item is positive. Such conditional structures encode important dependencies between variables and should therefore be reflected in the similarity measure used by a classification algorithm.

While numerous kernels have been proposed for structured data [30, 34, 45], very few address the case of binary vectors organized in two-level hierarchies, nor do they incorporate interaction terms among sub-items. In this section, we introduce a kernel specifically designed for such data.

4.1. Data Structure and Notation

In an interview with a doctor, there are often so-called main questions. For each main question, there are secondary questions, which are only asked if the answer to the main question is positive. By formalizing this concept, the variable X_j represents the answer to the main question j. For each given X_j there are q_j responses to secondary issues noted by the sub-variables $Z_1^j, \dots, Z_{q_j}^j$. Thus:

- the random variables $X = (X_j, j = 1, \dots, p)$ define the answers to the main questions representing the symptoms and the socio-demographic variables;
- the random variables $Z = (Z_\ell^j, \ell = 1, \dots, q_j, j = 1, \dots, p)$ define the answers to the secondary questions representing the q_j sub-variables for each variable X_j;
- the random variable Y defines the cause of death (or diagnostic class).

These hierarchies are represented by a two-level tree structure, as shown in Figure 1. The first level represents the answers to main questions. The second level represents the sub-variables, that is to say, the answers to the secondary issues of each main question.

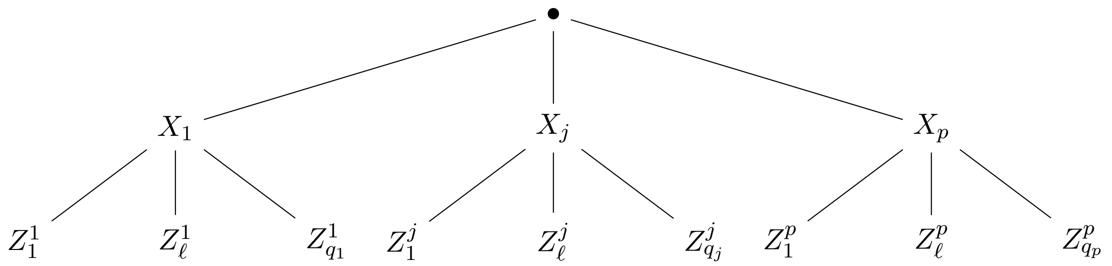

Figure 1. Example of a two-level tree structure linking main variables and sub-variables.

Following standard practice in verbal autopsy analysis [14, 15, 20], the main variable is defined as

$$X_j = \max\{Z_1^j, \dots, Z_{q_j}^j\},$$

meaning that $X_j = 1$ if any of its sub-variables is positive.

4.2. Dissimilarity Decomposition

We recall the algebraic decomposition that motivates the definition of the kernel. For each j,

$$X_j = 1 - \prod_{\ell=1}^{q_j}(1 - Z_\ell^j) = \sum_{\ell=1}^{q_j}(-1)^{\ell-1}\sum_{k=1}^{\ell}\sum_{|i|=k} Z_{i_1}^j \cdots Z_{i_k}^j,$$

where $|i| = k$ denotes the size of the multi-index $i = (i_1, \dots, i_k)$.

For two individuals (x, z) and (x', z'), the squared Euclidean distance between the main vectors can be decomposed as:

$$\begin{aligned} \|x - x'\|^2 &= \sum_{j=1}^{p}\big[X_j - X'_j\big]^2 \\ &= \sum_{j=1}^{p}\left[\prod_{\ell=1}^{q_j}(1 - z_\ell^j) - \prod_{\ell=1}^{q_j}(1 - z_\ell'^j)\right]^2 \\ &= \sum_{j=1}^{p}\sum_{\ell=1}^{q_j}\sum_{k=1}^{\ell}\sum_{|i|=k} s_{kji}^2 + R, \end{aligned}$$

where

$$s_{kji} = z_{i_1}^j \cdots z_{i_k}^j - z_{i_1}'^j \cdots z_{i_k}'^j,$$

and R is the sum of cross terms (double products).

By defining

$$SC(z, z') = \sum_{j=1}^{p}\sum_{\ell=1}^{q_j}\sum_{k=1}^{\ell}\sum_{|i|=k} s_{kji}^2, \tag{5}$$

we obtain the decomposition

$$\|x - x'\|^2 = SC(z, z') + R. \tag{6}$$

A dissimilarity measure combining main and sub-level information is defined, for all $\gamma \in [0, 1]$, by:

$$D\big((x, z), (x', z')\big) = (1 - \gamma)\|x - x'\|^2 + (2\gamma - 1)\, SC(z, z'). \tag{7}$$

Let

$$D_x(x, x') = \|x - x'\|^2, \quad D_z(z, z') = SC(z, z'),$$

so that (7) can be written more compactly as

$$D\big((x, z), (x', z')\big) = (1 - \gamma)D_x(x, x') + (2\gamma - 1)D_z(z, z').$$

4.3. Construction of the Hierarchical Kernel

Using the kernel construction method proposed in [46], we introduce the kernel:

$$\kappa_{\text{SGH}}\left((x,z),\left(x',z'\right)\right)=\kappa_x\left(x,x'\right)^{1-\gamma}\kappa_z\left(z,z'\right)^{2\gamma-1}, \tag{8}$$

where:

- $\kappa_x(x,x')=\exp\left(S(x,x')/(2\sigma_x^2)\right)$ is a kernel on main variables based on a similarity measure S between binary vectors (for instance Jaccard or Tversky similarity [47–52]);
- $\kappa_z(z,z')=\exp\left(SC_{(r)}(z,z')/(2\sigma_r^2)\right)$ is a kernel on the sub-variables based on a truncated version of SC.

The truncated interaction term of order r is defined as

$$\begin{aligned}SC_{(r)}(z,z')&=\sum_{j=1}^{p}\sum_{k=1}^{r}(q_j+1-k)\sum_{|i|=k}s_{kji}^2\\&=\sum_{j=1}^{p}sc_{(r,j)},\end{aligned}$$

with

$$sc_{(r,j)}=\sum_{k=1}^{r}(q_j+1-k)\sum_{|i|=k}\left(z_{i_1}^j\cdots z_{i_k}^j-z_{i_1}'^j\cdots z_{i_k}'^j\right)^2.$$

By combining these components, the hierarchical kernel of interactions of order r is thus

$$\kappa_{\text{SGH}}\left((x,z),\left(x',z'\right)\right)=\kappa_x\left(x,x'\right)^{1-\gamma}\kappa_z\left(z,z'\right)^{2\gamma-1}.$$

For some values of γ, it appears that the standard RBF kernel can be recovered in special cases. If $\kappa_x=\kappa_{\text{RBF}}$, then:

- $\gamma=\frac{1}{2}\Rightarrow\kappa_{\text{SGH}}\left((x,z),(x',z')\right)=\kappa_{\text{RBF}}(x,x')$,
- $\gamma=1$ and $r=1\Rightarrow\kappa_{\text{SGH}}\left((x,z),(x',z')\right)=\kappa_{\text{RBF}}(z,z')$,
- $\gamma=\frac{2}{3}$ and $r=1\Rightarrow\kappa_{\text{SGH}}\left((x,z),(x',z')\right)=\kappa_{\text{RBF}}\left((x\cup z),(x'\cup z')\right)$.

4.4. Justification and Relation to Existing Work

The KSGH formulation is designed as a structured generalization of the standard RBF kernel. It collapses to an RBF kernel in the non-hierarchical case, providing a strong theoretical baseline while allowing higher-order interactions to be incorporated via r.

Our approach differs from hierarchical kernels proposed for images [36], text [35], or general structured objects [11, 12], in that it:

(i) specifically targets hierarchical binary questionnaire data,

(ii) models both hierarchy and interaction terms,

(iii) integrates naturally into the pgpDA classification framework,

(iv) allows fine-tuning of the relative contribution of each hierarchical level through γ.

5. Experiments

This section evaluates the performance of the proposed hierarchical interaction kernel when embedded in the pgpDA classification framework and compares it with standard methods and a deep-learning benchmark.

Our objectives are:

(i) to assess how the interaction order r and hierarchical weight γ influence classification accuracy,

(ii) to compare our approach with existing classification methods commonly used for verbal autopsy data,

(iii) to evaluate the usefulness of the kernel features for deep neural networks.

5.1. Application to Verbal Autopsy Data

Verbal autopsy is a standardized method used to infer the probable cause of death in settings where medical certification is incomplete or absent [2, 3]. A list of p possible symptoms is established and the collected data $X = (X_1, \dots, X_p)$ consist of the absence or presence (encoded as 0 or 1) of each symptom on the deceased person. The probable cause of death is assigned by a physician and encoded as a qualitative random variable Y. We refer the reader to [1–3, 14–16, 18–21, 53–59] for details on verbal autopsy methods and automatic procedures for assigning causes of death from verbal autopsy data.

In this study, we focus on data measured on deceased persons during the period from 1985 to 2010 in three IRD (Research Institute for Development) sites (Niakhar, Bandafassi and Mlomp) in Senegal. The dataset includes:

- $n = 2500$ individuals (deceased persons),
- $K = 18$ classes (causes of death),
- $p = 100$ binary variables (symptoms and socio-demographic variables),

organized into main items and sub-items according to the questionnaire structure.

5.2. Methodology

The pgpDA models presented depend on the choice of hyperparameters

$$\omega = (\gamma, \alpha, \sigma_x, \sigma_r),$$

where:

- γ: the weighting parameter between the main variables and secondary variables (hierarchical parameter),
- α: the weighting parameter for the presence and absence of the main variables in the kernel $\kappa_x(x, x')$,
- σ_x: the smoothing parameter for the kernel of the main variables,
- σ_r: the smoothing parameter for the interaction kernel on the secondary variables.

A double cross-validation technique is employed. The total sample of size n is randomly divided $M = 50$ times into a training sample $\mathcal{L}_m$ of size τn and a test sample $\mathcal{T}_m$ of size $(1-\tau)n$, with $\tau \in (0, 1)$ representing a proportion, and $m = 1, \ldots, M$.

The parameter α is fixed at 0.5 to recover an RBF-type kernel for the main variables, and σ_x is set at 1.5, following [17].

The parameter σ_r is selected via cross-validation on the training set: consecutively, 100 individuals are randomly removed from the training sample 5 times, and σ_r is estimated by maximizing the classification accuracy rate on the 100 removed individuals.

The overall classification accuracy rate is estimated on the test sample by repeating the entire procedure 50 times. Thus, for each training sample $\mathcal{L}_m$, the optimal hyperparameter $\widehat{\sigma}_{r_m}$ is estimated through 5-fold cross-validation for $m = 1, \ldots, M$.

Furthermore, the overall optimal hyperparameter $\widehat{\sigma}_r$ is calculated as the empirical mode of the set $\{\widehat{\sigma}_{r_1}, \ldots, \widehat{\sigma}_{r_M}\}$. The average classification accuracy is computed for both the training samples $\mathcal{L}_m, m = 1, \ldots, M$ and the test samples $\mathcal{T}_m, m = 1, \ldots, M$.

For each level of interaction r, σ_r is selected via cross-validation and retained for the subsequent interaction level. To address computational time constraints, the maximum number of interactions is fixed at $r = 3$.

5.3. Results for First-Order Interactions

We first restrict interactions to the first order ($r = 1$). Table 1 reports the results.

For $\gamma = 0.5$ and $\gamma = 1$, the classification rates correspond respectively to those associated with the main variables and the secondary variables considered separately. For $\gamma = 0.67$, the classification rate matches that of an RBF kernel on the concatenated variables, while incorporating a structure into the data yields slight improvements for $\gamma = 0.7$ and 0.8.

Table 1: Parameter values and correct classification rates for $\gamma \in [0.5, 1]$ and $r = 1$.

γ	α	σ_x	σ_1	Train CCR (%)	Test CCR (%)
0.50	0.50	1.5	4.25	76.21	67.44
0.60	0.50	1.5	1.75	83.50	74.33
0.67	0.50	1.5	1.75	84.20	74.92
0.70	0.50	1.5	1.75	84.53	75.25
0.80	0.50	1.5	3.75	84.32	74.95
0.90	0.50	1.5	3.25	83.15	73.72
1.00	0.50	1.5	1.50	71.36	61.54

5.4. Results for Second-Order Interactions

Table 2 summarizes the parameter values and classification rates for a second-order interaction level ($r = 2$).

The optimal values of σ_2 are selected via cross-validation, while the values of σ_1 are those calculated in Table 1.

By considering second-order interactions, the kernel shows an improvement in results. For certain values of γ, the classification rates exceed 75%, and for $\gamma = 0.7$, the classification rate reaches 77%.

Table 2: Parameter values and correct classification rates for $\gamma \in [0.5, 1]$ and $r = 2$.

γ	α	σ_x	σ_1	σ_2	Train CCR (%)	Test CCR (%)
0.50	0.5	1.5	4.25	19	76.21	67.44
0.60	0.5	1.5	1.75	13	85.77	76.48
0.67	0.5	1.5	1.75	19	86.50	76.95
0.70	0.5	1.5	1.75	19	86.63	77.00
0.80	0.5	1.5	3.75	19	84.94	75.76
0.90	0.5	1.5	3.25	18	83.97	74.50
1.00	0.5	1.5	1.50	18	75.09	64.91

5.5. Results for Third-Order Interactions

Table 3 summarizes the parameter values and classification rates for a third-order interaction level ($r = 3$).

The optimal values of σ_3 are selected via cross-validation; the values of σ_1 and σ_2 are those calculated in Tables 1 and 2.

For $r = 3$ and $\gamma \in \{0.6, 0.67, 0.7\}$, the classification rates slightly exceed 77.00%, with a maximum of 77.19% for $\gamma = 0.67$.

Table 3: Parameter values and correct classification rates for $\gamma \in [0.5, 1]$ and $r = 3$.

γ	α	σ_x	σ_1	σ_2	σ_3	Train CCR (%)	Test CCR (%)
0.50	0.5	1.5	4.25	19	27	76.21	67.44
0.60	0.5	1.5	1.75	13	22	86.59	77.07
0.67	0.5	1.5	1.75	19	31	86.93	**77.19**
0.70	0.5	1.5	1.75	19	36	86.93	77.14
0.80	0.5	1.5	3.75	19	44	85.10	75.57
0.90	0.5	1.5	3.25	18	44	83.01	73.21
1.00	0.5	1.5	1.50	18	29	74.72	64.59

5.6. Summary Across Interaction Levels

Table 4 summarizes the classification rates depending on the level of interaction and the value of γ. We note that the classification rates associated with $r = 3$ are higher than those for $r = 1$ and $r = 2$. For $\gamma = 0.5$, the classification rate is invariant with respect to the order of interaction, since for this value the kernel reduces to a kernel on main variables only.

Table 4: Summary of correct classification rates for $\gamma \in [0.5, 1]$ and interaction levels $r = 1, 2, 3$.

	$r = 1$		$r = 2$		$r = 3$	
γ	Train	Test	Train	Test	Train	Test
0.50	76.21	67.44	76.21	67.44	76.21	67.44
0.60	83.50	74.33	85.77	76.48	86.59	77.07
0.67	84.20	74.92	86.50	76.95	86.93	**77.19**
0.70	84.53	75.25	86.63	77.00	86.93	77.14
0.80	84.32	74.95	84.94	75.76	85.10	75.57
0.90	83.15	73.72	83.97	74.50	83.01	73.21
1.00	71.36	61.52	75.09	64.91	74.72	64.59

Overall, higher-order interactions (up to $r = 3$) consistently enhance test accuracy, and intermediate values of γ (around 0.60–0.70) provide the best trade-off.

5.7. Deep Learning Benchmark Using KSGH Kernel Features

To evaluate the usefulness of the proposed hierarchical kernel beyond the pgpDA framework, we conducted an additional benchmark using a neural network classifier trained on KSGH kernel features. The objective of this experiment is twofold: (i) to assess whether the hierarchical interactions encoded in the KSGH kernel remain useful when combined with high-capacity deep models, and (ii) to provide a comparison with modern deep learning techniques.

A fully connected neural network was trained on the Gram matrix $K_{\text{train}} = \kappa_{\text{SGH}}((x_i, z_i), (x_j, z_j))$ computed from real verbal autopsy data. The network consists of two hidden layers (64 and 32 neurons), batch normalisation, ReLU activations, dropout regularisation (0.3), and a multi-label sigmoid output. The model was trained for 100 epochs with Adam (learning rate 10^{-3}).

The experiment was repeated across the same grid of hierarchical parameters as in pgpDA:

$$\gamma \in \{0.50, 0.60, 0.67, 0.70, 0.80\}, \qquad r \in \{1, 2, 3\}.$$

The results are summarized in Table 5. Accuracies range from 94.6% to 97.5%, confirming that KSGH provides highly informative, structure-aware representations of symptom data.

Table 5: Summary of correct classification rates (Test) for $\gamma \in [0.5, 0.8]$ and interaction levels $r = 1, 2, 3$.

γ	$r = 1$	$r = 2$	$r = 3$
0.50	0.9667	0.9708	0.9458
0.60	0.9667	0.9542	0.9625
0.67	0.9583	0.9583	0.9667
0.70	0.9708	0.9625	**0.9750**
0.80	0.9500	0.9667	0.9542

Several observations can be made:

- The neural network achieves consistently high accuracies (95–97%), confirming that KSGH provides highly informative, structure-aware representations of symptom data.
- As with pgpDA, intermediate values of γ (around 0.60 to 0.70) lead to the best results, while extreme values ($\gamma = 0.50$ or $\gamma = 0.80$) are less stable.
- The highest accuracy is obtained for $(\gamma, r) = (0.70, 3)$ with an accuracy of 97.50%, indicating that deeper interaction modelling ($r = 3$) is beneficial even for deep learning models.
- The consistency between pgpDA and deep learning in terms of optimal (γ, r) values clearly demonstrates the robustness of the hierarchical interaction modelling introduced by KSGH.

5.8. Validation on Simulated Data

To assess the performance of the proposed kernel-based approach in a controlled environment, we first conducted experiments on simulated data specifically designed to reproduce a hierarchical questionnaire structure. The dataset was generated according to the following specifications:

- $p = 3$ binary main variables $X = (X_1, X_2, X_3)$,

- for each main variable X_j, $q_j = 3$ associated binary sub-questions Z_{j1}, Z_{j2}, Z_{j3},
- a total sample size of $n = 500$ independent individuals.

The main variables X_j were sampled from independent Bernoulli distributions with fixed probability

$$\mathbb{P}(X_j = 1) = 0.5.$$

Each associated sub-question variable Z_{jk} was then generated conditionally on its parent variable X_j according to:

$$\mathbb{P}(Z_{jk} = 1 \mid X_j = 1) = 0.8, \qquad \mathbb{P}(Z_{jk} = 1 \mid X_j = 0) = 0.2.$$

This controlled dependency structure enforces a positive correlation between each main feature and its descendants, reflecting a realistic hierarchical pattern commonly observed in medical questionnaires.

We compared the performance of a standard SVM classifier equipped with a classical RBF kernel against an SVM with our KSGH kernel. The results showed that KSGH systematically outperforms the standard RBF kernel on data known to be hierarchical. Furthermore, performance increased with the interaction order r, with the best results obtained for $r = 3$, confirming the benefit of explicitly modeling interactions among sub-variables.

6. Conclusion

This work was motivated by the need to incorporate the hierarchical structure of medical questionnaire data—particularly verbal autopsy interviews—into statistical learning methods. We introduced a new kernel designed specifically for binary predictors organized in a two-level tree structure, where main items and their associated sub-items capture clinically meaningful dependencies. The proposed hierarchical interaction kernel integrates both the presence of hierarchical relationships and the interactions among sub-variables of arbitrary order, thereby providing a richer representation of symptom patterns.

Embedded within the pgpDA classification framework, the kernel yields a flexible and parsimonious model that avoids the numerical instabilities commonly encountered in kernel-based Gaussian process mixture approaches. The experimental results on a large Senegalese verbal autopsy dataset demonstrate consistent improvements over standard kernels and widely used machine learning methods, such as SVMs, Random Forests, and the Tariff method. Performance increases with the interaction order, confirming the diagnostic relevance of symptom combinations and hierarchical structure.

Deep learning experiments further confirm that the kernel encodes robust and informative structure: accuracies above 97% were obtained for the best configurations, highlighting the kernel's potential for hybrid kernel–deep architectures. The agreement between pgpDA and deep learning regarding the optimal parameter region (γ, r) reinforces the relevance and robustness of the proposed kernel.

Several research directions emerge from this study. First, extending the kernel to accommodate *mixed data types* (binary, ordinal, and continuous variables) would broaden its applicability to more general medical and epidemiological datasets. Second, the hierarchical formulation could be adapted to *graph-structured symptom networks*, providing a bridge between kernel methods and probabilistic graphical models. Finally, integration into deep kernel learning architectures could further enhance representation power, especially in high-dimensional or weakly supervised settings.

Overall, the proposed hierarchical interaction kernel provides a principled, interpretable, and effective approach to modeling structured binary data, with substantial potential for future extensions in medical diagnosis and beyond.

Acknowledgements

The author is grateful to the demographic surveillance teams in Niakhar, Bandafassi and Mlomp, and to the physicians involved in cause-of-death assignment, for providing the verbal autopsy data used in this study.

References

[1] M. Anker. The effect of misclassification error on reported cause-specific mortality fractions from verbal autopsy. *International Journal of Epidemiology*, 26(5):1090–1096, 1997.

[2] Organisation Mondiale pour la Santé. Normes d'autopsies verbales: Etablissements de la cause de décès. *OMS*, 2009.

[3] B.C. Reeves and M. Quigley. A review of data-derived methods for assigning causes of death from verbal autopsy data. *International Journal of Epidemiology*, 26(5):1080–1089, 1997.

[4] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[5] F. Bach, G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

[6] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12(null):2211–2268, July 2011.

[7] G. Lanckriet, T. De Bie, N. Cristianini, M. Jordan, and W.S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[8] Jérôme Mariette and Nathalie Villa-Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6):1009–1015, 10 2017.

[9] Chi-Shian Dai and Jun Shao. Kernel regression utilizing heterogeneous datasets. *Statistical Theory and Related Fields*, 8(1):51–68, 2024.

[10] F. Bach. Hierarchical kernel learning. 2010.

[11] Ingo Steinwart, Philipp Thomann, and Nico Schmid. Learning with hierarchical gaussian kernels, 2016.

[12] Wentao Huang, Houbao Lu, and Haizhang Zhang. Hierarchical kernels in deep kernel learning. *Journal of Machine Learning Research*, 24(391):1–30, 2023.

[13] Wentao Huang, Houbao Lu, and Haizhang Zhang. Hierarchical kernels in deep kernel learning. *J. Mach. Learn. Res.*, 24(1), March 2024.

[14] G. King, Y. Lu, et al. Verbal autopsy methods with multiple causes of death. *Statistical Science*, 23(1):78–91, 2008.

[15] P. Byass, D. Chandramohan, S.J. Clark, et al. Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global Health Action*, 5, 2012.

[16] C.J. Murray, R. Lozano, A.D. Flaxman, P. Serina, D. Phillips, A. Stewart, S.L. James, A. Vahdatpour, C. Atkinson, M.K. Freeman K, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC Medicine*, 12(1):5, 2014.

[17] S.N. Sylla, S. Girard, A.K. Diongue, A. Diallo, and C. Sokhna. A classification method for binary predictors combining similarity measures and mixture models. *Dependence Modeling*, 3:1090–1096, 2015.

[18] A. Desgrées du Loû, G. Pison, B. Samb, and J.F. Trape. L'évolution des causes de décès d'enfants en Afrique : une étude de cas au Sénégal avec la méthode d'autopsie verbale. *Population*, pages 845–882, 1996.

[19] G. Duthé, S.H.D.Faye, E. Guyavarch, P. Arduin, A.M. Kanté, A. Diallo, R. Laurent, A. Marra, and G. Pison. Changement de protocole dans la méthode d'autopsie verbale et mesure de la mortalité palustre en milieu rural sénégalais. *Bulletin de la Société de Pathologie Exotique*, 103(5):327–332, 2010.

[20] P. Byass, E. Fottrell, D.L Huong, Y. Berhane, T. Corrah, K. Kahn, and L. Muhe. Refining a probabilistic model for interpreting verbal autopsy data. *Scandinavian journal of public health*, 34(1):26–31, 2006.

[21] A.D. Flaxman, A. Vahdatpour, S. Green, S.L. James, et al. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*, 9(1):1, 2011.

[22] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis.* Cambridge University Press, 2004.

[23] B. Schölkopf and A.J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* The MIT Press, 2002.

[24] T. Hofmann, B. Schölkopf, and A.J. Smola. Kernel methods in machine learning. *The Annals of Atatistics*, pages 1171–1220, 2008.

[25] B. Scholkopf and A.J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* The Mit Press, 2001.

[26] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 775–782. ACM, 2007.

[27] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Computer Vision, 2009 IEEE 12th International Conference*, pages 221–228. IEEE,

2009.

[28] F. Suard, A. Rakotomamonjy, and A. Bensrhair. Model selection in pedestrian detection using multiple kernel learning. In *Intelligent Vehicle Symposium*, pages 13–14, 2007.

[29] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference*, pages 1–8. IEEE, 2007.

[30] D. Haussler. Convolution kernels on discrete structures. Technical report, Citeseer, 1999.

[31] M. Collins and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*, pages 625–632, 2001.

[32] H. Kashima and T. Koyanagi. Kernels for semi-structured data. volume 2, pages 291–298.

[33] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002.

[34] T. Gärtner. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1):49–58, 2003.

[35] Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, page 364–371. IEEE, December 2017.

[36] Kamran Kowsari, Rasoul Sali, Lubaina Ehsan, William Adorno, Asad Ali, Sean Moore, Beatrice Amadi, Paul Kelly, Sana Syed, and Donald Brown. Hmic: Hierarchical medical image classification, a deep learning approach. *Information*, 11(6):318, June 2020.

[37] Yanwei Cui, Laetitia Chapel, and Sébastien Lefèvre. Scalable bag of subpaths kernel for learning on hierarchical image representations and multi-source remote sensing data classification. *Remote Sensing*, 9(3):196, 2017.

[38] Lu Bai, Lixin Cui, and Edwin R. Hancock. A hierarchical transitive-aligned graph kernel for un-attributed graphs. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 1327–1336. PMLR, 2022.

[39] Xinlei Wang, Junchang Xin, Zhongyang Wang, Luxuan Qu, Jiani Li, and Zhiqiong Wang. Graph kernel of brain networks considering functional similarity measures. *Computers in Biology and Medicine*, 171:108148, 2024.

[40] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Statistics and Computing*, pages 1–20, 2014.

[41] M. Fauvel, C. Bouveyron, and S. Girard. Parsimonious Gaussian process models for the classification of hyperspectral remote sensing images. *Geoscience and Remote Sensing Letters, IEEE*, 12(12):2423–2427, 2015.

[42] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston,

MA, 2009.

[43] Jung Hun Oh, Rena Elkin, Anish Kumar Simhal, Jiening Zhu, Joseph O Deasy, and Allen Tannenbaum. Optimal transport for kernel gaussian mixture models, 2023.

[44] R.B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[45] H. Kashima and Y. Tsuboi. Kernel-based discriminative learning algorithms for labeling sequences, trees, and graphs. In *Proceedings of the twenty-first international conference on Machine learning*, page 58. ACM, 2004.

[46] Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel banach spaces, 2021.

[47] P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vandoise Sci Nat*, 37:547–579, 1901.

[48] A. Tversky. Feature of similarity. *Psychological Review*, 84:327–352, 1977.

[49] V. Batagelj and M. Bren. Comparing resemblance measures. *Journal of Classification*, 12:73–90, 1995.

[50] Z. Hubalek. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57(4):669–689, 1982.

[51] D.A. Jackson, K.M. Somers, and H.H. Harvey. Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist*, 133(3):436–453, 1989.

[52] P. Willett. Similarity-based approaches to virtual screening. *Biochemical Society Transactions*, 31(3):603–606, 2003.

[53] D. Chandramohan, P. Setel, and M. Quigley. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *International Journal of Epidemiology*, 30(3):509–514, 2001.

[54] B. Snow and K. Marsh. How useful are verbal autopsies to estimate childhood causes of death? *Health policy and planning*, 7(1):22–29, 1992.

[55] C.J. Murray, A.D. Lopez, D.M. Feehan, S.T Peter, and G. Yang. Validation of the symptom pattern method for analyzing verbal autopsy data. 4(11):e327.

[56] P.W. Setel, C. Rao, Y. Hemed, D.R. Whiting, and G. Yang. Core verbal autopsy procedures with comparative validation results from two countries. *PLoS Med*, 3(8):e268, 2006.

[57] B.A Lopman, R.V. Barnabas, J.T. Boerma, G. Chawira, K. Gaitskell, T. Harrop, P. Mason, et al. Creating and validating an algorithm to measure aids mortality in the adult population using verbal autopsy. *Plos Med*, 3(8):e312.

[58] M. Fantahun, E. Fottrell, Y. Berhane, S. Wall, U. Högberg, and P. Byass. Assessing a new approach to verbal autopsy interpretation in a rural ethiopian community: the interva model. *Bulletin of the World Health Organization*, 84(3):204–210, 2006.

[59] S.L. James, A.D. Flaxman, C.J. Murray, et al. Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(1):31, 2011.