

## THE ANALYSIS FROM NONLINEAR DISTANCE METRIC TO KERNEL-BASED PRESCRIPTION PREDICTION SYSTEM

DER-CHEN CHANG<sup>1,2</sup>, OPHIR FRIEDER<sup>1</sup>, CHI-FENG HUNG<sup>3</sup>, HAO-REN YAO<sup>1\*</sup>

<sup>1</sup>*Department of Computer Science, Georgetown University, Washington, D.C., USA*

<sup>2</sup>*Department of Mathematics & Statistics, Georgetown University, Washington, D.C., USA*

<sup>3</sup>*School of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan*

**Abstract.** The distance metric and its nonlinear variant play a substantial role in machine learning, particularly yoso in building kernel functions. Often, the Euclidean distance with a radial basis function (RBF) is used to construct a RBF kernel for nonlinear classification. However, domain implications periodically constrain the distance metrics. Specifically, within the domain of drug efficacy prediction, distance measures must account for time that varies based on disease duration, short to chronic. Recently, a distance-derived graph kernel approach was commercially licensed for drug prescription efficacy prediction. The analysis of the distance functions used therein, namely the Euclidean and cosine distance measures and their respective derived graph kernels, is provided. Theoretically, we provide a formulation of our efforts and demonstrate how both the Euclidean and cosine distance induce space and discuss the difference from geometric perspectives. The aforementioned approach is likewise empirically evaluated using a million-plus patient subset of a life-spanning, real-world, electronic health record database. Diseases are characterized as either short in duration or chronic and either common, hence balanced data, or relatively rare, hence imbalanced. Empirically, the system accurately predicted the efficacy of prescriptions for both balanced and imbalanced and short-term and chronic diseases, with at least one of the measures used being statistically significantly superior to conventional prediction methods. Succinctly, for short-term, balanced diseases, the Euclidean and cosine measures were generally statistically equivalent. For short-term, imbalanced diseases however, the Euclidean measure was superior to the cosine measure, at times and not infrequently, statistically significantly so. For chronic, balanced diseases, Euclidean was slightly superior to the cosine measure, but they were statistically equivalent. In contrast, for chronic, imbalanced diseases, the cosine measure was consistently statistically significantly superior to the Euclidean measure. These findings indicate the need for both measures depending on the use case. Our empirical findings match our theoretical underpinnings.

**Keywords.** Machine Learning; Graph Kernel; Deep Metric Learning; Geometry.

### 1. INTRODUCTION

Distance metrics and their nonlinear variants play a fundamental role in machine learning tasks. They measure the degree of linear or nonlinear similarity between objects, grouping similar objects into  $k$  different groups (e.g.,  $k$ -means clustering) or assigning class labels based

---

\*Corresponding author.

E-mail address: hao-ren@ir.cs.georgetown.edu (H.R. Yao)

Received February 3, 2021; Accepted February 24, 2021.

on their nearest neighbors (e.g.,  $k$ -nearest neighbor classification) [1]. There are multiple choices of distance functions to solve specific problems. Generally, Euclidean distance is used for the majority of classification problems, whereas, cosine distance is more suitable for document classification [2, 3]. Distance metrics can also be part of the kernel function design [4]. In [5], the authors used the Euclidean distance as part of a graph kernel to capture time difference similarity between patients. In [6], the authors demonstrated the positive definiteness for a distance substitution kernel. Usually, the kernel function  $k$  for two data objects  $x$  and  $y$  and user defined parameter  $\alpha$  is defined as:

$$k_t(x, y) = C_t e^{-\alpha \|x-y\|},$$

where  $\|x-y\|$  is the Euclidean distance between  $x$  and  $y$ , and  $C_t$  is a constant depending on time  $t$ .

Distance metric learning, on the other hand, tries to find the optimal distance metric or the embedding space given a task specific objective for a set of data objects [7]. Under this scope, semantically similar objects are encouraged to be closer to each other and further apart for non-similar objects. The prevalence of deep learning has largely populated the development of the deep metric learning, where nonlinear observation can be captured, specifically in unsupervised representation learning for images (e.g., computer vision [8, 9]). It also motivated the deep kernel learning [10, 11]. In [12], a deep metric learning based graph kernel is proposed to solve the outcome prediction problem in complex chronic disease treatment planning, where the cosine distance was shown superior to the Euclidean distance measure.

In practice, both distance metrics and kernel functions are used to solve complex real-world problems, especially in medicine. A problem of interest is the drug prescription efficacy prediction [13, 14]. Accurate predictive models for drug prescription improve healthcare. They can further reduce medication errors and identify possible drug prescription pathways to pursue for clinical personnel. In [15], a framework was proposed to predict success and failure outcomes of a given drug prescription for antibiotic treatment-based disease. The approach is further extended to overcome biased data distribution [5] and chronic disease treatment plan [12]. Moreover, Euclidean and cosine distance measures exhibit different performance behaviors. In [12], the cosine distance was found superior to the Euclidean measure under highly data imbalanced chronic disease, however, the Euclidean distance was used for short-term disease in [5]. To further investigate such differences, we establish a unified framework, which integrates two model structures proposed in the aforementioned works, to conduct rigorous empirical evaluation on all diseases investigated in previous studies [5, 12], in addition to a theoretical discussion. The aforementioned prescription efficacy prediction approaches are now under commercial licensing.

Our contributions are as follows:

- We propose a scalable unified framework for prescription efficacy prediction.
- We evaluate performance using 10-fold cross validation on large-scale, real-world, electronic health record dataset that includes common and rare, short and chronic illnesses.
- We investigate the difference between Euclidean and cosine distance on learned embedding space.
- We provide a theoretical explanation from geometry perspective, generalizing Euclidean and cosine distance.

## 2. PRELIMINARIES

Our previous efforts [5, 12, 15] present a graph kernel-based system for outcome prediction of drug prescription, particularly the success or failure treatment, on short-term and chronic diseases. In [5], a Multiple Graph Kernel Fusion (MGKF) was proposed to overcome noise effect on short-term disease. A deep graph kernel learning approach, e.g., Cross-Global Attention Graph Kernel Network (Cross-Global), is proposed in [12] to handle long-term chronic disease. In short, we initially determine success and failure patients for the target disease treatment as training data within a user-defined time quantum, where a set of medical events are extracted between this time period. Then, we construct a patient graph, a graphical representation of the patient EHRs, given the extracted medical events. Finally, we perform binary graph classification as prediction through a graph kernel and a kernel-based classifier. We detail each part of the prediction framework in this section.

**2.1. Outcome Selection.** We define a failure drug prescription or treatment plan for a disease diagnosis if certain events occur within a predefined time period, otherwise a success. In short-duration diseases, we observe the similar or identical type of disease diagnosis, while in chronic disease, the observed target will be severe complication defined via medication guideline. We name a disease as short-term if it only considers a single medication with immediate outcome observation and recent medical history (e.g., 2 months prior to the diagnostic). For a chronic disease, we consider a multiple medication treatment plan with long-term outcome observation and medical history (e.g., 10 years prior to the diagnostic). We refer readers to MGKF [5] and Cross-Global [12] for greater detail.

**2.2. Patient Graph.** A subset of patient’s EHRs is formulated as directly acyclic graph where a node represents each medical event and an edge with time difference (e.g., days) used as a weight connects two consecutive medical events. Patient demographic information, such as gender and age, is included by connecting to the first medical event with age as an edge weight<sup>1</sup>. We define a patient graph here as in [5] and [12]:

**Definition 2.1 (Patient Graph).** Given  $n$  medical events, set  $M = \{(m_1, t_1), \dots, (m_n, t_n)\}$  represents a patient’s EHR with  $m_i$  denoting a medical event such as diagnosis, and  $t_i$  denoting the time for  $m_i$ . The patient graph  $P_g = (V, E)$  of events  $M$  is a weighted directed acyclic graph with its vertices  $V$  containing all events  $m_i \in M$  and edges  $E$  containing all pairs of consecutive events  $(m_i, m_j)$ . The edge weight from node  $i$  to node  $j$  is defined as  $W_{ij} = t_j - t_i$  which defines the time interval between  $m_i, m_j$ .

**2.3. Graph Kernel.** For kernel-based binary graph classification, building a pairwise kernel matrix between patient graphs is the first step. The graph kernel computes the similarity between pairs of graph. It is also a positive definite or semidefinite kernel defined for graphs, which performs an inner product implicitly by mapping data point from an input space to the Hilbert space. It can also be treated as a similarity measurement between two data objects (e.g, graphs). We point readers to these articles [16] for a more in-depth graph kernel discussion and [4] for a better understanding and design principle of graph kernel and its associated

<sup>1</sup>To simplify model assumption, we only use gender and age as demographic information.

feature maps. In [5, 15], several graph kernels were proposed to solve the drug prescription outcome prediction problem as patient graph classification. Please refer [5, 15] for more in-depth descriptions on kernel definitions.

**2.4. Prediction Framework.** We then formulate a binary graph classification problem on the resulting patient graph by using a kernelized Support Vector Machine (K-SVM) [15], Multiple Graph Kernel Fusion (MGKF) [5], and Cross-Global Attention Graph Kernel Network (Cross-Global) [12].

As mentioned in Section 1, cosine distance is superior to its Euclidean counterpart under highly imbalanced chronic disease (Cross-Global), while both the Euclidean and cosine distance measures achieve high prediction performance under short-term diseases (MGKF). We now question how they relate to each other. It is not suitable to directly compare Cross-Global and MGKF since the following:

- Different datasets<sup>2</sup> are used in MGKF and Cross-Global.
- Different model structures and optimization perspectives. In MGKF, optimization aims at generating optimal kernel fusion, while it turns out to find optimal graph embedding under Cross-Global.
- Different data balance and imbalance ratio between short-term and chronic disease.

To fairly compare MGKF and Cross-Global with Euclidean and cosine distance under short-term and chronic disease, a unified framework is required. Here, we extend and generalize previous efforts to differentiate the behavior of Euclidean and cosine distance, in addition to the theoretical discussion. A unified framework for graph-kernel based drug prescription outcome prediction is presented to conduct a rigorous empirical evaluation on all diseases in previous works on a very large-scale real-world EHRs.

### 3. DISCUSSION FROM THE GEOMETRIC POINT OF VIEW

**3.1. Riemannian and SubRiemannian geometries.** To discuss differences between Euclidean and cosine distance, we first establish some mathematical properties with those distance under geometry point of view. In fact, we may consider this problem in a more general setting.

Let  $\mathcal{X} = \{X_1, \dots, X_m\}$  be  $m$  linearly independent vector fields on an  $n$ -dimensional real manifold  $\mathcal{M}_n$  with  $m \leq n$  of the tangent bundle  $T\mathcal{M}_n$ . To find a good kernel function to describe the diffusion (energy flow) between two points in  $\mathcal{M}_n$ , we need to solve the heat equation associate to the the sum of square of  $X_j$ 's:

$$\Delta_{\mathcal{X}} = \frac{1}{2} \sum_{j=1}^m X_j^2.$$

When  $m = n$ , the operator  $\Delta_{\mathcal{X}}$  is elliptic. Assume that

$$X_j \perp X_k, \quad 1 \leq j, k \leq n, \quad j \neq k, \quad \text{and} \quad \|X_j\| = 1.$$

In this case, we have a natural volume element which yields the adjoint vector fields  $X_j^*$  for  $1 \leq j \leq m$ . More precisely, for  $\phi, \psi \in C_0^\infty(\mathcal{M}_n)$ ,

$$\int_{\mathcal{M}_n} (X_j \phi) \psi = \int_{\mathcal{M}_n} \phi (X_j^* \psi),$$

---

<sup>2</sup>Different database provider.

whence  $-\frac{1}{2} \sum_{j=1}^m X_j^* X_j$  is the classical Laplace-Beltrami operator whose second order part agrees with the operator  $\Delta_{\mathcal{X}}$ . This suggests us that we may use the given differential operator  $\Delta_{\mathcal{X}}$  to introduce a geometry on  $\mathcal{M}_n$  which may help us to solve  $\Delta_{\mathcal{X}}$  and hence the heat equation. Hence, for  $n \geq 2$ , the solving kernel for the heat operator  $\frac{\partial}{\partial t} - \Delta_{\mathcal{X}}$  takes the form

$$\mathcal{P}_t(x, x_0) = \frac{1}{(2\pi t)^{\frac{n}{2}}} e^{-\frac{d^2(x, x_0)}{2t}} (a_0 + a_1 t + a_2 t^2 + \dots).$$

The  $a_j$ 's are functions of  $x$  and  $x_0$ . Here  $d(x, x_0)$  represents the induced Riemannian distance between the points  $x_0$  and  $x$  in  $\mathcal{M}_n$ . Moreover,  $+\dots$  stands for negligible error. Furthermore,

$$\frac{\partial}{\partial t} \left( \frac{d^2(x, x_0)}{2t} \right) + \frac{1}{2} \sum_{j=1}^n \left( X_j \frac{d^2(x, x_0)}{2} \right)^2 = \frac{\partial}{\partial t} \left( \frac{d^2(x, x_0)}{2t} \right) + \frac{1}{2} H \left( X_1 \left( \frac{d^2}{2t} \right), \dots, X_n \left( \frac{d^2}{2t} \right) \right) = 0,$$

i.e.,  $\frac{d^2(x, x_0)}{2t}$  is a solution of the *Hamilton-Jacobi equation*. Here  $H$  is the Hamiltonian function associated with  $\Delta_{\mathcal{X}}$ .

The simplest example will be the Euclidean distance  $d(x, y) = \|x - y\|$  and the kernel  $\mathcal{P}_t(x, x_0)$  is the Gaussian (see [17]). In this paper, we are going to use another non-trivial example of Riemannian metric. Given a large sample space, we first embed those sample  $s$  in an  $n$ -dimensional sphere  $\mathcal{M}_n = \mathbb{S}^{n-1}$ . Given two points  $x_1$  and  $x_2$  in  $\mathcal{M}_n$ , we define the “distance”  $d(x_1, x_2) = 1 - \cos \left( \frac{\langle x_1, x_2 \rangle}{\|x_1\| \|x_2\|} \right)$ . Here  $\|x\|$  is the Euclidean distance between the point  $x$  and the origin. This is so-called *cosine metric*. In other words,  $x_1$  and  $x_2$  are on the same sphere. Hence,  $x_1$  and  $x_2$  are located on a “big circle” which is determined by the center  $x_0$  of the sphere and these two points. Instead of measuring the arc-length (which maybe huge), we consider the angle  $\theta$  between  $x_0 \vec{x}_1$  and  $x_0 \vec{x}_2$ . This metric provides better estimates for the kernel in applications of drug prescription prediction system for long term disease.

When  $m < n$ , the operator is non-elliptic. In this case, the subspace  $\mathcal{D} = \text{span}\{X_1, \dots, X_m\}$  is called the “horizontal subspace” of  $T\mathcal{M}_n$ , and the vectors  $X \in \mathcal{D}_p$  are called horizontal vectors at  $p$ . Sometimes, we call the distribution  $\mathcal{D}$  the *horizontal distribution*. The sections in the horizontal bundle  $(\mathcal{D}, \mathcal{M}_n)$  are called horizontal vector fields. They are smooth assignments  $\mathcal{M}_n \ni p \rightarrow X_p \in \mathcal{D}_p$ . The set of the horizontal vector fields on  $\mathcal{M}_n$  will be denoted by  $\Gamma(\mathcal{D})$ . If  $U$  is an open subset of  $\mathcal{M}_n$ , the set of horizontal vector fields on  $U$  will be denoted by  $\Gamma(\mathcal{D}, U)$ . We call the complement of  $\Gamma(\mathcal{D}, U)$  the “missing direction” at  $U$ .

Now we encounter new problems since  $\mathcal{D} \neq T\mathcal{M}_n$ , and we cannot find arc-length in general. We overcome this difficulty by assuming bracket generating property: “the horizontal vector fields  $\mathcal{X}$  and their brackets span  $T\mathcal{M}_n$ ”, then Chow’s theorem [18] to conclude that given any two points  $A, B \in \mathcal{M}_n$ , there is a piecewise  $C^1$  horizontal curve  $\gamma: [0, 1] \rightarrow \mathcal{M}_n$  such that

$$\gamma(0) = A, \quad \gamma(1) = B,$$

and

$$\dot{\gamma}(s) = \sum_{k=1}^m a_k(s) X_k.$$

This yields a distance, and therefore a geometry, which we shall call *subRiemannian geometry*. SubRiemannian geometry was first discussed in the field of thermodynamics around 1800s. Carnot discovered the principle of an engine in 1824 involving two isotherms and two adiabatic processes, Jule studied adiabatic processes, and Clausius formulated the existence of the

entropy in the second law of thermodynamics in 1854. In 1909 Carathéodory made the point regarding the relationship between the connectivity of two states by adiabatic processes and nonintegrability of a distribution, which is defined by the one form of work. Chow proved the general global connectivity in 1934 which was used in studying of partial differential equations. There are significant differences between Riemannian and subRiemannian geometries. However, this geometry can be applied in many situations in our daily life. For more details, readers can read the book by Calin and Chang [19].

A subRiemannian structure over a manifold  $\mathcal{M}_n$  is a pair  $(\mathcal{D}, \langle \cdot, \cdot \rangle)$ , where  $\mathcal{D}$  is a bracket generating distribution and  $\langle \cdot, \cdot \rangle$  a fibre inner product defined on  $\mathcal{D}$ . The length of the horizontal curve  $\gamma$  is

$$\ell(\gamma) := \int_0^\tau \sqrt{\langle \dot{\gamma}(s), \dot{\gamma}(s) \rangle} ds.$$

The shortest length  $d_{cc}(A, B)$  is called a *Carnot-Carathéodory distance* between  $A, B \in \mathcal{M}_n$  which is given by

$$d_{cc}(A, B) := \inf \ell(\gamma),$$

where the infimum is taken over all absolutely continuous horizontal curves joining  $A$  and  $B$  [20].

Here we mention two examples of subRiemannian geometry. Let  $X_1 = \frac{\partial}{\partial x}$  and  $X_2 = x^m \frac{\partial}{\partial y}$  be the the Grushin vector fields [21, 22] in  $\mathbf{R}^2$  which satisfy bracket generating property, *i.e.*,

$$[X_1, [X_1, \dots, [X_1, X_2] \dots]] = m! \frac{\partial}{\partial y}.$$

These vector fields can be used to described parallel parking and even self-driven cars [23].

**3.2. Horizontal Connectivity.** In outcome prediction task of drug prescription, one of the main difficulties is to overcome the distinguishing features under short-term and long-term disease progression. Moreover, for long term diseases, we need to avoid some low-effective (or even useless or dangerous) drugs. The answer to this question not only helps us better characterize the embedding space inferred by Euclidean and subRiemannian distances, but also leads to different optimization formulations. Mathematically, the first task is to address the following question: *Given any two points on a topologically connected subRiemannian manifold, under what conditions can we join them by a horizontal curve?* In the outcome prediction task of drug prescription, we must distinguish features under short-term and long-term disease progression. The answer to this question not only helps us to better characterize the embedding space inferred by Euclidean and cosine distance, but also lead to useful optimization formulation under their embedding properties.

To answer this question, we need to prove the following two results. Readers can find more detailed discussions in the book by Calin and Chang [19].

**Proposition 3.1.** Let  $U \subset \mathbf{R}^n$  be an open set and  $\mathcal{D} = \text{span}\{X_1, \dots, X_m\}$  be a differentiable distribution on  $U$ . Then for any point  $p \in U$  there is a manifold  $V_p^m$  such that

- (1).  $p \in V_p^m$ ;
- (2).  $\dim p \in V_p^m = m$ ;
- (3). any two points of  $V_p^m$  can be joined by a piecewise horizontal curve.

*Proof.* Let  $X_j = \sum_{k=1}^n a_{jk} \partial_{x_k}$  be the vector fields in local coordinates. Consider the ODE system

$$\frac{dx_\ell(t)}{dt} = -F_\ell(x(t), u), \quad \ell = 1, \dots, n, \quad (3.1)$$

where  $F_\ell(x(t), u) = \sum_{j=1}^m a_{j\ell}(x(t))u_j$  with  $(u_1, \dots, u_m) \in \mathbf{R}^m$ , is a system with  $m$  parameters.

The solutions of (3.1) are horizontal curves with controls  $u_j$ . Let  $x_\ell(0) = x_{\ell 0}$  be the initial conditions of system (3.1). Standard theorems of ODE system provide the existence and local uniqueness of the solutions, which can be expressed by

$$x_\ell(t) = \phi(t; \mathbf{x}_0; u) = \phi_\ell(t; x_{10}, \dots, x_{n0}; u_1, \dots, u_m),$$

for  $|t| < \varepsilon$ , with  $\phi_\ell(0; \mathbf{x}_0; u) = \mathbf{x}_0$ . Since the vector components  $a_{j\ell}$  are differentiable, a general theorem states that the functions  $\phi_\ell$  are twice differentiable with respect to  $t$  and locally continuous differentiable with respect to  $\mathbf{x}_0$ .

Since system (3.1) is autonomous, a simple application of the chain rule shows that the functions  $\phi_\ell$  verify the relations

$$\phi_\ell(tt_0; \mathbf{x}_0; u) = \phi_\ell(t_0; \mathbf{x}_0; tu) = \psi_\ell(\mathbf{x}_0; tu),$$

where  $|t_0| < \varepsilon$  and  $|tu| < \varepsilon$ .

Applying the theorem on differentiability with respect to a parameter to system (3.1) yields that  $\phi_\ell$  are continuous differentiable with respect to  $u_\ell$  if  $|u_\ell| < \delta$  with sufficiently small  $\delta > 0$ .

If we let  $tu_\ell = \lambda_\ell$ , for  $\ell = 1, \dots, m$ , then the formulas

$$x_\ell = \psi_\ell(\mathbf{x}_0; \lambda_1, \dots, \lambda_m), \quad \ell = 1, \dots, n,$$

for  $|\lambda_j| < \varepsilon\delta$  define a  $m$ -dimensional manifold  $V_p^m$  passing through the point  $(x_{10}, \dots, x_{n0})$ . To finish the proof we will need to show that the rank of Jacobian  $\frac{\partial \psi_\ell}{\partial \lambda_j}$  is maximum, i.e., equal to  $m$ . This is equivalent with the fact that the vector fields

$$\frac{\partial \psi_1}{\partial \lambda_1}, \frac{\partial \psi_1}{\partial \lambda_2}, \dots, \frac{\partial \psi_n}{\partial \lambda_m}$$

are linearly independent. Since  $\lambda_\ell = tu_\ell$ , it suffices to show that

$$\frac{\partial \phi_1}{\partial u_1}, \frac{\partial \phi_1}{\partial u_2}, \dots, \frac{\partial \phi_n}{\partial u_m}$$

are linearly independent. Since

$$\phi(t; \mathbf{x}_0; u) = \exp\left(t \sum_{k=1}^m u_k X_k\right),$$

it follows that

$$\frac{\partial \phi}{\partial u_\ell} = tX_\ell, \quad \ell = 1, \dots, m,$$

which are linearly independent vector fields for  $t \neq 0$ . It follows that

$$\text{rank}\left(\frac{\partial \psi_\ell}{\partial \lambda_k}\right) = m.$$

The proof of this proposition is therefore complete.  $\square$



**Proposition 3.2.** Let  $\mathcal{D}$  be a nonintegrable distribution. Assume that through each point of the domains  $\mathcal{R} \subset U$  passes a  $\mathcal{D}$ -connected  $m$ -dimensional manifold  $V^m$  defined by the equations

$$x_j = f_j(\xi_1, \dots, \xi_m; c_1, \dots, c_{n-m}), \quad j = 1, \dots, n, \quad (3.2)$$

where  $f_j$  are continuous differentiable functions on a domain  $\mathcal{R}^* \subset \mathbf{R}^n$ , such that

$$\frac{\partial(f_1, \dots, f_n)}{\partial(\xi_1, \dots, \xi_m; c_1, \dots, c_{n-m})} \neq 0 \quad \text{on } \mathcal{R}^*.$$

Then there is a domain  $\mathcal{R}_1 \subset \mathcal{R}^*$  such that

- (1). for all  $\mathcal{R}_1$ , there is a  $\mathcal{D}$ -connected  $(m+1)$ -dimensional manifold  $V_p^{m+1}$  passing through  $p$ ;
- (2). the functions that define the manifolds  $V_p^{m+1}$  on  $\mathcal{R}_1$  have the same properties as the functions  $f_j$ 's in (3.2).

*Proof.* Let  $\mathcal{D} = \text{span}\{X_1, \dots, X_m\}$  be the horizontal distribution and

$$\mathcal{I} = \{\theta_1, \dots, \theta_{n-m}\}$$

be the extrinsic ideal associated with  $\mathcal{D}$ . Since the distribution  $\mathcal{D}$  is not integrable, the Pfaff system  $\mathcal{I}$  is not integrable; i.e., it cannot have integral manifolds of dimension  $m$ .

*Proof of Statement (1).* For any  $p \in \mathcal{R}$ , there is a horizontal vector  $X_p \in \mathcal{D}_p$  such that  $X_p \notin T_p V_p^m$ , i.e., not tangent to the manifold  $V_p^m$ .

The proof of (1) is by contradiction. Let  $p \in \mathcal{R}$  be a fixed point. Assume that any horizontal vector field  $X$  about  $p$  is tangent to the manifold  $V_p^m$ . Then

$$X_q \in T_q V_p^m, \quad \forall q \in V_p^m.$$

Therefore  $\mathcal{D}_q \subset T_q V_p^m$ , and since  $\dim(\mathcal{D}_q) = \dim(T_q V_p^m) = m$ , it follows that the inclusion is in fact identity; i.e.,  $\mathcal{D}_q = T_q V_p^m$  for all  $q \in V_p^m$ . Since the one-forms  $\theta_j$  vanish on  $\mathcal{D}_q$ , it follows that  $T_q V_p^m$  is an integral  $m$ -plane for the Pfaff system  $\mathcal{I}$  and hence  $V_p^m$  is an integral manifold for  $V_p^m$  is an integral manifold  $\mathcal{I}$ , which is a contradiction, because  $\mathcal{I}$  is not integrable. Hence we prove the assertion (1),

Let  $p_0 \in \mathcal{R}$  be a point with coordinates  $(x_{10}, x_{20}, \dots, x_{+n0})$  and  $u = X_{p_0}$  be the vector given by (1); i.e.,  $u \in \mathcal{D}_{p_0}$  and  $u \notin T_{p_0} V_{p_0}^m$ . Let  $(u_1, \dots, u_m) \in \mathbf{R}^m$  be such that

$$u = \sum_{k=1}^m u_k X_k(p_0).$$

The numbers  $u_k$  will be kept constant for the rest of the proof.

*Proof of Statement (2).* The matrix

$$M = \begin{bmatrix} \frac{\partial f_1}{\partial \lambda_1} & \cdots & \frac{\partial f_n}{\partial \lambda_1} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_1}{\partial \lambda_m} & \cdots & \frac{\partial f_n}{\partial \lambda_m} \\ u_1 X_1(p_0) & \cdots & u_m X_m(p_0) \end{bmatrix}$$

has rank  $m+1$  at the point  $p_0$ .

The first  $m$  rows of the matrix  $M$  are the components of the coordinate vector fields  $\frac{\partial f}{\partial \lambda_j}$  on the manifold  $V_{p_0}^m$ , which are tangent to  $V_{p_0}^m$ , linearly independent, and span the tangent space



$T_{p_0}V_{p_0}^m$ . The last row of  $M$  has the component of the vector  $u$ , which is transversal to  $T_{p_0}V_{p_0}^m$ , so all  $m+1$  vectors are linearly independent at  $p_0$  and hence  $\text{rank}(M) = m+1$ .

Since all the elements of the matrix  $M$  are continuous functions of the coordinates of the point  $p_0$ , while  $u_j$  are still kept constant, there is a subdomain  $\mathcal{R}' \subset \mathcal{R}$  such that  $p_0 \in \mathcal{R}'$  and  $\text{rank}(M) = m+1$  on  $\mathcal{R}'$ .

From the non-vanishing Jacobian condition  $\frac{\partial f}{\partial(\lambda, c)} \neq 0$  on  $\mathcal{R}'$ , it follows that the following  $n$  vector fields

$$\frac{\partial f}{\partial \lambda_1}, \dots, \frac{\partial f}{\partial \lambda_m}, \frac{\partial f}{\partial c_1}, \dots, \frac{\partial f}{\partial c_{n-m}} \quad (3.3)$$

are linearly independent on  $\mathcal{R}'$ .

From the preceding discussion, the following  $m+1$  vector fields

$$\frac{\partial f}{\partial \lambda_1}, \dots, \frac{\partial f}{\partial \lambda_m}, X \quad (3.4)$$

are linearly independent on  $\mathcal{R}'$ . We can complete system (3.4) with  $n - (m+1)$  elements of set (3.3), say

$$\frac{\partial f}{\partial \lambda_1}, \dots, \frac{\partial f}{\partial \lambda_m}, X, \frac{\partial f}{\partial c_1}, \dots, \frac{\partial f}{\partial c_{n-m-1}} \quad (3.5)$$

are linearly independent on  $\mathcal{R}'$ .

In the following, we shall deal with the construction of a  $(m+1)$ -dimensional manifold passing through  $p_0$ , which depends on  $n - (m+1)$  parameters. In equation (3.2), consider the parameters  $c_{n-m}$  frozen. Let  $x_{j0}$  be the coordinates on this new manifold  $V^m$ . Then

$$x_{j0} = g_j(\lambda_1, \lambda_2, \dots, \lambda_m; c_1, \dots, c_{n-m-1}), \quad (3.6)$$

where  $g_j$  is continuous differentiable with respect to  $\lambda_\ell$  and  $c_j$  and

$$\frac{\partial g_\ell}{\partial \lambda_\sigma} = \frac{\partial f_\ell}{\partial \lambda_\sigma}, \quad \frac{\partial g_\ell}{\partial c_k} = \frac{\partial f_\ell}{\partial c_k}, \quad \sigma = 1, \dots, m, \quad k = 1, \dots, n-m-1.$$

The equation of the integral curves of the vector field  $X$  on  $\mathcal{R}'$  are given by

$$x_j = \phi_j(t; x_{10}, x_{20}, \dots, x_{n,0}; u_1, \dots, u_m), \quad j = 1, \dots, n. \quad (3.7)$$

We shall construct a  $(m+1)$ -dimensional manifold by pushing the manifold  $V^m$  in the direction of the integral curves of  $X$ . This can be done by substituting the variables  $x_{j0}$  given by (3.6) into the expressions provided by (3.7). Let  $\lambda_{m+1} = t$  be the  $m+1$  variable. We obtain

$$\begin{aligned} x_\ell &= \phi_\ell(t; \mathbf{x}_0; u_1, u_2, \dots, u_m) \\ &= \phi_\ell(\lambda_{m+1}; g_\ell(\lambda_1, \dots, \lambda_m; c_1, \dots, c_{n-m-1}); u_1, u_2, \dots, u_m) \\ &= F_\ell(\lambda_1, \dots, \lambda_{m+1}; c_1, \dots, c_{n-m-1}) \quad \ell = 1, \dots, n, \end{aligned}$$

where  $u_j$  are kept constant.  $F_\ell$  are continuous differentiable functions of  $\lambda_1, \dots, \lambda_{m+1}, c_1, \dots, c_{n-m-1}$ .

To show that the equations

$$x_\ell = F_\ell(\lambda_1, \dots, \lambda_{m+1}; c_1, \dots, c_{n-m-1}) \quad (3.8)$$

defines a manifold of dimension  $m+1$ , we need to show that

$$\text{rank}\left(\frac{\partial(F_1, \dots, F_n)}{\partial(\lambda_1, \dots, \lambda_{m+1})}\right) = m+1 \quad (3.9)$$

on some neighborhood of  $p_0$  included in  $\mathcal{R}'$ .

Applying the chain rule yields

$$\begin{aligned}\frac{\partial F_\ell}{\partial \lambda_\alpha} &= \frac{\partial \phi_\ell}{\partial x_{j_0}} \frac{\partial x_{j_0}}{\partial \lambda_\alpha} = \frac{\partial \phi_\ell}{\partial x_{j_0}} \frac{\partial f_j}{\partial \lambda_\alpha} \quad \alpha = 1, \dots, m, \\ \frac{\partial F_\ell}{\partial \lambda_{m+1}} &= \frac{\partial \phi_\ell}{\partial t} = u_j X_j, \\ \frac{\partial F_\ell}{\partial c_\beta} &= \frac{\partial \phi_\ell}{\partial x_{j_0}} \frac{\partial x_{j_0}}{\partial c_\beta} = \frac{\partial \phi_\ell}{\partial x_{j_0}} \frac{\partial f_{j_0}}{\partial c_\beta} \quad \beta = 1, \dots, n - (m + 1).\end{aligned}$$

Since  $\det \left[ \frac{\partial \phi_\ell}{\partial x_{j_0}} \right] \neq 0$  on a neighborhood of  $p_0$ , using that vector fields (3.4) are linearly independent yields that

$$\frac{\partial F_\ell}{\partial \lambda_1}, \frac{\partial F_\ell}{\partial \lambda_2}, \dots, \frac{\partial F_\ell}{\partial \lambda_m}, \frac{\partial F_\ell}{\partial \lambda_{m+1}}$$

are linearly independent, which means that (3.9) holds.

Using that (3.5) are linearly independent on  $\mathcal{R}'$ , it follows that the vector fields

$$\frac{\partial F_\ell}{\partial \lambda_\alpha}, \frac{\partial F_\ell}{\partial \lambda_{m+1}}, \frac{\partial F_\ell}{\partial c_\beta}, \quad \alpha = 1, \dots, m, \quad \beta = 1, \dots, n - (m + 1),$$

are linearly independent on a subdomain  $\mathcal{R}_1 \subset \mathcal{R}'$ , which contains  $p_0$ . Therefore

$$\frac{\partial(F_1, \dots, F_n)}{\partial(\lambda_1, \dots, \lambda_{m+1}, c_1, \dots, c_{n-m-1})} \neq 0 \quad \text{on } \mathcal{R}_1,$$

and hence the functions  $F_\ell$  have the same properties as the functions  $f_j$  in (3.2).

In conclusion, through each point of  $\mathcal{R}_1$  passes a  $\mathcal{D}$ -connected  $m + 1$  manifold defined by equations (3.8), and each manifold depends on  $n - (m + 1)$  parameters. We finish the proof of this proposition.  $\square$

Now we are in a position to prove the local connectivity property. This result was proved by Teleman [24] for the Pfaff systems that do not contain integrable combinations in 1957. Here we shall prove it from the point of view of distributions.

**Theorem 3.3.** Let  $\mathcal{D}$  be a nonintegrable differentiable distribution of rank  $m$  on the open set  $U \subset \mathbf{R}^n$ . Then any domain  $U_1 \subset U$  contains a subdomain  $U_2 \subset U_1$  such that for any  $p, q \in U_2$ , there is a piecewise horizontal curve that joins the points  $p$  and  $q$ .

*Proof.* From Proposition 3.1, for any  $p \in U_1$ , there is a  $m$ -dimensional  $\mathcal{D}$ -connected manifold  $V_p^m$  passing through  $p$ . Applying Proposition 3.2  $n - m$  times yields a subdomain  $U_2 \subset U_1$  such that for all  $p \in U_2$ , there is an  $n$ -dimensional  $\mathcal{D}$ -connected manifold  $W_p^n$  passing through  $p$ .

Let  $p, q \in U_2$  be two arbitrary points. Let  $\gamma$  be a path joining  $p$  and  $q$  contained in  $U_2$  ( $\gamma$  not necessarily supposed to be a horizontal curve.) Since  $\sup_{x \in \gamma} W_x^n$  covers the compact set  $\text{Image}(\gamma)$ , there is a finite subcovering; i.e., we can choose  $\ell + 1$  points on  $\gamma$

$$x_0 = p, \quad x_1, x_2, \dots, x_{\ell-1}, \quad x_\ell = q$$

such that

$$\text{Image}(\gamma) \subset \bigcup_{k=0}^{\ell} W_{x_k}^n.$$

We can choose the points  $x_i$  such that any two consecutive points  $x_i$  and  $x_{i+1}$  belong to the same manifold  $W_{x_i}^n$ . Since the manifolds  $W_{x_i}^n$  are  $\mathcal{D}$ -connected, the points  $x_i$  and  $x_{i+1}$  can be joined by a horizontal curve. This way, the points  $p$  and  $q$  can be joined by a piecewise horizontal curve.  $\square$

**3.3. Subelliptic Heat Kernel.** Now we need to use Hamilton or Lagrange formalisms to construct the fundamental solution of the subelliptic heat operator. In other words, we are interested in finding the solving kernels for the operators  $\frac{\partial}{\partial t} - \Delta_{\mathcal{H}}$ . Inspired by the Gaussian, it is reasonable to expect the kernel has the form:

$$\mathcal{P}_t(x, x_0) = \frac{c}{t^\alpha} e^{-h(x, x_0)} \quad \text{for some suitable } \alpha.$$

See, e.g., [21], [22] and [25].

The *modified complex action* function  $h(x, x_0)$  can be written as  $\frac{f(x, x_0)}{2t}$  which plays the role of  $\frac{1}{2} \frac{d^2(x, x_0)}{2t}$  and satisfies the Hamilton-Jacobi equation

$$\frac{\partial h}{\partial t} + H\left(x_1, \dots, x_n, \frac{\partial h}{\partial x_1}, \dots, \frac{\partial h}{\partial x_n}\right) = 0.$$

In general, when we deal with a subelliptic heat operator, the heat kernel will depend on  $n - m$  parameters (or Lagrange multipliers)  $\tau = (\tau_1, \dots, \tau_{n-m})$ . Furthermore, after calculation, one may see that the action function  $h(x, x_0, \tau, t)$  can be written as

$$h(x, x_0, \tau, t) = \frac{1}{t} g(x, x_0, \tau t, 1).$$

We look for a heat kernel in the form  $\frac{c}{t^\alpha} e^{-h} = \frac{C}{t^\alpha} e^{-\frac{g}{t}}$ . The heat kernel should not depend on  $\tau$ . So we use an age old technique to get rid of  $\tau$  by summing over it. Since  $\tau$  are continuous parameters. Thus we shall look for a heat kernel in the following form:

$$\mathcal{P}_t(x, x_0) = \frac{1}{(2\pi t)^\alpha} \int_{\mathbf{R}^{n-m}} e^{-\frac{g(x, x_0, \lambda)}{2t}} V(x, x_0, \lambda) d\lambda.$$

Here  $\tau t = \lambda$  and  $V(x, x_0, \lambda)$  is so-called *volume element* which is an appropriate measure that makes the integral is convergent. Now we may apply properties of the heat kernel and reduce the problem to solving the *transport equation* to find  $V$ . Once we obtain  $V$ , then the index  $\alpha$  can be determined which depends on the *Hausdorff dimension* of the subRiemannian manifold  $\mathcal{M}_n$ . When  $m = n$ , the  $\mathcal{M}_n$  is an  $n$ -dimensional Riemannian manifold and  $\alpha = \frac{n}{2}$  where  $n$  is the topological dimension of the manifold. In the case, the volume element is just the zero section that will recover results in elliptic cases. For more details, we refer to [17] and [21].

#### 4. UNIFIED FRAMEWORK

To compare how distance metrics affect prediction performance, we present a unified framework for a graph kernel-based drug prescription prediction system in support of a rigorous empirical evaluation. We consider all disease and distance metric configurations for both data balance and imbalance ratios. Motivated by MGKF and Cross-Global, a hybrid model is formulated to leverage advantages from these two models. Following the same MGKF three-kernel architecture, namely a Weisfeiler-Lehman subtree kernel  $K_{wl}$  [26], Temporal topological kernel  $K_{tp}$  [5], and Vertex histogram kernel  $K_{vh}$  [27], we generate a fused kernel embedding with distance metric loss from Cross-Global as regularization.

Specifically, three pairwise kernel matrices via the aforementioned graph kernels are constructed. A single representation for a fused kernel embedding is generated through a deep neural network for successive classification. The distance regularization, achieved by contrastive loss [28], is integrated to combine the power of deep metric learning, and we force a kernel embedding to preserve an optimal distance property. Semantically similar embeddings are encouraged to be closer to each other, and dissimilar further apart in the kernel space. With this setting, kernel embedding is optimized jointly with classification loss and contrastive loss, deriving a single representation with multi-views and selected distance property. We discuss how embedding and prediction performance differs under different distance metrics in the next section.

Given a set  $T$  of  $N$  patients with their patient graphs  $P_i \in T$  where  $i \in \{1, \dots, N\}$  and associated class labels  $y \in R^{N \times 1}$  where  $y_i \in \{1, 0\}$ , we compute their pairwise kernel gram matrices  $K_{wl} \in R^{N \times N}$ ,  $K_{tp} \in R^{N \times N}$ , and  $K_{vh} \in R^{N \times N}$  by  $K_{wl}$ ,  $K_{tp}$ , and  $K_{vh}$  respectively. Let  $F_{emb}$  be a deep neural network parameterized by weight  $\theta$  and  $F_{sigmoid}$  as a single layer sigmoid function parameterized by  $w$ . We defined an unified framework as the following optimization problem:

$$\min_{\theta, w} \mathcal{L}_{contrastive} + \mathcal{L}_{crossentropy}, \quad (4.1)$$

$$\mathcal{L}_{contrastive} = \frac{1}{|N|} \sum_{i=1}^N \sum_{j=1}^N (1 - Y_{ij}) \max(0, \lambda - D_{ij})^2 + Y_{ij} D_{ij}, \quad (4.2)$$

$$\mathcal{L}_{crossentropy} = \frac{1}{|N|} \sum_{i=1}^N y_i \log(F_{sigmoid}_w(F_{emb\theta}(K_{wli}, K_{tpi}, K_{vhi}))), \quad (4.3)$$

where  $\lambda > 0$  is a constant margin threshold,  $Y_{ij} = 1$  if  $y_i = y_j$  else  $Y_{ij} = 0$ , and  $D_{ij}$  is a pairwise distance metric between kernel embedding  $F_{emb\theta}(K_{wli}, K_{tpi}, K_{vhi})$  calculated by (4.4) or (4.5).

Letting  $F_{emb\theta}(K_{wli}, K_{tpi}, K_{vhi}) = k_{embi}$  and  $F_{emb\theta}(K_{wlj}, K_{tpj}, K_{vhj}) = k_{embj}$ , we also have

$$d_{cosine}(k_{embi}, k_{embj}) = 1 - \frac{\langle k_{embi}, k_{embj} \rangle}{\|k_{embi}\| \cdot \|k_{embj}\|}, \quad (4.4)$$

$$d_{Euclidean}(k_{embi}, k_{embj}) = \|k_{embi} - k_{embj}\|_2, \quad (4.5)$$

where  $d_{cosine}$  is a cosine distance and  $d_{Euclidean}$  is the Euclidean distance. As usual,  $\langle \cdot \rangle$  is a standard inner product.  $d$  in (4.2) can be either  $d_{cosine}$  or  $d_{euclidean}$ .

The optimization problem in (4.1) can be solved by mini-batch Stochastic Gradient Descent (SGD). Once we find the optimal  $\theta$  and  $w$ , we can perform the prediction. Considering a

new incoming patient with patient graph  $P_i$ , we compute pairwise kernel matrices  $\mathbf{K}_{wl\hat{i}} \in R^{1 \times N}$ ,  $\mathbf{K}_{tp\hat{i}} \in R^{1 \times N}$ , and  $\mathbf{K}_{vh\hat{i}} \in R^{1 \times N}$  between  $P_i$  and all patient graphs in  $T$ . Then, we have the following decision function:

$$F_{\text{sigmoid}_w}(F_{\text{emb}\theta}(\mathbf{K}_{wl\hat{i}}, \mathbf{K}_{tp\hat{i}}, \mathbf{K}_{vh\hat{i}})) = \hat{y}, \quad (4.6)$$

where  $\hat{y}$  is the predicted class label (e.g., success or failure) of  $P_i$ .

The problem reduces to find good Riemannian or subRiemannian structures to handle a huge and complicated data set under certain constraints. In other words, we need to handle the related optimization problem (4.1) by finding horizontal vector fields and then construct solving kernel of the heat operator associate to the subelliptic operator  $\Delta_{\mathcal{X}}$ . In this paper, we consider Reproducing Kernel Hilbert Space (RKHS) with certain geometric properties derived from Euclidean or cosine distances.

TABLE 1. Dataset / Disease statistics

| Disease                 | Number of cases | Number of failure | Number of success | Failure-success ratio |
|-------------------------|-----------------|-------------------|-------------------|-----------------------|
| Urinary tract infection | 1,501,310       | 703,646           | 797,664           | 47%:53%               |
| Acute otitis media      | 151,522         | 72,264            | 79,258            | 48%:52%               |
| Pneumonia               | 95,796          | 37,724            | 58,072            | 39%:61%               |
| Acute cystitis          | 733,119         | 301,902           | 431,217           | 41%:59%               |
| Hypertension            | 235,695         | 104,936           | 130,759           | 45%:55%               |
| Hyperlipidemia          | 123,380         | 26,043            | 97,337            | 21%:79%               |
| Diabetes                | 131,997         | 34,414            | 97,583            | 26%:74%               |

## 5. DATASET AND EVALUATION PROTOCOL

To investigate how distance metric relates to kernel embedding and prediction, we conduct a rigorous empirical evaluation with our proposed unified framework under different data balance-imbalance ratio on a very large-scale real-world EHRs, a subset of the Taiwanese National Health Insurance Research Database (NHIRD)<sup>3</sup>.

Our sample of the NHIRD contains a 20-plus year, complete, medical history for over one-million randomly sampled patients. The database is provided by Taiwan’s National Health Insurance Administration and the Ministry of Health and Welfare. Data are composed of registration files and original claim data for reimbursement to hospitals that participate in the National Health Insurance (NHI) program. The International Classification of Diseases, 9th Revision, Clinical Modification (ICD9-CM) code indicates the disease diagnosed. A unique identifier is used per drug and can be further linked to the Anatomical Therapeutic Chemical (ATC) code. For privacy purposes, the NHIRD contains no patient personal information such as name, contact information, and exact birth day (e.g., only with birth year and month). Also all identification numbers for patients and hospitals are de-identified in an attempt to prevent possible

<sup>3</sup><https://nhird.nhri.org.tw/en/>

information leak. Institutional Review Board (IRB) approvals for our research were granted by all associated institutions.

We select four short-term diseases and three most prevalent chronic diseases in Taiwan. We follow our previous efforts setting an observation window for each type of disease. Refer Table 1 for complete disease list, data statistics, and outcome observation setup. To validate the claim, cosine distance is superior under data imbalance in chronic disease, in Cross-Global [12], and examine such a case on short-term disease in MGKF [5], we prepare balance and imbalance data. For balanced one, we downsample the size of majority cases to minority cases designating rare diseases, while keeping 70 percent of majority cases and 30 percents of minority cases for imbalanced one. The pairwise t-test with a p-value set to 0.01 is used to reject the null-hypothesis to measure the statistical significance for comparisons.

We compare Euclidean (Euclidean) and cosine (Cosine) distance on our unified framework. In  $F_{emb}$ , we set 5000 dimensions for the first embedding layer of each kernel and 50 dimensions for kernel fusion layer, as a two-layer Multi-layer perceptron (MLP). We set 50 dimensions for sigmoid classifier  $F_{sigmoid}$ . During training, we use the Adamax optimizer [29] with a fixed learning rate 0.0001 and setup 64 batch size for 1000 epochs with early stopping criteria on batch loss. Two machine learning models are included as baselines for comparison purposes, e.g., Support Vector Machine (SVM) and Logistic Regression (LR) with all regularization constant set up to 1 (e.g.,  $C = 1$ ). All patients are represented as documents containing all medical codes from all visits, and transfer to low-dimensional embedding via Paragraph to Vector [30] with embedding size 512. Accuracy (ACC), Macro F1-score (F1), and the area under the receiver operating characteristic curve (AUC) are used as our evaluation metrics. All models are developed by Tensorflow and scikit-learn packages using Python. The experiments are executed on an Intel Core i9 CPU with 64GB memory and one Nvidia Titan-RTX GPU with 24GB memory<sup>4</sup>. Accuracy comparisons with other learning models, including a diversity of the latest neural configurations, are found in [5, 12]

## 6. MAIN RESULT

Tables<sup>5,6</sup> 2, 3, 4, and 5 show evaluation results under short-term diseases with data balance-imbalance, under different models. Chronic diseases are reported in Tables<sup>5,6</sup> 6, 7, and 8. Under a balanced setting, both Euclidean and cosine distance are relatively similar in their short-term and chronic disease evaluations. Euclidean distance even outperforms cosine distance in 5 out of 7 diseases, implying that Euclidean distance can achieve favorable results when data variation is small, no matter for short/long term disease progression. When it comes to an imbalance setting, the Euclidean distance measure is superior to the cosine measure for all short-term diseases. This confirms our premise that Euclidean distance is applicable to local problems, namely short disease progression. On the other hand, cosine distance is preferable in imbalance long-term chronic disease, which outperforms Euclidean distance especially in F1 score (e.g., an indicator to model performance under imbalance data set). It is worth noting that the evaluation margin between Euclidean and cosine distance is pretty large in all chronic diseases. The degree of

<sup>4</sup>We do not perform hyper-parameters tuning for all models.

<sup>5</sup> Shaded regions indicate statistical equivalence (light) and significant difference (dark) of the Euclidean and cosine measures.

<sup>6</sup> A \* designation indicates statistical significance over all baselines (SVM and LR) with p-value at 0.01.

outcome variation (e.g., co-morbidity) of long-term chronic disease patient group is larger than patient group in short-term disease, which reflects that Euclidean is more applicable under a low-variation data set. The comparison to the baseline models validates our unified framework. Note that, we did not perform any hyper-parameters tuning nor customize to any specific disease group. The purpose of this evaluation was strictly to investigate possible conclusions on model behavior under different distance metrics.

TABLE 2. Evaluation Results for Urinary tract infection.

| Urinary tract infection |  |  |  |
|-------------------------|--|--|--|
| Balanced                |  |  |  |
| Model                   | ACC                                    | F1                                     | AUC                                    |
| Euclidean               | *0.6220 $\pm$ 0.0212                   | *0.6186 $\pm$ 0.0229                   | *0.6220 $\pm$ 0.0212                   |
| Cosine                  | <b>*0.6243 <math>\pm</math> 0.0284</b> | <b>*0.6216 <math>\pm</math> 0.0283</b> | <b>*0.6243 <math>\pm</math> 0.0284</b> |
| SVM                     | 0.5047 $\pm$ 0.0257                    | 0.5046 $\pm$ 0.0258                    | 0.5047 $\pm$ 0.0257                    |
| LR                      | 0.5210 $\pm$ 0.0240                    | 0.4874 $\pm$ 0.0300                    | 0.4988 $\pm$ 0.0296                    |
| Imbalanced              |  |  |  |
| Euclidean               | <b>*0.6280 <math>\pm</math> 0.0469</b> | *0.5465 $\pm$ 0.0290                   | *0.5895 $\pm$ 0.0249                   |
| Cosine                  | *0.6165 $\pm$ 0.0532                   | <b>*0.5632 <math>\pm</math> 0.0356</b> | <b>*0.6194 <math>\pm</math> 0.0182</b> |
| SVM                     | 0.5023 $\pm$ 0.0266                    | 0.5051 $\pm$ 0.0266                    | 0.5053 $\pm$ 0.0265                    |
| LR                      | 0.5208 $\pm$ 0.0240                    | 0.4872 $\pm$ 0.0260                    | 0.4896 $\pm$ 0.0297                    |

TABLE 3. Evaluation Results for Acute otitis media.

| Acute otitis media |  |  |  |
|--------------------|--|--|--|
| Balanced           |  |  |  |
| Model              | ACC                                    | F1                                     | AUC                                    |
| Euclidean          | <b>*0.6245 <math>\pm</math> 0.0200</b> | <b>*0.6224 <math>\pm</math> 0.0218</b> | <b>*0.6245 <math>\pm</math> 0.0200</b> |
| Cosine             | *0.6138 $\pm$ 0.0183                   | *0.6097 $\pm$ 0.0185                   | *0.6137 $\pm$ 0.0185                   |
| SVM                | 0.5023 $\pm$ 0.0204                    | 0.5021 $\pm$ 0.0203                    | 0.5023 $\pm$ 0.0203                    |
| LR                 | 0.5011 $\pm$ 0.0211                    | 0.5010 $\pm$ 0.0211                    | 0.5011 $\pm$ 0.0212                    |
| Imbalanced         |  |  |  |
| Euclidean          | <b>*0.6570 <math>\pm</math> 0.0203</b> | *0.5453 $\pm$ 0.0342                   | *0.6037 $\pm$ 0.0324                   |
| Cosine             | *0.6238 $\pm$ 0.0306                   | <b>*0.5554 <math>\pm</math> 0.0258</b> | <b>*0.6042 <math>\pm</math> 0.0346</b> |
| SVM                | 0.5165 $\pm$ 0.0196                    | 0.4803 $\pm$ 0.0212                    | 0.4899 $\pm$ 0.0246                    |
| LR                 | 0.5170 $\pm$ 0.0177                    | 0.4804 $\pm$ 0.0201                    | 0.4898 $\pm$ 0.0237                    |

We further compare the embedding space induced by Euclidean and cosine distance measure under their original data-imbalance condition (refer to Table 1). We apply t-SNE [31] algorithm to perform 2D visualization for all diseases under balance and imbalance conditions on the learned kernel embedding from  $F_{emb}$  in 4. We randomly divide each disease into training and testing in an 80:20 ratio and employ the same training procedure described in 5. Figure 1 and 2 show the kernel embedding on the training and testing sets with the Euclidean and cosine distance measures under short-term and long-term chronic diseases. In each Figure, the ability to handle out-of-sample (unseen) embedding is depicted with testing set in the sub-title. We



TABLE 4. Evaluation Results for Pneumonia.

| Pneumonia  |   |   |   |
|------------|---|---|---|
| Balanced   |   |   |   |
| Model      | ACC                                     | F1                                      | AUC                                     |
| Euclidean  | *0.6013 $\pm$ 0.0279                    | * <b>0.5922 <math>\pm</math> 0.3112</b> | *0.6013 $\pm$ 0.0279                    |
| Cosine     | * <b>0.6023 <math>\pm</math> 0.0211</b> | *0.5918 $\pm$ 0.0263                    | * <b>0.6023 <math>\pm</math> 0.0211</b> |
| SVM        | 0.4976 $\pm$ 0.0127                     | 0.4975 $\pm$ 0.0127                     | 0.4976 $\pm$ 0.0126                     |
| LR         | 0.4979 $\pm$ 0.0130                     | 0.4978 $\pm$ 0.0130                     | 0.4979 $\pm$ 0.0129                     |
| Imbalanced |   |   |   |
| Euclidean  | * <b>0.6398 <math>\pm</math> 0.0688</b> | *0.5626 $\pm$ 0.0423                    | *0.6028 $\pm$ 0.0270                    |
| Cosine     | *0.6255 $\pm$ 0.0470                    | * <b>0.5712 <math>\pm</math> 0.0209</b> | * <b>0.6220 <math>\pm</math> 0.0250</b> |
| SVM        | 0.5430 $\pm$ 0.0243                     | 0.5070 $\pm$ 0.0246                     | 0.5179 $\pm$ 0.0268                     |
| LR         | 0.5430 $\pm$ 0.0243                     | 0.5074 $\pm$ 0.0242                     | 0.5186 $\pm$ 0.0261                     |

TABLE 5. Evaluation Results for Acute cystitis.

| Acute cystitis |   |   |   |
|----------------|---|---|---|
| Balanced       |   |   |   |
| Model          | ACC                                     | F1                                      | AUC                                     |
| Euclidean      | * <b>0.6143 <math>\pm</math> 0.0189</b> | * <b>0.6087 <math>\pm</math> 0.0245</b> | * <b>0.6143 <math>\pm</math> 0.0189</b> |
| Cosine         | *0.6095 $\pm$ 0.0182                    | *0.6068 $\pm$ 0.0199                    | *0.6095 $\pm$ 0.0182                    |
| SVM            | 0.5049 $\pm$ 0.0231                     | 0.5048 $\pm$ 0.0231                     | 0.5049 $\pm$ 0.0231                     |
| LR             | 0.5037 $\pm$ 0.0228                     | 0.5037 $\pm$ 0.0228                     | 0.5037 $\pm$ 0.0228                     |
| Imbalanced     |   |   |   |
| Euclidean      | * <b>0.6353 <math>\pm</math> 0.0346</b> | *0.5607 $\pm$ 0.0231                    | *0.5763 $\pm$ 0.0325                    |
| Cosine         | *0.6280 $\pm$ 0.0405                    | * <b>0.5632 <math>\pm</math> 0.0201</b> | * <b>0.5839 <math>\pm</math> 0.0267</b> |
| SVM            | 0.5235 $\pm$ 0.0227                     | 0.4871 $\pm$ 0.0219                     | 0.4965 $\pm$ 0.0233                     |
| LR             | 0.5230 $\pm$ 0.0248                     | 0.4860 $\pm$ 0.0232                     | 0.4957 $\pm$ 0.0242                     |

TABLE 6. Evaluation Results for Hypertension.

| Hypertension |   |   |   |
|--------------|---|---|---|
| Balanced     |   |   |   |
| Model        | ACC                                     | F1                                      | AUC                                     |
| Euclidean    | * <b>0.7315 <math>\pm</math> 0.0126</b> | * <b>0.7305 <math>\pm</math> 0.0131</b> | * <b>0.7315 <math>\pm</math> 0.0126</b> |
| Cosine       | *0.7290 $\pm$ 0.0131                    | *0.7282 $\pm$ 0.0139                    | *0.7290 $\pm$ 0.0131                    |
| SVM          | 0.4956 $\pm$ 0.0165                     | 0.4954 $\pm$ 0.0165                     | 0.4956 $\pm$ 0.0165                     |
| LR           | 0.4961 $\pm$ 0.0168                     | 0.4959 $\pm$ 0.0168                     | 0.4961 $\pm$ 0.0167                     |
| Imbalanced   |   |   |   |
| Euclidean    | *0.7036 $\pm$ 0.005                     | 0.4255 $\pm$ 0.0180                     | 0.5065 $\pm$ 0.0086                     |
| Cosine       | * <b>0.7497 <math>\pm</math> 0.0236</b> | * <b>0.6525 <math>\pm</math> 0.0500</b> | * <b>0.6455 <math>\pm</math> 0.0418</b> |
| SVM          | 0.5163 $\pm$ 0.0223                     | 0.4785 $\pm$ 0.0228                     | 0.4871 $\pm$ 0.0253                     |
| LR           | 0.5170 $\pm$ 0.0227                     | 0.4791 $\pm$ 0.0229                     | 0.4876 $\pm$ 0.0254                     |

TABLE 7. Evaluation Results for Hyperlipidemia.

| Hyperlipidemia |                         |                         |                         |
|----------------|-------------------------|-------------------------|-------------------------|
| Balanced       |                         |                         |                         |
| Model          | ACC                     | F1                      | AUC                     |
| Euclidean      | <b>*0.7478 ± 0.0193</b> | <b>*0.7469 ± 0.0197</b> | <b>*0.7478 ± 0.0193</b> |
| Cosine         | *0.7470 ± 0.0185        | *0.7461 ± 0.0187        | *0.7470 ± 0.0185        |
| SVM            | 0.4891 ± 0.0172         | 0.4890 ± 0.0171         | 0.4891 ± 0.0172         |
| LR             | 0.4888 ± 0.0181         | 0.4887 ± 0.0181         | 0.4888 ± 0.0181         |
| Imbalanced     |                         |                         |                         |
| Euclidean      | *0.7323 ± 0.0408        | 0.5312 ± 0.1482         | *0.5797 ± 0.1011        |
| Cosine         | <b>*0.7565 ± 0.0356</b> | <b>*0.6402 ± 0.0955</b> | <b>*0.6415 ± 0.0730</b> |
| SVM            | 0.5183 ± 0.0275         | 0.4821 ± 0.0264         | 0.4914 ± 0.0281         |
| LR             | 0.5193 ± 0.0272         | 0.4830 ± 0.0263         | 0.4923 ± 0.0281         |

TABLE 8. Evaluation Results for Diabetes.

| Diabetes   |                         |                         |                         |
|------------|-------------------------|-------------------------|-------------------------|
| Balanced   |                         |                         |                         |
| Model      | ACC                     | F1                      | AUC                     |
| Euclidean  | <b>*0.7085 ± 0.0151</b> | <b>*0.7064 ± 0.0156</b> | <b>*0.7085 ± 0.0151</b> |
| Cosine     | *0.6973 ± 0.0191        | *0.6958 ± 0.0191        | *0.6973 ± 0.0191        |
| SVM        | 0.4894 ± 0.0151         | 0.4892 ± 0.0151         | 0.4894 ± 0.0151         |
| LR         | 0.4886 ± 0.0148         | 0.4884 ± 0.0147         | 0.4886 ± 0.0148         |
| Imbalanced |                         |                         |                         |
| Euclidean  | *0.7025 ± 0.0075        | 0.4315 ± 0.0591         | 0.5104 ± 0.0311         |
| Cosine     | <b>*0.7353 ± 0.0195</b> | <b>*0.6096 ± 0.0684</b> | <b>*0.6114 ± 0.0495</b> |
| SVM        | 0.5168 ± 0.0164         | 0.4827 ± 0.0126         | 0.4939 ± 0.0133         |
| LR         | 0.5167 ± 0.0167         | 0.4828 ± 0.0125         | 0.4941 ± 0.0130         |

can see both Euclidean and cosine distance perfectly separate two outcome groups on training set in short-term disease from Figures 1(a), 1(b), 1(e), 1(f), 1(i), 1(j), 1(m), and 1(n). Their analogous out-of-sample embedding ability in Figures 1(c), 1(d), 1(g), 1(h), 1(k), 1(l), 1(o), and 1(p) demonstrate the similar effect between the two distance measures under short-term disease, correlating the evaluation results in tables 2, 3, 4, and 5. It also suggests that Euclidean distance is not sufficiently sensitive to highly variant features under long-term disease progressions in chronic disease. As shown in Figures 2(c), 2(g), and 2(k), the out-of-sample embedding ability of Euclidean distance under-performs as compared to its cosine distance counterpart in Figures 2(d), 2(h), and 2(l), although it can perfectly fit the training set in Figures 2(a), 2(e), and 2(i). Cosine distance exhibits certain regularization ability to avoid model overfitting on the training set in chronic disease (refer to Figures 2(b), 2(f), and 2(j)), resulting in better performance on out-of-sample embedding, which also coincides the evaluation results in Tables 6, 7, and 8. It further shows that cosine distance benefits from more complex data representations and Euclidean distance can only be applied to local problems with low data variation, which may not be sufficient to express complex feature characteristics.

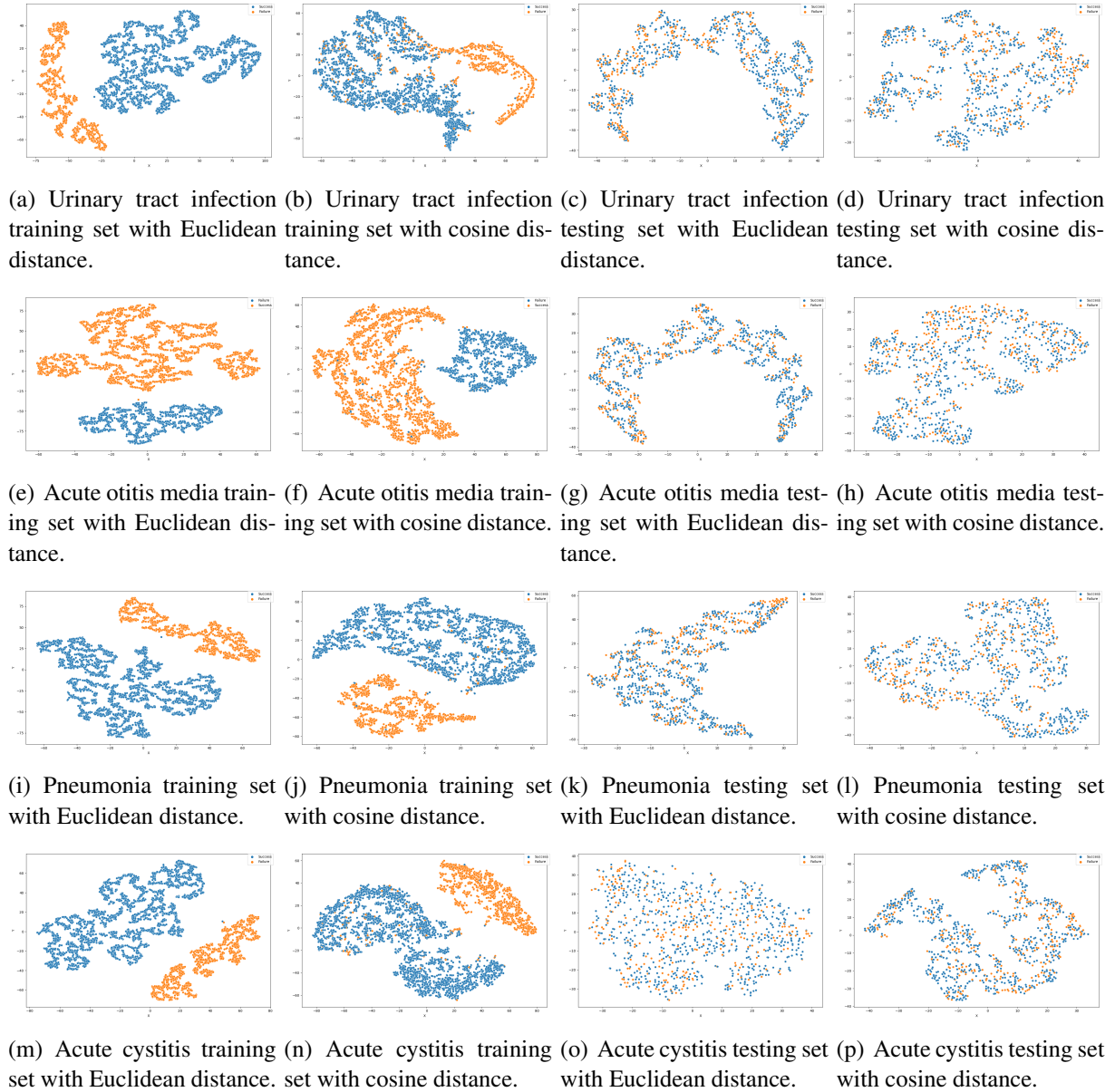


FIGURE 1. 2D visualization of kernel embedding by tSNE with different distance metrics under short-term diseases. We fit our unified framework on training set and evaluate out-of-sample embedding ability on testing set. Blue dot denotes failure case, and orange dot denotes success case.

## 7. CONCLUSION

Distance metrics and their nonlinear variant play a crucial role in machine learning based real-world problem solving. We demonstrated how Euclidean and cosine distance measures differ not only theoretically but also in real-world medical application, namely, outcome prediction of drug prescription. Euclidean distance exhibits favorable properties in the local geometry

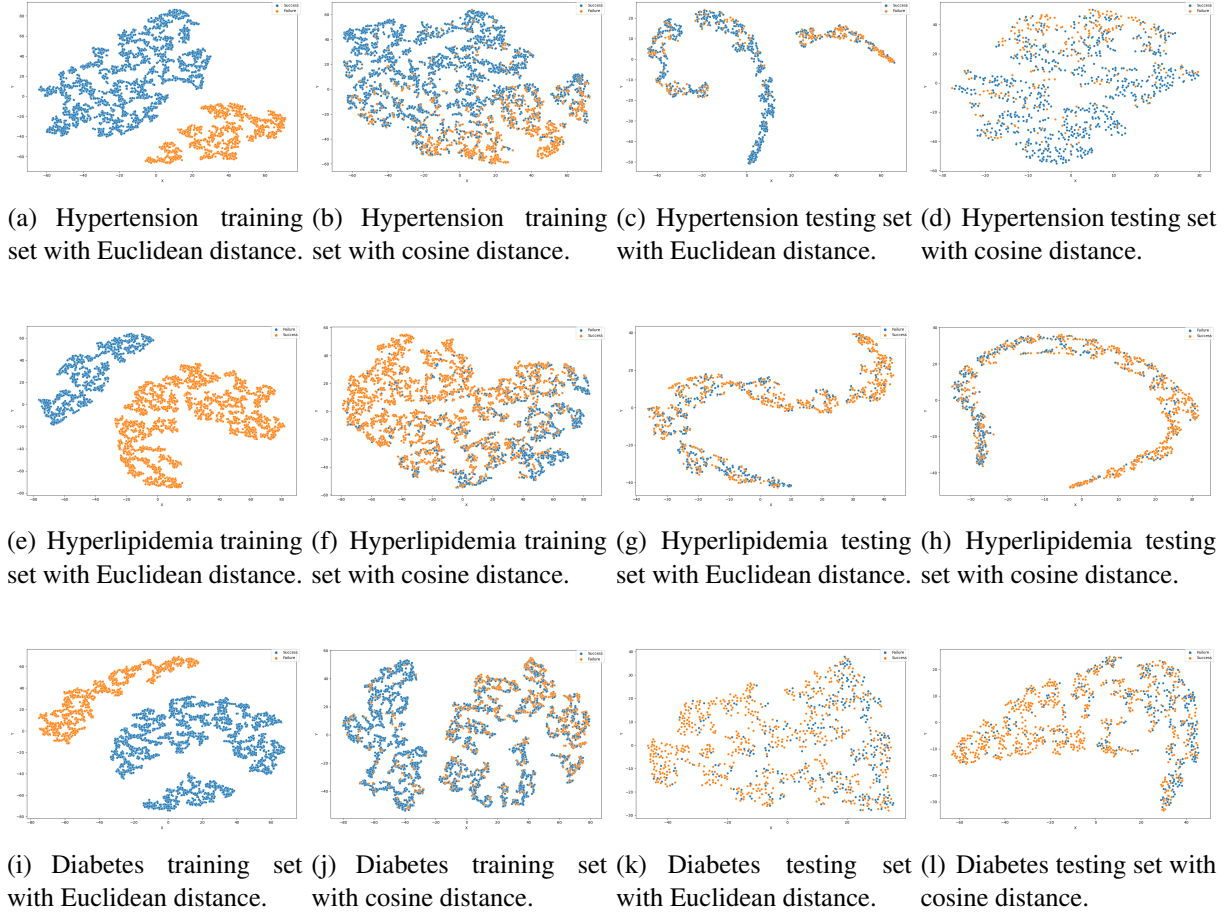


FIGURE 2. 2D visualization of kernel embedding by tSNE with different distance metrics under long-term chronic diseases. We fit our unified framework on training set and evaluate out-of-sample embedding ability on testing set. Blue dot denotes failure case, and orange dot denotes success case.

problem. To this regard, Euclidean distance can be applied under short-term disease with low-variation outcome observation. Moreover, when presenting to highly variant chronic disease, it is preferable to use cosine distance. These different geometric properties lead to different sub-manifolds in the original embedded space, and hence, to different optimizing nonlinear kernel embedding frameworks. We first established the geometric properties that we needed in these frameworks. From these properties interpreted their differences in certain perspectives. Our evaluation on real-world, large-scale electronic health records and embedding space visualization empirically validated our approach.

### Acknowledgments

The authors are grateful to the reviewers for useful suggestions which improved the contents of this paper. The first author was partially supported by an NSF grant DMS-1408839 and a McDevitt Endowment Fund at Georgetown University.

## REFERENCES

- [1] J. Han, M. Kamber, J. Pei, Data mining concepts and techniques third edition, The Morgan Kaufmann Series in Data Management Systems 5 (2011), 83-124.
- [2] A. Singhal, Modern information retrieval: A brief overview, IEEE Data Eng. Bull. 24 (2001), 35-43.
- [3] A. S. Ruocco, O. Frieder, Clustering and classification of large document bases in a parallel environment, J. Amer. Soc. Info. Sci. 48 (1997), 932-943.
- [4] D.C. Chang, O. Frieder, H.R. Yao, On bochner's theorem and its application to graph kernels, J. Nonlinear Convex Anal. 19 (2018), 2135-2151.
- [5] H.R. Yao, D.C. Chang, O. Frieder, W. Huang, T.S. Lee, Multiple Graph Kernel Fusion Prediction of Drug Prescription, Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 103-112, 2019.
- [6] B. Haasdonk, C. Bahlmann, Learning with Distance Substitution Kernels, Joint Pattern Recognition Symposium, Springer, Berlin, Heidelberg, 2004.
- [7] B. Kulis, Metric learning: A survey, Foundations and Trends in Machine Learning 5 (2012), 287-364.
- [8] J. Hu, J. Lu, Y.P. Tan, Discriminative deep metric learning for face verification in the wild, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1875-1882, 2014.
- [9] X. Liu, B.V.K. Vijaya Kumar, J. You, P. Jia, Adaptive deep metric learning for identity-aware facial expression recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 522-531, 2017.
- [10] M. Jiu, H. Sahbi, Nonlinear deep kernel learning for image annotation, IEEE Trans. Image Process. 26 (2017), 1820-1832.
- [11] A.G. Wilson, Z. Hu, R. Salakhutdinov, E.P. Xing, Deep kernel learning, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, pp. 370-378, 2016.
- [12] H.R. Yao, D.C. Chang, O. Frieder, W. Huang, I.C. Liang, C.F. Hung, Cross-global attention graph kernel network prediction of drug prescription, Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2020, United States, 2020.
- [13] J. K. Aronson, Medication errors: what they are, how they happen, and how to avoid them, QJM: An International Journal of Medicine 102 (2009), 513-521.
- [14] H.A. Jackson, J. Cashy, O. Frieder, A.J. Schaeffer, Data mining derived treatment algorithms from the electronic medical record improve theoretical empirical therapy for outpatient urinary tract infections, J. Urology 186 (2011), 2257-2262.
- [15] H.R. Yao, D.C. Chang, O. Frieder, W. Huang, T.S. Lee, Graph Kernel prediction of drug prescription, 2019 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2019 - Proceedings, University of Illinois at Chicago/Chicago, 8834676, 2019.
- [16] N.M. Kriege, F.D. Johansson, C. Morris, A survey on graph kernels, Appl. Network Sci. 5 (2020), 1-42.
- [17] O. Calin, D.C. Chang, Geometric Mechanics on Riemannian Manifolds: Applications to Partial Differential Equations, Applied and Numerical Analysis, 29, Birkhäuser, Boston, Massachusetts, 2004.
- [18] W.L. Chow, Über Systeme van Linearen partiellen Differentialgleichungen erster Ordnung, Math. Ann. 117 (1939), 98-105.
- [19] O. Calin, D.C. Chang, Sub-Riemannian Geometry, General Theory and Examples, Encyclopedia of Mathematics and Its Applications, 126, Cambridge University Press, 2009.
- [20] O. Carathéodory, Untersuchungen über die Grundlagen der Thermodynamik, Math. Anal. 67 (1909), 93-161.
- [21] O. Calin, D.C. Chang, K. Furutani, C. Iwasaki, Heat Kernels for Elliptic and Sub-elliptic Operators: Methods and Techniques, Applied and Numerical Analysis, 47, Birkhäuser, Boston, Massachusetts, 2010.
- [22] D.C. Chang, Y. Li, Heat Kernels for a family of Grushin operators, Methods Appl. Anal. 21 (2014), 291-312.
- [23] D.C. Chang, K.P. Lin, Stephen S.T. Yau, "Schrödinger equation with quartic potential and nonlinear filtering problem, Proceedings of the 48th IEEE Conference on Decision and Control, Shanghai, 8089-8094, 2009.
- [24] C. Teleman, Asupra Sistemelor Mecanice Neonolome, Anal. Univ. "C.I. Purhon", Bucuresti, Seria St. Naturii, 13 (1957), 45-52.
- [25] D.C. Chang, I. Markina, A. Vasilév, Geodesics and distance respecting Hopf fibration of  $n$ -spheres, J. Geometry Phys. 61 (2011), 986-1000.

- [26] N. Shervashidze, P. Schweitzer, E.J. Van Leeuwen, K. Mehlhorn, K.M. Borgwardt, Weisfeiler-lehman graph kernels, *J. Machine Learning Res.* 12 (2011), 2539-2561.
- [27] M. Sugiyama, K.M. Borgwardt, Halting in random walk kernels, *Adv. Neural Info. Process. Sys.* 2015 (2015), 1639-1647.
- [28] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 1735-1742, New York, 2006.
- [29] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, 2015.
- [30] Q. Le, T. Mikolov, Distributed representations of sentences and documents, *31st International Conference on Machine Learning, ICML 2014*, Vol. 4, pp. 2931-293, 2014.
- [31] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Machine Learn. Res.* 9 (2008), 2579-2625.