Transport Generative Models in Pattern Analysis and Recognition

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Mohammad Shifat E Rabbi

August 2023

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Mohammad Shifat E Rabbi

This Dissertation has been read and approved by the examing committee:

Advisor: Gustavo Rohde, Ph.D.

Advisor:

Committee Member: Craig Meyer, Ph.D.

Committee Member: Jason Papin, Ph.D.

Committee Member: Scott Acton, Ph.D.

Committee Member: John Ozolek, M.D.

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

August 2023

© 2023

Mohammad Shifat-E-Rabbi

All Rights Reserved

Abstract

Transport Generative Models in Pattern Analysis and Recognition

Mohammad Shifat-E-Rabbi

There exists a growing need for computational models for pattern analysis and recognition in numerous scientific and technological fields, including computer vision, biology, and healthcare. Although generic feature approximation and deep learning approaches have been widely used in this aspect, they suffer from limitations in robustness, generalizability, and interpretability. Moreover, they are computationally expensive, require a vast amount of training data, and are vulnerable to out-of-distribution samples. In this study, we introduce a transport-based modeling approach for solving pattern analysis and recognition problems. Our approach focuses on modeling data obtained from processes involving some kind of transport of mass or intensity of pixels, tissue, or molecules, such as tissue growth, cell division, carcinogenesis, and gene expression. We postulate that data classes obtained from such processes can be represented as instances of an unknown template under the effect of unknown spatial deformations. Using this hypothesis, we aim to demonstrate that our classification and modeling approach can solve problems involving segmented data in closed-form. We show that our proposed method has the potential to achieve better accuracy, generalizability, interpretability, and data efficiency compared to existing approaches. Moreover, our method is designed to be simple and computationally efficient, potentially making it a more practical solution for real-world applications.

In order to accomplish our research objectives, we introduce a novel transport-based data

generative model for image classification and develop a new supervised image classification method applicable to a broad class of image deformation models. We formulate and derive the mathematical properties of the data generative model and solve the classification problem in closed-form using transport-based embeddings. Additionally, we demonstrate how the method can learn data invariances without the need for data augmentation. Furthermore, we extend the aforementioned framework to formulate transport-based embeddings for the classification of high-dimensional distributions, which can be applied in a variety of applications. Our approach is not only simple to implement, but also non-iterative, computationally efficient, data-efficient, and possesses out-of-distribution generalization. Lastly, we introduce a transport-based morphometry framework for modeling nuclear structures of digital pathology images in cancer and use this framework to explore the existence of shared nuclear structure biomarkers across different cancer types. We show that our model can reveal meaningful information within and across various tissue types by identifying morphological differences among them. We show that our framework can provide quantitative measurements for comparisons across diverse datasets and cancer types that can potentially enable numerous cancer studies, technologies, and clinical applications and help elevate the role of nuclear morphometry into a more quantitative science.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to begin by expressing my deepest gratitude to my parents for their unwavering love, support, and encouragement throughout my academic journey. Their belief in me has been a constant source of inspiration, and I am grateful for their sacrifices, guidance, and belief in me. I am forever thankful for their persistent support, which has enabled me to complete this important milestone in my life.

I would also like to extend my heartfelt gratitude to my advisor, Prof. Gustavo Rohde, for his invaluable guidance, mentorship, and support throughout my research. His expertise, encouragement, and constructive feedback have been instrumental in shaping my research and helping me achieve my goals. I am grateful for the countless hours he spent providing feedback on my work and his patience in guiding me through the challenges of research. Furthermore, his diligent work ethics, distinguished philosophy towards science and education, relentless passion for knowledge dissemination, and exceptional problem-solving skills have greatly influenced my growth as a scientific thinker and aided me in achieving success in my doctoral program. His big heart and kind mind not only helped me through difficult phases in my PhD but also taught me the importance of empathy and humanity in academia. I am forever grateful for his mentorship and honored to have worked with such an exceptional scientist.

I would like to sincerely express my gratitude to my esteemed PhD committee members, Prof. Craig Meyer, Prof. Jason Papin, Prof. Scott Acton, and Prof. John Ozolek, for their invaluable feedback, encouragement, and guidance throughout the process of completing my dissertation. Their vast knowledge and expertise have been instrumental in shaping my research and helping me

achieve my academic goals. Their constructive and insightful suggestions have been invaluable in improving my work and enhancing my ability to think critically about my research questions. I am deeply grateful for the time and dedication they devoted to ensuring that my dissertation was of the highest quality, and their unwavering support and encouragement have been a source of motivation throughout my academic journey.

I would like to thank my teachers, labmates, co-authors, and collaborators for their support, guidance, and encouragement throughout my academic journey. Their expertise, feedback, and support have been instrumental in shaping my research and helping me achieve my goals. I am grateful for their constant encouragement and inspiration. Specifically, I would like to express my gratitude to my co-authors of the published paper [1], for their valuable contributions. I also acknowledge Drs. Watnik and Nichols for the permission to use the OAM dataset shown in Table 2.3.

I am grateful for the love, support, and encouragement from my friends and family, which has been a constant source of motivation and inspiration throughout my academic journey. Additionally, I want to extend my thanks to the University staff, whose tireless efforts were critical in helping me navigate the academic system and overcome the various challenges that arose. Finally, I acknowledge the support from NIH grants GM130825, GM090033, NSF grant 1759802, and CSBC grant U54-CA274499 for my PhD.

Dedication

I would like to dedicate this work to my beloved family (my parents, sister, and other close relatives), teachers, colleagues, friends, and collaborators whose steadfast support have played a pivotal role in shaping my personal and academic development. Their relentless encouragement and guidance have been a constant source of inspiration, instilling in me the drive to pursue excellence in all my endeavors.

Chapter 1: Introduction and Background

Image classification methods occupy a predominant place in data sciences, given their inherent link to numerous medical imaging, computer vision, and computational biology applications [1, 2, 3]. Automated image classification methods have been utilized to detect cancer from microscopy images of tumor specimens [4, 5], detect and quantify atrophy from magnetic resonance images of the human brain [3, 6], identify and authenticate a person from cell phone camera images [7], and many other applications in computer vision, medical imaging, automated driving, and related fields. In many image classification problems, image classes can be thought of being an instance of a template observed under a set of spatial deformations. For example, consider the classes of the MNIST dataset [8]. Each image in a class can be considered as an image of a prototype digit with a transformation applied to it (such as translation, scaling, shear, rotation, higher-order deformations, and others). Other instances of this category of image classification problems include detecting the protein localization patterns within a cell [9], classifying the nuclear structures from the fluorescence measurements of a population of cells [5], and profiling the distribution of gray matter within a brain as depicted through MRI [3], among many other examples.

Beyond the analysis of two-dimensional images, the analysis of high-dimensional data distributions has also become a crucial component in various fields of computer vision and computational biology. This includes the analysis of high-content and high-throughput cytometry data, which facilitates the understanding and profiling of the phenotypic and functional characteristics of millions of individual cells [10, 11]. Additionally, it includes the classification of set structured data obtained from a diverse range of scanning technologies, such as LiDAR or photogrammetry [12, 13]. Sampling a continuous probability density function over an N-dimensional space is another technique for obtaining high-dimensional point-set distributions [14]. Furthermore, the modeling of a vast quantity of histopathology data obtained via microscopes can also be a prominent ap-

plication in this field [5]. Recent advances in measurement technology have provided access to vast quantities of high-dimensional data, making it possible to conduct computational analyses of multiple parameters [15]. High-dimensional analysis techniques are being utilized to characterize metastatic breast cancers [16, 17], identify brain macrophage development [18], and profile COVID-19 using images of platelet aggregates [1], among other applications.

In addition to the aforementioned classification methods for 2-dimensional images and high-dimensional distributions, pattern analysis methods can be employed in many other modeling problems. An exemplary application is the utilization of pattern recognition, machine learning, computer vision, and mathematical modeling techniques in real-world digital pathology applications. Alterations in nuclear morphology have been a staple in the pathologists' repertoire of diagnostic tools since the inception of microscopic examination of tissue [19, 20, 21, 22]. Nuclear morphology is usually determined by the microscopic structure and degree of chromatin condensation, which are regulated by interactions between the cell and its local microenvironment [23, 21]. Defects in the coupling of the nucleus to the cytoskeleton can lead to genomic instability and the transformation from a benign to a malignant cell, altering chromosomal organization [19, 24]. In cytopathology, nuclear morphological parameters, such as increased nuclear size, increased nuclear-to-cytoplasmic ratio, irregularities of the nuclear membrane, and abnormalities in chromatin organization, provide crucial visual clues for pathologists in diagnosis and patient management decisions [19, 21, 20, 22, 25]. Traditionally, pathologists determine the malignant potential of tissue specimens based on morphology through visual microscopic inspection [26, 27]. However, recent advances in computer-aided digital pathology and computational pattern recognition methods have led to several successful experimental applications in cancer detection [28], staging [29], prognosis prediction [30], drug discovery [31], and cell biology [32, 33]. These methods have the potential to perform standardized, efficient, and automated large-scale analyses of nuclear structure with the aim of providing a quantitative method for evaluating relationships between nuclear morphological changes and cellular discovery [25, 34].

Existing approaches

Over the past few decades, image classification methods have evolved from feature engineering-based methods relying on hand-tailored numerical features [35, 36] to hierarchical (deep) convolutional neural network-based (CNN) methods utilizing a series of computational layers [8, 2]. CNNs have recently emerged as leading classification methods for several reasons [37, 38, 39]. They provide a framework for end-to-end learning bypassing the feature engineering process, often decreasing the time and expenses related to bringing classification systems into production [2]. They obtain high accuracy in several image classification tasks [38, 39] and offer feasibility to be implemented in parallel utilizing graphical processing units (GPU) [40], among other improvements and conveniences over feature engineering methods. Recent improvements in computing power [41], the availability of annotated data [40], and open-source software [42] have also contributed to the usability of CNNs. However, it is broadly understood that CNN-based methods require large amounts of data for training [43], are computationally expensive [44], time-consuming [45], require careful parameter and hyper-parameter tuning [46], and are often vulnerable against out-of- distribution samples and attacks [47, 48].

High-dimensional data analysis is also typically achieved through the use of either deep neural networks, predetermined numerical features, or a combination of both [49, 2, 39]. While these methods are useful in some automation tasks, they may not be suitable for scientific endeavors due to their lack of transparency, interpretability, and physical units [50, 48]. Furthermore, their generalizability beyond the training data and mathematical understanding are often limited [51]. Therefore, alternative methods that offer more interpretability and transparency, such as physics-based or other explainable models, might be explored for high-dimensional data analysis in scientific applications.

The current practice of computational digital pathology is also predominantly based on heuristic feature engineering [52] and end-to-end feature learning [33], whereby analytical features are either predetermined or learned from the data. End-to-end deep learning methods using convolu-

tional neural networks (CNN) have obtained high classification accuracy in several experimental applications [53, 54], can be implemented in parallel using graphical processing units (GPUs) [55, 40], and are able to learn from large quantities of annotated data [40]. However, CNN methods are frequently limited by their lack of connection with an underlying physical process that can provide a scientific rationale for their use. A limited knowledge of their internal workings makes it difficult to distinguish settings when they do and do not work [56]. This predisposes CNNs to unpredictably producing misleading and inaccurate results, severely restricting their interpretability and generalizability [50, 48]. Confidence in their results, from a mechanistic point of view, is critical for safe and effective translation to clinical or scientific use. Using these methods, it is evidently challenging to develop a model capable of quantitative nuclear morphological analysis that contributes to understanding of the relationship between the structure and function of tumor cells and the biology of cancer [50, 32]. Consequently, nuclear morphological studies have not yet become a quantitative science, despite being a prime target for cancer research.

While quantitative studies exist in other branches of cancer bioinformatics, including genomics and proteomics [22, 57, 58], methods to build a reliable and understandable analytical model to quantify malignant transformation solely using nuclear morphological features (other than mitotic figures) have been lacking [59, 60]. In addition to the aforementioned limitations of end-to-end methods, current computational systems lack robustness to adversarial information. Slight changes in image data (e.g., different pathology staining protocols) can cause systems to make confident but erroneous predictions. [50, 61, 62]. This limits the accuracy of comparisons between datasets when performing meta-analyses to produce scientifically meaningful results. Previous approaches to overcome such variability and integrate information across datasets, including transfer learning [63] and z-score normalization [64], have been found to lack reliability and generalizability [65, 66], rendering their application in biomedical inference and diagnosis limited. Consequently, it has not yet been possible to describe similarities or differences across cancer types, combine datasets to enhance clinical practice and cell biology applications, (e.g., drug discovery and biomarker discovery) [67, 68], identify organ-specific, cancer-specific, or shared malignant signatures in nuclear

morphology, or test correlations between nuclear morphological signatures and gene expression, drug efficacy, or treatment response [22, 57, 69].

A less commonly employed method in modeling and classification involves representing an observed image or high-dimensional distribution as the transformation of another. To accomplish this, "morphing" models have been developed to capture transformations among two or more inputs [70]. Recently, a new class of general "transport" models has been introduced that describes an input distribution as a smooth, nonlinear, invertible transformation of a reference distribution [71, 72]. The estimation of such models from observed data is facilitated by a set of transport-based transforms [71, 73]. Unlike most numerical feature-based methods, these transform operations are invertible, making them a mathematical representation method for input images or distributions. These transforms, such as Linear Optimal Transport (LOT) and Radon-Cumulative Distribution Transform (R-CDT), developed in [71, 73] are linked to the optimal transport theory [72, 74]. In previous studies, LOT and R-CDT models have been combined with linear classifiers like Fisher discriminant analysis and support vector machines, along with their respective kernel techniques [71, 75, 76]. While this approach has been successful in some applications [76], it has failed to achieve state-of-the-art classification results in certain other applications (refer to Fig. 3 from [32]).

Objectives and contributions of our study

Our study aims to propose a new set of methods for solving several pattern analysis and recognition problems, where the data at hand can be described with a transport-based data generative model. We first introduce a novel classification method for a particular type of image classification problem where image classes are formed with unknown templates under the effect of unknown deformations. To achieve this, we utilize transport-based embeddings and demonstrate that such classification problems can be solved in closed-form using our technique. Our proposed method exhibits competitive accuracy performance compared with state-of-the-art methods in both low and high data regimes, requiring minimal labeled data for training and being computationally efficient. Moreover, our approach is robust under out-of-distribution scenarios, implying that the

model can generalize to previously unseen data. Additionally, we leverage the mathematical properties of the model to provide approximations for known transformations, enabling the model to learn those transformations automatically without the need for data augmentation. We further extend our framework to formulate transport-based embeddings for classifying high-dimensional distributions, which can be applied to a broad range of applications. Our experimental results demonstrate that the transport-based embeddings approach enables the development of a simple, efficient, mathematically coherent, and robust classification method with high accuracy using real-world data.

Our final objective is to address a digital pathology problem by proposing a mathematical method that models nuclear chromatin structure and morphology using routinely processed and imaged tissues in clinical settings. By considering normalized intensity values as relative measurements of chromatin density, our technique models the relative intensity observed in each pixel within a nucleus relative to a template (i.e. average) nucleus. The technique thus preserves the entire information content of each nucleus image within a biologically meaningful, transport-based, representation. Statistical analyses are then employed to summarize chromatin transport-based variabilities observed within and across datasets, as well as to elucidate meaningful discriminating information between relative malignancy levels within and across cancer types. We demonstrate our transport-based morphometry (TBM) technique can not only detect and interpret meaningful malignancy levels within each of the four cancer tissue types (liver, thyroid, lung, and skin), but also to detect and interpret persistent discriminating information along the spectrum from benign and malignant categories across these different cancer types, even when imaged using different protocols, resolutions, and staining patterns. We believe these proof of concept calculations can be used as preliminary evidence that our proposed technique can provide the quantitative measurements necessary to enable meaningful comparisons across a wide range of datasets. In combination with interesting emerging datasets (such as the human protein atlas [77], the cancer genome atlas [78]), we believe that our techniques can elevate the role of nuclear morphometry for use in cancer studies, technologies, and clinical applications in the emerging use of digital pathology tools to aid

the pathologists, and help to render nuclear morphology studies into a more quantitative science.

Thesis overview

In this dissertation, we introduce a new set of methods for solving various pattern analysis and recognition problems, which can be described using a transport-based data generative model. In the first chapter, we introduce the transport-based transforms and embeddings, which form the basis of the proposed methods. Chapters 2 and 3 focus on the classification of a certain type of image and high-dimensional distribution classes, respectively. These classes are formed with unknown templates under the effect of unknown spatial deformations, and are classified using transport-based embeddings. In Chapter 4, we demonstrate the application of our method in a digital pathology context by quantifying nuclear structures of digital pathology images across cancers using transport-based morphometry. Finally, Chapter 5 offers concluding remarks and suggests avenues for future research. Mathematical derivations and other related details are presented in the appendices.

1.1 The Cumulative Distribution Transform (CDT)

The CDT [79] is an invertible nonlinear 1D signal transform from the space of smooth probability densities to the space of diffeomorphisms. The CDT morphs a given input signal, defined as a probability density function (PDF), into another PDF in such a way that the Wasserstein distance between them is minimized. More formally, let $s(x), x \in \Omega_s$ and $r(x), x \in \Omega_r$ define a given signal and a reference signal, respectively, which we consider to be appropriately normalized such that $s > 0, r > 0$, and $\int_{\Omega_s} s(x)dx = \int_{\Omega_r} r(x)dx = 1$. The forward CDT transform[1] of $s(x)$ with respect to $r(x)$ is given by the strictly increasing function $\widehat{s}(x)$ that satisfies

$$\int_{-\infty}^{\widehat{s}(x)} s(u)du = \int_{-\infty}^{x} r(u)du$$

[1] We are using a slightly different definition of the CDT than in [79]. The properties of the CDT outlined here hold in both definitions.

As described in detail in [79], the CDT is a nonlinear and invertible operation, with the inverse being

$$s(x) = \frac{d\widehat{s}^{-1}(x)}{dx} r\left(\widehat{s}^{-1}(x)\right), \text{ and } \widehat{s}^{-1}(\widehat{s}(x)) = x$$

Moreover, like the Fourier transform [80] for example, the CDT has a number of properties which will help us render signal and image classification problems easier to solve.

Property 1.1-A *(Composition)*: Let $s(x)$ denote a normalized signal and let $\widehat{s}(x)$ be the CDT of $s(x)$. The CDT of $s_g = g's \circ g$ is given by

$$\widehat{s}_g = g^{-1} \circ \widehat{s} \tag{1.1}$$

Here, $g \in \mathcal{T}$ is an invertible and differentiable function (diffeomorphism), $g' = dg(x)/dx$, and '$\circ$' denotes the composition operator with $s \circ g = s(g(x))$. For a proof, see Appendix A.1.

The CDT composition property implies that, variations in a signal caused by applying $g(x)$ to the independent variable will change only the dependent variable in CDT space. In essence, this property asserts that variations along both the independent and dependent axes in the original signal space are translated into changes entirely along the dependent axis in CDT space.

Property 1.1-B *(Embedding)*: CDT induces an isometric embedding between the space of 1D signals with the 2-Wasserstein metric and the space of their CDT transforms with a weighted-Euclidean metric [71][79], i.e.,

$$W_2^2(s_1, s_2) = \left\|(\widehat{s}_1 - \widehat{s}_2)\sqrt{r}\right\|^2_{L^2(\Omega_r)}, \tag{1.2}$$

for all signals s_1, s_2. That is to say, if we wish to use the Wasserstein distance as a measure of similarity between $s_1,\ s_2$, we can compute it as simply a weighted Euclidean norm in CDT space.

For a proof, see Appendix A.2.

The property above naturally links the CDT and Wasserstein distances for PDFs. Wasserstein [74] distances are linked to optimal transport and have been used in a variety of applications in signal and image processing and machine learning (see [72] for a recent review).

1.2 The Radon transform

The Radon transform of an image $s(\mathbf{x}), \mathbf{x} \in \mathbf{\Omega}_s \subset \mathbb{R}^2$, which we denote by $\widetilde{s} = \mathscr{R}(s)$, is defined as

$$\widetilde{s}(t,\theta) \;=\; \int_{\Omega_s} s(\mathbf{x})\delta(t - \mathbf{x}\cdot\xi_\theta)d\mathbf{x} \tag{1.3}$$

Here, t is the perpendicular distance of a line from the origin and $\xi_\theta = [\cos(\theta), \sin(\theta)]^T$, where θ is the angle over which the projection is taken.

Furthermore, using the Fourier Slice Theorem [81][82], the inverse Radon transform $s = \mathscr{R}^{-1}(\widetilde{s})$ is defined as

$$s(\mathbf{x}) \;=\; \int_0^\pi \int_{-\infty}^{\infty} \widetilde{s}(\mathbf{x}\cdot\xi_\theta - \tau, \theta)w(\tau)d\tau d\theta, \tag{1.4}$$

where w is the ramp filter (i.e.,$(\mathscr{F}w)(\xi) = |\xi|, \forall\xi$) and $\mathscr{F}$ is the Fourier transform.

Property 1.2-A (*Intensity equality*): Note that

$$\int_{\Omega_s} s(\mathbf{x})d\mathbf{x} = \int_{-\infty}^{\infty} \widetilde{s}(t,\theta)dt, \qquad \forall\theta \in [0,\pi] \tag{1.5}$$

which implies that $\int_{-\infty}^{\infty} \widetilde{s}(t,\theta_i)dt = \int_{-\infty}^{\infty} \widetilde{s}(t,\theta_j)dt$ for any two choices $\theta_i, \theta_j \in [0,\pi]$.

1.3 Radon Cumulative Distribution Transform (R-CDT)

The CDT framework was extended for 2D patterns (images as normalized density functions) through the sliced-Wasserstein distance in [71], and was denoted as R-CDT. The main idea behind the R-CDT is to first obtain a family of one dimensional representations of a two dimensional probability measure (e.g., an image) through the Radon transform and then apply the CDT over the t dimension in Radon transform space. More formally, let $s(\mathbf{x})$ and $r(\mathbf{x})$ define a given image and a reference image, respectively, which we consider to be appropriately normalized. The forward R-CDT of $s(\mathbf{x})$ with respect to $r(\mathbf{x})$ is given by the measure preserving function $\widehat{s}(t,\theta)$ that satisfies

$$\int_{-\infty}^{\widehat{s}(t,\theta)} \widetilde{s}(u,\theta)du = \int_{-\infty}^{t} \widetilde{r}(u,\theta)du, \quad \forall\theta \in [0,\pi] \tag{1.6}$$

As in the case of the CDT, a transformed signal in R-CDT space can be recovered via the following inverse formula [71],

$$s(\mathbf{x}) = \mathscr{R}^{-1}\left(\frac{\partial \widehat{s}^{-1}(t,\theta)}{\partial t}\widetilde{r}\left(\widehat{s}^{-1}(t,\theta),\theta\right)\right)$$

As with the CDT, the R-CDT has a couple of properties outlined below which will be of interest when classifying images.

Property 1.3-A *(Composition)*: Let $s(\mathbf{x})$ denotes an appropriately normalized image and let $\widetilde{s}(t,\theta)$ and $\widehat{s}(t,\theta)$ be the Radon transform and the R-CDT transform of $s(\mathbf{x})$, respectively. The R-CDT transform of $s_{g^\theta} = \mathscr{R}^{-1}\left(\left(g^\theta\right)'\widetilde{s}\circ g^\theta\right)$ is given by

$$\widehat{s}_{g^\theta} = (g^\theta)^{-1}\circ\widehat{s}, \tag{1.7}$$

where $\left(g^\theta\right)' = dg^\theta(t)/dt$, $\widetilde{s}\circ g^\theta := \widetilde{s}(g^\theta(t),\theta)$, and $(g^\theta)^{-1}\circ\widehat{s} = (g^\theta)^{-1}(\widehat{s}(t,\theta))$. Here for a fixed θ, g^θ can be thought of an increasing and differentiable function with respect to t. The above equation

hence follows from the composition property for 1D CDT. For a proof, see Appendix A.3.

The R-CDT composition property implies that, variations along both independent and dependent axis directions in an image, caused by applying $g^\theta(t)$ to the independent t variable of its Radon transform, become changes solely along the dependent variable in R-CDT space.

Property 1.3-B *(Embedding)*: R-CDT induces an isometric embedding between the space of images with sliced-Wasserstein metric and the space of their R-CDT transforms with a weighted-Euclidean metric, i.e.,

$$SW_2^2(s_1, s_2) = \left\| (\widehat{s}_1 - \widehat{s}_2)\sqrt{\bar{r}} \right\|^2_{L^2(\Omega_{\bar{r}})} \tag{1.8}$$

for all images s_1 and s_2. For a proof, see Appendix A.4.

As the case with the 1D CDT shown above, the property above naturally links the R-CDT and sliced Wasserstein distances for PDFs and affords us a simple means of computing similarity among images [71]. We remark that throughout this chapter we use the notation $\widehat{s}$ for both CDT or R-CDT transforms of a signal or image s with respect to a fixed reference signal or image r, if a reference is not specified.

1.4 Linear Optimal Transport (LOT) embeddings

The fundamental principle of optimal transport theory relies on quantifying the amount of effort (measured as the product of mass and distance) required to rearrange one distribution to another, which gives rise to the Wasserstein metric between distributions. Here we present a linearized version of this metric, as outlined in [73], which is constructed formally through a tangent space approximation of the underlying manifold.

Following the construction in [73], we define the linear optimal transport transform for probability measures in $\mathcal{P}_2(\mathbb{R}^L)$, which is the set of absolutely continuous measures with bounded finite second moments and densities [2]. For simplicity, let us fix a reference measure σ as the Lebesgue

[2]Any $\mu \in \mathcal{P}_2(\mathbb{R}^L)$ has the following two properties (i) bounded second moment, i.e. $\int \|x\|^2 d\mu(x) < \infty$; (ii) absolute continuity with respect to the Lebesgue measure on $\mathbb{R}^L$ with bounded density, i.e., μ has a density function

measure on a convex compact set of $\mathbb{R}^L$. Thanks to Brenier's theorem [83], there is a unique minimizer T_σ^μ to the following optimal transportation problem

$$\min_{T_\sharp\sigma=\mu} \int_{\mathbb{R}^L} \|x - T(x)\|^2 d\sigma(x), \tag{1.9}$$

where the push-forward (transport) relation $T_\sharp\sigma = \mu$ is defined via $\mu(B) = \sigma(T^{-1}(B))$ for any measurable set $B \subseteq \mathbb{R}^L$. The linear optimal transport (LOT) transform is given by the following correspondence

$$\mu \mapsto T_\sigma^\mu, \tag{1.10}$$

where each probability measure μ is identified with the optimal transport map $T_\sigma^\mu : \mathbb{R}^L \to \mathbb{R}^L$ from a fixed reference σ to μ, which lies in a linear space. This square-root of the minimum is called the Wasserstein-2 distance between σ and μ [84]. The LOT metric between two probability distributions $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^L)$ is [3]

$$d_{\mathrm{LOT}}(\mu, \nu) := \|T_\sigma^\mu - T_\sigma^\nu\|_\sigma. \tag{1.11}$$

For simplicity, we denote $\widehat{\mu}$ as the LOT transform of μ, i.e., $\widehat{\mu} = T_\sigma^\mu$ where σ is fixed.

1.4.1 LOT for point-sets

For the analysis of discrete point-set data, a discrete version of the Linear Optimal Transport (LOT) embedding is required. In this particular case, both the reference σ and target μ are chosen as discrete probability measures, represented by point-sets in $\mathbb{R}^L$. A point-set in a L-dimensional space is a finite set of points in $\mathbb{R}^L$. A point-set Ω_s with N points can be thought as the image of an injective map $s : \{1, \cdots, N\} \to \mathbb{R}^L$ [4]. Given a point-set Ω_s with N points, we define a discrete

f_μ defined on $\mathbb{R}^L$ with $\|f_\mu\|_\infty < \infty$.

[3] Note that $\|T\|_\sigma := \left(\int_{\mathbb{R}^L} \|T(x)\|^2 d\sigma(x) \right)^{1/2}$.

[4] Note that a point-set may be associated with many injective maps, e.g. the image sets of $s \circ \gamma$ and s are the same for any permutation γ.

probability distribution associated with the point-set as

$$P_s := \frac{1}{N} \sum_{\mathbf{x} \in \Omega_s} \delta_{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{s(i)}. \tag{1.12}$$

Given a diffeomorphism $g \in \mathcal{T}_L$, the push-forward distribution of P_s under g is given as

$$g_{\#} P_s := \frac{1}{N} \sum_{\mathbf{x} \in \Omega_s} \delta_{g(\mathbf{x})} = \frac{1}{N} \sum_{i=1}^{N} \delta_{g(s(i))} = P_{g \circ s}. \tag{1.13}$$

Let $\mathcal{F}_{\mathcal{N},\mathcal{L}}$ denote the collection of injective maps from $\{1, \cdots, N\}$ to $\mathbb{R}^L$. Given $s, r \in \mathcal{F}_{\mathcal{N},\mathcal{L}}$, the optimal transportation (Wasserstein-2) distance between associated distributions P_s and P_r can be obtained by solving the linear programming problem given below:

$$d_W^2(P_s, P_r) = \min_{\pi \in \mathbb{R}^{N \times N}} \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{ij} |s(i) - r(j)|^2 \tag{1.14}$$

where $\pi_{ij} \geq 0$, and $\sum_{i=1}^{N} \pi_{ij} = \sum_{j=1}^{N} \pi_{ij} = 1/N$ for all $i, j = 1, \cdots, N$. Let us fix some $r \in \mathcal{F}_{\mathcal{N},\mathcal{L}}$ and use P_r as a reference. It turned out that any minimizer matrix π^* to the optimal transport problem in (1.14) is a permutation matrix[84]. In other words, there is a permutation $\sigma_s^* : \{1, \cdots, N\} \to \{1, \cdots, N\}$ such that

$$\pi_{ij}^* = \begin{cases} 1/N & \text{if } j = \sigma_s^*(i) \\ 0 & \text{otherwise} \end{cases}.$$

Hence with r being fixed, an optimal transport map between P_r and P_s can be determined by σ_s^* and s. The LOT transform for P_s is defined as [73] [5]

$$\widehat{P}_s := s \circ \sigma_s^*, \tag{1.15}$$

[5]Note one can write $s \circ \sigma_s^* = \left[s(\sigma_s^*(1)), \cdots, s(\sigma_s^*(N)) \right]^T$. Note also that σ_s^* may not be unique in general, we follow the implementation in [73] to estimate one of them.

and the LOT distance between two point-set measures is

$$d_{\text{LOT}}(P_s, P_q) := ||\widehat{P}_s - \widehat{P}_q||, \tag{1.16}$$

where $s, q \in \mathcal{F}_{\mathcal{N},\mathcal{L}}$.

Chapter 2: Image classes formed with unknown templates under the effect of unknown deformations can be classified in closed-form using the transport-based embeddings

Part A: Formulating a data generative model for image classification and developing a framework for a data-efficient, computationally efficient, generalizable, and robust classification method.

We introduce a new supervised image classification method applicable to a broad class of image deformation models. The method makes use of the Radon Cumulative Distribution Transform (R-CDT) for image data, whose mathematical properties are exploited to express the image data in a form that is more suitable for machine learning. While certain operations such as translation, scaling, and higher-order transformations are challenging to model in native image space, we show the R-CDT can capture some of these variations and thus render the associated image classification problems easier to solve. The method — utilizing a nearest-subspace algorithm in the R-CDT space – is simple to implement, non-iterative, has no hyper-parameters to tune, is computationally efficient, label efficient, and provides competitive accuracies to state-of- the-art neural networks for many types of classification problems. In addition to the test accuracy performances, we show improvements (with respect to neural network-based methods) in terms of computational efficiency (it can be implemented without the use of GPUs), number of training samples needed for training, as well as out-of-distribution generalization.

2.1 Generative model and problem statement

We begin with a discussion of a generative model-based problem statement for the type of classification problems we discuss in this chapter. We note that in many applications we are concerned

with classifying image or signal patterns that are instances of a certain prototype (or template) observed under some often unknown deformation pattern. Consider the problem of classifying handwritten digits (e.g. the MNIST dataset [8]). A good model for each class in such a dataset is to assume that each observed digit image can be thought of as being an instance of a template (or templates) observed under some (unknown) deformation or similar variation or confound. For example, a generative model for the set of images of the digit 1 could be a fixed pattern for the digit 1, but observed under different translations – the digit can be positioned randomly within the field of view of the image. Alternatively, the digit could also be observed with different sizes, or slight deformations. The generative models stated below for 1D and 2D formalize these statements.

Example 1 (1D generative model with translation) *Consider a 1D signal pattern denoted as $\varphi^{(k)}$ (the superscript (k) here denotes the class in a classification problem), observed under a random translation parameter μ. In this case, we can mathematically represent the situation by defining the set of all possible functions $g(x) = x - \mu$, with μ being a random variable whose distribution is typically unknown. A random observation (randomly translated) pattern can be written mathematically as $g'(x)\varphi^{(k)}(g(x))$. Note that in this case $g'(x) = 1$, and thus the generative model simply amounts to random translation of a template pattern. Fig. 2.1 depicts this situation.*

The example above (summarized in Fig. 2.1) can be expressed in more general form. Let $\mathcal{G} \subset \mathcal{T}$ denotes a set of 1D spatial transformations of a specific kind (e.g. the set of affine transformations). We then use these transformations to provide a more general definition for a mass (signal intensity) preserving generative data model.

Definition 2.1.1 (1D generative model)**.** *Let $\mathcal{G} \subset \mathcal{T}$. The 1D mass (signal intensity) preserving generative model for the k^{th} class is defined to be the set*

$$\mathbb{S}^{(k)} \quad = \quad \{s_j^{(k)} | s_j^{(k)} = g_j' \varphi^{(k)} \circ g_j, \forall g_j \in \mathcal{G}\}. \tag{2.1}$$

The notation $s_j^{(k)}$ here is meant to denote the j^{th} signal from the k^{th} class. The derivative term

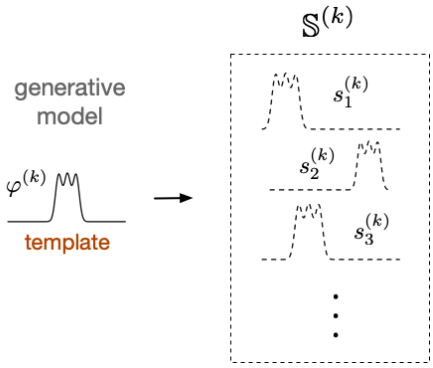

Figure 2.1: Generative model example. A signal generative model can be constructed by applying randomly drawn confounding spatial transformations, in this case translation ($g(x) = x - \mu$), to a template pattern from class (k), denoted here as $\varphi^{(k)}$. The notation $s_j^{(k)}$ here is meant to denote the j^{th} signal from the k^{th} class.

g'_j preserves the normalization of signals. This extension allows us to define and discuss problems where the confound goes beyond a simple translation model.

With the definition of the 2-Dimensional Radon transform from section 1.2, we are now ready to define the 2-dimensional definition of the generative data model we use throughout the chapter:

Definition 2.1.2 (2D generative model)**.** *Let $\mathcal{G} \subset \mathcal{T}$ be our set of confounds. The 2D mass (image intensity) preserving generative model for the k^{th} class is defined to be the set*

$$\mathbb{S}^{(k)} = \left\{ s_j^{(k)} | s_j^{(k)} = \mathscr{R}^{-1}\left(\left(g_j^{\theta} \right)' \widetilde{\varphi}^{(k)} \circ g_j^{\theta} \right), \forall g_j^{\theta} \in \mathcal{G} \right\}. \tag{2.2}$$

We note that the generative model above can yield a non convex set, depending on the choice of template function $\varphi^{(k)}$ and confound category $\mathcal{G}$. Note that we use the same notation $\mathbb{S}^{(k)}$ for both 1D and 2D versions of the set. The meaning each time will be clear from the context.

We are now ready to define a mathematical description for a generative model-based problem statement using the definitions above:

Definition 2.1.3 (Classification problem)**.** *Let $\mathcal{G} \subset \mathcal{T}$ and $\mathcal{G}$ define our set of confounds, and let $\mathbb{S}^{(k)}$ be defined as in equation* (2.1) *(for signals) or equation* (2.2) *(for images). Given training*

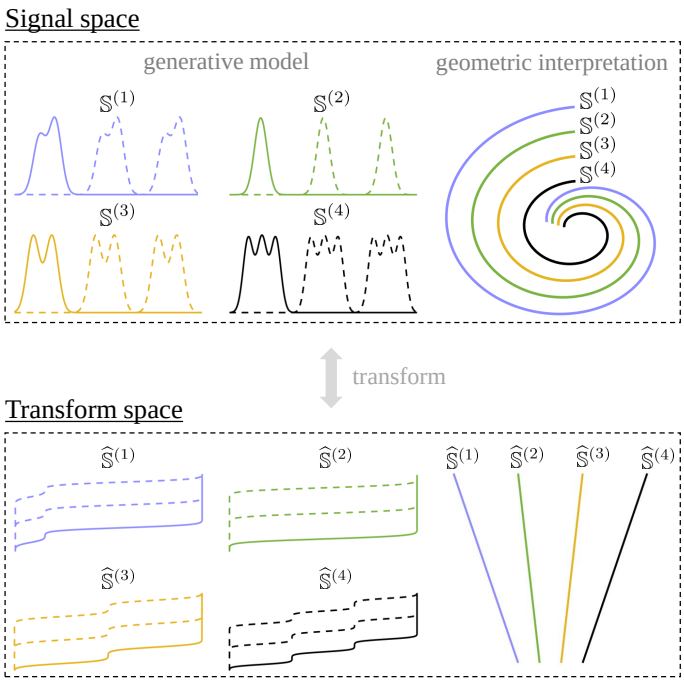

Figure 2.2: Generative model for signal classes in signal (top panel) and transform (bottom panel) spaces. Four classes are depicted on the left: $\mathbb{S}^{(1)}, \mathbb{S}^{(2)}, \mathbb{S}^{(3)}, \mathbb{S}^{(4)}$, each with three example signals shown. The top panel: it shows the signal classes in their corresponding native signal spaces. For each class, three example signals are shown under different translations. The right portion of the top panel shows the geometry of these four classes forming nonlinear spaces. The bottom panel: it depicts the situation in transform (CDT, or R-CDT) space. The left portion of the bottom panel shows the corresponding signals in transform domain, while the right portion shows the geometry of the signal classes forming convex spaces.

samples $\{s_1^{(1)}, s_2^{(1)}, \cdots\}$ *(class 1),* $\{s_1^{(2)}, s_2^{(2)}, \cdots\}$ *(class 2),* $\cdots$ *as training data, determine the class* (k) *of an unknown signal or image* s*.*

It is important to note that the generative model discussed yields nonconvex (and hence nonlinear) signal classes (see Fig. 2.2, top panel). We express this fact mathematically as: for arbitrary $s_i^{(k)}$ and $s_j^{(k)}$ we have that $\alpha s_i^{(k)} + (1-\alpha) s_j^{(k)}$, for $\alpha \in [0,1]$, may not necessarily be in $\mathbb{S}^{(k)}$. The situation is similar for images (the 2D cases). Convexity, on the other hand, means the weighted sum of samples *does* remain in the set; this property greatly simplifies the classification problem as will be shown in the next section.

2.2 Proposed solution

We postulate that the CDT and R-CDT introduced earlier can be used to drastically simplify the solution to the classification problem posed in definition 2.1.3. While the generative model discussed above generates nonconvex (hence nonlinear) signal and image classes, the situation can change by transforming the data using the CDT (for 1D signals) or the R-CDT (for 2D images). We start by analyzing the one dimensional generative model from definition 2.1.1.

Employing the composition property of the CDT (see Section 1.1) to the 1D generative model stated in equation (2.1) we have that

$$\widehat{s}_j^{(k)} = g_j^{-1} \circ \widehat{\varphi}^{(k)} \tag{2.3}$$

and thus

$$\widehat{\mathbb{S}}^{(k)} \quad = \quad \{\widehat{s}_j^{(k)} | \widehat{s}_j^{(k)} = g_j^{-1} \circ \widehat{\varphi}^{(k)}, \forall g_j \in \mathcal{G}\}.$$

Thus we have the following lemma:

Lemma 2.2.1. *If $\mathcal{G} \subset \mathcal{T}$ is a convex group, the set $\widehat{\mathbb{S}}^{(k)}$ is convex.*

Proof. Let $\varphi^{(k)}$ be a template signal defined as a PDF. For $g_j \in \mathcal{G}$, let $s_j^{(k)} = g_j'(\varphi^{(k)} \circ g_j)$. Then using the composition property of CDT, we have that $\widehat{s}_j^{(k)} = g_j^{-1} \circ \widehat{\varphi}^{(k)}$. Hence $\widehat{\mathbb{S}}^{(k)} = \{g_j^{-1} \circ \widehat{\varphi}^{(k)} \mid g_j \in \mathcal{G}\}$. Since $\mathcal{G}$ is a convex group, $\mathcal{G}^{-1}$ is convex, and it follows that $\widehat{\mathbb{S}}^{(k)}$ is convex.

□

Remark 2.2.2. *Let $\mathbb{S}^{(k)}$ and $\mathbb{S}^{(p)}$ represent two generative models. If $\mathbb{S}^{(k)} \cap \mathbb{S}^{(p)} = \varnothing$, then $\widehat{\mathbb{S}}^{(k)} \cap \widehat{\mathbb{S}}^{(p)} = \varnothing$.*

This follows from the fact that the CDT is a one-to-one map between the space of probability density functions and the space of 1D diffeomorphisms. As such the CDT operation is one to one, and therefore there exists no $\widehat{s}_j^{(k)} = \widehat{s}_i^{(p)}$.

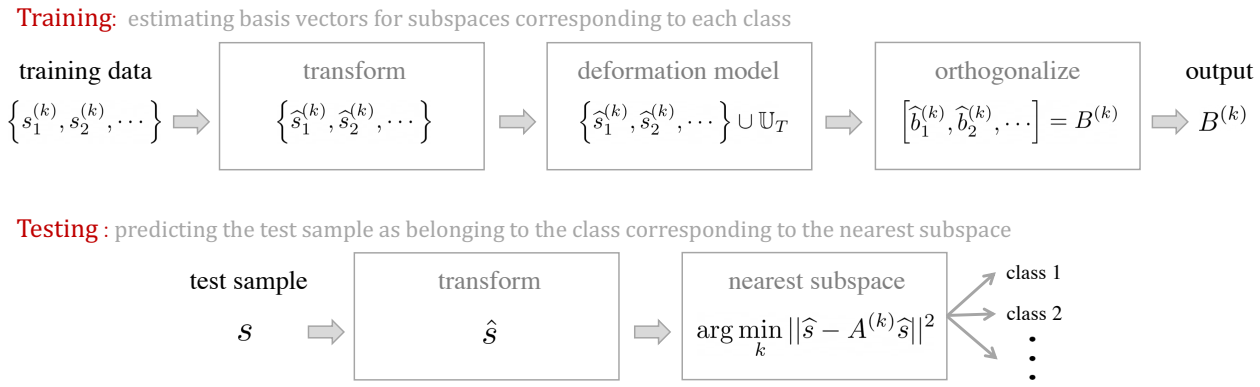

Figure 2.3: The training and testing process of the proposed classification model. Training: First, obtain the transform space representations of the given training samples of a particular class (k). Then, enrich the space by adding the deformation spanning set $\mathbb{U}_T$ (see text for definition). Finally, orthogonalize to obtain the basis vectors which span the enriched space. Testing: First, obtain the transform space representation of a test sample s. Then, the class of s is estimated to be the class corresponding to the subspace $\widehat{\mathbb{S}}_E^{(k)}$ which has the minimum distance $d^2(\widehat{s}, \widehat{\mathbb{S}}_E^{(k)})$ from $\widehat{s}$ (see text for definitions). Here, $A^{(k)} = B^{(k)}B^{(k)^T}$.

Lemma 2.2.1 above implies that if the set of spatial transformations formed by taking elements of $\mathcal{G}$ and inverting them (denoted as $\mathcal{G}^{-1}$) is convex, then the generative model will be convex in signal transform space. The situation is depicted in Fig. 2.2. The top part shows a four class generative model that is nonlinear/non-convex. When examined in transform space, however, the data geometry simplifies in a way that signals can be added together to generate other signals in the same class – the classes become convex in transform space.

The analysis above can be extended to the case of the 2D generative model (definition 2.1.2) through the R-CDT. Employing the composition property of the R-CDT (see Section 1.3) to the 2D generative model stated in equation (2.2) we have that

$$\widehat{\mathbb{S}}^{(k)} = \{\widehat{s}_j^{(k)} | \widehat{s}_j^{(k)} = \left(g_j^\theta\right)^{-1} \circ \widehat{\varphi}^{(k)}, \forall g_j^\theta \in \mathcal{G}\}. \tag{2.4}$$

Lemma 2.2.1 and Remark 2.2.2 hold true in the 2-dimensional R-CDT case as well. Thus, if $\mathcal{G}_R$ is a convex group, the R-CDT transform simplifies the data geometry in a way that image classes become convex in the R-CDT transform space. Fig. 2.4(a) depicts the situation.

We use this information to propose a simple non-iterative training algorithm (described in

more detail in Section 2.2.1) by estimating a projection matrix that projects each (transform space) sample onto $\widehat{\mathbb{V}}^{(k)}$, for all classes $k = 1, 2, \cdots$, where $\widehat{\mathbb{V}}^{(k)}$ denotes the subspace generated by the convex set $\widehat{\mathbb{S}}^{(k)}$ as follows:

$$\widehat{\mathbb{V}}^{(k)} = \text{span}\left(\widehat{\mathbb{S}}^{(k)}\right) = \left\{ \sum_{j \in J} \alpha_j \widehat{s}_j^{(k)} \mid \alpha_j \in \mathbb{R}, J \text{ is finite} \right\}. \tag{2.5}$$

Fig. 2.4(b) provides a pictorial representation of $\widehat{\mathbb{V}}^{(k)}$.

Lemma 2.2.3. *Let $\mathbb{S}^{(k)}$, $k = 1, 2, ...$, be generative classes with a common confound set $\mathcal{G}$ such that for any $f \notin \mathcal{G}$, $f'\varphi^{(k)} \circ f \notin \mathbb{S}^{(k)}$. If $\mathcal{G}$ is a convex group that also includes scaling, $\widehat{\mathbb{S}}^{(k)} \cap \widehat{\mathbb{S}}^{(p)} = \emptyset$, and*

$$\alpha\, id + (1 - \alpha) h \notin \mathcal{G} \tag{2.6}$$

$\forall$ increasing function $h \notin \mathcal{G}$ and $0 < \alpha < 1$ (here id denotes the identity function, $f(x) = x$), then $\widehat{\mathbb{S}}^{(k)} \cap \widehat{\mathbb{V}}^{(p)} = \emptyset$.

Proof. For a proof, see Appendix 2.12.1. □

Lemma 2.2.3 above states that given certain assumptions, the convex space for a particular class does not overlap with the subspace corresponding to a different class. A corollary from Lemma 2.2.3 is that $\alpha\widehat{s}_i^{(k)} + (1 - \alpha)\widehat{s}_j^{(p)} \notin \widehat{\mathbb{S}}^{(k)} \cup \widehat{\mathbb{S}}^{(p)}$ for all $\widehat{s}_i^{(k)} \in \widehat{\mathbb{S}}^{(k)}$ and $\widehat{s}_j^{(p)} \in \widehat{\mathbb{S}}^{(p)}$ with $0 < \alpha < 1$ (see Appendix 2.12.1). Intuitively speaking, the generative classes generated by $\mathcal{G}$ are "thin" in the transform space. Lemma 2.2.3 holds true for the 2-dimensional R-CDT case as well.

There are a number examples of $\mathcal{G}$ that satisfy the assumption in equation (2.6). For example, if $\mathcal{G}$ is the set of translation functions, any strict convex combination (i.e., for $0 < \alpha < 1$) of a function other than translation and the identity function, is not a translation function either. One can also verify that there are other sets of functions that also satisfy the assumption, e.g., the set of increasing affine functions, the set of diffeomorphisms that have a common fixed point, etc.

It follows from Lemma 2.2.3 that, if the test sample was generated according to the generative model for one of the classes, then there will exist exactly one class (k) for which $d^2(\widehat{s}, \widehat{\mathbb{S}}^{(k)}) = d^2(\widehat{s}, \widehat{\mathbb{V}}^{(k)}) = 0$. It also follows, $d^2(\widehat{s}, \widehat{\mathbb{V}}^{(p)}) > 0$ when $k \neq p$ [1]. Here $d^2(\cdot, \cdot)$ is the Euclidean distance between $\widehat{s}$ and the nearest point in $\widehat{\mathbb{S}}^{(k)}$ or $\widehat{\mathbb{V}}^{(k)}$.

As far as a test procedure for determining the class of some unknown signal or image s, under the assumption that $\widehat{\mathbb{S}}^{(k)} \cap \widehat{\mathbb{V}}^{(p)} = \varnothing$, it then suffices to measure the distance between $\widehat{s}$ and the nearest point in each subspace $\widehat{\mathbb{V}}^{(k)}$ corresponding to the generative model $\widehat{\mathbb{S}}^{(k)}$. Therefore, under the assumption that the testing sample at hand s was generated according to one of the (unknown) classes as described in definition 2.1.3, the class of the unknown sample can be decoded by solving

$$\arg\min_{k} d^2(\widehat{s}, \widehat{\mathbb{V}}^{(k)}). \tag{2.7}$$

Finally, note that due to property 1.3-B we also have that

$$d^2(\widehat{s}, \widehat{\mathbb{S}}^{(k)}) = \min_{g^\theta} SW_2^2 \left(s, \mathscr{R}^{-1} \left(\left(g^\theta \right)' \widetilde{\varphi}^{(k)} \circ g^\theta \right) \right)$$

with $g^\theta \in \mathcal{G}$. In words, the R-CDT nearest subspace method proposed in equation (2.7) can be considered to be equivalent to a nearest (in the sense of the sliced-Wasserstein distance) subset method in image space, with the subset given by the generative model stated in definition 2.1.2. Fig. 2.4 shows a system diagram outlining the main computational modeling steps in the proposed method.

2.2.1 Training algorithm

Using the principles and assumptions laid out above, the algorithm we propose estimates the subspace $\widehat{\mathbb{V}}^{(k)}$ corresponding to the transform space $\widehat{\mathbb{S}}^{(k)}$ given sample data $\{s_1^{(k)}, s_2^{(k)}, \cdots\}$. Naturally, the first step is to transform the training data to obtain $\{\widehat{s}_1^{(k)}, \widehat{s}_2^{(k)}, \cdots\}$. We then approximate

[1] Rigorously speaking, if $\widehat{\mathbb{V}}^{(p)}$ is a closed subspace, then $d^2(\widehat{s}, \widehat{\mathbb{V}}^{(p)}) > 0$ if and only if $\widehat{s} \notin \widehat{\mathbb{V}}^{(p)}$. In practice, $\widehat{\mathbb{V}}^{(p)}$ will be a finite dimensional space and hence the closedness condition is satisfied.

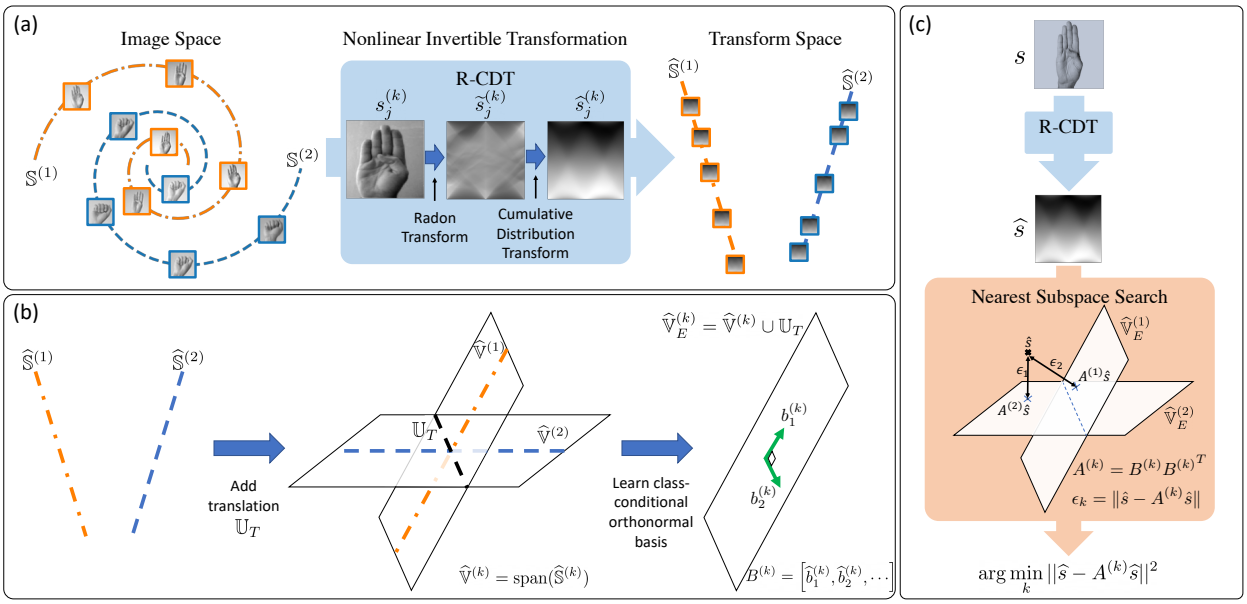

Figure 2.4: System diagram outlining the proposed Radon cumulative distribution transform subspace modeling technique for image classification. (a) R-CDT - a nonlinear, invertible transformation: The R-CDT transform simplifies the data space; (b) Generative modeling - subspace learning: the simplified data spaces can be modeled as linear subspaces; (c) Classification pipeline: the classification method consists of the R-CDT transform followed by a nearest subspace search in the R-CDT space.

$\widehat{\mathbb{V}}^{(k)}$ as follows:

$$\widehat{\mathbb{V}}^{(k)} = \text{span}\left\{\widehat{s}_1^{(k)}, \widehat{s}_2^{(k)}, \cdots\right\}.$$

Given the composition properties for the CDT and R-CDT, it is also possible to enrich $\widehat{\mathbb{V}}^{(k)}$ in such a way that it will automatically include the samples undergoing some specific deformations without explicitly training with those samples under said deformation. The spanning sets corresponding to two such deformations, image domain translation and isotropic scaling, are derived below:

i) Translation: let $g(\mathbf{x}) = \mathbf{x} - \mathbf{x_0}$ be the translation by $\mathbf{x}_0 \in \mathbb{R}^2$ and $s_g(\mathbf{x}) = |\det Jg| s \circ g = s(\mathbf{x} - \mathbf{x_0})$. Note that Jg denotes the Jacobian matrix of g. Following [71] we have that $\widehat{s}_g(t,\theta) = \widehat{s}(t,\theta) + \mathbf{x}_0^T \xi_\theta$ where $\xi_\theta = [\cos(\theta), \sin(\theta)]^T$. We define the spanning set for translation in transform domain as $\mathbb{U}_T = \{u_1(t,\theta), u_2(t,\theta)\}$, where $u_1(t,\theta) = \cos\theta$ and

$u_2(t, \theta) = \sin\theta$.

ii) Isotropic scaling: let $g(\mathbf{x}) = \alpha\mathbf{x}$ and $s_g(\mathbf{x}) = |Jg|s \circ g = \alpha^2 s(\alpha\mathbf{x})$, which is the normalized dilatation of s by α where $\alpha \in \mathbb{R}_+$. Then according to [71], $\widehat{s}_g(t, \theta) = \widehat{s}(t, \theta)/\alpha$, i.e. a scalar multiplication. Therefore, an additional spanning set is not required here and thereby the spanning set for isotropic scaling becomes $\mathbb{U}_D = \varnothing$.

Note that the spanning sets are not limited to translation and isotropic scaling only. Other spanning sets might be defined as before for other deformations as well. However, deformation spanning sets other than translation and isotropic scaling are not used here and left for future exploration. We also emphasize that there is a crucial difference between enriching the subspace and data augmentation. Our method only requires one spanning set (set of basis functions) corresponding to a deformation type. The method can use it to learn the instances under that deformation type. On the contrary, the data augmentation method requires having those instances in the training set, which can be thousands of data samples.

In light of the above discussion, we define the enriched space $\widehat{\mathbb{V}}_E^{(k)}$ as follows:

$$\widehat{\mathbb{V}}_E^{(k)} = \text{span}\left(\left\{\widehat{s}_1^{(k)}, \widehat{s}_2^{(k)}, \cdots\right\} \cup \mathbb{U}_T\right) \tag{2.8}$$

where $\mathbb{U}_T = \{u_1(t, \theta), u_2(t, \theta)\}$, with $u_1(t, \theta) = \cos\theta$ and $u_2(t, \theta) = \sin\theta$. We remark that although the R-CDT transform (1.6) is introduced in a continuous setting, numerical approximations for both the Radon and CDT transforms are available for discrete data, i.e., images in our applications [71]. Here we utilize the computational algorithm described in [79] to estimate the CDT from observed, discrete data. Using this algorithm, and given an image s, $\widehat{s}$ is computed on a chosen grid $[t_1, ..., t_m] \times [\theta_1, ..., \theta_n]$ and reshaped as a vector in $\mathbb{R}^{mn}$.[2] Also the elements in $\mathbb{U}_T$ were computed on the above grid and reshaped to obtain a set of vectors in $\mathbb{R}^{mn}$.

Finally, the proposed training algorithm includes the following steps: for each class k

1. Transform training samples to obtain $\left\{\widehat{s}_1^{(k)}, \widehat{s}_2^{(k)}, \cdots\right\}$

[2]The same grid is chosen for all images. m, n are positive integers .

2. Orthogonalize $\left\{\widehat{s}_1^{(k)}, \widehat{s}_2^{(k)}, \cdots\right\} \cup \mathbb{U}_T$ to obtain the set of basis vectors $\left\{b_1^{(k)}, b_2^{(k)}, \cdots\right\}$, which spans the space $\widehat{\mathbb{V}}_E^{(k)}$ (see equation (2.19)). Use the output of orthogonalization procedure to define the matrix $B^{(k)}$ that contains the basis vectors in its columns as follows:

$$B^{(k)} = \left[b_1^{(k)}, b_2^{(k)}, \cdots\right]$$

The training algorithm described above is summarized in Fig. 2.3.

2.2.2 Testing algorithm

The testing procedure consists of applying the R-CDT transform followed by a nearest subspace search in the R-CDT space. Let us consider a testing image s whose class is to be predicted by the classification model described above. As a first step, we apply R-CDT on s to obtain the transform space representation $\widehat{s}$. We then estimate the distance between $\widehat{s}$ and the subspace model for each class by $d^2(\widehat{s}, \widehat{\mathbb{V}}_E^{(k)}) \sim \|\widehat{s} - B^{(k)}B^{(k)^T}\widehat{s}\|^2$. Note that $B^{(k)}B^{(k)^T}$ is an orthogonal projection matrix onto the space generated by the span of the columns of $B^{(k)}$ (which form an orthogonal basis). To obtain this distance, we must first obtain the projection of $\widehat{s}$ onto the nearest point in the subspace $\widehat{\mathbb{V}}_E^{(k)}$, which can be easily computed by utilizing the orthogonal basis $\left\{b_1^{(k)}, b_2^{(k)}, \cdots\right\}$ obtained in the training algorithm. Although the pseudo-inverse formula could be used, it is advantageous in testing to utilize an orthogonal basis for the subspace instead. The class of $\widehat{s}$ is then estimated to be

$$\arg\min_k \|\widehat{s} - A^{(k)}\widehat{s}\|^2.$$

where, $A^{(k)} = B^{(k)}B^{(k)^T}$. Fig. 2.3 shows a system diagram outlining these steps.

2.3 Computational experiments

2.3.1 Experimental setup

Our goal is to study the classification performance of the method outlined above with respect to state of the art techniques (deep CNN's), and in terms of metrics such as classification accuracy, computational complexity, and amount of training data needed. Specifically, for each dataset we study, we generated train-test splits of different sizes from the original training set, trained the models on these splits, and reported the performances on the original test set. For a train split of a particular size, its samples were randomly drawn (without replacement) from the original training set, and the experiments for this particular size were repeated 10 times. All algorithms saw the same train-test data samples for each split. Apart from predictive performances, we also measured different models' computational complexity, in terms of total number of floating point operations (FLOPs).

A particularly compelling property of the proposed approach is that the R-CDT subspace model can capture different sizes of deformations (e.g. small translations vs. large translations) without requiring that all such small and large deformations be present in the training set. In other words, our model generalizes to data distributions that were previously unobserved. This is a highly desirable property particularly for applications such as the optical communication under turbulence problem described below, where training data encompassing the full range of possible deformations are limited. This property will be explored in section 2.4.

Given their excellent performance in many classification tasks, we utilized different kinds of neural network methods as a baseline for assessing the relative performance of the method outlined above. Specifically, we tested three neural network models: 1) a shallow CNN model consisting of two convolutional layers and two fully connected layers (based on PyTorch's official MNIST demonstration example), 2) the standard VGG11 model [37], and 3) the standard Resnet18 model [38]. All these models were trained for 50 epochs, using the Adam [85] optimizer with learning rate of 0.0005. When the training set size was less than or equal to 8, a validation set

was not used, and the test performance was measured using the model after the last epoch. When the training set had more than 8 samples we used 10% of the training samples for validation, and reported the test performance based on the model that had the best validation performance. In addition to the deep learning-based methods, we also implemented several other linear and nonlinear classification methods: kNN, linear SVM, kernel SVM (with RBF kernel), and nearest subspace classifiers [86] used on the HOG [87], SIFT [88], wavelet [89], and raw image features (See Appendix 2.12.3). To make a fair comparison, we used the same number of training images for all methods.

The proposed method was trained and tested using the methods explained in section 2.2. The orthogonalization of $\widehat{\mathbb{V}}_E^{(k)}$ was performed using singular value decomposition (SVD). The matrix of basis vectors $B^{(k)}$ was constructed using the left singular vectors obtained by the SVD of $\widehat{\mathbb{V}}_E^{(k)}$. The number of the basis vectors was chosen in such a way that the sum of variances explained by all the selected basis vectors in the k-th class captures 99% of the total variance explained by all the training samples in the k-th class. A 2D uniform probability density function was used as the reference image for R-CDT computation (see equation (1.6)).

2.3.2 Datasets

To demonstrate the comparative performance of the proposed method, we identified seven datasets for image classification: Chinese printed characters, MNIST, Affine-MNIST, optical OAM, sign language, OASIS Brain MRI, and CIFAR10 image datasets. The Chinese printed character dataset with 1000 classes was created by adding random translations and scalings to the images of 1000 printed Chinese characters. The MNIST dataset contains images of ten classes of handwritten digits which was collected from [8]. The Affine-MNIST dataset was created by adding random translations and scalings to the images of the MNIST dataset. The optical orbital angular momentum (OAM) communication dataset was collected from [76]. The dataset contains images of 32 classes of multiplexed oribital angular momentum beam patterns for optical communication which were corrupted by atmospheric turbulence. The sign language dataset was collected from

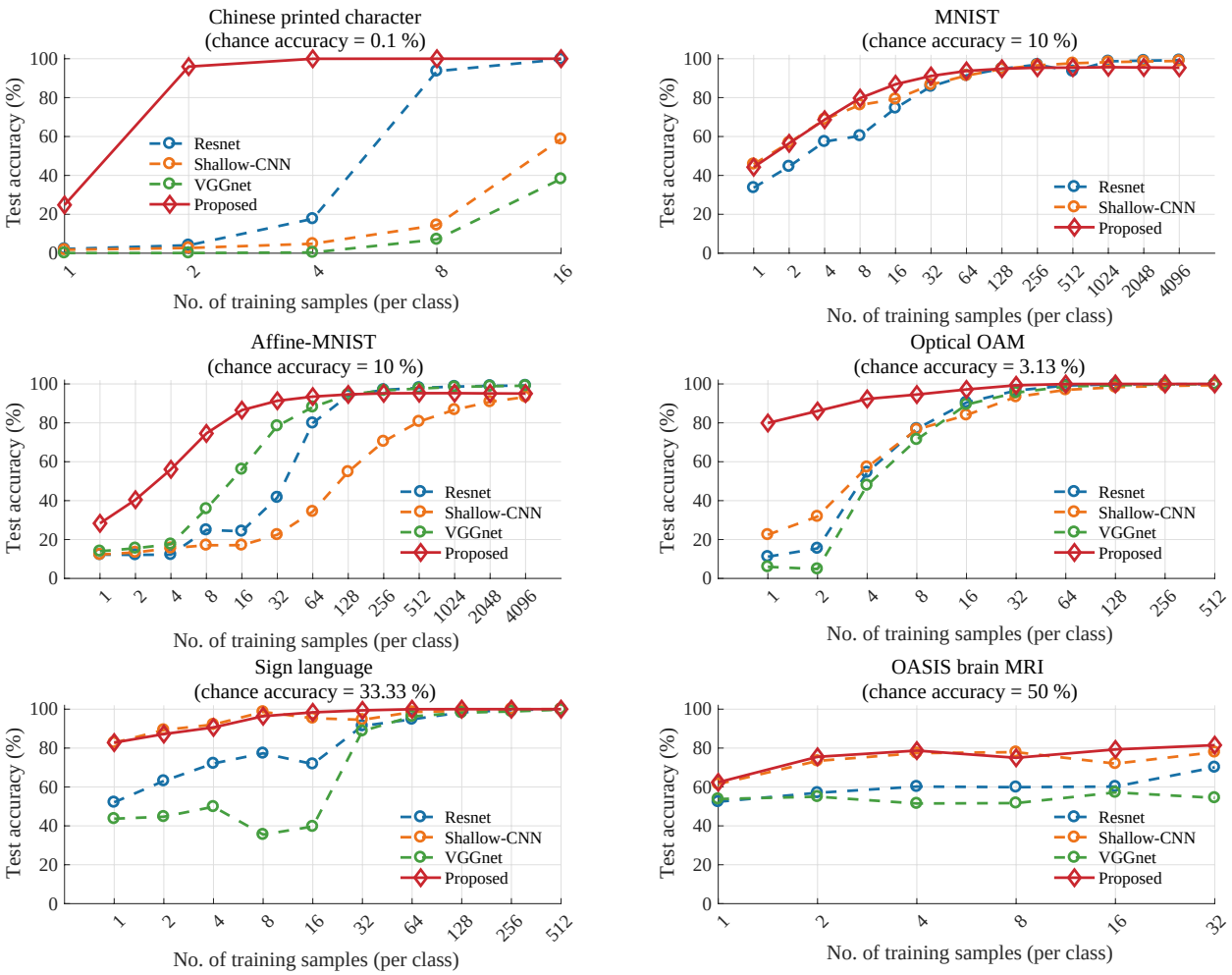

Figure 2.5: Percentage test accuracy of different methods as a function of the number of training images per class.

[90] which contains images of hand gestures. Normalized HOGgles images [91] of first three classes of the original RGB hand gesture images were used. Finally, the OASIS brain MRI image dataset was collected from [92]. The 2D images from the middle slices of the the original 3D MRI data were used in this dissertation. Besides these six datasets, we also demonstrated the results on the natural images of the gray-scale CIFAR10 dataset [93].

2.4 Results

In this section, we compare our method with the deep learning-based methods. The comparison with other linear and nonlinear classification methods is provided in Appendix 2.12.3.

2.4.1 Test accuracy

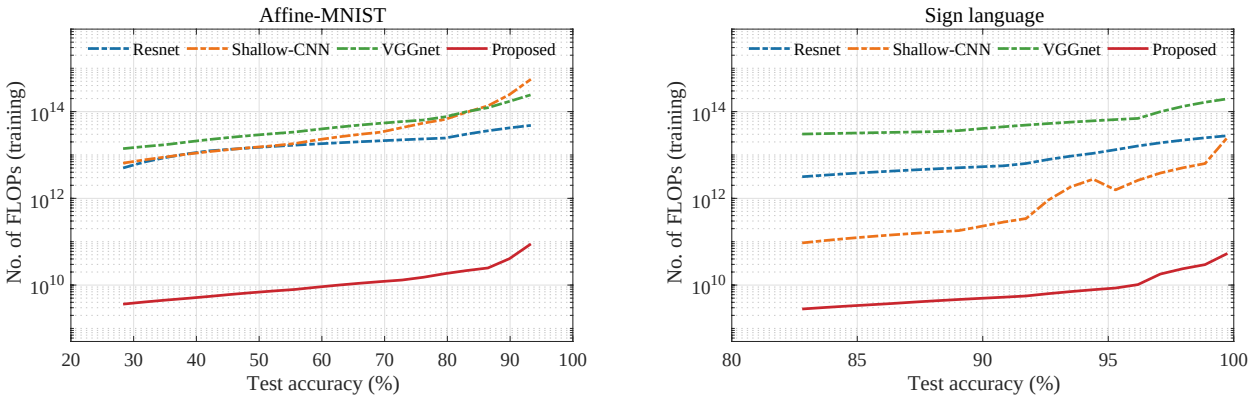

Figure 2.6: The total number of floating point operations (FLOPs) required by the methods to attain a particular test accuracy in the MNIST dataset (left) and the sign language dataset (right).

The average test accuracy values of the methods tested on Chinese printed character, MNIST, Affine-MNIST, optical OAM, sign language, and OASIS brain MRI image datasets for different number of training samples per class are shown in Fig. 2.5. Note that we did not use VGG11 in the MNIST dataset because the dimensions of MNIST images (28×28) are too small for VGG11.

Overall, the proposed method outperforms other methods when the number of training images per class is low (see Fig. 2.5). For some datasets, the improvements are strikingly significant. For example, in the optical OAM dataset, and for learning from only one sample per class, our method provides an absolute improvement in test accuracy of $\sim$ 60% over the CNN-based techniques. Also, the proposed method offers comparable performance to its deep learning counterparts when increasing the number of training samples.

Furthermore, in most cases, the accuracy vs. training size curves have a smoother trend in the proposed method as compared with that of CNN-based learning. The standard deviation of test accuracy of the proposed method is also lower than the other methods in most of the cases (see Appendix 2.12.2). Moreover, the accuracy vs. training curves of the neural network architectures significantly vary as a function of the choice of the dataset. For example, Shallow-CNN outperforms Resnet in MNIST dataset while it underperforms Resnet in Affine-MNIST dataset in terms of test accuracy. Again, while outperforming VGG11 in the sign language dataset, the Resnet

architecture underperforms VGG11 in the Affine-MNIST dataset.

2.4.2 Computational efficiency

Fig. 2.6 presents the number of floating point operations (FLOPs) required in the training phase of the classification models in order to achieve a particular test accuracy value. We used the Affine-MNIST and the sign language datasets in this experiment.

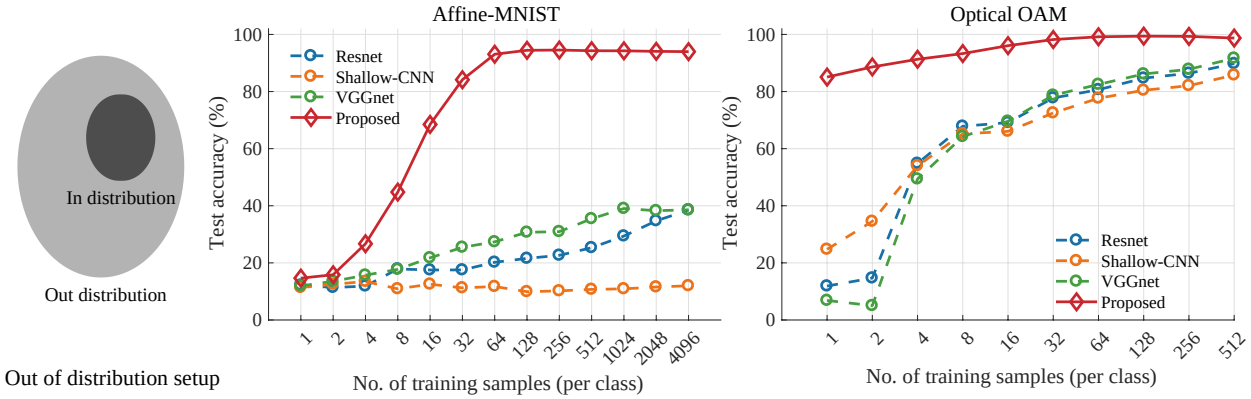

Figure 2.7: Computational experiments under the out-of-distribution setup. The out-of-distribution setup consists of disjoint training ('in distribution') and test ('out distribution') sets containing different sets of magnitudes of the confounding factors (see the left panel). Percentage test accuracy of different methods are measured as a function of the number of training images per class under the out-of-distribution setup (see the middle and the right panel).

The proposed method obtains test accuracy results similar to that of the CNN-based methods with $\sim$ 50 to $\sim$ 10,000 times savings in computational complexity, as measured by the number of FLOPs (see Fig. 2.6). The reduction of the computational complexity is generally larger when compared with a deep neural network, e.g., VGG11. The number of FLOPs required by VGG11 is $\sim$ 3,000 to $\sim$ 10,000 times higher than that required by the proposed method, whereas Shallow-CNN is $\sim$ 50 to $\sim$ 6,000 times more computationally expensive than the proposed method in terms of number of FLOPs. Note that, we have included the training FLOPs only in Fig. 2.6. We also calculated the number of FLOPs required in the testing phase. For all the methods, the number of test FLOPs per image is approximately 5 orders of magnitude ($\sim 10^5$) lower than the number of training FLOPs. The testing FLOPs of the proposed method depend on the number of training samples. Despite this fact, the number of test FLOPs required by the CNN-based methods in our

experiments is $\sim$ 5 to $\sim$ 100 times more than the maximum number of test FLOPs required by the proposed method. These plots are not shown for brevity.

2.4.3 Out-of-distribution testing

In this experiment, we varied the magnitude of the confounding factors (e.g., translation) to generate a gap between training and testing distributions that allows us to test the out-of-distribution performance of the methods. Formally, let $\mathcal{G} \subset \mathcal{T}$ define the set of confounding factors. Let us consider two disjoint subsets of $\mathcal{G}$, denoted as $\mathcal{G}_{in}$ and $\mathcal{G}_{out}$, such that $\mathcal{G}_{in} \subset \mathcal{G}$ and $\mathcal{G}_{out} = \mathcal{G} \backslash \mathcal{G}_{in}$. Using the generative model in equation (2.2) the 'in distribution' image subset $\mathbb{S}_{in}^{(k)}$ and the 'out distribution' image subset $\mathbb{S}_{out}^{(k)}$ are defined using the two disjoint confound subsets $\mathcal{G}_{in}$ and $\mathcal{G}_{out}$ as follows:

$$\mathbb{S}_{in}^{(k)} = \left\{ s_j^{(k)} | s_j^{(k)} = \mathscr{R}^{-1} \left(\left(g_j^{\theta} \right)' \widetilde{\varphi}^{(k)} \circ g_j^{\theta} \right), \forall g_j^{\theta} \in \mathcal{G}_{in} \right\}$$
$$\mathbb{S}_{out}^{(k)} = \left\{ s_j^{(k)} | s_j^{(k)} = \mathscr{R}^{-1} \left(\left(g_j^{\theta} \right)' \widetilde{\varphi}^{(k)} \circ g_j^{\theta} \right), \forall g_j^{\theta} \in \mathcal{G}_{out} \right\}$$

We defined the 'in distribution' image subset $\mathbb{S}_{in}^{(k)}$ as the generative model for the training set and the 'out distribution' image subset $\mathbb{S}_{out}^{(k)}$ as the generative model for the test set in this modified experimental setup (see the left panel of Fig. 2.7).

We measured the accuracy of the methods on the Affine-MNIST and the optical OAM datasets under the modified experimental setup. The Affine-MNIST dataset for the modified setup was generated by applying random translations and scalings to the original MNIST images in a controlled way so that the confound subsets $\mathcal{G}_{in}$ and $\mathcal{G}_{out}$ do not overlap. The 'in distribution' image subset $\mathbb{S}_{in}^{(k)}$ consisted of images with translations by not more than 7 pixels and scale factors varying between $0.9 \sim 1.2$. On the other hand, images with translations by more than 7 pixels and scale factors varying between $1.5 \sim 2.0$ were used to generate the 'out distribution' image subset $\mathbb{S}_{out}^{(k)}$. For the optical OAM dataset, the images at turbulence level 5 (low turbulence) [76] were included in the 'in distribution' subset $\mathbb{S}_{in}^{(k)}$ and those at turbulence level 10 and 15 (medium and high turbulence) were included in the 'out distribution' subset $\mathbb{S}_{out}^{(k)}$. The average test accuracy results for

different training set sizes under the out-of-distribution setup are shown in Fig. 2.7.

The proposed method outperforms the other methods by a greater margin than before under this modified experimental scenario (see Fig. 2.7). For the Affine-MNIST dataset, the test accuracy values of the proposed method are $\sim$ 2 to $\sim$ 85% higher than that of the CNN-based methods. For the optical OAM dataset, the accuracy values of the proposed method are $\sim$ 7 to $\sim$ 85% higher than those of the CNN-based methods (see Fig. 2.7).

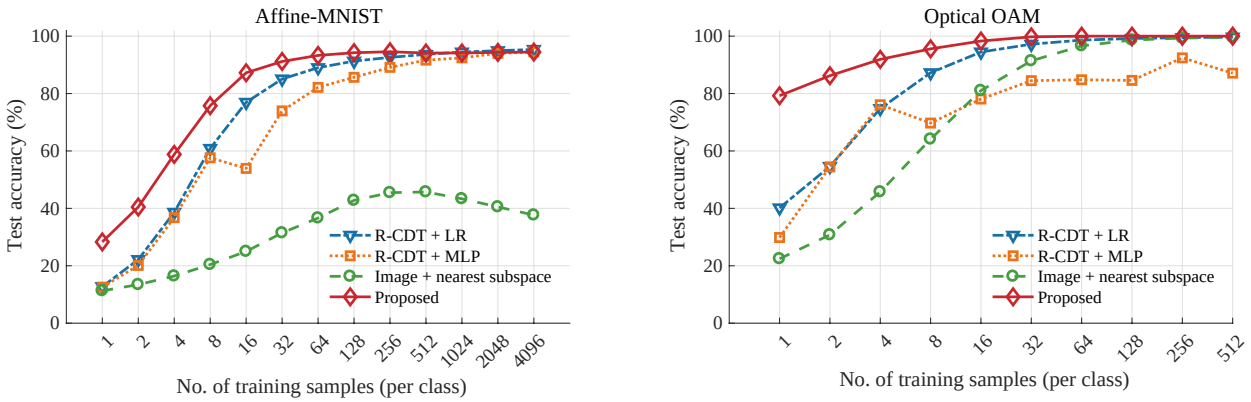

Figure 2.8: Comparison of the percentage test accuracy results obtained in the three ablation studies conducted (using the MLP-based and LR classifiers in the R-CDT space and the nearest subspace classifier in image space) with that of the proposed method.

As compared with the general experimental setup (Fig. 2.5), the test accuracy results of all the methods mostly reduce under this challenging modified experimental setup (Fig. 2.7). The average reduction of test accuracy of the proposed method under the modified setup is also significantly lower than that of the CNN-based methods. For the Affine-MNIST dataset, the average reduction of test accuracy for the proposed method is $\sim$ 10%. Whereas, the reduction of test accuracy for the CNN-based methods are $\sim$ 36% – 42%. Similarly, for the optical OAM dataset, the average reduction of accuracy are $\sim$ 0% and $\sim$ 9% – 12% for the proposed method and the CNN-based methods, respectively.

2.4.4 Ablation study

To observe the relative impact of different components of our proposed method, we conducted three ablation studies using the Affine-MNIST and the optical OAM datasets. In the first two

studies, we replaced the nearest subspace-based classifier used in our proposed method with a multilayer perceptron (MLP) [94] and a logistic regression (LR) classifier [95], respectively, and measured the test accuracy of these modified models. In the third study, we replaced the R-CDT transform representations with the raw images. We measured the test accuracy of the nearest subspace classifier used with the raw image data. The percentage test accuracy results obtained in these modified experiments are illustrated in Fig. 2.8 along with the results of the proposed method for comparison. The proposed method outperforms all these modified models in terms of test accuracy (see Fig. 2.8).

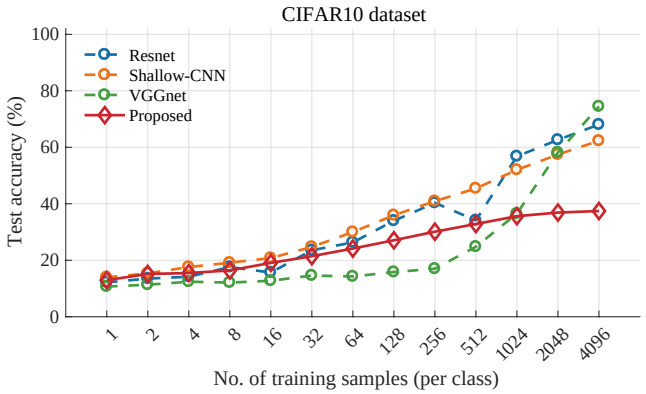

Figure 2.9: Percentage test accuracy results in the CIFAR10 dataset. The natural images in the CIFAR10 dataset might not conform to the underlying generative model, and therefore, the proposed method doesnot perform well in the CIFAR10 dataset.

2.4.5 An example where the proposed method fails

There are examples of image classification problems (e.g. natural image dataset) where the proposed method does not perform well. One such example of this kind of dataset is CIFAR10 dataset. To demonstrate this point, we measured the test accuracies of different methods on the gray-scale CIFAR10 dataset (see Fig. 2.9). It can be seen that, the highest accuracy of the proposed method is lower than the CNN-based methods. All of the CNN-based methods used outperform the proposed method in the gray-scale CIFAR10 dataset in terms of maximum test accuracy.

2.5 Discussion

The results show that the proposed method offers a label efficient and computationally efficient solution to a broad category of image classification problems providing competitive accuracies to state-of-the-art neural networks, as well as other methods. The method also performs well under certain challenging practical scenarios, e.g., the out-of-distribution setting. However, the proposed method does not perform well in certain other problems where data at hand do not conform to the generative model in equation (2.2).

The method is suitable for classification problems where an image class can be viewed as a single "template" image that has been altered by one or more confounding factors to produce the other images in the class. If these alterations can be appropriately modeled as a set of smooth, non-linear transformations (see equation (2.2)), then different image classes become easily separable in the transformed (R-CDT) space via the properties outlined in Lemmas 2.2.1 and 2.2.3. These properties also allow for the approximation of image classes as convex subspaces in the R-CDT space, providing a more appropriate data model for the nearest subspace method. The resulting classifier can then be expected to enjoy the accuracy and computational efficiency we predict.

Test accuracy

Results shown with 6 example datasets suggest the proposed method obtains competitive accuracy figures as compared with state of the art techniques such as CNNs (as well as other methods shown in Appendix 2.12.3) as long as the data at hand conform to the generative model in equation (2.2). Moreover, in these examples, the nearest R-CDT subspace method was shown to be more data efficient: generally speaking, it can achieve higher accuracy with fewer training samples.

Computational efficiency

The proposed method obtains accuracy figures similar to that of the CNN-based methods with ~ 50 to $\sim 10,000$ times reduction of the computational complexity. Such a drastic reduction of

computation can be achieved due to the simplicity and non-iterative nature of the proposed solution. As opposed to the neural networks where GPU implementations are imperative, the proposed method can efficiently be implemented in a CPU and greatly simplify the process of obtaining an accurate classification model for the set of problems that are well modeled by our problem statement defined in definition 2.1.3. We note that the proposed method can also be implemented in parallel using a GPU, where the computation of image projections (and subsequent CDT), the computation of inner products for each projection, etc., can be performed in parallel. This will further enhance the method's efficiency. We also note that one can further balance between accuracy and computational complexity of the proposed method by varying the number of Radon projections. However, the gain in accuracy or computational complexity obtained via this procedure can be much smaller than the improvements we already found in our method.

Out-of-distribution testing

The accuracy results of the CNN-based methods drastically fall under the out-of-distribution setting whereas the proposed method maintains its test accuracy performance. Based on the above findings we infer that the proposed method can be suitable for both interpolation (predicting the classes of data samples within the known distribution) and extrapolation (predicting the classes of data samples outside the known distribution) when the data conforms to the generative model expressed in definition 2.1.2.

The proposed method performs well under the out-of-distribution setting because it does not only learn the data; it actually learns the underlying generative model. Specifically, the method learns the type of deformations that might have generated the dataset by using very few training examples. Then it can detect other magnitudes of that deformation type – including those outside the training distribution. More precisely, the R-CDT subspace represents a mathematical group of deformations (i.e., diffeomorphisms), in the sense that composition of these deformations is automatically captured by the learned subspace. Hence, observing a few deformations (e.g., a few translations) in the training set would lead to a learned subspace that represents all possible

deformations obtained via the compositions of seen deformations (e.g., all possible translations). The CNN-based methods underperform in this setting as their generic formulation does not learn the deformation types.

The out-of-distribution setting for image classification also bears practical significance. For example, consider the problem of classifying the OAM beam patterns for optical communications (see the optical OAM dataset in Fig. 2.5). As these optical patterns traverse air with often unknown air flow patterns, temperature, humidity, etc., exact knowledge of the turbulence level that generated a test image may not always be at hand. Therefore, it is practically infeasible to train the classification model with images at the same turbulence level as the test data. The out-of-distribution setup is more practical under such circumstances.

Ablation study

Based on the ablation study results, we conclude that the proposed method of using the nearest subspace classifier in the R-CDT domain is more appropriate for the category of classification problems we are considering. Data classes in original image domain do not generally form a convex set and therefore and in these instances the subspace model is not appropriate in image domain. The subspace model is appropriate in the R-CDT domain as the R-CDT transform provides a linear data geometry. Considering the subspace model in the R-CDT space also enhances the generative nature of the proposed classification method by implicitly including the data points from the convex combination of the given training data points. Use of a discriminative model for classification (e.g., MLP, LR, etc.) with the R-CDT domain representations of images does not have that advantage.

When are $\mathcal{G}^{-1}$ and $\mathcal{G}^{-1}$ convex?

Given the performance in terms of accuracy and complexity, the R-CDT subspace model presented above seems to be an appropriate model for many applications. However, that is not always the case, as the results with the CIFAR10 dataset show. It is thus natural to ask for what types of

problems will the proposed method work well.

The definitions expressed in Definition 2.1.1 and 2.1.2 define the generative model for the data classes used in our classification problem statement 2.1.3. As part of the solution to the classification problem, it was proved in Lemma 2.2.1 that so long as $\mathcal{G}^{-1}$ or $\mathcal{G}^{-1}$ (the inverse of the transportation subset of functions) is convex, $\widehat{\mathbb{S}}^{(k)}$ is convex, and that is a precondition for the proposed classification algorithm summarized in Fig. 2.3 to solve the classification problem stated in Definition 2.1.3. A natural question to ask is when, or for what types of transportation functions is this condition met? Certain simple examples are easy to describe. For example, when $\mathcal{G}$ or $\mathcal{G}$ denotes the set of translations in 1 or 2D, then $\mathcal{G}^{-1}$ or $\mathcal{G}^{-1}$ can be shown to be convex. Furthermore, when $\mathcal{G}$ or $\mathcal{G}$ refers to the set of scalings of a function, then $\mathcal{G}^{-1}$ or $\mathcal{G}^{-1}$ can be shown to be convex. When $\mathcal{G}$ or $\mathcal{G}$ contains a set of fixed points, i.e. when $g(t_i) = t_i$, then $\mathcal{G}^{-1}$ or $\mathcal{G}^{-1}$ can be shown to be convex. Our hypothesis is that the 6 problems we tested the method on conform to the generative model specifications at least in part, given that classification accuracies significantly higher than chance are obtained with the method. A careful mathematical analysis of these and related questions is the subject of present and future work.

Limitation: An example where the proposed method fails

The fundamental assumption of the proposed method is that the data at hand conform to an underlying generative model (equation (2.2)). If the dataset does not conform to the generative model, the proposed method may not perform well. The CIFAR10 dataset (Fig. 2.9) is an example where the data classes might not follow the generative model. The proposed method underperforms the CNN-based methods in the case of the CIFAR10 dataset.

2.6 Conclusions

We introduced a new algorithm for supervised image classification. The algorithm builds on prior work related to the Radon Cumulative Distribution Transform (R-CDT) [71] and classifies a given image by measuring the the distance between the R-CDT of that image and the linear

subspaces $\widehat{\mathbb{V}}^{(k)}$, $k = 1, 2, \cdots, N_{\text{classes}}$ estimated from the linear combination of the transformed input training data. As distances between two images in the R-CDT space equate to the sliced Wasserstein distances between the inverse R-CDT of the same points, the classification method can be interpreted as a 'nearest' Sliced Wasserstein distance method between the input image and other images in the generative model $\mathbb{S}^{(k)}$ for each class k.

The model was demonstrated to solve a variety of real-world classification problems with accuracy figures similar to or better than the state-of-the-art neural networks (including a shallow method, VGG11 [37], and a Resnet18 [38]; see Fig. 2.5), in both low and high data regimes. The proposed method also outperformed the methods we used other than neural networks (see Appendix 2.12.3). Furthermore, the proposed model was shown to outperform the neural networks by a large margin in some specific practical scenarios, e.g., training with very few training samples and testing with 'out of distribution' test sets. The method is also extremely simple to implement, non-iterative, and it does not require tuning of hyperparameters. Finally, as far as training is concerned the method was also demonstrated to be significantly less demanding in terms of floating point operations relative to different neural network methods.

We remark that six datasets we used to evaluate our method are real image datasets with noise, corruption, and degradation. Theoretical developments on the effect of noise over the CDT were made in a recent study [96]. It was shown that the distribution for the CDT values can be approximated as Gaussian at high SNR (approximately greater than zero) if the input signal is corrupted by additive Gaussian noise [96]. The R-CDT has an additional linear operation (i.e., the Radon transform), and therefore, should also be approximated as Gaussian. In that case, it is possible to show that the nearest subspace, taking into account the noise co-variance, provides the maximum likelihood solution. However, these stochastic developments are beyond the scope of this dissertation.

We note, however, that the method above is best suited for problems that are well modeled by the generative model definition provided in Section 2.1. The definition is naturally tailored towards modeling images which are segmented (foreground extracted). Examples shown here

include classifying written Chinese characters, MNIST numerical digits, optical communication patterns, sign language hand shapes, and brain MRIs. We also note that the model does not account for many other variations present in many important image classification problems. Specifically, the proposed model does not account for occlusions, introduction of other objects in the scene, or variations which cannot be modeled as a mass (intensity) preserving transformation on a set of templates. Computational examples using the CIFAR10 dataset demonstrate that indeed the proposed model lags far behind, in terms of classification accuracy, the standard deep learning classification methods to which it was compared.

Finally, we note that numerous adaptations of the method are possible. We note that the linear subspace method (in the R-CDT space) described above can be modified to utilize other assumptions regarding the set that best models each class. While certain classes of problems may benefit from a simple linear subspace method as described above, where all linear combinations are allowed, other classes may be composed by the union of non-orthogonal subspaces. We also remark that it is possible to extend the proposed method and use it on 3D images with the application of the 3D Radon transform; in fact, this process is generalizable to d-dimensional data, where $d \geq 2$. The generative model can also be adapted for RGB images as the wave equation that gives rise to the model also holds for each channel of an RGB image independently [97]. Furthermore, note that we focus on supervised learning in this dissertation. The method can however be adapted to be used in the context of unsupervised learning also (subspace clustering, for example). The exploration of this and other modifications and extensions of the method are left for future work.

Part B: Utilizing the framework in Part A to develop a method of mathematically learning data invariances without requiring data augmentation for classification problems with limited data.

Deep convolutional neural networks (CNNs) are broadly considered to be state-of-the-art generic end-to-end image classification systems. However, they are known to underperform when training data are limited and thus require data augmentation strategies that render the method computationally expensive and not always effective. Rather than using a data augmentation strategy to encode invariances as typically done in machine learning, here we propose to mathematically augment a nearest subspace classification model in sliced-Wasserstein space by exploiting certain mathematical properties of the Radon Cumulative Distribution Transform (R-CDT), a recently introduced image transform. We demonstrate that for a particular type of learning problem, our mathematical solution has advantages over data augmentation with deep CNNs in terms of classification accuracy and computational complexity, and is particularly effective under a limited training data setting. The method is simple, effective, computationally efficient, non- iterative, and requires no parameters to be tuned.

2.7 Problem statement

One attractive property of the method in part A (R-CDT-NS) is that the class generative models become linear subspaces in R-CDT domain, and can be built from available training data, or mathematical knowledge of image transformations that are known to be present in a particular problem. However, the method in part A does not entirely exploit the possibilities to utilize the mathematical knowledge of image transformations, which were limited to a few simple transformations such as translation and isotropic scaling only. Part of the difficulty with prescribing spatial transformations that go beyond scaling and translation in the R-CDT domain is the fact that the mathematical formulations of these in the R-CDT domain are not known. Here we aim to provide the necessary mathematical approximation of affine transformations in the R-CDT domain, and incorporate them in the R-CDT domain, thus greatly enhancing the performance of the method when few training

samples are available, while maintaining the overall mathematical structure that allows the method to perform well when numerous training examples are available. We aim to utilize the R-CDT-NS framework and thereby propose an improved end-to-end machine learning system for the category of classification problems outlined before. We aim to provide mathematical approximations for the affine set of spatial transformations in the Radon CDT domain, which we then use to extend the method proposed in Part A. Here we begin by reiterating the generative model and problem statement specific to the types of image classification problems under consideration.

Generative model: Let $\mathcal{G} \subset \mathcal{T}_I$ be a set of smooth one-to-one transformations involving non-singular affine transformations up to a certain degree and other possible non-linear deformations. The mass (image intensity) preserving generative model for the k-th image class is defined to be the set

$$\mathbb{S}^{(k)} = \left\{ s_j^{(k)} | s_j^{(k)} = |\text{det} J g_j | \varphi^{(k)} \circ g_j, \forall g_j \in \mathcal{G} \right\}, \tag{2.9}$$

where, $\varphi^{(k)}$ and $s_j^{(k)}$ denote the template and the j-th image, respectively, from the k-th image class and $\text{det} J g_j$ denotes the determinant of the Jacobian matrix of g_j. In our discussion, it is useful to state an equivalent Radon-space definition of the generative model in equation (2.9). Let $\mathcal{G} \subset \mathcal{T}$ be the set of smooth deformations in the Radon space. The equivalent Radon-space generative model for the k-th image class is defined to be the set

$$\mathbb{S}^{(k)} = \left\{ s_j^{(k)} | s_j^{(k)} = \mathscr{R}^{-1} \left(\left(g_j^{\theta} \right)' \tilde{\varphi}^{(k)} \circ g_j^{\theta} \right), \forall g_j^{\theta} \in \mathcal{G} \right\}, \tag{2.10}$$

where, $\tilde{\varphi}^{(k)}$ denotes the Radon transform of the template $\varphi^{(k)}$ and $\mathscr{R}^{-1}(\cdot)$ denotes the inverse Radon transform operator. With the above definition of the data generative model, we formally define the above category of image classification problem as follows:

Classification problem: Let $\mathcal{G} \subset \mathcal{T}$ (or $\mathcal{G} \subset \mathcal{T}_I$) be the set of smooth deformations, and let

the set of image classes $\mathbb{S}^{(k)}$ be defined as in equation (2.10) (or equation (2.9)). Given training samples $\{s_1^{(1)}, s_2^{(1)}, \cdots\}$ (class 1), $\{s_1^{(2)}, s_2^{(2)}, \cdots\}$ (class 2), $\cdots$ as training data, determine the class (k) of an unknown image s.

It was demonstrated in Part A and also in [98] that image classes following the generative model in equation (2.10) yield nonconvex data geometry, causing the above classification problem to be difficult to solve and necessitating nonlinear classifiers. The work in [98] utilized the property of Radon Cumulative Distribution Transform (R-CDT) [71] to simplify the classification problem and provide a non-iterative solution. Here, we briefly explain the solution provided in [98] as follows:

The solution in [98] begins by applying the R-CDT on the generative model in equation (2.10). The R-CDT space generative model then becomes

$$\widehat{\mathbb{S}}^{(k)} = \{\hat{s}_j^{(k)} | \hat{s}_j^{(k)} = \left(g_j^{\theta}\right)^{-1} \circ \hat{\varphi}^{(k)}, \forall g_j^{\theta} \in \mathcal{G}\}. \tag{2.11}$$

It was shown that if $\mathcal{G} \subset \mathcal{T}$ is a convex group then $\widehat{\mathbb{S}}^{(k)}$ is convex [98]. Also, if $\mathbb{S}^{(k)} \cap \mathbb{S}^{(p)} = \varnothing$, then $\hat{\mathbb{S}}^{(k)} \cap \hat{\mathbb{S}}^{(p)} = \varnothing$. The method in [98] then proposes a non-iterative training algorithm by estimating subspaces $\widehat{\mathbb{V}}^{(k)}$ where $\widehat{\mathbb{V}}^{(k)}$ denotes the subspace generated by the convex set $\widehat{\mathbb{S}}^{(k)}$ as follows:

$$\hat{\mathbb{V}}^{(k)} = \text{span}\left(\widehat{\mathbb{S}}^{(k)}\right) = \left\{\sum_{j \in J} \alpha_j \hat{s}_j^{(k)} \mid \alpha_j \in \mathbb{R}, J \text{ is finite}\right\}. \tag{2.12}$$

It was also shown in [98] that, under certain assumptions, $\widehat{\mathbb{S}}^{(k)} \cap \hat{\mathbb{V}}^{(p)} = \varnothing$, for $k \neq p$, i.e., the subspaces $\widehat{\mathbb{V}}^{(k)}$ do not overlap with the data class of another class. The method also explained a framework to prescribe invariance with respect to the translation operation by enhancing the subspace $\widehat{\mathbb{V}}^{(k)}$ with a spanning set corresponding to the translation deformation as follows:

$$\hat{\mathbb{V}}_A^{(k)} = \text{span}\left(\widehat{\mathbb{S}}^{(k)} \cup \mathbb{U}_T\right) \tag{2.13}$$

where $\mathbb{U}_T = \{u_1(t,\theta), u_2(t,\theta)\}$, with $u_1(t,\theta) = \cos\theta$ and $u_2(t,\theta) = \sin\theta$, denotes the spanning set corresponding to the translation deformation type and $\widehat{\mathbb{V}}_A^{(k)}$ denotes the enhanced subspace. Finally, the class of an unknown test sample s was determined by solving

$$\arg\min_k d^2(\hat{s}, \hat{\mathbb{V}}_A^{(k)}). \tag{2.14}$$

The method proposed in Part A is characterized by two prime aspects of it. First, once the training examples containing a specific type of deformation are available, the method can learn other instances of that deformation type from data. Secondly, the method can encode invariances with respect to a few specific deformations in the model without explicit data augmentation when the training samples are low in number or do not contain those specific deformations. For more details, refer to [98].

2.8 Proposed solution

Here we expand upon the latter aspect of the method in Part A (or [98]). The mathematically prescribed invariances in [98] were limited to a few simple spatial transformations such as translations and isotropic scalings (scaling by the same magnitudes in both x and y directions of the image). We expand the extent of the method and mathematically encode invariances with respect to a more complicated deformation set: the set of affine deformations (e.g., translation, both isotropic and anisotropic scaling, both horizontal and vertical shear, and rotation). The detailed explanations of the deformation types used to encode invariances and the corresponding methodologies are explained as follows:

2.8.1 Deformation modeling

Translation

Let $g(\mathbf{x}) = \mathbf{x} - \mathbf{x_0}$ be the translation by $\mathbf{x}_0 \in \mathbb{R}^2$ and $s_g(\mathbf{x}) = |\det Jg| s \circ g = s(\mathbf{x} - \mathbf{x_0})$. Note that Jg denotes the Jacobian matrix of g. Following [71] we have that $\hat{s}_g(t,\theta) = \hat{s}(t,\theta) + \mathbf{x}_0^T \xi_\theta$

where $\xi_\theta = [\cos(\theta), \sin(\theta)]^T$. Therefore, as in [98], we define the spanning set for translation as $\mathbb{U}_T = \{u_T^{(1)}(t,\theta), u_T^{(2)}(t,\theta)\}$, where $u_T^{(1)}(t,\theta) = \cos\theta$ and $u_T^{(2)}(t,\theta) = \sin\theta$. The spanning set $\mathbb{U}_T$ (and the other spanning sets described below, defined for other deformation types) is then used to enhance the subspace $\widehat{\mathbb{V}}^{(k)}$ and encode invariance to the corresponding deformation type. The methodology used to obtain the enhanced subspace $\widehat{\mathbb{V}}_A^{(k)}$ is explained in more detail in section 2.8.2.

Isotropic scaling

Let $g(\mathbf{x}) = a\mathbf{x}$ and $s_g(\mathbf{x}) = |Jg|s \circ g = a^2 s(a\mathbf{x})$, which is the normalized dilatation of s by a where $a \in \mathbb{R}_+$. Then according to [71], $\hat{s}_g(t,\theta) = \hat{s}(t,\theta)/a$, i.e. a scalar multiplication. Therefore, as in [98], an additional spanning set is not required here as the subspace containing $\hat{s}(t,\theta)$ naturally contains its scalar multiplication. Therefore, the spanning set for isotropic scaling is defined as $\mathbb{U}_D = \varnothing$.

Anisotropic scaling

Let $g(\mathbf{x}) = \breve{\mathcal{D}}\mathbf{x}$ with $\breve{\mathcal{D}} = \begin{bmatrix} 1/a, & 0 \\ 0, & 1/b \end{bmatrix}$, $a \neq b$, and $s_g(\mathbf{x}) = |Jg|s \circ g = \frac{1}{ab} s(\breve{\mathcal{D}}\mathbf{x})$, which is the normalized anisotropic dilatation of s by a, b where $a, b \in \mathbb{R}_+$. We postulate that $\hat{s}_g(t,\theta)$ can be approximated as $a\hat{s}(t,\theta) + \alpha a \sin^2\theta \hat{s}(t,\theta)$, $b/a = 1+\alpha$, for $a \leq b$, and as $b\hat{s}(t,\theta) + \beta b \cos^2\theta \hat{s}(t,\theta)$, $a/b = 1+\beta$, for $a > b$.

Proof. Consider two functions $s_g(x,y)$ and $s(x,y)$ such that $s_g(x,y) = |Jg|s \circ g = \frac{1}{ab} s(\breve{\mathcal{D}}\mathbf{x}) = \frac{1}{ab} s(x/a, y/b)$, for some $a, b > 0$, $a \neq b$. By direct computation (see Appendix 2.12.4), we have that

$$\hat{s}_g(t,\theta) = \gamma \hat{s}(t,\theta'), \tag{2.15}$$

where $\gamma = \sqrt{a^2\cos^2\theta + b^2\sin^2\theta}$ and $\theta' = \tan^{-1}\left(\frac{b}{a}\tan\theta\right)$. To formulate the approximation for

$\hat{s}_g(t,\theta)$, we need the next two lemmas to estimate γ and $|\theta' - \theta|$.

Lemma 2.8.1. *For $a \leq b$ (i.e., $\alpha \geq 0$), $|\theta' - \theta| \leq \frac{1}{2}(\alpha + \alpha^2)$, and for $a > b$ (i.e., $-1 < \alpha < 0$), $|\theta' - \theta| \leq \frac{1}{2}(|\alpha| + \frac{\alpha^2}{(1+\alpha)^2})$.*

For a proof of Lemma 2.8.1, see Appendix 2.12.4. If $|\theta' - \theta| \leq \epsilon$, where ϵ is a small number [3], we can approximate θ' as θ. Then using equation (2.15), we have that

$$\hat{s}_g(t,\theta) = \gamma \hat{s}(t,\theta') \approx \gamma \hat{s}(t,\theta) \tag{2.16}$$

Lemma 2.8.2. *For $a \leq b$ (i.e., $\alpha \geq 0$), $\gamma = (1+\alpha \sin^2\theta)a + O(\alpha^2)$, and for $a > b$ (i.e., $-1 < \alpha < 0$, $\beta > 0$), $\gamma = b(1 + \beta\cos^2\theta) + O(\beta^2)$.*

For a proof of Lemma 2.8.2, see Appendix 2.12.4. From equation (2.16) and Lemma 2.8.2, $\hat{s}_g(t,\theta) \approx \gamma\hat{s}(t,\theta) = a\hat{s}(t,\theta) + \alpha a \sin^2\theta \hat{s}(t,\theta) + O(\alpha^2)$ for $a \leq b$ and $\hat{s}_g(t,\theta) \approx \gamma\hat{s}(t,\theta) = b\hat{s}(t,\theta) + \beta b \cos^2\theta \hat{s}(t,\theta) + O(\beta^2)$ for $a > b$. In summary, to model anisotropic scalings of s, we use the set $\hat{E} = \{\cos^2\theta\hat{s}, \sin^2\theta\hat{s}\}$ as enrichment to the training subspace in the transform space. Therefore, we define the spanning set for anisotropic scaling as $\mathbb{U}_{\check{D}} = \{u^{(1)}_{\check{D}}(t,\theta), u^{(2)}_{\check{D}}(t,\theta)\}$, where $u^{(1)}_{\check{D}}(t,\theta) = (\cos^2\theta)\hat{s}(t,\theta)$ and $u^{(2)}_{\check{D}}(t,\theta) = (\sin^2\theta)\hat{s}(t,\theta)$.

□

Shear

Let $g_1(\mathbf{x}) = \mathcal{H}_1\mathbf{x}$ with $\mathcal{H}_1 = \begin{bmatrix} 1, & -h \\ 0, & 1 \end{bmatrix}$, $g_2(\mathbf{x}) = \mathcal{H}_2\mathbf{x}$ with $\mathcal{H}_2 = \begin{bmatrix} 1, & 0 \\ -v, & 1 \end{bmatrix}$, and $s_{g_1}(\mathbf{x}) = |Jg_1|s \circ g_1 = s(\mathcal{H}_1\mathbf{x})$, $s_{g_2}(\mathbf{x}) = |Jg_2|s \circ g_2 = s(\mathcal{H}_2\mathbf{x})$, which are the normalized horizontal and vertical shears of s, respectively, by h and v where $h \neq v$; $h, v \in \mathbb{R}$. We postulate that $\hat{s}_{g_1}(t,\theta)$ can

[3] Note that by Lemma 2.8.1, ϵ can be estimated using α. A practical example for the choice of ϵ is provided in Appendix 2.12.4. For classification purposes, a larger value of ϵ (hence α) is allowed. In addition, as the R.H.S. of the inequalities in Lemma 2.8.1 are the upper bounds for $|\theta' - \theta|$, for a fixed α, $|\theta' - \theta|$ might be much smaller than these bounds. It can be one of the reasons why a larger α can be chosen in practice (as in Table 2.2 of parameters, a much larger range of α is used in the dataset, and the classification accuracy is not compromised). We did not derive how large ϵ (hence α) is allowed, but our approximations work well for practical choices of α (see Table 2.2).

be approximated as $\hat{s}(t,\theta) + \frac{1}{2}(h\sin(2\theta) + h^2\cos^2\theta)\hat{s}(t,\theta)$ and $\hat{s}_{g_2}(t,\theta)$ can be approximated as $\hat{s}(t,\theta) + \frac{1}{2}(v\sin(2\theta) + v^2\sin^2\theta)\hat{s}(t,\theta)$.

Proof. Let us first consider the case of horizontal shear with two functions $s_{g_1}(x,y)$ and $s(x,y)$ such that for some $h \in \mathbb{R}$, $s_{g_1}(x,y) = |Jg_1| s \circ g_1 = s(\mathcal{H}_1\mathbf{x}) = s(x-hy, y)$. By direction computation (see Appendix 2.12.5), we have that

$$\hat{s}_{g_1}(t,\theta) = \gamma\hat{s}(t,\theta'), \tag{2.17}$$

where $\gamma = \sqrt{1 + h^2\cos^2\theta + h\sin(2\theta)}$ and $\theta' = \tan^{-1}(\tan\theta + h)$. To formulate the approximation for $\hat{s}_{g_1}(t,\theta)$, we need the next two lemmas to estimate γ and $\theta' - \theta$.

Lemma 2.8.3. *For $h \geq 0$, $|\theta' - \theta| \leq h + \frac{1}{2}h^2$, and for $h < 0$, $|\theta' - \theta| \leq |h| + h^2$.*

For a proof of Lemma 2.8.3, see Appendix 2.12.5. If $|\theta' - \theta| \leq \epsilon$, where ϵ is a small number [4], we can approximate θ' as θ. Then using equation (2.17), we have

$$\hat{s}_{g_1}(t,\theta) = \gamma\hat{s}(t,\theta') \approx \gamma\hat{s}(t,\theta). \tag{2.18}$$

Lemma 2.8.4. $|\gamma - 1| \leq \frac{1}{2}(h + h^2) + \frac{1}{8}(h + h^2)^2$.

For a proof of Lemma 2.8.4, see Appendix 2.12.5. From equation (2.18) and Lemma 2.8.4, $\hat{s}_{g_1}(t,\theta) \approx \gamma\hat{s}(t,\theta) = \hat{s}(t,\theta) + \frac{1}{2}(h\sin(2\theta) + h^2\cos^2\theta)\hat{s}(t,\theta) + O(h^2)$. In summary, to model small horizontal shearing of s, we add the following additional spanning set $\hat{E} = \{\big(h^2\cos^2\theta + h\sin(2\theta)\big)\hat{s}\}$ (for small h) as enrichment to the training subspace in the transform space. Similarly, to model small vertical shearing of s, we add the following additional spanning set $\hat{E} =$

[4] Note that by Lemma 2.8.3, ϵ can be estimated using h. A practical example for the choice of ϵ is provided in Appendix 2.12.5. For classification purposes, a larger value of ϵ (hence h) is allowed. In addition, as the R.H.S. of the inequalities in Lemma 2.8.3 are the upper bounds for $|\theta' - \theta|$, for a fixed h, $|\theta' - \theta|$ might be much smaller than these bounds. It can be one of the reasons why a larger h can be chosen in practice (as in Table 2.2 of parameters, a much larger range of h is used in the dataset, and the classification accuracy is not compromised). We did not derive how large ϵ (hence h) is allowed, but our approximations work well for practical choices of h (see Table 2.2).

$\{\left(v^2 \sin^2\theta + v\sin(2\theta)\right)\hat{s}\}$ (for small v) as enrichment to the training subspace in the transform space. The proof for the vertical shear is available in Appendix 2.12.5. Therefore, the spanning set for shear is defined as $\mathbb{U}_H = \{u_H^{(1)}(t,\theta), u_H^{(2)}(t,\theta)\}$, where $u_H^{(1)}(t,\theta) = (v^2\sin^2\theta + v\sin 2\theta)\hat{s}(t,\theta)$ and $u_H^{(2)}(t,\theta) = (h^2\cos^2\theta + h\sin 2\theta)\hat{s}(t,\theta)$. □

Rotation

Let $g(\mathbf{x}) = \mathcal{R}\mathbf{x}$ with $\mathcal{R} = \begin{bmatrix} \cos\theta_0, & -\sin\theta_0 \\ \sin\theta_0, & \cos\theta_0 \end{bmatrix}$, and $s_g(\mathbf{x}) = |Jg| s \circ g = s(\mathcal{R}\mathbf{x})$, which is the normalized rotation of s by θ_0 where $\theta_0 \in [0,\pi]$. Following [71] we have that $\hat{s}_g(t,\theta) = \hat{s}(t,\theta-\theta_0)$, i.e., rotation in image space results in a circular translation in angle θ in the R-CDT space, whereas our previous discussion pertains to a fixed θ. Here, we encode rotation invariance in an alternate manner, in the testing phase of the method. If data contains a rotation confound and if a test image s belongs to the class (k) then $d^2\left(\mathbb{P}_\alpha\hat{s}, \hat{\mathbb{V}}^{(k)}\right) = 0$ and $d^2\left(\mathbb{P}_\alpha\hat{s}, \hat{\mathbb{V}}^{(l)}\right) > 0; \quad k \neq l$ where, $\mathbb{P}_\alpha$ is a fixed permutation matrix that causes circular translation in θ by $\alpha \in [0,\pi]$ (which eventually causes rotation in the native image space). Therefore, the class of s can be decoded by solving $\arg\min_k \min_\alpha d^2\left(\mathbb{P}_\alpha\hat{s}, \hat{\mathbb{V}}^{(k)}\right)$.

2.8.2 Training algorithm

Using the principles laid out above, the algorithm we propose estimates the enhanced subspace $\widehat{\mathbb{V}}_A^{(k)}$ corresponding to the transform space $\widehat{\mathbb{S}}^{(k)}$ given sample data $\{s_1^{(k)}, s_2^{(k)}, \cdots\}$. Naturally, the first step is to transform the training data to obtain $\{\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots\}$. We then approximate $\widehat{\mathbb{V}}_A^{(k)}$ as follows:

$$\widehat{\mathbb{V}}_A^{(k)} = \text{span}\left(\left\{\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots\right\} \cup \mathbb{U}_A\right) \tag{2.19}$$

where $\mathbb{U}_A = \mathbb{U}_T \cup \mathbb{U}_D \cup \mathbb{U}_{\check{D}} \cup \mathbb{U}_H$ denotes the combined spanning set corresponding to the translation, isotropic/anisotropic scaling, and horizontal/vertical shear deformation types; the spanning set $\mathbb{U}_A$ is used in equation (2.19) to obtain invariances to these deformation types. As mentioned

Table 2.1: Training algorithm

Algorithm: Training procedure of the proposed method
Input: Training images $\{s_1^{(k)}, s_2^{(k)}, \cdots\}$.
Output: The matrix of basis vectors $B^{(k)}$.
for each class k:
– Transform training data to obtain $\left\{\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots\right\}$.
– Orthogonalize $\left\{\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots\right\} \cup \mathbb{U}_A$ to obtain the set of basis vectors $\left\{b_1^{(k)}, b_2^{(k)}, \cdots\right\}$, which spans the space $\widehat{\mathbb{V}}_A^{(k)}$ (see equation (2.19)).
– Use the output of orthogonalization procedure to define the matrix $B^{(k)}$ containing the basis vectors in its columns as follows: $B^{(k)} = \left[b_1^{(k)}, b_2^{(k)}, \cdots\right]$.

before, the invariance to the rotation deformation is obtained alternately in the testing phase of the method (see section 2.8.3 for details). Here, $\mathbb{U}_T = \{u_T^{(1)}(t,\theta), u_T^{(2)}(t,\theta)\}$, with $u_T^{(1)}(t,\theta) = \cos\theta$ and $u_T^{(2)}(t,\theta) = \sin\theta$, corresponds to the translation deformation type; $\mathbb{U}_D = \varnothing$ corresponds to isotropic scaling; $\mathbb{U}_{\breve{D}} = \{u_{\breve{D}}^{(1)}(t,\theta), u_{\breve{D}}^{(2)}(t,\theta)\}$, with $u_{\breve{D}}^{(1)}(t,\theta) = (\cos^2\theta)\left\{\hat{s}_1^{(k)}(t,\theta), \hat{s}_2^{(k)}(t,\theta), \cdots\right\}$ and $u_{\breve{D}}^{(2)}(t,\theta) = (\sin^2\theta)\left\{\hat{s}_1^{(k)}(t,\theta), \hat{s}_2^{(k)}(t,\theta), \cdots\right\}$, corresponds to anisotropic scaling; and $\mathbb{U}_H = \{u_H^{(1)}(t,\theta), u_H^{(2)}(t,\theta)\}$, with $u_H^{(1)}(t,\theta) = (v^2\sin^2\theta + v\sin 2\theta)\left\{\hat{s}_1^{(k)}(t,\theta), \hat{s}_2^{(k)}(t,\theta), \cdots\right\}$ and $u_H^{(2)}(t,\theta) = (h^2\cos^2\theta + h\sin 2\theta)$
$\left\{\hat{s}_1^{(k)}(t,\theta), \hat{s}_2^{(k)}(t,\theta), \cdots\right\}$, corresponds to the vertical and horizontal shear deformations [5].

‘

Next, we orthogonalize $\left\{\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots\right\} \cup \mathbb{U}_A$ to obtain the set of basis vectors $\left\{b_1^{(k)}, b_2^{(k)}, \cdots\right\}$, which spans the space $\widehat{\mathbb{V}}_A^{(k)}$. We then use the basis vectors to define the following matrix:

$$B^{(k)} = \left[b_1^{(k)}, b_2^{(k)}, \cdots\right] \tag{2.20}$$

The proposed training algorithm is outlined as in table 2.1. Fig. 2.10 shows a system diagram

[5]Note that the spanning set $\left\{(v^2\sin^2\theta + v\sin 2\theta)\hat{s}\right\}_{|v|\le\epsilon}$ (for vertical shear, see equation (2.19)) in transform domain for some ϵ is not linear in the sense that it is not in the span$\left\{(v_0^2\sin^2\theta + v_0\sin 2\theta)\hat{s}\right\}$ for some fixed small v_0. The situation for horizontal shear is similar. However, it is not hard to show that $\mathbb{U}_H \in \text{span}\,(\mathbb{U}_A)$ for all possible $v, h \in \mathbb{R}$. Indeed it can be shown that span$\left\{\cos^2\theta\hat{s}, \sin^2\theta\hat{s}, \sin(2\theta)\hat{s}\right\}$ = span$\left\{\cos^2\theta\hat{s}, \sin^2\theta\hat{s}, \{(v^2\sin^2\theta + v\sin 2\theta)\hat{s}\}_{v\in\mathbb{R}}, \{(h^2\cos^2\theta + h\sin 2\theta)\hat{s}\}_{h\in\mathbb{R}}\right\}$ = span$\left\{\cos^2\theta\hat{s}, \sin^2\theta\hat{s}, (v_0^2\sin^2\theta + v_0\sin 2\theta)\hat{s}, (h_0^2\cos^2\theta + h_0\sin 2\theta)\hat{s}\right\}$ for some fixed v_0, h_0. Therefore, in our numerical implementation choosing small fixed v_0, h_0 in for $\mathbb{U}_H$ (and hence for $\mathbb{U}_A$) will not change the subspace $\widehat{\mathbb{V}}_A^{(k)}$.

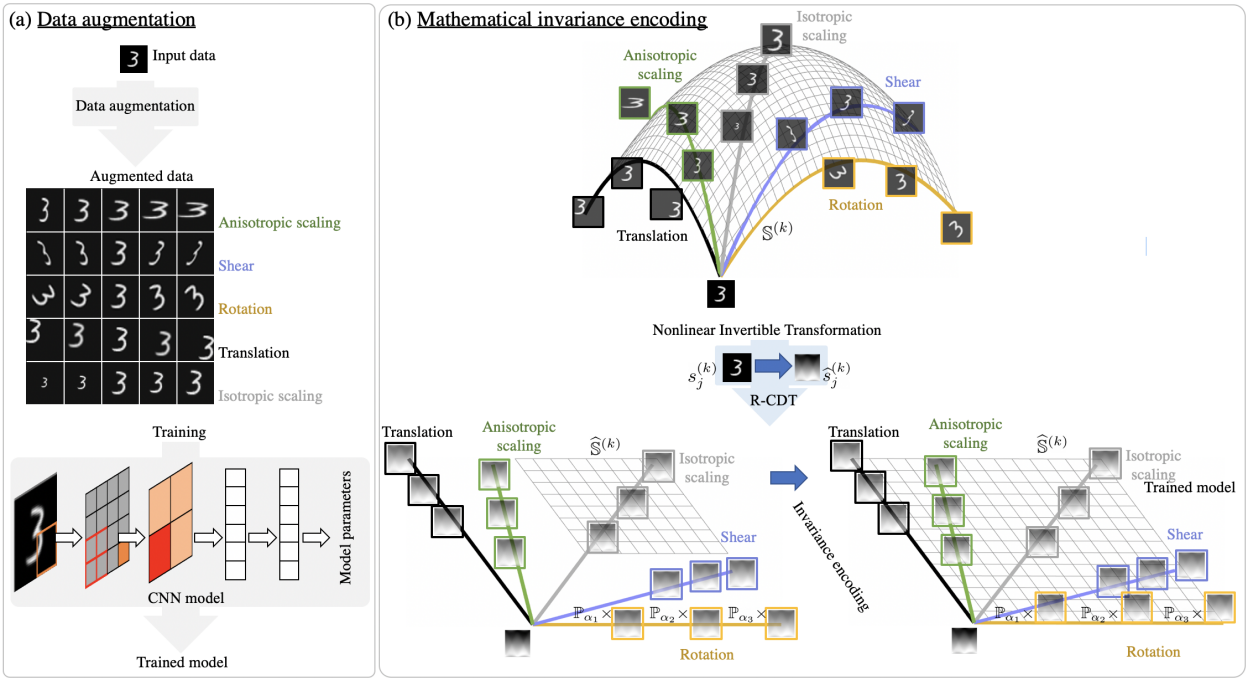

Figure 2.10: System diagrams outlining the data augmentation-based methods and the proposed method. (a) Data augmentation-based methods augment the training set by artificially applying known transformations to the original training set. (b) The proposed invariance encoding method models the underlying data space (represented by grey grid lines) corresponding to known transformations to learn invariances to those transformations. The R-CDT renders data space convex and enables it to be modeled with a linear subspace. The invariance encoding framework expands the subspace to incorporate invariances to desired transformations.

outlining the main computational modeling steps in the proposed method.

2.8.3 Testing algorithm

The testing procedure consists of applying the R-CDT transform followed by a nearest subspace search in the R-CDT space. Let us consider a test image s whose class is to be predicted by the classification method described above. As a first step, we apply the R-CDT on s to obtain the transform space representation $\hat{s}$. We then estimate the distance $d^2(\mathbb{P}_\alpha \hat{s}, \widehat{\mathbb{V}}_A^{(k)}) \sim \|\mathbb{P}_\alpha \hat{s} - B^{(k)} B^{(k)^T} \mathbb{P}_\alpha \hat{s}\|^2$. Here, $\mathbb{P}_\alpha$ is a fixed permutation matrix that causes circular translation in θ by $\alpha \in [0, \pi]$. Note that $B^{(k)} B^{(k)^T}$ is an orthogonal projection matrix onto the space generated by the span of the columns of $B^{(k)}$ (which form an orthogonal basis). The class of $\hat{s}$ is then estimated

to be

$$\arg\min_{k}\min_{\alpha}\|\mathbb{P}_{\alpha}\hat{s}-A^{(k)}\mathbb{P}_{\alpha}\hat{s}\|^{2};\quad \text{where, } A^{(k)}=B^{(k)}B^{(k)^{T}}.$$

2.9 Results

2.9.1 Simulated experiment

We compared the proposed method with conventional classification methods with respect to classification accuracy, data efficiency, computational efficiency, and out-of-distribution robustness. In this respect, we have identified four state-of-the-art methods with data augmentation: MNISTnet [99] (a shallow CNN model based on PyTorch's official example), the standard VGG11 model [37], the standard Resnet18 model [38], and the standard k-nearest neighbors (kNN) classifier model [98]. The CNN-based methods were implemented using Adam optimizer with 50 epochs and a learning rate of 0.0005. Singular value decomposition was used to obtain the basis vectors of the proposed method. The least possible number of basis vectors was chosen so that the sum of variances explained by the selected basis vectors in any class is at least 99% of the total variance in that class, and subspaces corresponding to all classes have the same dimensionality. We generated training splits of different sizes from the original training set, trained the models on these splits, and reported the performances on the original test set. For a training split of a particular data size, its samples were randomly drawn (without replacement) from the original training set. The experiments for a particular size were repeated ten times, and the mean values of the results (i.e., test accuracy, computational complexity, etc.) were reported. The upper and lower bounds of the test accuracy estimates were also reported in terms of mean and standard deviation ($\mu \pm \sigma$) in Appendix 2.12.6. All methods saw the same set of training and test images.

To evaluate the comparative performance of the methods, we selected the following datasets: MNIST, Affine-MNIST, OMNIGLOT, Brain MRI, Sign language, OAM (under the regular and the out-of-distribution setup), and FMNIST datasets. The handwritten character images of the MNIST and OMNIGLOT datasets were selected from [8] and [101], respectively. The 2D images from the

Table 2.2: Parameters for additional deformations added to the data

	Translation (x_0, y_0)	Isotropic scaling a	Anisotropic scaling (a, b)	Shear h	Rotation θ_0
Synthetic	constrained by FoV	$0.5 - 2.0$	$(0.5 - 2.0, 0.5 - 2.0)$	± 0.6	$\pm 40^\circ$
AffNIST (reg), FMNIST, Brain MRI	constrained by FoV	$0.5 - 2.0$	$(0.5 - 2.0, 0.5 - 2.0)$	± 0.3	$\pm 20^\circ$
MNIST, OMNIGLOT, OAM, Sign Lang	$(-, -)$	$-$	$(-, -)$	$-$	$-$
AffNIST (out)	training: $(0, 0)$ testing: $(\pm 20, \pm 20)$	training: 1 testing: $0.5 - 2.0$	training: $(1, 1)$ testing: $(0.5 - 2.0, 0.5 - 2.0)$	training: 0 testing: ± 0.3	training: 0° $\pm 20^\circ$
Official Affine MNIST [100]	constrained by FoV	$0.8 - 1.2$	$(0.8 - 1.2, 0.8 - 1.2)$	± 0.2	$\pm 20^\circ$

middle slices of the 3D MRI data of the Brain MRI dataset and normalized HOGgles images of hand gestures of the Sign language dataset were collected from [98]. The fashion object images (trouser, pullover, bag, ankle boot) of the FMNIST dataset were collected from [102]. The optical communication images (orbital angular momentum beam patterns) of the OAM dataset under the influence of various atmospheric turbulence levels were collected from [103]. We tested the methods on the OAM dataset under two experimental setups: the regular setup, where the training and test sets contain images at the same turbulence level, and the out-of-distribution setup, where the training and test sets contain images at different turbulence levels. We randomly selected 8 images from the OMNIGLOT dataset for training, and the rest were used for testing. The Affine-MNIST dataset was created using random Affine transformations (translation, isotropic/anisotropic scaling, shear, rotation) to both training and test sets of the MNIST dataset. To increase the classification complexity, random affine transformations were also added with the Brain MRI and FMNIST datasets. We also created a synthetic dataset using randomly selected ten classes of the OMNIGLOT dataset [101]. The training set of the synthetic dataset contains a randomly selected 1 image per class, and the test set contains 200 instances of the same image but observed under random affine transformations (translation, isotropic/anisotropic scaling, shear, and rotation). Fig. 2.11(a) illustrates the single image of the training set and a few sample images of the test set of a random class of the synthetic dataset. Parameters for the additional deformations added to the datasets, in addition to their natural deformations, are presented in Table 2.2. Parameters for deformations used in the widely used official Affine MNIST dataset [100] are also included in Table 2.2

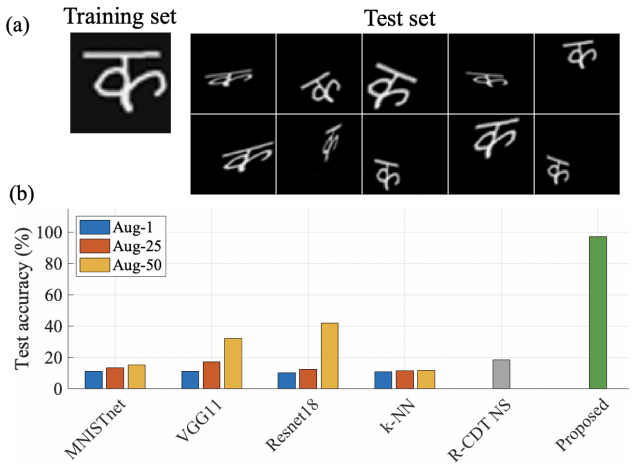

Figure 2.11: The accuracy of the methods on the synthetic dataset. (a) Training and test sets of a random class of the synthetic dataset. (b) The percentage test accuracy of methods. Aug-1, Aug-25, and Aug-50 indicate that the corresponding methods were trained using both original and augmented set where the sizes of the augmented set were 1, 25, and 50 times the size of the original training set, respectively. The R-CDT-NS and the proposed method did not use any augmented images.

for reference. When the translation is constrained by the field of view (FoV), any parameter value is allowed as long as the object stays within the field of view of the image.

2.9.2 Effectiveness and data-efficiency

The proposed method provides an effective and data-efficient solution in the synthetic dataset (see Fig. 2.11(b)). Note that the MNISTnet, VGG11, Resnet18, and kNN models were trained using the original training set in addition to the augmented training set generated from the original training set. The sizes of the augmented training set used were 1, 25, and 50 times the size of the original training set. We also emphasize that the R-CDT-NS method and the proposed method did not use augmented images; they were trained only using the original training set. The proposed method provides test accuracy close to 100% using no augmented data, whereas the other methods reach accuracy up to 40% using 50 times more augmented images (see Fig. 2.11(b)).

The classification accuracy, for different training set sizes, for the real datasets are shown in Fig. 2.12 and Table 2.3. Note that, in the real datasets also, the MNISTnet, VGG11, Resnet18, and

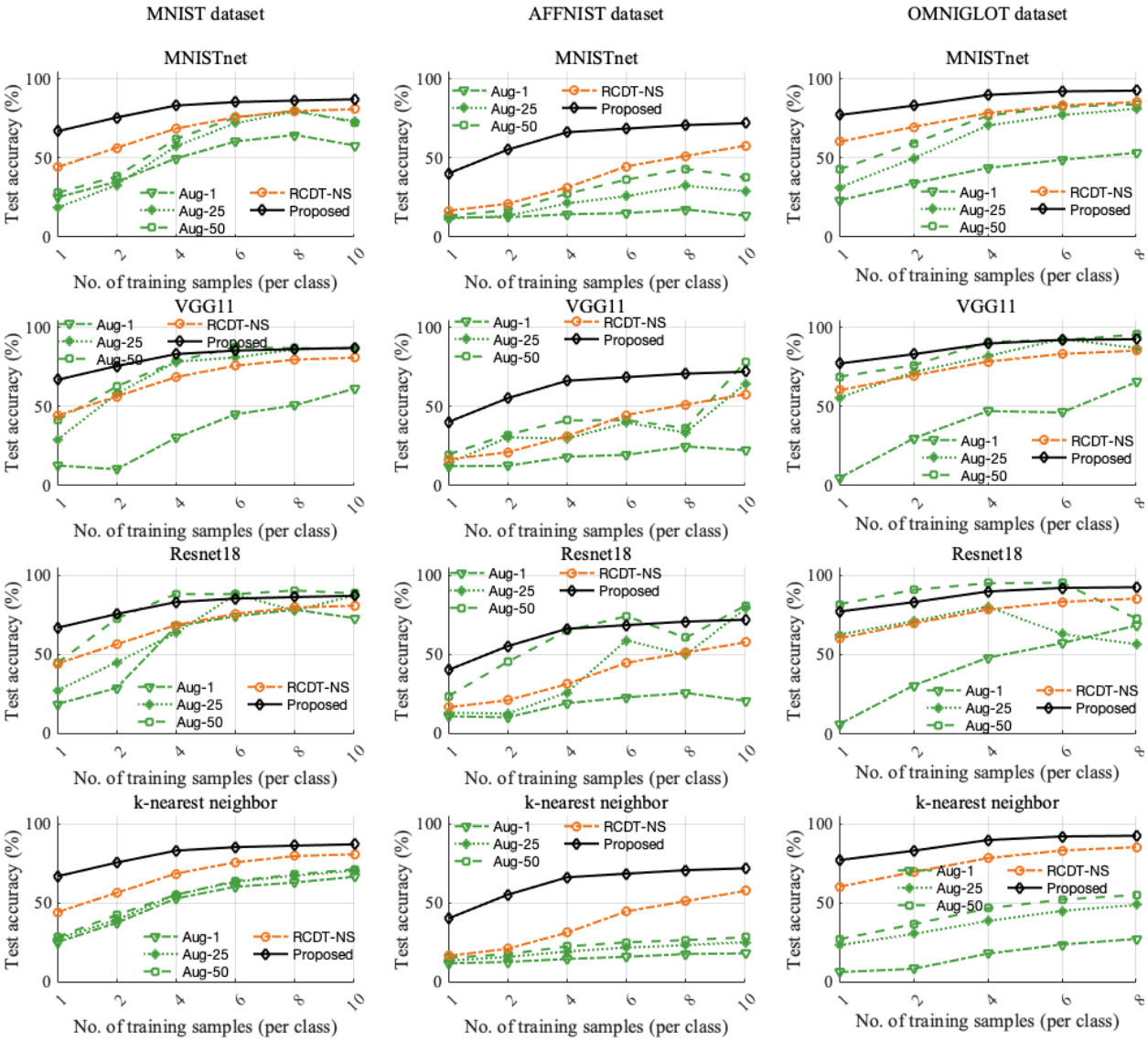

Figure 2.12: The accuracy of the methods as a function of the number of training samples on the MNIST, AFFNIST, and OMNIGLOT datasets. Aug-1, Aug-25, and Aug-50 indicate that the corresponding methods were trained using both original and augmented set where the sizes of the augmented set were 1, 25, and 50 times the size of the original training set, respectively. The R-CDT-NS and the proposed method did not use any augmented images.

kNN models were trained using the original training set in addition to the augmented training set where the sizes of the augmented training set used were 1, 25, and 50 times the size of the original training set. Also, note that the R-CDT-NS method and the proposed method were implemented without using any augmented images. Results in the real datasets show that the proposed method

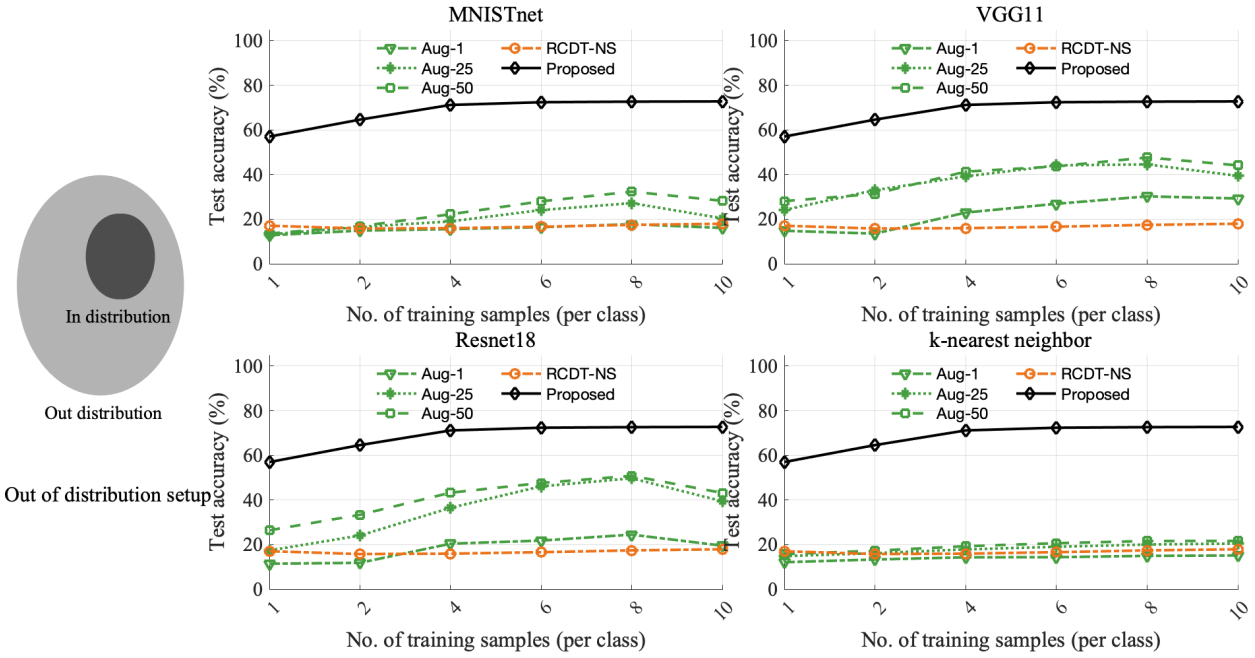

Figure 2.13: Experimental results under the out-of-distribution setup, which is characterized by the disjoint training ('in distribution') and test ('out distribution') sets containing different sets of magnitudes of the spatial transformations (see the left panel). The percentage test accuracy values of different methods are measured as a function of the number of training images per class.

provides accuracy better than or equivalent to the other methods without using any augmented images, where the other methods used up to 50 times more augmented images (see Fig. 2.12). The proposed method outperforms the other methods by a more significant margin at the low-training sample end. Increasing the augmentation size improves the performance of the other methods, but it significantly increases the other methods' computational burden, as will be clarified in the next section.

2.9.3 Out-of-distribution robustness

To compare the effectiveness of the proposed method under the out-of-distribution setting, we generated a gap between the training and test sets with respect to the magnitudes of the deformations. Formally, if $\mathcal{G}_{in} \subset \mathcal{G}$ denotes the deformation set of the 'in-distribution', then $\mathcal{G}_{out} = \mathcal{G} \backslash \mathcal{G}_{in}$ was used as the deformation set for the 'out-distribution' (see Fig. 2.13). Then we trained the models using the 'in-distribution' data and tested using the 'out-distribution' data. We

performed two out-of-distribution experiments. In one experiment, we used the MNIST dataset as the 'in-distribution' training set and the Affine-MNIST dataset as the 'out-distribution' test set (see Fig. 2.13), and in the other out-of-distribution experiment, we used the OAM images at low turbulence levels as the 'in-distribution' training set and those at medium and high turbulence levels as the 'out-distribution' test set as in [98] (see Table 2.3). The range of deformation magnitudes used in the augmentation set was also chosen to be different from the Affine-MNIST dataset in this out-of-distribution experiment. The results show that the proposed method outperforms the other methods by an even more significant margin under the challenging out-of-distribution setup (see Fig. 2.13 and Table 2.3). Under this setup, the proposed method maintains similar accuracy figures in the Affine-MNIST and OAM test data compared with the standard experimental setup (i.e., Affine-MNIST in Fig. 2.12 and OAM (regular) in Table 2.3). On the other hand, the other methods decline in accuracy significantly under the out-of-distribution setup compared with the standard experimental setup (see Affine-MNIST and OAM results in Figs. 2.12, 2.13 and Table 2.3).

2.9.4 Computational efficiency

To compare the computational efficiency of the methods, we computed the number of the floating-point operations (FLOPs) [98] in the training phase of the methods (see the FLOPs vs. percentage accuracy results for the MNIST dataset in Fig. 2.14). The results show that the proposed method requires up to 6 orders of magnitude ($1,000,000$ times) less computational cost than the other methods to achieve the same test accuracy. As in the previous experiments, we augmented the training set for the MNISTnet, VGG11, Resnet18, and kNN methods and did not augment the training set for the R-CDT-NS and the proposed method. The size of the augmentation set used in this experiment was 50 times more than the original training set. The computational complexity of other methods would potentially reduce if the augmentation set size were reduced, but that would also aggravate their classification accuracy (see Fig. 2.12).

Table 2.3: Accuracy of the methods (%) on images with complex foregrounds

	Training set size (per class) = 1					
	MNISTnet (Aug-1 / Aug-50)	VGG11 (Aug-1 / Aug-50)	Resnet18 (Aug-1 / Aug-50)	k-NN (Aug-1 / Aug-50)	R-CDT NS (Aug-0)	Proposed (Aug-0)
Brain MRI	48.70 / 51.30	49.60 / 56.10	49.10 / 54.00	50.50 / 50.70	48.70	**57.20**
Sign Lang	68.97 / 74.29	42.53 / 46.58	40.78 / 57.78	30.85 / 32.65	83.00	**87.35**
OAM (reg)	5.25 / 27.18	3.02 / 16.56	5.77 / 45.98	17.31 / 19.40	80.58	**81.51**
OAM (out)	4.86 / 28.21	3.87 / 15.73	5.28 / 45.85	20.70 / 21.88	**85.55**	85.20
FMNIST	34.02 / 32.68	33.14 / 33.69	26.45 / 34.26	25.90 / 35.52	34.78	**57.90**
	Training set size (per class) = 5					
	MNISTnet (Aug-1 / Aug-50)	VGG11 (Aug-1 / Aug-50)	Resnet18 (Aug-1 / Aug-50)	k-NN (Aug-1 / Aug-50)	R-CDT NS (Aug-0)	Proposed (Aug-0)
Brain MRI	54.80 / 55.40	52.80 / 52.00	50.60 / 56.30	47.20 / 49.40	47.80	**62.20**
Sign Lang	91.05 / 91.59	44.66 / 79.93	47.49 / 91.46	92.01 / 92.40	93.21	**96.10**
OAM (reg)	43.61 / 68.48	29.37 / 26.16	72.62 / 81.84	34.03 / 38.92	92.69	**93.71**
OAM (out)	38.78 / 56.24	25.41 / 41.32	67.70 / 58.67	37.69 / 41.83	91.20	**91.44**
FMNIST	43.97 / 50.92	31.87 / 65.46	31.67 / 65.22	38.69 / 42.95	54.31	**82.64**
	Training set size (per class) = 10					
	MNISTnet (Aug-1 / Aug-50)	VGG11 (Aug-1 / Aug-50)	Resnet18 (Aug-1 / Aug-50)	k-NN (Aug-1 / Aug-50)	R-CDT NS (Aug-0)	Proposed (Aug-0)
	MNISTnet	VGG11	Resnet18	k-NN	R-CDT NS	Proposed
Brain MRI	50.70 / 54.10	50.40 / 57.00	50.60 / 60.00	49.00 / 50.30	52.10	**62.30**
Sign Lang	83.82 / 88.52	43.30 / 75.52	39.61 / 87.87	95.77 / 95.74	97.47	**98.26**
OAM (reg)	54.41 / 79.40	69.38 / 87.85	88.50 / 95.43	45.46 / 51.23	95.82	**97.34**
OAM (out)	44.27 / 60.72	57.28 / 75.06	72.38 / 79.27	46.83 / 51.28	92.76	**94.16**
FMNIST	37.95 / 58.62	42.42 / 74.49	37.45 / 78.38	46.55 / 49.66	75.68	**86.14**

2.10 Discussion

The results above show that our method's mathematically prescribed invariance encoding technique can reasonably model the specific deformation set under consideration, i.e., the affine set. The proposed method offers high classification accuracy using significantly less training data and computation without explicitly using any augmented images. The method is robust under challenging experimental setups such as out-of-distribution testing cases.

Test accuracy and data efficiency

Results in synthetic and real data show that, so long as the data at and conform to the generative model stated in equation (2.10) (or equation (2.9)), the proposed method can classify images with high accuracy without explicitly using any augmented images. While the proposed and RCDT-NS methods were implemented using no augmented images, the other methods we compared to were implemented using low to high numbers of augmented images. The other methods significantly underperform while using a low number of augmented images (see Figs. 2.11(b), 2.12, 2.13, and

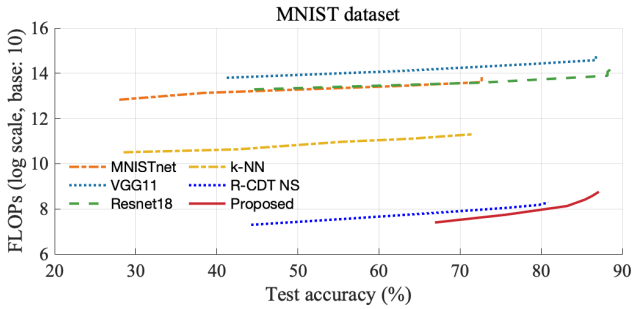

Figure 2.14: The computational complexity of the methods as measured by the total number of floating-point operations (FLOPs) to attain a particular test accuracy in the MNIST dataset.

Table 2.3). The other methods, in some cases, offer equivalent classification accuracy to the proposed method while using a significantly high number of augmented images (see Fig. 2.12 and Table 2.3). However, this approach significantly increases the computational burden of the other methods (see Fig. 2.14). Here we emphasize that, though the proposed method was implemented without using any data augmentation here, data augmentation could also be used in the proposed method as in the other methods. In that case, the accuracy of the proposed method could potentially improve even further. We note that in some datasets in Table 2.3 (e.g., Brain MRI, FMNIST), the classification performances of all methods are relatively lower than that in the other datasets. Classification problems involving these datasets might be more complex than the other ones, causing relatively low classification performances of all the methods in these datasets. However, in these datasets also, the proposed method outperforms the other methods.

Computational efficiency

The proposed method provides up to 6 orders of magnitude (1,000,000 times) savings in computational cost (see Fig. 2.14) as measured by the number of floating-point operations (FLOPs) over data augmented deep learning alternatives. Such an improvement in computational efficiency could be achieved due to the simple and non-iterative nature of the proposed solution. The MNISTnet, VGG11, Resnet18, and kNN methods require a high number of augmented images to achieve reasonable accuracy figures and require iterative optimization procedures to reach a solution, con-

tributing to high computational costs. Computational costs of these methods could be reduced by reducing the number of augmented images used, but that would also adversely affect their classification accuracy (see Fig. 2.12).

Out-of-distribution robustness

The proposed method maintains high classification accuracy while the accuracy figures of the other methods fall drastically under the challenging out-of-distribution experimental setup (see Fig. 2.13 and Table 2.3). These results suggest that the proposed method provides a better generalization of the underlying data distribution resulting in robust classification performance. The reason for better accuracy under the out-of-distribution setup is that the proposed method does not only learn the deformations present in the given data; it actually learns the underlying data model. More specifically, it learns the type of deformation (such as translation, scaling, shear, and others) present in the data. It thereby can detect the presence of different magnitudes of these deformation types. The type of the deformation can be learned from a very few training samples containing those deformations as well as from the mathematically prescribed invariances proposed in this dissertation.

2.11 Conclusion

This dissertation proposes an enhanced end-to-end classification system with a mathematical framework to attain invariances to a set of given image transformations. The proposed method is pertinent to a specific category of image classification problems where image classes can be thought of being an instance of a template observed under a set of spatial deformations. If these deformations are appropriately modeled as a collection of smooth, one-to-one, and nonlinear transformations (see equations (2.9) and (2.10) of the dissertation), then the image classes become easily separable in the transform space (i.e., the R-CDT space) via the properties mentioned in [98]. These properties also allow for the approximation of image classes as convex subspaces in the R-CDT space, providing a more suitable data model for the nearest subspace method. The resulting

classifier can then be expected to provide high accuracy, computational efficiency, and out-of-distribution robustness, as we found in the experiments. A large number of image classification problems can be formulated this way and thus can benefit from our proposed solution. Heuristically, any classification problem for which one image in a class can be constructed from another by a smooth rearrangement of pixel intensities is an appropriate fit for the generative model. Obvious examples are affine transformations (translation, scaling, shear, etc.). Less obvious examples are distortions to images in an optical communication channel resulting from the influence of a transparent medium (e.g., turbulence, see [103]) or morphological changes in the gray matter of MRI images under the influence of a disease (see [3]).

The proposed mathematical solution attains high classification accuracy (compared with state-of-the-art end-to-end systems), especially at the low data regime. The method was also demonstrated to significantly improve the computational cost of classification (up to 1,000,000 times reduction in the computational cost can be attained). The method is mathematically coherent, understandable, non-iterative, requires no hyper-parameters to tune, and is simple enough to be implemented without GPU support. However, the proposed method can also be implemented in parallel using a GPU, which should further enhance the method's efficiency. The method also demonstrated robustness in challenging experimental scenarios, e.g., the out-of-distribution setup. The method performs well under the out-of-distribution setup because it learns the underlying generative model of the image classes. More specifically, the method learns the type of deformations that might have generated the dataset by using very few training examples.

We obtain superior performance by expanding upon the recently published R-CDT-NS classification method [98], which can also be interpreted as the 'nearest' sliced-Wasserstein distance method. The R-CDT-NS method [98] was demonstrated to show equivalent or better classification accuracy at both low and high data regimes. In this dissertation, we improved upon the performance of the R-CDT-NS method at the low data regime without altering its previously superior performance at the high data regime. The performance improvement at the low data regime was achieved by improving the invariance prescribing framework of the R-CDT-NS method. In the

proposed method, we encode invariances with respect to a more complicated deformation set than the previous paper [98]: the affine deformations, i.e., translation, isotropic/anisotropic scaling, horizontal/vertical shear, and rotation in the sliced-Wasserstein space. Though images under the effect of these deformations are challenging to classify in native image space, the R-CDT subspace can capture these variations and thus simplify the associated classification problems. We mathematically derive approximate basis vectors corresponding to these deformations and use them to enhance the R-CDT subspace to encode invariances instead of augmenting the individual training images. As a result, the method can learn a specific deformation type using a few basis vectors without requiring to use thousands of augmented images representing that deformation.

Finally, we note that the method is well-suited for the problems where the data at hand conform to the generative model stated in equation (2.10). One example where the data do not follow the generative model is classification problems involving natural unsegmented images (e.g., CIFAR10, imagenet datasets). However, some datasets (such as the OAM and Sign language datasets) contain unsegmented images, and the proposed method still outperforms the other methods in these datasets. In addition, our method can potentially be extended to be suitably applied to natural unsegmented images with more complex backgrounds. However, it would require redefining the problem statement and the generative model. One step forward in this direction is a few recent papers [104, 105] that consider images as a collection of patches. When an image is considered as a collection of patches, it might be possible to adaptively assign lower weights to the background and discard them automatically. However, these analyses require reformulation of the problem, which we leave to future work. Another potential solution is to use an object detection and segmentation method along with our proposed classification method. We also note that we used some approximations and assumptions in the derivation for the spanning sets of shear and anisotropic scaling. However, we showed that these approximations work reasonably for practical purposes, as we have seen in the results provided above. In our experiments, we chose much bigger deformation parameters than the official Affine MNIST [100] and obtained good classification results. We did not derive how much deformation is allowed and leave the robustness analysis to more extreme

deformations for future work.

2.12 Appendix

2.12.1 Proof of Lemma 2.2.3

Let $\mathbb{S}^{(k)}, k = 1, 2, ...,$ be the generative classes with a common confound set $\mathcal{G}$ such that any $f \notin \mathcal{G}$, $f'\varphi^{(k)} \circ f \notin \mathbb{S}^{(k)}$.[6]

Proposition: $\widehat{\mathbb{S}}^{(k)} \cap \widehat{\mathbb{V}}^{(p)} = \varnothing, \ \ \forall\, k \neq p.$

Assumptions:

1. $\mathbb{S}^{(k)} \cap \mathbb{S}^{(p)} = \varnothing.$
2. $\{f(x) = ax | a > 0\} \subseteq \mathcal{G}.$
3. $\mathcal{G}$ is a convex group.
4. $\forall$ increasing function $h \notin \mathcal{G}$ and $0 < \alpha < 1, \ \alpha\, id + (1-\alpha)h \notin \mathcal{G}$ (id denotes the identity function, $f(x) = x$).

Proof. Before we prove the main claim, let us start by stating and proving the following claim:

Claim (1): $\forall\, \widehat{s}_i^{(k)} \in \widehat{\mathbb{S}}^{(k)}$ and $\widehat{s}_j^{(p)} \in \widehat{\mathbb{S}}^{(p)}$ and $0 < \alpha < 1$,

$$\alpha \widehat{s}_i^{(k)} + (1-\alpha)\widehat{s}_j^{(p)} \notin \widehat{\mathbb{S}}^{(k)} \cup \widehat{\mathbb{S}}^{(p)}.$$

Proof of Claim (1): Let us prove by contradiction and assume that the claim is not true. Then,

[6] This condition is automatically satisfied if $\varphi^{(k)} > 0$ on $\mathbb{R}$ but may not hold in general if $\varphi^{(k)}$ is supported on a finite interval.

given $\alpha \in (0,1)$

$$\begin{aligned} &\alpha\widehat{s}_i^{(k)} + (1-\alpha)\widehat{s}_j^{(p)} \in \widehat{\mathbb{S}}^{(k)}. \\ \implies &\alpha\widehat{s}_i^{(k)} + (1-\alpha)\widehat{s}_j^{(p)} = g^{-1} \circ \widehat{\varphi}^{(k)}, \end{aligned} \tag{2.21}$$

for some $g \in \mathcal{G}$.

Then, $\exists\ h \notin \mathcal{G}$, where $h \circ \widehat{s}_i^{(k)} = \widehat{s}_j^{(p)}$. Using this fact in equation (2.21) we have that,

$$\begin{aligned} &\alpha\widehat{s}_i^{(k)} + (1-\alpha)h \circ \widehat{s}_i^{(k)} = g^{-1} \circ \widehat{\varphi}^{(k)} \\ \implies &(\alpha\ id + (1-\alpha)h) \circ g_i^{-1} \circ \widehat{\varphi}^{(k)} = g^{-1} \circ \widehat{\varphi}^{(k)};\ g_i \in \mathcal{G} \\ \implies &f^{-1} \circ \widehat{\varphi}^{(k)} = g^{-1} \circ \widehat{\varphi}^{(k)} \end{aligned} \tag{2.22}$$

where $f^{-1} = (\alpha\ id + (1-\alpha)h) \circ g_i^{-1}$. Note that by assumption (4), $\alpha\ id + (1-\alpha)h \notin \mathcal{G}$. Since $g_i \in \mathcal{G}$ and $\mathcal{G}$ is a group, it follows that $f^{-1} \notin \mathcal{G}$ and hence $f \notin \mathcal{G}$. By the assumption that for any $f \notin \mathcal{G}$, $f'\varphi^{(k)} \circ f \notin \mathbb{S}^{(k)}$ (or equivalently $f^{-1} \circ \widehat{\varphi}^{(k)} \notin \widehat{\mathbb{S}}^{(k)}$), it follows that the LHS of (2.22) does not belong to $\mathbb{S}^{(k)}$, which is a contradiction since the RHS of (2.22) belongs to $\mathbb{S}^{(k)}$. Therefore,

$$\alpha\widehat{s}_i^{(k)} + (1-\alpha)\widehat{s}_j^{(p)} \notin \widehat{\mathbb{S}}^{(k)}.$$

Similarly, we can show that

$$\alpha\widehat{s}_i^{(k)} + (1-\alpha)\widehat{s}_j^{(p)} \notin \widehat{\mathbb{S}}^{(p)}.$$

In other words,

$$\alpha\widehat{s}_i^{(k)} + (1-\alpha)\widehat{s}_j^{(p)} \notin \widehat{\mathbb{S}}^{(k)} \cup \widehat{\mathbb{S}}^{(p)}.$$

Therefore, Claim (1) is true.

Main claim:

$$\widehat{\mathbb{S}}^{(k)} \cap \widehat{\mathbb{V}}^{(p)} = \varnothing, \;\; \forall\, k \neq p$$

Proof of the main claim: Let us prove by contradiction and assume that the main claim is not true. Then, $\exists\, \beta_j \in \mathbb{R}$ for some $g \in \mathcal{G}$ such that

$$\sum_{j \in J} \beta_j \widehat{s}_j^{(p)} = g^{-1} \circ \widehat{\varphi}^{(k)} \tag{2.23}$$

Let us consider the case when $\beta_j > 0$ for all $j \in J$. Note that the LHS of equation (2.23) is a member of $\widehat{\mathbb{S}}^{(p)}$. To see this, we note that by assumption (2) and Lemma 2.2.1, any convex combination of elements in $\widehat{\mathbb{S}}^{(p)}$ lies in $\widehat{\mathbb{S}}^{(p)}$, i.e., $\sum\limits_{j \in J} \frac{\beta_j}{\sum\limits_{j \in J} \beta_j} \widehat{s}_j^{(p)} \in \widehat{\mathbb{S}}^{(p)}$. By assumption (3) and the composition property of the CDT, we have that $\alpha^{-1} \circ \widehat{s}^{(p)} \in \widehat{\mathbb{S}}^{(p)}$ for any $\alpha > 0$ and $\widehat{s}^{(p)} \in \widehat{\mathbb{S}}^{(p)}$. Letting $\alpha = (\sum\limits_{j \in J} \beta_j)^{-1}$ and $\widehat{s}^{(p)} = \frac{1}{\sum\limits_{j \in J} \beta_j} \sum\limits_{j \in J} \beta_j \widehat{s}_j^{(p)}$, we have that $\sum\limits_{j \in J} \beta_j \widehat{s}_j^{(p)} \in \widehat{\mathbb{S}}^{(p)}$. Since the RHS of equation (2.23) lies in $\widehat{\mathbb{S}}^{(k)}$, it follows that equation (2.23) cannot hold when $\beta_j > 0$ for all $j \in J$ as $\widehat{\mathbb{S}}^{(p)} \cap \widehat{\mathbb{S}}^{(k)} = \varnothing$ (by assumption (1) and Remark 1). On the other hand, equation (2.23) cannot hold when $\beta_j < 0$ for all $j \in J$ since the LHS of (2.23) would be a strictly decreasing function while the RHS is a strictly increasing function. Now, let us define the following:

$$J_+ = \left\{ j \in J | \beta_j > 0 \right\}; \;\; J_- = \left\{ j \in J | \beta_j < 0 \right\}$$

Equation (2.23) then can be written as

$$\begin{aligned} \frac{1}{2} \sum_{j \in J_+} \beta_j \widehat{s}_j^{(p)} + \frac{1}{2} \sum_{j \in J_-} \beta_j \widehat{s}_j^{(p)} &= \frac{1}{2} g^{-1} \circ \widehat{\varphi}^{(k)} \\ \frac{1}{2} \sum_{j \in J_+} \beta_j \widehat{s}_j^{(p)} &= \frac{1}{2} \sum_{j \in J_-} \left(-\beta_j\right) \widehat{s}_j^{(p)} + \frac{1}{2} g^{-1} \circ \widehat{\varphi}^{(k)} \end{aligned} \tag{2.24}$$

Now as $\beta_j|_{j\in J_+} > 0$ and $\left(-\beta_j\right)|_{j\in J_-} > 0$, by assumption (2), $\sum_{j\in J_+}\beta_j\widehat{s}_j^{(p)} \in \widehat{S}^{(p)}$ and $\sum_{j\in J_-}\left(-\beta_j\right)\widehat{s}_j^{(p)} \in \widehat{S}^{(p)}$. Also, $g^{-1}\circ\widehat{\varphi}^{(k)} \in \widehat{\mathbb{S}}^{(k)}$. Now,

$$\begin{aligned}
&\text{LHS of equation (2.24)}\\
&= \frac{1}{2}\sum_{j\in J_+}\beta_j\widehat{s}_j^{(p)} \in \widehat{\mathbb{S}}^{(p)}\\
&\text{RHS of equation (2.24)}\\
&= \frac{1}{2}\sum_{j\in J_-}\left(-\beta_j\right)\widehat{s}_j^{(p)} + \left(1-\frac{1}{2}\right)g^{-1}\circ\widehat{\varphi}^{(k)} \notin \widehat{\mathbb{S}}^{(k)}\cup\widehat{\mathbb{S}}^{(p)}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(by using Claim (1))}
\end{aligned}$$

which is a contradiction. Therefore, there exists no $\beta_j \in \mathbb{R}$ such that

$$\sum_{j\in J}\beta_j\widehat{s}_j^{(p)} = g^{-1}\circ\widehat{\varphi}^{(k)}$$

which implies, the main claim is true, i.e., $\widehat{\mathbb{S}}^{(k)}\cap\widehat{\mathbb{V}}^{(p)} = \varnothing,\ \ \forall\ k \neq p$. Note that, $\widehat{\mathbb{S}}^{(k)}$ here does not contain the origin because the generative models in equations (2.1) and (2.2) do not allow for zero elements. □

2.12.2 Standard deviation of test accuracy

Table 2.4: Standard deviation of percentage test accuracy in different datasets.

Chinese printed character dataset

	No. of training samples (per class)				
	1	2	4	8	16
Resnet	0.08	0.21	2.45	4.34	0.17
Shallow-CNN	0.04	0.06	0.17	0.82	2.21
VGGnet	0	0	0.87	20.32	41.54
Proposed	0.21	0.28	0.04	0	0

MNIST dataset

	No. of training samples (per class)												
	1	2	4	8	16	32	64	128	256	512	1024	2048	4096
Resnet	3.29	4.05	3.03	9.13	8.04	2.01	1.85	0.95	0.45	0.75	0.12	0.15	0.06
Shallow-CNN	4.08	6.89	2.64	1.12	3.90	0.96	1.42	0.49	0.31	0.23	0.09	0.09	0.07
Proposed	5.25	7.97	4.27	1.21	1.48	0.50	0.33	0.20	0.16	0.08	0.11	0.08	0.07

Affine-MNIST dataset

	No. of training samples (per class)												
	1	2	4	8	16	32	64	128	256	512	1024	2048	4096
Resnet	1.45	1.08	0.64	1.08	6.48	3.86	3.95	1.56	0.27	0.21	0.16	0.21	0.09
Shallow-CNN	1.18	1.31	1.09	0.67	2.58	2.06	2.95	1.65	1.03	0.38	0.45	0.33	0.27
VGGnet	2.59	2.99	3.17	4.67	4.78	2.97	1.35	0.99	0.45	0.33	0.18	0.17	0.12
Proposed	3.27	5.29	2.31	2.30	1.33	0.59	0.38	0.22	0.15	0.1	0.08	0.08	0.08

Optical OAM dataset

	No. of training samples (per class)									
	1	2	4	8	16	32	64	128	256	512
Resnet	1.71	4.31	2.60	1.39	0.84	0.78	0.22	0.05	0.04	0.16
Shallow-CNN	2.80	1.03	2.64	1.72	4.29	0.81	0.45	0.10	0.18	0.12
VGGnet	1.64	1.63	13.30	13.11	2.81	1.97	0.77	0.44	0.12	0.05
Proposed	2.40	1.73	0.66	0.54	0.28	0.09	0.02	0.01	0.01	0.01

Sign language dataset

	No. of training samples (per class)									
	1	2	4	8	16	32	64	128	256	512
Resnet	9.08	15.14	9.24	12.26	11.80	7.02	4.94	2.03	0.76	0.05
Shallow-CNN	9.87	4.49	3.07	1.62	5.93	7.58	1.62	1.22	0.03	0
VGGnet	8.83	15.48	16.35	19.79	1.76	5.67	3.76	1.22	1.39	0.27
Proposed	12.26	9.68	6.85	4.18	1.73	0.78	0.12	0	0	0

OASIS brain MRI dataset

	No. of training samples (per class)					
	1	2	4	8	16	32
Resnet	4.40	11.58	11.69	12.09	12.51	7.96
Shallow-CNN	18.12	17.42	8.28	5.49	12.68	6.37
VGGnet	5.07	5.12	4.06	4.50	11.99	10.02
Proposed	7.56	5.43	3.56	2.96	2.26	0.85

2.12.3 Comparisons with methods other than the neural networks

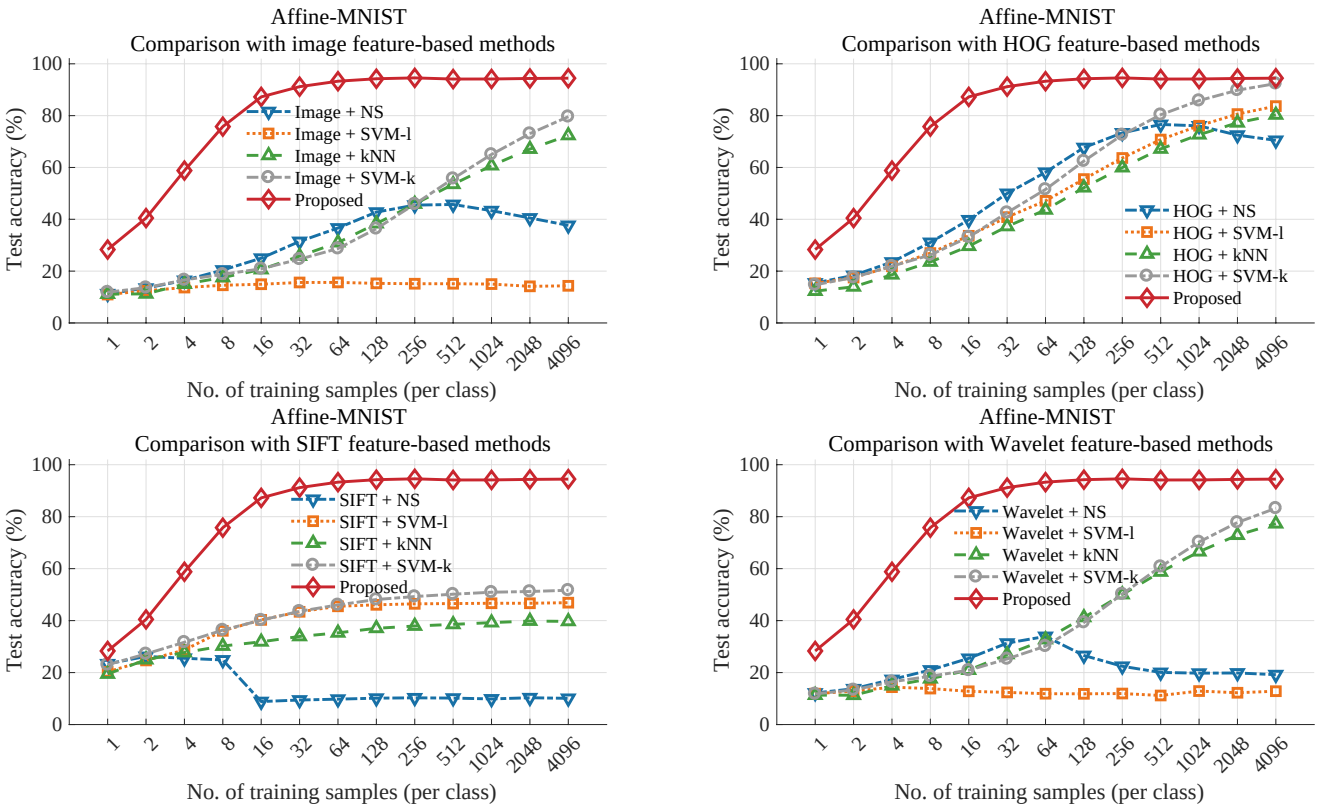

Figure 2.15: Comparison of the percentage test accuracy results of the proposed method with the results obtained by applying the nearest subspace (NS), linear support vector machine (SVM-l), k-nearest neighbor (kNN), and kernel support vector machine (SVM-k) classifiers on the raw image, HOG, SIFT, and wavelet features of the Affine-MNIST dataset.

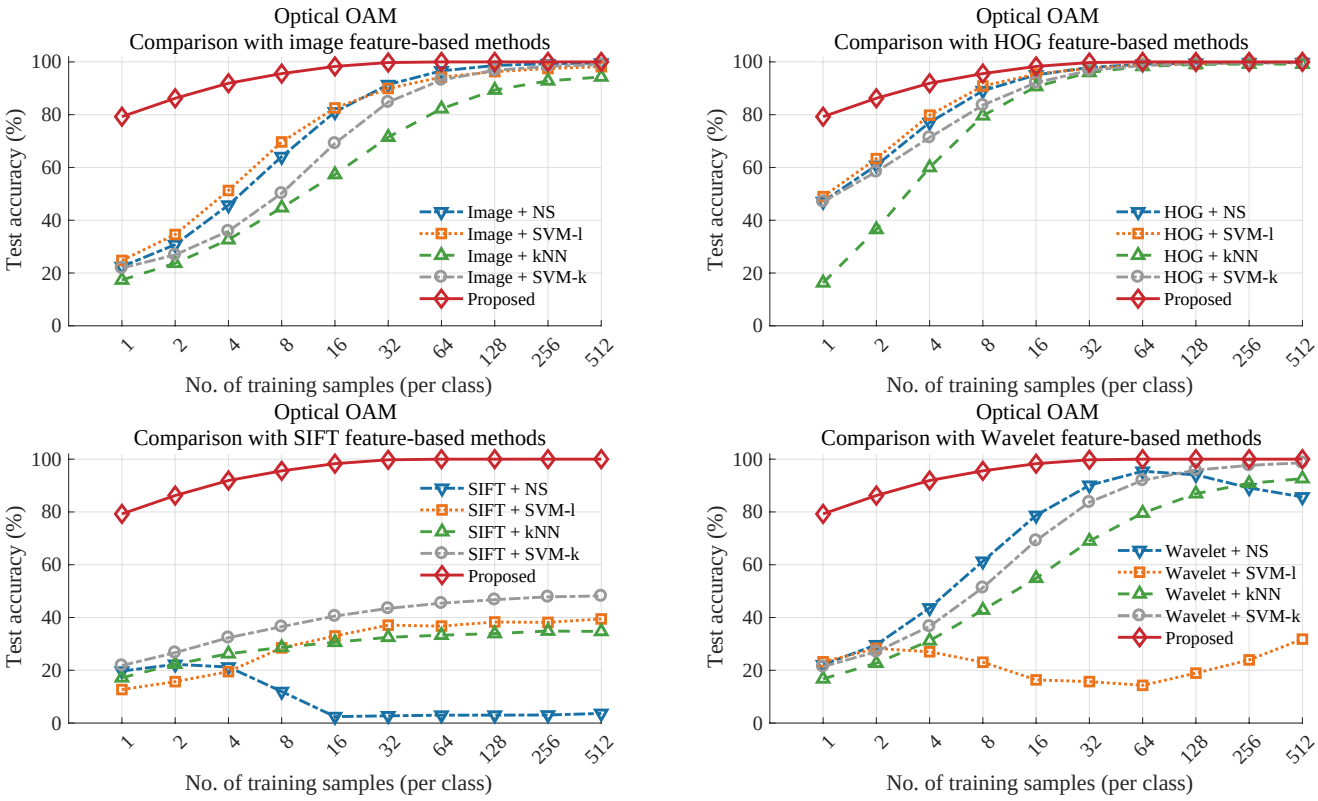

Figure 2.16: Comparison of the percentage test accuracy results of the proposed method with the results obtained by applying the nearest subspace (NS), linear support vector machine (SVM-l), k-nearest neighbor (kNN), and kernel support vector machine (SVM-k) classifiers on the raw image, HOG, SIFT, and wavelet features of the Optical OAM dataset.

2.12.4 Anisotropic scaling

Let $g(\mathbf{x}) = \breve{\mathcal{D}}\mathbf{x}$ with $\breve{\mathcal{D}} = \begin{bmatrix} 1/a, & 0 \\ 0, & 1/b \end{bmatrix}$. Consider two functions $s_g(x, y)$ and $s(x, y)$ such that $s_g(x, y) = |Jg|s \circ g = \frac{1}{ab}s(\breve{\mathcal{D}}\mathbf{x}) = \frac{1}{ab}s(x/a, y/b)$, for some $a, b > 0$, $a \neq b$, which is the normalized anisotropic dilatation of s by a, b where $a, b \in \mathbb{R}_+$.

Proof of equation (2.15) **of the dissertation**

By definition of the Radon transform and then applying the change of variables formula with $x' = x/a$ and $y' = y/b$, we have that

$$\begin{aligned} \tilde{s}_g(t, \theta) &= \frac{1}{ab}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(\frac{x}{a}, \frac{y}{b})\delta(t - x\cos\theta - y\sin\theta)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x', y')\delta(t - ax'\cos\theta - by'\sin\theta)dx'dy'. \end{aligned} \tag{2.25}$$

Using the co-area formula and letting $\gamma = \sqrt{a^2\cos^2\theta + b^2\sin^2\theta}$, we have that

$$\begin{aligned} \tilde{s}_g(t, \theta) &= \frac{1}{\gamma}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x', y')\delta(\frac{t}{\gamma} - x'\frac{a\cos\theta}{\gamma} - y'\frac{b\sin\theta}{\gamma})dx'dy' \\ &= \frac{1}{\gamma}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x', y')\delta(\frac{t}{\gamma} - x'\cos\theta' - y'\sin\theta')dx'dy', \end{aligned} \tag{2.26}$$

where $\theta' = \tan^{-1}\left(\frac{b}{a}\tan\theta\right)$. Hence $\tilde{s}_g(t, \theta) = \frac{1}{\gamma}\tilde{s}\left(\frac{t}{\gamma}, \theta'\right)$. Applying the scaling property of R-CDT

$$\hat{s}_g(t, \theta) = \gamma\hat{s}(t, \theta') \tag{2.27}$$

Proof of Lemma 2.8.1

For illustration purposes, we assume without loss of generality $\theta \in [0, \frac{\pi}{2})$ and $a \leq b$ (i.e., $\alpha \geq 0$) in the following derivations. Other cases are similar. Using Taylor's formula for $\tan^{-1}(x)$

around $x = \tan\theta$, we have that

$$\theta' = \tan^{-1}(\tan\theta + \alpha\tan\theta) = \theta + \frac{\alpha\tan\theta}{1+\tan^2\theta} - \frac{\xi}{(1+\xi^2)^2}(\alpha\tan\theta)^2, \tag{2.28}$$

where $\xi \in [\tan\theta, (1+\alpha)\tan\theta]$. Since $0 \le \frac{|\xi|}{1+\xi^2} \le \frac{1}{2}$ and $\frac{1}{1+\xi^2} \le \frac{1}{1+\tan^2\theta}$ for $\xi \ge \tan\theta \ge 0$, we have that $|\theta' - \theta| \le (\alpha\sin\theta\cos\theta + \frac{1}{2}\alpha^2\sin^2\theta)$. With the observation that $|\sin\theta\cos\theta| \le \frac{1}{2}$, it is also easy to derive from above a bound of $|\theta' - \theta|$ independent of θ:

$$|\theta' - \theta| \le \frac{1}{2}(\alpha + \alpha^2). \tag{2.29}$$

Similarly one can show that for $-1 < \alpha < 0$,

$$|\theta' - \theta| \le |\alpha\sin\theta\cos\theta| + \frac{1}{2}\frac{\alpha^2\tan^2\theta}{1+(1+\alpha)^2\tan^2\theta}) \le \frac{1}{2}(|\alpha| + \frac{\alpha^2}{(1+\alpha)^2}). \tag{2.30}$$

Proof of Lemma 2.8.2

Here, we aim to show an approximation for γ for $a \le b$. Observing that $\gamma = a\cos\theta\sqrt{1+(1+\alpha)^2\tan^2\theta} = a\cos\theta$

$\sqrt{1+\tan^2\theta+(2\alpha+\alpha^2)\tan^2\theta}$ and using Taylor's formula for $\sqrt{x}$ around $x = 1+\tan^2\theta$, we have that

$$\gamma = a\cos\theta\big(\sqrt{1+\tan^2\theta} + \frac{(2\alpha+\alpha^2)\tan^2\theta}{2\sqrt{1+\tan^2\theta}} - \frac{(2\alpha+\alpha^2)^2\tan^4\theta}{8(\xi)^{3/2}}\big), \tag{2.31}$$

where $\xi \in [1+\tan^2\theta, 1+(1+\alpha)^2\tan^2\theta]$. Observing that $\sqrt{1+\tan^2\theta} = \frac{1}{\cos\theta}$ and $\frac{1}{\xi} \le \frac{1}{1+\tan^2\theta}$ for $\xi \in [1+\tan^2\theta, (1+\alpha)^2\tan^2\theta]$, we obtain that

$$\begin{aligned}|\gamma - a| &\le a\cos\theta\Big((\alpha+\frac{\alpha^2}{2})\cos\theta\tan^2\theta + \frac{(2\alpha+\alpha^2)^2}{8}\cos^3\theta\tan^4\theta\Big)\\ &= a(\alpha+\frac{\alpha^2}{2})\sin^2\theta + a\frac{(2\alpha+\alpha^2)^2}{8}\sin^4\theta.\end{aligned} \tag{2.32}$$

Ignoring higher order terms of α, we have the following approximation

$$\gamma = a + \alpha a \sin^2\theta + O(\alpha^2) = (1 + \alpha \sin^2\theta)a + O(\alpha^2). \tag{2.33}$$

Analogously, if $a > b$, we let $\frac{a}{b} = 1 + \beta$ for some $\beta > 0$ and by similar arguments we have that

$$|\gamma - b| \leq b(\beta + \frac{\beta^2}{2}) \cos^2\theta + b\frac{(2\beta + \beta^2)^2}{8} \cos^4\theta. \tag{2.34}$$

Ignoring higher order terms of β in (2.34), we have the following approximation

$$\gamma = b(1 + \beta \cos^2\theta) + O(\beta^2). \tag{2.35}$$

A practical example for the choice of ϵ

Let us consider a practical example where we choose ϵ to be the numerical difference of consecutive angles used in Radon transform computation. If we choose 45 uniform angles between 0 and π, the difference of consecutive angles are $\frac{\pi}{45} \approx 0.07$. Choosing α between 0 and .12 guarantee that $|\theta' - \theta| \leq .068$, and similarly, choosing $\alpha \in [-.12, 0)$, we have that $|\theta' - \theta| \leq .07$, which makes the difference between θ' and θ smaller than the numerical difference of consecutive angles used in Radon transform computation. Note that one can choose a different number of uniform angles other than 45, and that might allow for a more relaxed choice of α. For classification purposes, a larger value of ϵ (hence α) is allowed. In addition, as the R.H.S. of the inequalities in Lemma 4.1 are the upper bounds for $|\theta' - \theta|$, for a fixed α, $|\theta' - \theta|$ might be much smaller than these bounds. It can be one of the reasons why a larger α (hence ϵ) can be chosen in practice.

2.12.5 Horizontal and vertical shear

Shear-horizontal

Let $g_1(\mathbf{x}) = \mathcal{H}_1\mathbf{x}$ with $\mathcal{H}_1 = \begin{bmatrix} 1, & -h \\ 0, & 1 \end{bmatrix}$. Consider two functions $s_{g_1}(x, y)$ and $s(x, y)$ such that $s_{g_1}(x, y) = |Jg_1|s \circ g_1 = s(\mathcal{H}_1\mathbf{x}) = s(x - hy, y)$, for some h, which is the normalized horizontal shear of s by h where $h \in \mathbb{R}$.

Proof of equation (2.17) **of the dissertation**

By definition of the Radon transform and the change of variables formula with $x' = x - hy$, $y' = y$, we have that

$$\begin{aligned} \tilde{s}_{g_1}(t, \theta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x - hy, y)\delta(t - x\cos\theta - y\sin\theta)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x', y')\delta(t - x'\cos\theta - y'(\sin\theta + h\cos\theta))dx'dy'. \end{aligned} \tag{2.36}$$

Using the co-area formula and the scaling properties of the Dirac delta function and letting $\gamma = \sqrt{1 + h^2\cos^2\theta + h\sin(2\theta)}$, we have

$$\tilde{s}_{g_1}(t, \theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\gamma} s(x', y')\delta(\frac{t}{\gamma} - \frac{x'\cos\theta - y'(\sin\theta + h\cos\theta))}{\gamma}dx'dy'. \tag{2.37}$$

Let $\theta' = \tan^{-1}(\frac{\sin\theta + h\cos\theta}{\cos\theta}) = \tan^{-1}(\tan\theta + h)$, then $\cos\theta' = \frac{\cos\theta}{\gamma}$ and $\sin\theta' = \frac{\sin\theta + h\cos\theta}{\gamma}$. Hence

$$\tilde{s}_{g_1}(t, \theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\gamma} s(x', y')\delta(\frac{t}{\gamma} - x'\cos\theta' - y'\sin\theta')dx'dy'. \tag{2.38}$$

By the definition of Radon transform, we see that

$$\tilde{s}_{g_1}(t, \theta) = \frac{1}{\gamma}\tilde{s}\left(\frac{t}{\gamma}, \theta'\right); \gamma = \sqrt{1 + h^2\cos^2\theta + h\sin(2\theta)},\ \theta' = \tan^{-1}(\tan\theta + h). \tag{2.39}$$

Applying the scaling property of R-CDT, we have that

$$\hat{s}_{g_1}(t,\theta) = \gamma \hat{s}(t,\theta'). \tag{2.40}$$

Proof of Lemma 2.8.3

For illustration purposes, we assume without loss of generality $\theta \in [0, \frac{\pi}{2})$ and $h \geq 0$ in the following derivations. Other cases are similar. Using Taylor's formula for $\tan^{-1}(x)$ around $x = \tan\theta$, we have that

$$\theta' = \tan^{-1}(\tan\theta + h) = \theta + \frac{h}{1+\tan^2\theta} - \frac{\xi h^2}{(1+\xi^2)^2}, \tag{2.41}$$

where $\xi \in [\tan\theta, \tan\theta + h]$. Since $0 \leq \frac{|\xi|}{1+\xi^2} \leq \frac{1}{2}$ and $\frac{1}{1+\xi^2} \leq \frac{1}{1+\tan^2\theta}$ for $\xi \geq \tan\theta \geq 0$, we have that

$$|\theta' - \theta| \leq (h + \frac{1}{2}h^2)\cos^2\theta. \tag{2.42}$$

With the observation that $|\cos\theta| \leq 1$, it is easy to see that

$$|\theta' - \theta| \leq h + \frac{1}{2}h^2. \tag{2.43}$$

Similarly one can show that for $h < 0$,

$$|\theta' - \theta| \leq |h| + h^2. \tag{2.44}$$

Proof of Lemma 2.8.4

Here, we show an approximation for γ. Apply Taylor's formula for $\sqrt{x}$ around $x = 1$ to γ, we have that

$$\begin{aligned}\gamma &= \sqrt{1 + h^2\cos^2\theta + h\sin(2\theta)} \\ &= 1 + \frac{1}{2}\big(h^2\cos^2\theta + h\sin(2\theta)\big) - \frac{\big(h^2\cos^2\theta + h\sin(2\theta)\big)^2}{8\xi^{3/2}},\end{aligned} \tag{2.45}$$

where $\xi \in [1, 1 + h^2\cos^2\theta + h\sin(2\theta)]$. Ignoring higher order terms we have that for $h \geq 0$

$$\gamma = 1 + \frac{1}{2}\big(h\sin(2\theta) + h^2\cos^2\theta\big) + O(h^2), \tag{2.46}$$

with

$$|\gamma - 1| \leq \frac{1}{2}(h + h^2) + \frac{1}{8}(h + h^2)^2. \tag{2.47}$$

One can derive similar approximations of γ for $h < 0$.

A practical example for the choice of ϵ

Now, let us consider a practical example where we choose ϵ to be the numerical difference of consecutive angles used in Radon transform computation. If we choose 45 uniform angles between 0 and π, choosing h between 0 and .067 guarantee that $|\theta' - \theta| \leq \frac{\pi}{45}$. It is easy to see that if $-.065 \leq h < 0$, then $|\theta' - \theta| \leq \frac{\pi}{45}$. In summary, with $|h| \leq .065$, the difference between θ' and θ is smaller than the numerical difference of consecutive angles used in Radon transform computation. Note that one can choose a different number of uniform angles other than 45, and that might allow for a more relaxed choice of h. For classification purposes, a larger value of ϵ (hence h) is allowed. In addition, as the R.H.S. of the inequalities in Lemma 4.3 are the upper bounds for $|\theta' - \theta|$, for a fixed h, $|\theta' - \theta|$ might be much smaller than these bounds. It can be one of the reasons why a larger h (hence ϵ) can be chosen in practice.

Shear-vertical

Let $g_2(\mathbf{x}) = \mathcal{H}_2\mathbf{x}$ with $\mathcal{H}_2 = \begin{bmatrix} 1, & 0 \\ -v, & 1 \end{bmatrix}$. Consider two functions $s_{g_2}(x, y)$ and $s(x, y)$ such that $s_{g_2}(x, y) = |Jg_2|s \circ g_2 = s(\mathcal{H}_2\mathbf{x}) = s(x, y - vx)$, for some v, which is the normalized vertical shear of s by v where $v \in \mathbb{R}$.

By similar arguments as horizontal shear, we have that

$$\tilde{s}_{g_2}(t, \theta) = \frac{1}{\gamma}\tilde{s}\left(\frac{t}{\gamma}, \theta'\right);\ \gamma = \sqrt{1 + v^2 \sin^2\theta + v\sin(2\theta)},$$
$$\theta' = \cot^{-1}(\cot\theta + v) \tag{2.48}$$

Applying the scaling property of R-CDT, we have that

$$\hat{s}_{g_2}(t, \theta) = \gamma\hat{s}(t, \theta'). \tag{2.49}$$

For illustration purposes, we assume without loss of generality $\theta \in [0, \frac{\pi}{2})$ and $v \geq 0$ in the following derivations. Other cases are similar. Using Taylor's formula for $\tan^{-1}(x)$ around $x = \tan\theta$, we have that $\cot^{-1}(x)$ around $x = \cot\theta$, we have that

$$\theta' = \cot^{-1}(\cot\theta + v) \tag{2.50}$$
$$= \theta - \frac{v}{1 + \cot^2\theta} + \frac{\xi v^2}{(1 + \xi^2)^2}, \tag{2.51}$$

where $\xi \in [\cot\theta, \cot\theta + v]$. Since $0 \leq \frac{|\xi|}{1+\xi^2} \leq \frac{1}{2}$ and $\frac{1}{1+\xi^2} \leq \frac{1}{1+\cot^2\theta}$ for $\xi \geq \cot\theta \geq 0$, we have that

$$|\theta' - \theta| \leq (v + \frac{1}{2}v^2)\sin^2\theta. \tag{2.52}$$

With the observation that $|\cos\theta| \leq 1$, it is easy to see that

$$|\theta' - \theta| \leq v + \frac{1}{2}v^2. \tag{2.53}$$

Similarly one can show that for $h < 0$,

$$|\theta' - \theta| \leq |v| + v^2. \tag{2.54}$$

If $|\theta' - \theta| \leq \epsilon$, where ϵ is a small number, we can approximate θ' as θ. Then using equation (2.49), we have

$$\hat{s}_{g_2}(t, \theta) = \gamma \hat{s}(t, \theta') \approx \gamma \hat{s}(t, \theta). \tag{2.55}$$

Next we show an approximation for γ. Apply Taylor's formula for $\sqrt{x}$ around $x = 1$ to γ, we have that

$$\begin{aligned} \gamma &= \sqrt{1 + v^2 \sin^2\theta + v \sin(2\theta)} \\ &= 1 + \frac{1}{2}\big(v^2 \sin^2\theta + v \sin(2\theta)\big) - \frac{\big(v^2 \sin^2\theta + v \sin(2\theta)\big)^2}{8\xi^{3/2}}, \end{aligned} \tag{2.56}$$

where $\xi \in [1, 1 + v^2 \sin^2\theta + v \sin(2\theta)]$. Ignoring higher order terms we have that for $v \geq 0$

$$\gamma = 1 + \frac{1}{2}\big(v \sin(2\theta) + v^2 \sin^2\theta\big) + O(v^2), \tag{2.57}$$

with

$$|\gamma - 1| \leq \frac{1}{2}(v + v^2) + \frac{1}{8}(v + v^2)^2. \tag{2.58}$$

One can derive similar approximations of γ for $v < 0$. Hence $\hat{s}_{g_2}(t, \theta) \approx \gamma \hat{s}(t, \theta) = \hat{s}(t, \theta) + \frac{1}{2}(v \sin(2\theta) + v^2 \sin^2\theta)\hat{s}(t, \theta) + O(v^2)$. In summary, to model small vertical shearing of s, we add the following additional spanning set $\hat{E} = \{\big(v^2 \sin^2\theta + v \sin(2\theta)\big)\hat{s}\}$ (for small v) as enrichment to the training subspace in the transform space.

2.12.6 Percentage test accuracy of the methods ($\mu \pm \sigma$) in different datasets

Synthetic dataset

No. of training samples per class = 1	
MNISTnet (Aug-1 / Aug-25 / Aug-50)	11 ± 1.3 / 13 ± 1.1 / 15 ± 1.4
VGG11 (Aug-1 / Aug-25 / Aug-50)	11 ± 1.8 / 17 ± 5.2 / 32 ± 11.9
Resnet18 (Aug-1 / Aug-25 / Aug-50)	10 ± 0.3 / 12 ± 4.2 / 42 ± 6.3
k-NN (Aug-1 / Aug-25 / Aug-50)	11 ± 0.7 / 11 ± 0.7 / 12 ± 1.1
R-CDT NS (Aug - none)	18 ± 0.0
Proposed (Aug - none)	97 ± 0.0

MNIST dataset

	No. of training samples (per class)		
	1	2	4
MNISTnet (Aug-1 / Aug-25 / Aug-50)	25 ± 4.6 / 19 ± 5.1 / 28 ± 5.1	35 ± 5.0 / 33 ± 4.9 / 38 ± 5.4	50 ± 2.3 / 58 ± 7.2 / 62 ± 6.1
VGG11 (Aug-1 / Aug-25 / Aug-50)	13 ± 4.0 / 29 ± 8.3 / 41 ± 5.5	10 ± 2.0 / 58 ± 6.7 / 63 ± 8.5	30 ± 5.6 / 78 ± 11.4 / 79 ± 13.0
Resnet18 (Aug-1 / Aug-25 / Aug-50)	19 ± 4.3 / 27 ± 10.6 / 45 ± 5.5	29 ± 7.2 / 45 ± 9.8 / 73 ± 3.9	68 ± 1.9 / 64 ± 27.6 / 88 ± 2.6
k-NN (Aug-1 / Aug-25 / Aug-50)	25 ± 4.3 / 27 ± 4.9 / 28 ± 6.7	38 ± 4.3 / 40 ± 6.8 / 43 ± 4.6	53 ± 3.1 / 55 ± 4.4 / 55 ± 4.9
R-CDT NS (Aug - none)	44 ± 5.3	56 ± 8.0	69 ± 4.3
Proposed (Aug - none)	67 ± 4.5	76 ± 4.2	83 ± 1.8
	No. of training samples (per class)		
	6	8	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	60 ± 3.3 / 72 ± 3.0 / 76 ± 1.5	65 ± 3.6 / 79 ± 3.4 / 80 ± 2.4	58 ± 8.9 / 73 ± 7.5 / 73 ± 8.3
VGG11 (Aug-1 / Aug-25 / Aug-50)	45 ± 5.9 / 81 ± 12.4 / 88 ± 2.8	51 ± 10.9 / 86 ± 7.7 / 87 ± 5.6	61 ± 11.2 / 87 ± 4.1 / 87 ± 6.2
Resnet18 (Aug-1 / Aug-25 / Aug-50)	74 ± 2.7 / 88 ± 6.7 / 88 ± 11.4	78 ± 1.5 / 78 ± 23.8 / 91 ± 3.5	73 ± 10.3 / 87 ± 4.1 / 89 ± 4.2
k-NN (Aug-1 / Aug-25 / Aug-50)	60 ± 1.9 / 63 ± 2.7 / 64 ± 2.1	63 ± 2.2 / 67 ± 1.9 / 68 ± 2.3	67 ± 2.7 / 70 ± 2.5 / 71 ± 2.1
R-CDT NS (Aug - none)	76 ± 2.0	80 ± 1.2	81 ± 2.3
Proposed (Aug - none)	85 ± 1.2	86 ± 1.9	87 ± 1.2

AFFNIST (regular) dataset

	No. of training samples (per class)		
	1	2	4
MNISTnet (Aug-1 / Aug-25 / Aug-50)	12 ± 2.0 / 12 ± 0.9 / 14 ± 1.1	13 ± 1.5 / 13 ± 1.9 / 17 ± 3.2	14 ± 1.4 / 21 ± 2.2 / 27 ± 2.7
VGG11 (Aug-1 / Aug-25 / Aug-50)	12 ± 2.3 / 14 ± 4.3 / 20 ± 6.3	13 ± 2.7 / 31 ± 12.5 / 32 ± 9.7	18 ± 3.1 / 30 ± 15.7 / 41 ± 21.5
Resnet18 (Aug-1 / Aug-25 / Aug-50)	11 ± 1.2 / 13 ± 3.5 / 24 ± 7.0	10 ± 1.3 / 13 ± 3.1 / 45 ± 10.2	19 ± 1.6 / 26 ± 23.1 / 65 ± 19.8
k-NN (Aug-1 / Aug-25 / Aug-50)	12 ± 1.8 / 14 ± 2.4 / 15 ± 2.1	13 ± 1.1 / 16 ± 1.6 / 18 ± 1.4	15 ± 1.0 / 19 ± 1.8 / 23 ± 2.2
R-CDT NS (Aug - none)	17 ± 2.1	21 ± 2.3	31 ± 2.7
Proposed (Aug - none)	40 ± 4.8	55 ± 4.4	66 ± 2.6
	No. of training samples (per class)		
	6	8	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	15 ± 1.7 / 26 ± 2.8 / 36 ± 2.3	18 ± 1.7 / 32 ± 4.8 / 43 ± 3.7	13 ± 2.7 / 29 ± 5.4 / 38 ± 8.0
VGG11 (Aug-1 / Aug-25 / Aug-50)	19 ± 6.5 / 40 ± 24.4 / 41 ± 21.6	25 ± 7.5 / 34 ± 17.8 / 36 ± 20.1	22 ± 7.6 / 64 ± 13.0 / 78 ± 4.4
Resnet18 (Aug-1 / Aug-25 / Aug-50)	23 ± 2.6 / 59 ± 28.2 / 74 ± 21.5	26 ± 2.4 / 50 ± 33.4 / 61 ± 31.9	21 ± 7.4 / 79 ± 4.2 / 81 ± 5.2
k-NN (Aug-1 / Aug-25 / Aug-50)	16 ± 1.5 / 22 ± 1.2 / 25 ± 1.8	18 ± 1.2 / 23 ± 1.0 / 27 ± 1.4	18 ± 1.3 / 25 ± 1.4 / 28 ± 1.9
R-CDT NS (Aug - none)	45 ± 2.8	51 ± 2.3	58 ± 3.4
Proposed (Aug - none)	69 ± 2.2	71 ± 2.6	72 ± 1.8

OMNIGLOT dataset

	No. of training samples (per class)		
	1	2	4
MNISTnet (Aug-1 / Aug-25 / Aug-50)	23 ± 2.6 / 31 ± 2.8 / 43 ± 5.6	34 ± 2.5 / 50 ± 3.9 / 59 ± 4.0	44 ± 2.0 / 71 ± 2.9 / 77 ± 2.4
VGG11 (Aug-1 / Aug-25 / Aug-50)	5 ± 2.6 / 56 ± 10.8 / 69 ± 5.5	30 ± 7.1 / 72 ± 13.7 / 76 ± 9.7	47 ± 7.4 / 82 ± 11.4 / 91 ± 3.9
Resnet18 (Aug-1 / Aug-25 / Aug-50)	6 ± 1.1 / 63 ± 29.3 / 82 ± 3.1	30 ± 3.0 / 71 ± 36.0 / 91 ± 1.7	48 ± 3.0 / 80 ± 31.3 / 95 ± 1.1
k-NN (Aug-1 / Aug-25 / Aug-50)	6 ± 2.3 / 23 ± 3.3 / 27 ± 2.2	8 ± 1.3 / 30 ± 2.7 / 37 ± 3.0	18 ± 1.7 / 39 ± 1.5 / 47 ± 2.2
R-CDT NS (Aug - none)	60 ± 4.9	70 ± 2.7	78 ± 2.5
Proposed (Aug - none)	77 ± 2.3	83 ± 2.3	90 ± 1.9
	No. of training samples (per class)		
	6	8	-
MNISTnet (Aug-1 / Aug-25 / Aug-50)	49 ± 2.0 / 77 ± 1.2 / 82 ± 2.2	53 ± 2.5 / 81 ± 2.2 / 84 ± 2.0	-
VGG11 (Aug-1 / Aug-25 / Aug-50)	46 ± 12.7 / 93 ± 1.6 / 92 ± 3.3	65 ± 7.3 / 87 ± 22.5 / 95 ± 1.4	-
Resnet18 (Aug-1 / Aug-25 / Aug-50)	57 ± 3.6 / 63 ± 41.9 / 95 ± 1.0	68 ± 2.9 / 56 ± 45.3 / 73 ± 32.4	-
k-NN (Aug-1 / Aug-25 / Aug-50)	24 ± 1.7 / 45 ± 2.4 / 52 ± 3.5	27 ± 1.0 / 49 ± 2.9 / 55 ± 1.2	-
R-CDT NS (Aug - none)	83 ± 2.9	85 ± 1.8	-
Proposed (Aug - none)	92 ± 1.3	93 ± 1.6	-

AFFNIST (out-of-distribution) dataset

	No. of training samples (per class)		
	1	2	4
MNISTnet (Aug-1 / Aug-25 / Aug-50)	13 ± 2.0 / 12 ± 1.3 / 14 ± 2.3	13 ± 1.5 / 15 ± 2.1 / 21 ± 2.4	15 ± 0.7 / 23 ± 1.2 / 28 ± 1.3
VGG11 (Aug-1 / Aug-25 / Aug-50)	14 ± 2.3 / 20 ± 5.3 / 28 ± 7.7	14 ± 2.4 / 39 ± 5.9 / 33 ± 12.2	25 ± 5.1 / 48 ± 14.8 / 51 ± 13.9
Resnet18 (Aug-1 / Aug-25 / Aug-50)	12 ± 0.7 / 17 ± 4.4 / 33 ± 3.7	11 ± 0.1 / 24 ± 8.8 / 48 ± 3.8	19 ± 1.9 / 45 ± 12.4 / 63 ± 2.6
k-NN (Aug-1 / Aug-25 / Aug-50)	12 ± 1.4 / 12 ± 1.5 / 13 ± 1.7	12 ± 1.0 / 13 ± 0.9 / 15 ± 1.7	13 ± 0.5 / 15 ± 1.5 / 17 ± 1.3
R-CDT NS (Aug - none)	17 ± 2.4	16 ± 2.3	16 ± 1.8
Proposed (Aug - none)	57 ± 4.1	65 ± 4.3	71 ± 1.6
	No. of training samples (per class)		
	6	8	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	16 ± 0.8 / 27 ± 3.0 / 37 ± 2.7	17 ± 1.1 / 34 ± 2.4 / 41 ± 2.3	16 ± 1.1 / 28 ± 3.4 / 32 ± 6.6
VGG11 (Aug-1 / Aug-25 / Aug-50)	25 ± 7.1 / 55 ± 11.6 / 58 ± 13.1	33 ± 4.4 / 53 ± 13.1 / 59 ± 25.3	35 ± 3.3 / 58 ± 7.4 / 58 ± 8.0
Resnet18 (Aug-1 / Aug-25 / Aug-50)	19 ± 1.5 / 57 ± 16.5 / 70 ± 1.6	22 ± 1.0 / 67 ± 3.8 / 74 ± 1.5	20 ± 3.2 / 56 ± 3.3 / 61 ± 12.7
k-NN (Aug-1 / Aug-25 / Aug-50)	14 ± 1.1 / 17 ± 1.2 / 19 ± 1.3	14 ± 0.7 / 18 ± 1.3 / 21 ± 0.8	14 ± 0.9 / 20 ± 1.0 / 21 ± 1.0
R-CDT NS (Aug - none)	17 ± 2.3	17 ± 1.8	18 ± 2.2
Proposed (Aug - none)	72 ± 1.9	73 ± 1.6	73 ± 2.4

Brain MRI dataset

	No. of training samples (per class)		
	1	5	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	49 ± 4.9 / 51 ± 4.0	55 ± 4.2 / 55 ± 9.8	51 ± 2.4 / 54 ± 5.0
VGG11 (Aug-1 / Aug-25 / Aug-50)	50 ± 5.0 / 56 ± 10.4	53 ± 6.0 / 52 ± 11.2	50 ± 2.5 / 57 ± 9.9
Resnet18 (Aug-1 / Aug-25 / Aug-50)	49 ± 1.9 / 54 ± 8.2	51 ± 3.7 / 56 ± 9.2	51 ± 1.6 / 60 ± 9.3
k-NN (Aug-1 / Aug-25 / Aug-50)	50 ± 1.6 / 51 ± 3.7	47 ± 3.0 / 49 ± 3.5	49 ± 7.0 / 50 ± 4.2
R-CDT NS (Aug - none)	49 ± 6.1	48 ± 5.3	52 ± 6.3
Proposed (Aug - none)	57 ± 11.3	62 ± 8.0	62 ± 5.8

Sign Language dataset

	No. of training samples (per class)		
	1	5	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	69 ± 13.0 / 74 ± 11.0	91 ± 4.7 / 92 ± 7.6	84 ± 11.2 / 89 ± 7.9
VGG11 (Aug-1 / Aug-25 / Aug-50)	43 ± 9.2 / 47 ± 9.3	45 ± 10.3 / 80 ± 12.3	43 ± 6.2 / 76 ± 15.2
Resnet18 (Aug-1 / Aug-25 / Aug-50)	41 ± 7.5 / 58 ± 13.0	47 ± 10.6 / 91 ± 6.4	40 ± 12.8 / 88 ± 10.8
k-NN (Aug-1 / Aug-25 / Aug-50)	31 ± 0.0 / 33 ± 3.8	92 ± 3.3 / 92 ± 3.1	96 ± 2.6 / 96 ± 2.2
R-CDT NS (Aug - none)	83 ± 12.2	93 ± 3.0	97 ± 2.7
Proposed (Aug - none)	87 ± 8.0	96 ± 3.1	98 ± 1.6

OAM (regular) dataset

	No. of training samples (per class)		
	1	5	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	5 ± 2.2 / 27 ± 3.3	44 ± 1.5 / 68 ± 1.6	54 ± 4.0 / 79 ± 3.2
VGG11 (Aug-1 / Aug-25 / Aug-50)	3 ± 0.2 / 17 ± 9.1	29 ± 16.1 / 26 ± 21.6	69 ± 7.3 / 88 ± 9.2
Resnet18 (Aug-1 / Aug-25 / Aug-50)	6 ± 1.3 / 46 ± 4.0	73 ± 2.4 / 82 ± 27.4	88 ± 2.9 / 95 ± 2.0
k-NN (Aug-1 / Aug-25 / Aug-50)	17 ± 1.1 / 19 ± 1.0	34 ± 1.5 / 39 ± 1.3	45 ± 1.3 / 51 ± 1.1
R-CDT NS (Aug - none)	81 ± 2.3	93 ± 0.7	96 ± 0.3
Proposed (Aug - none)	82 ± 2.0	94 ± 0.6	97 ± 0.5

OAM (out-of-distribution) dataset

	No. of training samples (per class)		
	1	5	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	5 ± 2.4 / 28 ± 4.2	39 ± 2.0 / 56 ± 2.8	44 ± 3.4 / 61 ± 3.9
VGG11 (Aug-1 / Aug-25 / Aug-50)	4 ± 2.1 / 16 ± 10.3	25 ± 14.1 / 41 ± 25.2	57 ± 9.1 / 75 ± 4.9
Resnet18 (Aug-1 / Aug-25 / Aug-50)	5 ± 2.0 / 46 ± 2.1	68 ± 1.1 / 59 ± 29.4	72 ± 4.1 / 79 ± 3.4
k-NN (Aug-1 / Aug-25 / Aug-50)	21 ± 1.4 / 22 ± 1.3	38 ± 0.9 / 42 ± 0.7	47 ± 0.7 / 51 ± 0.8
R-CDT NS (Aug - none)	86 ± 1.2	91 ± 0.4	93 ± 0.5
Proposed (Aug - none)	85 ± 1.1	91 ± 0.6	94 ± 0.5

FMNIST dataset

	No. of training samples (per class)		
	1	5	10
MNISTnet (Aug-1 / Aug-25 / Aug-50)	34 ± 6.1 / 33 ± 4.1	44 ± 7.6 / 51 ± 4.0	38 ± 8.0 / 59 ± 11.5
VGG11 (Aug-1 / Aug-25 / Aug-50)	33 ± 8.6 / 34 ± 9.9	32 ± 8.4 / 65 ± 16.0	42 ± 11.2 / 74 ± 13.5
Resnet18 (Aug-1 / Aug-25 / Aug-50)	26 ± 2.2 / 34 ± 7.0	32 ± 4.5 / 65 ± 21.7	37 ± 8.5 / 78 ± 8.8
k-NN (Aug-1 / Aug-25 / Aug-50)	26 ± 3.1 / 36 ± 4.1	39 ± 5.0 / 43 ± 3.9	47 ± 2.2 / 50 ± 1.9
R-CDT NS (Aug - none)	35 ± 3.6	54 ± 4.6	76 ± 3.9
Proposed (Aug - none)	58 ± 9.8	83 ± 2.2	86 ± 2.3

Chapter 3: Transport-based embeddings for classifying high dimensional distributions

We introduce a new method for classifying high-dimensional distributions with potential applications in a number of fields. Our method employs the Radon Cumulative Distribution Transform (R-CDT) and the Linear Optimal Transform (LOT) to represent high-dimensional data as a linear embedding that is more suitable for machine learning. High-dimensional data are challenging to model, especially for classification under spatial deformations. However, the transforms we introduce can handle these variations by providing linear embeddings, resulting in a convex data space that simplifies classification problems. By utilizing a nearest-subspace algorithm and general machine learning techniques in the transform space, we develop a new classification approach that is label-efficient, requires no hyper-parameter tuning, and offers a more efficient and effective approach to classifying high-dimensional distributions. Our approach achieves competitive accuracies compared to state-of-the-art methods in various classification problems, while also enhancing out-of-distribution generalization beyond test accuracy performances. Furthermore, our method is mathematically coherent, simple to implement, and can be effectively executed without the need for GPU acceleration.

3.1 Problem statement

To address the challenge of high-dimensional distribution modeling, we employ two distinct methodologies. The first approach entails extending the transport-based frameworks for 1D (signals) and 2D (images) distributions to N-dimensions using the N-dimensional Radon transform. The second methodology involves utilizing the Linear Optimal Transform (LOT) technique to model high-dimensional distributions in the form of point-sets.

Let us consider a general high-dimensional distribution $s(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^L$ and a high-dimensional point-set distribution $P_s := \frac{1}{N}\sum_{\mathbf{x}\in\Omega_s} \delta_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{s(i)}$. Let us also define one-to-one diffeomorphisms $g(\mathbf{x});\ \mathbf{x} \in \mathbb{R}^L$, for distributions, and $g^{\theta}(t);\ t \in \mathbb{R}$, when they need to be parameterized by a projection angle θ. The set of all possible one-to-one diffeomorphisms from $\mathbb{R}$ to $\mathbb{R}$ and from $\mathbb{R}^L$ to $\mathbb{R}^L$ are denoted as $\mathcal{T}$ and $\mathcal{T}_L$, respectively. Finally, $\mathscr{R}(\cdot)$ and $\mathscr{R}^{-1}(\cdot)$ denote the N-dimensional Radon transform and inverse Radon transform operators, respectively. The generative model stated below formalizes the definition of a class of N-dimensional distributions.

Generative model: Let, $\mathcal{G}_L \subset \mathcal{T}_L$ be a set of smooth one-to-one transformations. The mass preserving generative model for the k-th distribution class is defined to be the set

$$\mathbb{S}^{(k)} = \left\{ s_j^{(k)} | s_j^{(k)} = |\det Jg_j| \varphi^{(k)} \circ g_j,\ \forall g_j \in \mathcal{G}_L \right\} \tag{3.1}$$

where, $\varphi^{(k)}$ and s_j^k denote the template and the j-th distribution, respectively, from the k-th class and $\det Jg_j$ denotes the determinant of the Jacobian matrix of g_j. In the N-dimensional R-CDT setting, the equivalent sliced-projection representation of the generative model is given as

$$\mathbb{S}^{(k)} = \left\{ s_j^{(k)} | s_j^{(k)} = \mathscr{R}^{-1}\left(\left(g_j^{\theta} \right)' \widetilde{\varphi}^{(k)} \circ g_j^{\theta} \right),\ \forall g_j^{\theta} \in \mathcal{G} \right\} \tag{3.2}$$

where, $\widetilde{\varphi}^{(k)}$ denotes the N-dimensional Radon transform of the template $\varphi^{(k)}$.

In the LOT setting, the representation of the mass-preserving generative model for the k-th class is defined to be the set

$$\mathbb{S}^{(k)} = \left\{ P_{s_j^{(k)}} | P_{s_j^{(k)}} = g_{j\#} P_{\varphi^{(k)}},\ \forall g_j \in \mathcal{G}_L \right\} \tag{3.3}$$

where $P_{\varphi^{(k)}}$ corresponds to the point-set distribution of the prototype template pattern for the k-th class, while $P_{s_j^{(k)}}$ represents the point-set distribution of the j-th sample from the k-th class in $\mathbb{S}^{(k)}$. With these definitions, we can now construct a formal mathematical description for the generative model-based problem statement for point-set classification.

Classification problem: *Let the set of distribution classes $\mathbb{S}^{(k)}$ are given as above. Given training distribution samples $\{s_1^{(1)}, s_2^{(1)}, \cdots\}$ (class 1), $\{s_1^{(2)}, s_2^{(2)}, \cdots\}$ (class 2), $\cdots$ as training data, determine the class of an unknown distribution s.*

Note that the generative models described above are generally non-convex, which poses challenges to machine learning techniques to classify them effectively. Also the generative model in equation (3.3) describes set-structured point-set data, which makes it challenging to compare point-sets due to the permutation-invariant nature of a set. In the subsequent sections, we present solutions to the above classification problem using the Radon Cumulative Distribution Transform (R-CDT) and Linear Optimal Transport (LOT) embeddings. By approximating the resulting convex spaces with subspaces, as has been done in previous works on image [98, 106], signal [107, 108], and gradient distribution [104] classification problems, we obtain effective solutions to our problem. In addition, we evaluate a general machine learning classifier (penalized linear discriminant analysis classifier [5]) in the R-CDT embedding space for the classification task.

3.2 Proposed solution

Here we attempt to simplify the classification problem above using the transport-based embeddings. In the N-dimensional R-CDT setting we first obtained a set of projections of the distributions along a set of directions parameterized by θ in the high-dimensional space. Next, we obtained the cumulative distribution (CDT) [79] transform of the projections. We postulate that, the classification problem can be simplified by applying the N-dimensional R-CDT to the distribution. The transform space generative model then becomes

$$\widehat{\mathbb{S}}^{(k)} = \left\{ \widehat{s}_j^{(k)} | \widehat{s}_j^{(k)} = \left(g_j^{\theta}\right)^{-1} \circ \widehat{\varphi}^{(k)},\ \forall g_j^{\theta} \in \mathcal{G} \right\} \tag{3.4}$$

The CDT is a map from the space of smooth PDFs to the space of diffeomorphisms, which can be defined as the inverse function of the cumulation of each individual PDF. The CDT en-

hances linear separability in data by removing certain nonlinearities, renders data space convex, and simplifies the classification problem above [79]. The CDT is an invertible, one-to-one, and differentiable map, which enables us to interpret the trained model by visualizing the class differences obtained by the model. We propose to utilize the linear separability property of transport transform and employ linear classifiers in the high-dimensional sliced-transform space. The solution above can be utilized in solving several generic high-dimensional classification problems. Here we illustrate an example application of detecting COVID-19 using a distribution of platelet aggregate images. We propose to extend and improve this framework in solving imaging cytometry applications also.

In the LOT setting, it turned out the linearization ability of LOT is closely related to the scope of the following so-called composition property [109, 110]

$$T_\sigma^{g_\sharp \mu} = g \circ T_\sigma^\mu, \tag{3.5}$$

where $g \in \mathcal{T}_L$, and $\mathcal{T}_L$ is the set of all diffeomorphisms from $\mathbb{R}^L$ to $\mathbb{R}^L$. In particular, given a convex $\mathcal{G} \subseteq \mathcal{T}_L$, the LOT embedding of deformed measures via maps in $\mathcal{G}$ become convex [1] if all $g \in \mathcal{G}$ satisfies the above composition property (3.5), which is shown more formally below.

Proposition 3.2.1 (Lemma A.2 in [110])**.** *Let $\mathcal{G} \subseteq \mathcal{T}_L$ be convex. Given $\mu \in \mathcal{P}_2(\mathbb{R}^L)$, define $\mathcal{G}_\sharp \mu := \{g_\sharp \mu : g \in \mathcal{G}\}$. If $\forall g \in \mathcal{G}$,* (3.5) *holds, then $\widehat{\mathcal{G}_\sharp \mu} := \{\widehat{\nu} : \nu \in \mathcal{G}_\sharp \mu\}$ is convex in the LOT transform domain.*

When the dimension $L \geq 2$, it is shown in [109] that g can only be "basic" transformations (more specifically, translations or isotropic scalings or their compositions) for the composition property (3.5) to hold for arbitrary μ's. Luckily, [110] proposes an approximate composition property for perturbations of the aforementioned basic transformations, the set of which we denote as $\mathcal{A} = \{h(x) = ax + b : a > 0, b \in \mathbb{R}^L\}$.

[1] Note in general $\mathcal{G}_\sharp \mu$ is not convex as $(\lambda_1 g_1 + \lambda_2 g_2)_\sharp \mu \neq \lambda_1 (g_1)_\sharp \mu + \lambda_2 (g_2)_\sharp \mu$.

Property 1 (Approximate composition, p.388 in [110][2]) Let $\epsilon \geq 0$ and $\mu \in \mathcal{P}_2(\mathbb{R}^L)$. Let $g \in \mathcal{T}_L$ such that $\|g - h\| \leq \epsilon$ for some $h \in \mathcal{A}$. Then there exists some δ such that

$$\|T_\sigma^{g_\sharp \mu} - g \circ T_\sigma^\mu\|_\sigma < \delta, \tag{3.6}$$

Remark: Using the $\widehat{\mu}$ notation for LOT transform of μ, we have

$$\|\widehat{g_\sharp \mu} - g \circ \widehat{\mu}\|_\sigma < \delta. \tag{3.7}$$

With the above approximate composition property, one can show the following approximate convexity analog of Proposition 3.2.1 using Lemma A.3, A.4 of [110]:

Proposition 3.2.2. *Let $\epsilon \geq 0$ and $\mathcal{G} \subseteq \mathcal{T}_L$ be convex such that for any $g \in \mathcal{G}$, there exists some $h \in \mathcal{A}$ such that $\|g - h\| \leq \epsilon$. Given $\mu \in \mathcal{P}_2(\mathbb{R}^L)$, we have $\widehat{\mathcal{G}_\sharp \mu} := \{\widehat{\nu} : \nu \in \mathcal{G}_\sharp \mu\}$ is 2δ-convex in the LOT transform domain, where δ is given in the above approximate composition property. In particular, for any $c \in [0, 1]$ and $\widehat{g_{1\sharp}\mu}, \widehat{g_{2\sharp}\mu} \in \widehat{\mathcal{G}_\sharp \mu}$ $(g_1, g_2 \in \mathcal{G})$,*

$$\|(1 - c)\widehat{g_{1\sharp}\mu} + c\widehat{g_{2\sharp}\mu} - \widehat{g_{c\sharp}\mu}\| < 2\delta, \tag{3.8}$$

where $g_c = (1 - c)g_1 + cg_2 \in \mathcal{G}$.

The LOT transform, which was previously described in Chapter 1, can significantly simplify the classification problem described earlier by providing a convex linear embedding for the set-structured point-set data. Let us first investigate the generative model in equation (3.3) in the LOT transform space.

$$\widehat{\mathbb{S}}^{(k)} = \left\{ \widehat{P}_{s_j^{(k)}} | \widehat{P}_{s_j^{(k)}} = g_{j\#} \widehat{P}_{\varphi^{(k)}},\ \forall g_j \in \mathcal{G}_L \right\} \tag{3.9}$$

In this context, $\widehat{P}s_j^{(k)}$ and $\widehat{P}\varphi^{(k)}$ refer to the LOT embedding of $P_{s_j^{(k)}}$ and $P_{\varphi^{(k)}}$, respectively, with

[2]This property is referred as δ-compatibility in [110].

respect to a reference structure P_r (see equation (1.15)). Based on the preliminary results presented above (Property 1, Proposition 3.2.2, and other results), it is possible to establish the convexity of the set $\widehat{\mathbb{S}}^{(k)}$ up to a certain bound, subject to certain constraints. Furthermore, we can show that when $\mathbb{S}^{(k)} \cap \mathbb{S}^{(p)} = \varnothing$, the intersection of $\widehat{\mathbb{S}}^{(k)}$ with $\widehat{\mathbb{S}}^{(p)}$ is empty. We use a standard machine learning pipeline with the standard training and testing procedures in the R-CDT setting, and the experimental details for this setting can be found in the Results section. Next, we describe the specific training and testing procedure that we adopted in the LOT-setting.

3.2.1 Training method in the LOT space

Based on the aforementioned theoretical discussions, we put forward a straightforward non-iterative training approach for the classification method. This involves computing a projection matrix that maps each sample in the LOT space onto the subspace $\widehat{\mathbb{V}}^{(k)}$ (as outlined in [98]), generated by the 2δ-convex set $\widehat{\mathbb{S}}^{(k)}$. Specifically, we estimate the projection matrix by applying the following procedure:

$$\widehat{\mathbb{V}}^{(k)} = \text{span}\left(\widehat{\mathbb{S}}^{(k)}\right) = \{\sum_{j\in J} \alpha_j \widehat{P}_{s_j^{(k)}} | \alpha_j \in \mathbb{R} \text{ is finite}\}. \tag{3.10}$$

Subject a given set of sample training data, denoted as $\{P_{s_1^{(k)}}, P_{s_2^{(k)}}, \cdots\}$, the first step in our proposed method is to apply a LOT transformation on them using a reference point $P_{r^{(k)}}$. This results in the generation of transformed samples, denoted as $\{\widehat{P}_{s_1^{(k)}}, \widehat{P}_{s_2^{(k)}}, \cdots\}$. The reference point $P_{r^{(k)}}$ is obtained by selecting a point-set at random from the training set, followed by the introduction of random perturbations. Subsequently, we estimate $\widehat{\mathbb{V}}^{(k)}$ using the following method:

$$\widehat{\mathbb{V}}^{(k)} = \text{span}\{\widehat{P}_{s_1^{(k)}}, \widehat{P}_{s_2^{(k)}}, \cdots\}. \tag{3.11}$$

The proposed method also provides a structure to mathematically encode invariances with respect to deformations that are known to be present in the data. In this chapter, we prescribe methods to encode invariances with respect to a set of affine transformations: translation, isotropic

and anisotropic scaling, and shear. Let a sample point-set in LOT space is given by $\widehat{s}(i) \in \mathbb{R}^L = ((\widehat{s}_1(1), \widehat{s}_2(1), \cdots, L\text{-terms}), (\widehat{s}_1(2), \widehat{s}_2(2), \cdots, L\text{-terms}), \cdots, N\text{-terms})$. First, the LOT space sample point-set is centered by subtracting the means over each of the L dimensions of the point-set. Next, deformation spanning sets [98, 106] are defined for each of the sample point-set. Detailed descriptions of the deformation types used for encoding invariances and the corresponding methodologies are explained as follows:

1. Translation: Let $g(\mathbf{x}) = \mathbf{x} - \mathbf{x_0}$ be the translation by $\mathbf{x}_0 \in \mathbb{R}^L$. The spanning set for translation is defined as $\mathbb{U}_T = \{\mathbb{U}_{T_1} \cup \mathbb{U}_{T_2} \cup \cdots \cup \mathbb{U}_{T_L}\}$, where $\mathbb{U}_{T_1} = ((1, 0, 0, \cdots), (1, 0, 0, \cdots), \cdots)$, $\mathbb{U}_{T_2} = ((0, 1, 0, \cdots), (0, 1, 0, \cdots), \cdots)$, and other relevant elements.

2. Isotropic scaling: Let $g(\mathbf{x}) = a\mathbf{x}$ be the normalized dilatation of s by a, where $a \in \mathbb{R}_+$. An additional spanning set for isotropic scaling is not required as the subspace containing $\widehat{s}_j^{(k)}$ naturally contains its scalar multiplication. The spanning set for isotropic is defined as $\mathbb{U}_{D_0} = \varnothing$.

3. Anisotropic scaling: Let $g(\mathbf{x}) = \breve{\mathcal{D}}\mathbf{x}$ with $\breve{\mathcal{D}} = \begin{bmatrix} 1/a_1, & 0, & \cdots \\ 0, & 1/a_2, & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$ be the normalized anisotropic dilatation of s, where $a_i \neq a_j$ and $a_i \in \mathbb{R}_+$. The spanning set for anisotropic scaling is defined as $\mathbb{U}_D = \{\mathbb{U}_{D_1} \cup \mathbb{U}_{D_2} \cup \cdots \cup \mathbb{U}_{D_L}\}$, where $\mathbb{U}_{D_1} = ((\widehat{s}_1(1), 0, 0, \cdots), (\widehat{s}_1(2), 0, 0, \cdots), \cdots)$, $\mathbb{U}_{D_2} = ((0, \widehat{s}_2(1), 0, \cdots), (0, \widehat{s}_2(2), 0, \cdots), \cdots)$, and other elements.

4. Shear: Let $g(\mathbf{x}) = \mathcal{H}\mathbf{x}$ with $\mathcal{H} = \begin{bmatrix} 1, & k, & \cdots \\ 0, & 1, & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$, be the normalized shear of s by k, where the shear matrix, $\mathcal{H}$, can also be constructed with the shear factor, k_i, located at other non-diagonal positions within the matrix and $k \in \mathbb{R}$. The spanning set for shear is defined as $\mathbb{U}_S = \{\mathbb{U}_{S_{1,1}} \cup \mathbb{U}_{S_{2,1}} \cup \cdots \cup \mathbb{U}_{S_{L,L-1}}\}$, where $\mathbb{U}_{S_{1,1}} = ((\widehat{s}_2(1), 0, 0, \cdots), (\widehat{s}_2(2), 0, 0, \cdots), \cdots)$, $\mathbb{U}_{S_{1,2}} = ((\widehat{s}_3(1), 0, 0, \cdots), (\widehat{s}_3(2), 0, 0, \cdots), \cdots)$, $\mathbb{U}_{S_{2,1}} = ((0, \widehat{s}_1(1), 0, \cdots), (0, \widehat{s}_1(2), 0, \cdots), \cdots)$, and others.

Finally, we can approximate $\widehat{\mathbb{V}}^{(k)}$ as follows:

$$\widehat{\mathbb{V}}^{(k)} = \text{span}\left(\{\widehat{P}_{s_1^{(k)}}, \widehat{P}_{s_2^{(k)}}, \cdots\} \cup \mathbb{U}_A\right), \tag{3.12}$$

where $\mathbb{U}_A = \mathbb{U}_T \cup \mathbb{U}_{D_0} \cup \mathbb{U}_D \cup \mathbb{U}_S$.

3.2.2 Testing method in the LOT space

To classify a given test sample P_s, we first apply the LOT transform to P_s to obtain its corresponding LOT space representation $\widehat{P}_{s,r^{(k)}}$ with respect to the reference $P_{r^{(k)}}$ (which was preselected duing the training phase). Assuming that the test samples originate from the generative model presented in equation (3.3) (or equation (3.9)), we can determine the class of an unknown test sample P_s using the following expression:

$$\arg\min_k d^2\left(\widehat{P}_{s,r^{(k)}}, \widehat{\mathbb{V}}^{(k)}\right) \tag{3.13}$$

where $d(\cdot,\cdot)$ is the distance between the test sample and the trained subspaces in the LOT transform space. We can estimate the distance between $\widehat{P}_{s,r^{(k)}}$ and the trained subspaces using $d^2\left(\widehat{P}_{s,r^{(k)}}, \widehat{\mathbb{V}}^{(k)}\right) \sim ||\widehat{P}_{s,r^{(k)}} - B^{(k)}B^{(k)T}\widehat{P}_{s,r^{(k)}}||^2_{L_2}$, where the matrix $B^{(k)}$ contains the basis vectors of the subspaces $\widehat{\mathbb{V}}^{(k)}$ arranged in its columns.

3.3 Results

3.3.1 Experimental setup

Our objective is to analyze how the proposed method performs compared to state-of-the-art approaches in terms of performance metrics such as classification accuracy, required training data, and robustness in out-of-distribution scenarios in limited training data setting. To achieve this, we created train-test splits of varying sizes from the original training set for each dataset under examination. We then trained the models using these splits and assessed their performance on

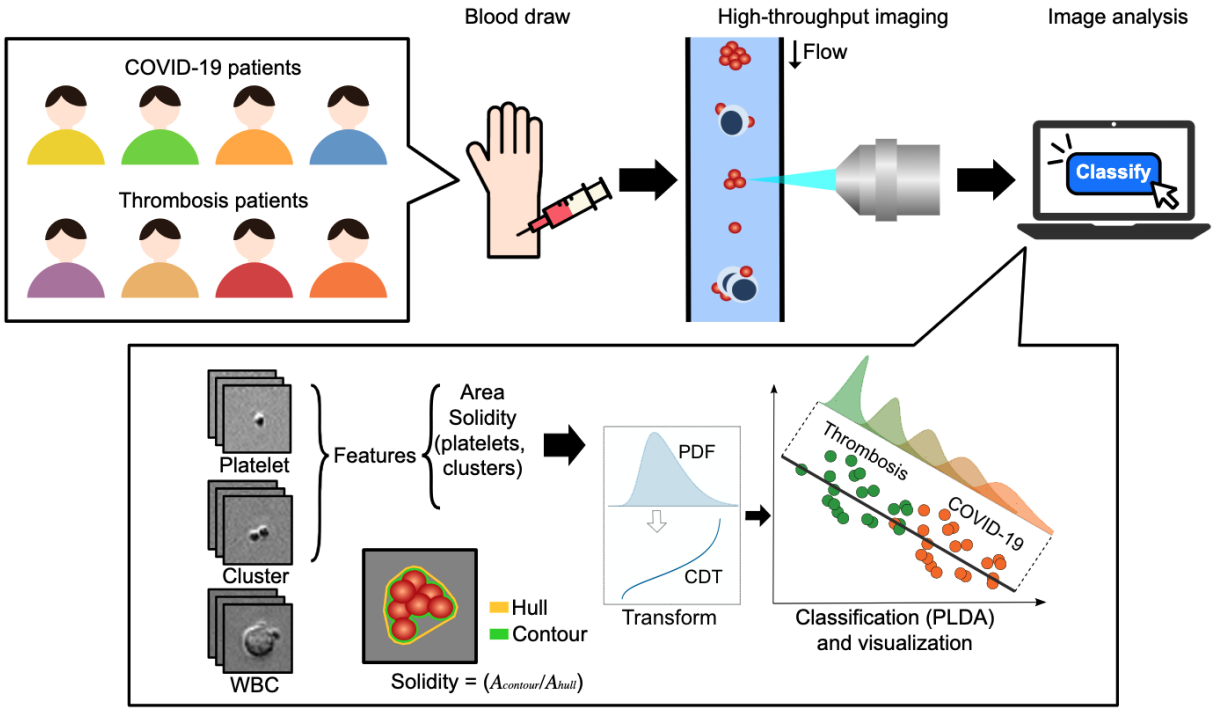

Figure 3.1: Conceptual schematic diagram of the COVID-19 detection workflow, including sample preparation, high-throughput imaging flow cytometry (IFC) measurement, computation of morphological features (area and solidity), and transport-based disease classification technique using N-dimensional R-CDT and PLDA.

the original test set. For our experiments with the LOT-based approach, we generated each train split by randomly selecting (without replacement) samples from the original training set. To ensure statistical significance, we repeated the experiments for each train split size ten times. The test split of our dataset consisted of the complete test dataset, which ensured a comprehensive evaluation of the proposed method. For our experiments using the R-CDT-based approach, we first selected 10 random measurements of both COVID-19 and non-COVID-19 thrombosis to create a testing dataset. Subsequently, we utilized the remaining 91 non-COVID-19 thrombosis and 171 COVID-19 measurements to train the classification model. To ensure the robustness of our approach, we repeated this process 1000 times and reported the mean results. The same train-test data samples were used for all algorithms in each split.

In order to assess the effectiveness of the LOT-based approach, we utilized several comparison methods. These included PointNet [111], DGCNN [112], and MLP in FSpool feature embedding

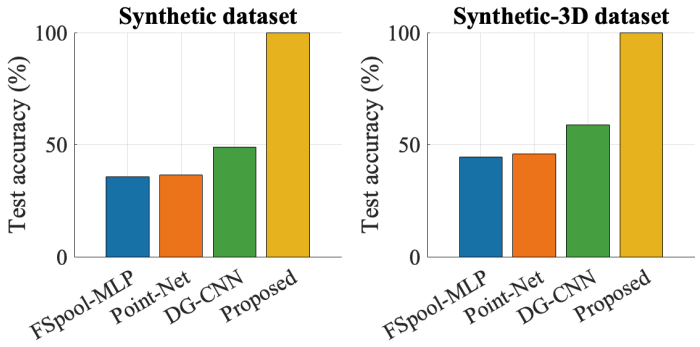

Figure 3.2: Performance comparison of various methods on synthetic datasets in terms of percentage test accuracy.

space [113]. We also conducted a comparative analysis with various conventional machine learning techniques across different set feature embedding spaces. These included logistic regression (LR), kernel support vector machine (k-SVM), multilayer perceptron (MLP), and nearest subspace (NS) classifier models [114] in GeM1, GeM2, GeM4 [115], COVpool [116, 117], and FSpool [113] embedding spaces. The performance of the proposed method was evaluated in relation to these baselines. We conducted these evaluations in addition to performing out-of-distribution experiments. In the proposed method, we selected the number of basis vectors for the subspaces $\widehat{\mathbb{V}}^{(k)}$ such that the total variance explained by the chosen basis vectors in the k-th class captured up to 99% of the total variance explained by that class.

To assess the relative performance of the methods, we evaluated them on several datasets, including Point cloud MNIST [118, 8], ModelNet [119], and ShapeNet [120] datasets. We additionally applied random translations, anisotropic scaling, and shear transformations to both the training and test sets of the datasets. For the ShapeNet dataset, we tested the methods under two experimental setups: the regular setup, where both the training and test sets contained point-sets at the same deformation level, and the out-of-distribution setup, where the training and test sets contained point-sets at different deformation levels.

In our R-CDT-based approach, we utilized the solution above to distinguish the COVID-19 patients from non-COVID-19 thrombosis patients using the images of platelet aggregates. Each patient corresponds to a set of platelet aggregate images and can be characterized by a

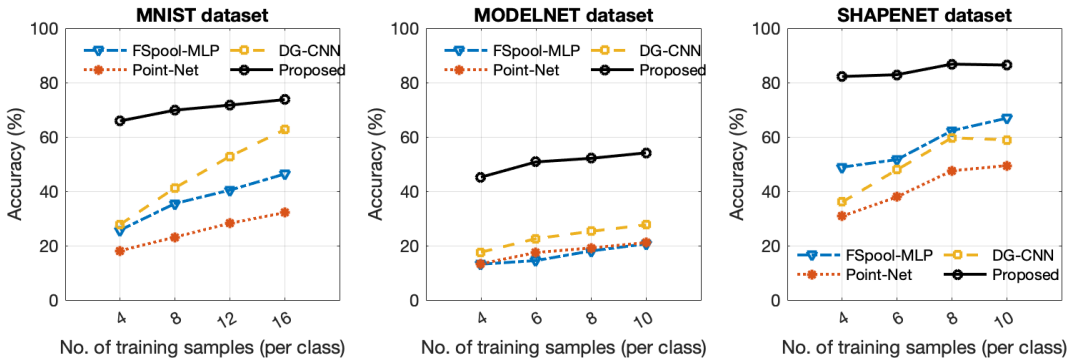

Figure 3.3: The relationship between the accuracy of different methods and the number of training samples evaluated on MNIST, ModelNet, and ShapeNet datasets.

high-dimensional distribution. Let us consider $\{x_1, x_2, \cdots\}_m^{(k)}$, $\{y_1, y_2, \cdots\}_m^{(k)}$, $\{z_1, z_2, \cdots\}_m^{(k)}$, and $\{w_1, w_2, \cdots\}_m^{(k)}$ to be the sets of the morphological feature measurements (here, $N = 4$) corresponding to the m-th subject of the k-th disease class (non-COVID-19 thrombosis/COVID-19). Here, x, y, z, and w denote the area of platelets, the solidity of platelets, the area of platelet clusters, and the solidity of platelet clusters, respectively. Let us also consider $\{s(x), s(y), s(z), s(w)\}_m^{(k)}$ to be the projections (sliced along the canonical axes) of the high-dimensional probability density functions (PDF) obtained from the morphological feature measurements using a kernel density estimation technique. The goal of the classification problem is to determine the class of a test set $\{s(x), s(y), s(z), s(w)\}$ corresponding to a subject with an unknown diagnosis. The first step is to obtain the transformed versions of the projections of the high-dimensional PDFs, denoted as $\{\widehat{s}(x), \widehat{s}(y), \widehat{s}(z), \widehat{s}(w)\}_m^{(k)}$, using the cumulative distribution transform (CDT) [79]. After CDT transformation, we employed principal component analysis to reduce data dimensionality using scikit-learn (v. 1.0.2). For classification, we employed the penalized linear discriminant analysis (PLDA) classifier [5], which differentiates between the classes of a given dataset by obtaining the most discriminant directions computed based on Fisher's linear discriminant in combination with penalized least-squares regression. We used the Python package PyTransKit (v. 0.2.3) to compute the CDTs and train the PLDA classifier and the Python package statsmodels (v. 0.13.2) to obtain the PDFs using the kernel density estimation technique. Fig. 3.1 shows a conceptual schematic diagram of the method.

3.3.2 Accuracy in synthetic case

We first evaluated the effectiveness of the proposed method by comparing it with other state-of-the-art techniques on two synthetic datasets. The synthetic datasets were generated by selecting one sample per class from the point cloud MNIST and ShapeNet datasets, followed by introducing random translations, anisotropic scaling, and shear transformations to each selected sample to generate training and test sets. Specifically, the training set consisted of two samples per class, while the test set comprised 25 samples per class. The obtained comparative results are displayed in Fig. 3.2. As observed, the proposed method substantially outperformed the other methods in this synthetic scenario.

3.3.3 Accuracy and efficiency in real datasets

We conducted the performance evaluation of the proposed method by comparing it with several state-of-the-art techniques, including PointNet, DGCNN, and MLP in FSpool feature embedding space, on the MNIST, ShapeNet, and ModelNet datasets. Fig. 3.3 presents the average test accuracy values obtained for different numbers of training samples per class. The results demonstrate that our proposed method outperformed the other methods across the range of training sample sizes used to train the models. Notably, the proposed method's accuracy vs. training size curves exhibited a smoother trend in most cases compared to the other methods.

3.3.4 Out-of-distribution robustness

To assess the effectiveness of the proposed method under the out-of-distribution setting, we introduced a gap between the magnitudes of deformations in the training and test sets. Specifically, we used $\mathcal{G}_{out}$ as the deformation set for the 'out-distribution' test set, while $\mathcal{G}_{in}$ was the deformation set for the 'in-distribution' training set. We trained the models using the 'in-distribution' data and tested using the 'out-distribution' data. For our out-of-distribution experiment, we used the ShapeNet dataset with small deformations as the 'in-distribution' training set and the ShapeNet dataset with larger deformations as the 'out-distribution' test set (see Fig. 3.4). The results showed

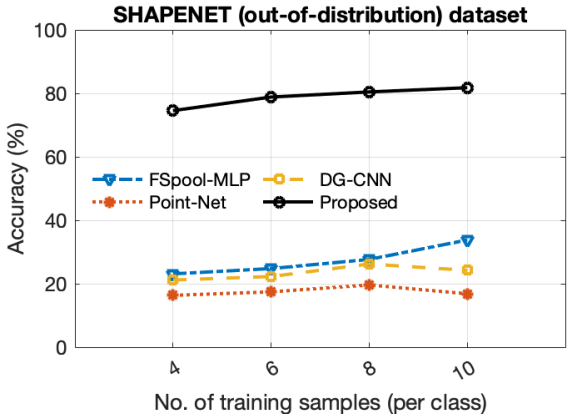

Figure 3.4: Performance assessment under an out-of-distribution experimental setup with non-overlapping training and test sets and varying degrees of spatial transformations. The accuracy of the methods was evaluated as a percentage of test accuracy and plotted against the number of training images per class.

that the proposed method outperformed the other methods by an even more significant margin under the challenging out-of-distribution setup, as shown in Fig. 3.4. Under this setup, the proposed method obtained accuracy figures closer to that in the standard experimental setup (i.e., ShapeNet in Fig. 3.3). On the other hand, the accuracy of the other methods declined significantly under the out-of-distribution setup compared to the standard experimental setup (see ShapeNet results in Figs. 3.3 and 3.4).

3.3.5 Comparison with set-embedding-based methods

We further evaluated the proposed method against various set embedding-based approaches in combination with classical machine learning methods. The study involved comparing the proposed method with different classifier techniques, including LR, k-SVM, MLP, and NS [114], that were employed with various set-to-vector embedding methods, such as GeM (1,2,4) [115], COVpool [116, 117], and FSpool [113]. Fig. 3.5 illustrates the percentage test accuracy results obtained from these modified experiments, along with the results of the proposed method for comparison. As shown in Fig. 3.5, the proposed method outperformed all these models in terms of test accuracy.

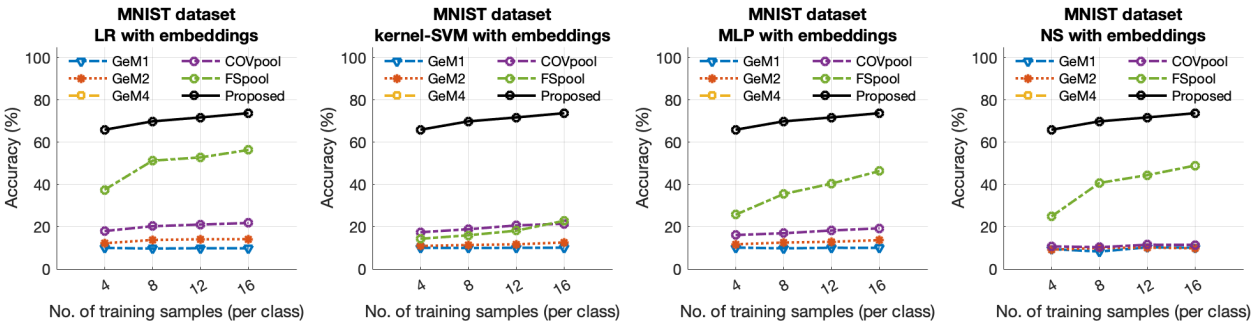

Figure 3.5: Comparative analysis of the percentage test accuracy results achieved by the proposed method and conventional machine learning techniques implemented across various feature embedding spaces.

3.3.6 Performance of the R-CDT-based approach

We compared our R-CDT-based method with a random forests-based (RF) classification method. Normalized confusion matrix shows the disease classification performance of the proposed and the Random Forest models on the on the testing dataset (average of 1000 iterations; see Fig. 3.6(b)). The proposed model reached an average testing accuracy of 75.79% whereas the random forests based model reached an average testing accuracy of 69.92%. To interpret the CDT-PLDA model, we used one model that was trained on the entire dataset and plotted the distribution profiles of the canonical projections along the most significant direction, as shown in Fig. 3.6(d). Blue and magenta curves indicate distribution shapes that are typical for non-COVID-19 thrombosis and COVID, respectively. Scatter plot of representative measurements are shown in Fig. 3.6(a). Fig. 3.6(c) shows the distribution of the feature with highest feature importance. Stars indicate statistical significance, determined via two-sided t-test ($p < 10^{-4}$). To obtain the PDF, we performed a kernel density estimation. The distributions were sampled over a uniform grid of $M = 5000$ points. Next, the PDF was transformed using CDT, which returns a feature vector of length M. To reduce data dimensionality, we used principal component analysis and selected the principal components such that the sum of variance explained was 99% of the total variance.

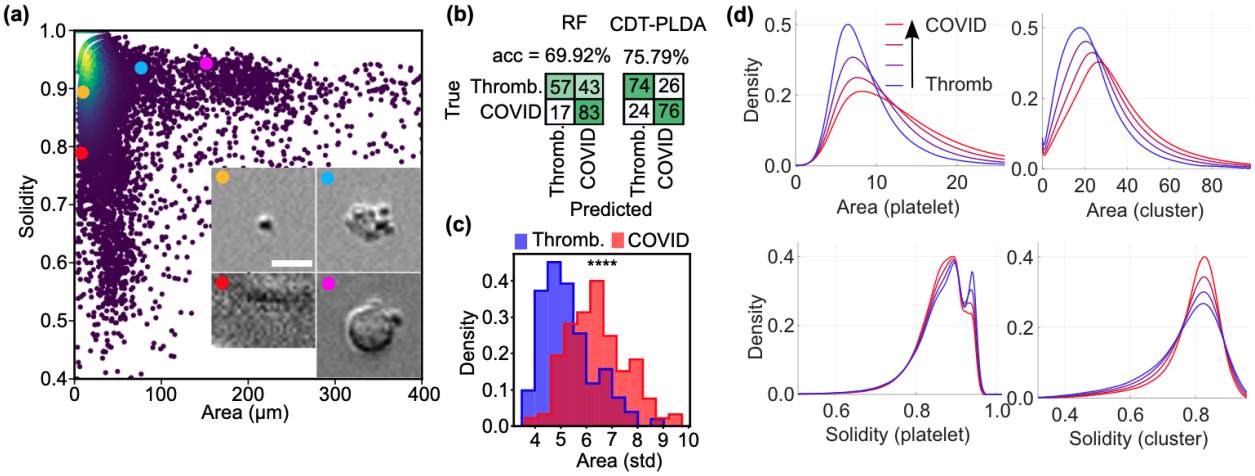

Figure 3.6: (a) Scatter plot of representative measurements. (b) The disease classification performance of Random Forest and the proposed models on the on the testing dataset (average of 1000 iterations). (c) The distribution of the feature with highest feature importance (standard deviation of area distribution). (d) Reconstructed distribution profiles along the most discriminant direction in the transport space.

3.4 Discussion

This dissertation presents a new method for classifying high-dimensional distributions using Radon-Cumulative Distribution Transform and Linear Optimal Transport Transform based models. Our method is appropriate for problems where the data at hand can be suitably represented as an instance of prototype template high-dimensional distribution patterns under the effect of smooth, nonlinear, and one-to-one transformations. Our results demonstrate that our method offers an effective and data-efficient solution for a wide range of high-dimensional distribution classification tasks, with competitive accuracy compared to current state-of-the-art techniques. Furthermore, our approach performs well even in challenging practical scenarios, such as out-of-distribution situations.

The outcomes achieved with various example datasets indicate that our proposed approach can deliver accuracy results comparable to state-of-the-art methods, provided that the data adheres to the generative model specified in equation (3.3). Additionally, our approach was shown to be more data-efficient in some cases, meaning that it can attain higher accuracy levels using fewer training samples.

Our proposed method maintains high classification accuracy, even in challenging out-of-distribution experimental conditions, as depicted in Fig. 3.4, whereas the accuracy figures of other methods decline sharply. These results indicate that our method provides a better overall representation of the underlying data distribution, resulting in robust classification performance. The key to achieving better accuracy under out-of-distribution conditions is that our method not only learns the deformations present in the data but also learns the underlying data model, including the type of deformation, such as translation, scaling, and shear, and their respective magnitudes. This deformation type can be learned from just a few training samples containing those deformations, as well as potentially from the mathematically prescribed invariances proposed in [106].

Our LOT-based approach, which utilizes the nearest subspace classifier in the LOT domain, is more suitable for classification problems in the above category compared with general set embedding methods in combination with classical machine learning classifiers, as demonstrated by its superior performance. Typically, point-set data classes in their original domain do not constitute embeddings, and commonly used set-to-vector representation techniques are inadequate in generating effective embeddings for them, as indicated by the results. This presents a significant challenge for any machine learning approach to perform effectively. However, the subspace model is appropriate in the LOT domain since the LOT transform provides a linear embedding and convex data geometry. Moreover, considering the subspace model in the LOT space improves the generative nature of our proposed classification method by implicitly including the data points from the convex combination of the provided training data points.

3.5 Conclusions

In this dissertation, we propose an enhanced end-to-end classification system designed for a specific category of high-dimensional distribution classification problems, where data classes are considered as instances of a template observed under a set of spatial deformations. If these deformations are appropriately modeled as a collection of smooth, one-to-one, and nonlinear transformations, then the data classes become easily separable in the transform space due to the properties

outlined in the dissertation. These properties sometimes enable the approximation of data classes as convex subspaces, resulting in a more suitable data model for the nearest subspace method. As we observed in our experiments, this approach yields high accuracy and robustness against out-of-distribution conditions. Numerous high-dimensional distribution classification problems can be formulated in this way, and therefore, our proposed solution has broad applicability.

Finally, we note that there can be many potential adaptations of the proposed method. For instance, the linear subspace method in the presented LOT space could be adjusted to incorporate alternative assumptions regarding the set that best represents each class. While some problems might benefit from a linear subspace method similar to the one described earlier, where all linear combinations are allowed, other problems may be require constraining the model using linear convex hulls. Additionally, investigating the sliced-Wasserstein distance using discrete CDT transform more elaborately (as proposed in [104]) in conjunction with subspace models is another promising avenue for future research.

One major component of the R-CDT based approach is obtaining a set of 1D projections of the high-dimensional distributions to apply the CDT transform on them. However, the choice of the projection directions is a open problem and might affect the method's performance. Currently, we proposed to obtain projections along the canonical axis directions. However, further exploration regarding the directions of projection might be necessary. High dimensional distribution analysis is computationally intensive and the computational complexity grows with the dimensionality of the distributions. Finally, we used morphological numerical features (area, solidity) for the current analysis. Utilization of transport-based features might also benefit the analysis.

Both our R-CDT and LOT-based approaches have yielded encouraging results in classifying high-dimensional distributions, laying a promising foundation for further exploration in this field. As the amount of 3D (or N-D) data continues to increase and accurate object recognition and scene understanding become more crucial, we believe that the integration of transport-based embeddings and mathematical modeling techniques such as subspaces, convex hulls, or machine learning classifiers in the transform space will play an increasingly significant role in the classification of high-

dimensional distributions. We anticipate that our proposed method will inspire further research in this direction and lead to novel developments in recognizing 3D (or N-D) objects or distributions.

Chapter 4: Quantifying nuclear structures of digital pathology images across cancers using transport-based morphometry

Alterations in nuclear morphology are useful adjuncts and even diagnostic tools used by pathologists in the diagnosis and grading of many tumors, particularly malignant tumors. Large datasets such as TCGA and the Human Protein Atlas, in combination with emerging machine learning and statistical modeling methods, such as feature extraction and deep learning techniques, can be used to extract meaningful knowledge from images of nuclei, particularly from cancerous tumors. Here we describe a new technique based on the mathematics of optimal transport for modeling the information content related to nuclear chromatin structure directly from imaging data. In contrast to other techniques, our method represents the entire information content of each nucleus relative to a template nucleus using a transport-based morphometry (TBM) framework. We demonstrate that the model is robust to different staining patterns and imaging protocols, and can be used to discover meaningful and interpretable information within and across datasets and cancer types. In particular, we demonstrate morphological differences capable of distinguishing nuclear features along the spectrum from benign to malignant categories of tumors across different cancer tissue types, including tumors derived from liver parenchyma, thyroid gland, lung mesothelium, and skin epithelium. We believe these proof of concept calculations demonstrate that the TBM framework can provide the quantitative measurements necessary for performing meaningful comparisons across a wide range of datasets and cancer types that can potentially enable numerous cancer studies, technologies, and clinical applications and help elevate the role of nuclear morphometry into a more quantitative science.

4.1 Problem insights

Image data of nuclei are obtained from physical tissue specimens with the aid of microscopes. Continuum mechanics can be used as a mathematical model for these images. Nuclei transform themselves during carcinogenesis according to laws expressed in partial differential equations, such as the continuity equation [72]. In other words, nuclear morphological alterations can be mathematically described as a continuous process of rearrangement of chromatin structures under the effect of biological processes. Consider the problem of modeling nuclear morphological alterations that occur during malignant transformation in a segmented nuclei image dataset (see Fig. 4.1). These morphological alterations can be the measurement of the rearrangement of chromatin structures, estimated as changes in the intensity measurements within nuclei images.

Given segmented nuclei image data, our goal here is to describe an approach to quantify nuclear structural changes that occur during malignant transformation that is robust to differences in staining pattern and imaging procedures. We then utilize this approach to synthesize data across multiple cancer tissue types to obtain nuclear morphological features of malignancy that are shared among cancers as a proof of concept calculation such that, like in genomics and proteomics, nuclear structure information can be used more generally than it is now. We utilize automatically segmented nuclei from histopathological images obtained from four tissue types (liver parenchyma, thyroid gland, lung mesothelium, and skin epithelium [121]) each imaged at different resolutions and with varied staining procedures (Feulgen, Diff-Quik, and Hematoxylin and Eosin). Each dataset contained specimens from two different histological cancer grades, which we assigned to the following classes: benign (or preneoplastic) and malignant. (see Fig. 4.2). Details of the definitions used for each class are presented in Appendix 4.5.1.

4.2 Proposed approach

Coupled with the concept of continuum mechanics mathematics to represent changes in nuclei images (as described in Section 4.1) and the principles of the optimal energy solution (i.e., optimal transport theory) [72], we propose an optimal mass transport-based approach to modeling nuclear morphological changes in malignancy whereby mass is represented as the image intensity [72]. With the notion of a reference image (e.g., prototype nucleus), we can apply optimal transport mathematics to represent the rearrangement of the intensity measurements of nuclear chromatin structures in a physically meaningful way. Let $s(x), x \in [0,1]^2$ represent an image of a segmented nucleus, which we model as a function $s : [0,1]^2 \rightarrow \mathbb{R}_+$. As commonly assumed in transmission and fluorescence microscopy, after appropriate preprocessing (see Appendix 4.5.2) the intensity $s(x)$ is approximately proportional to the amount of mass (in our case chromatin) present at pixel location x [122]. As the proportionality (calibration) constant is typically unknown in most routine clinical imaging procedures, we resort to normalizing it out of our analysis. That is, instead of analyzing each segmented nucleus $s(x)$ directly, given the absence of intensity calibration, we instead analyze $s(x)/\int_{[0,1]^2} s(x)dx$. Henceforth, when assume all images being analyzed have been normalized so they integrate (sum after discretization) to 1.

Now consider two nuclear images $s_1(x), s_0(y)$, with $x, y \in [0,1]^2$. We can define the "effort" (cost) of transporting normalized intensity $s_1(x)$ from location x to location y as $(x-y)^2 s_0(y)$ in units of $normalized\ intensity \times m^2$. Given a function that maps each coordinate from s_0 to s_1, $f(y) = x$, such that the entire normalized mass s_0 is transported to match s_1 we can define the total cost in re-arranging the normalized chromatin content from $s_1(x)$ onto $s_0(y)$ as

$$\int_{[0,1]^2} (f(y) - y)^2 s_0(y) dy \tag{4.1}$$

where the units are once again $normalized\ intensity \times m^2$. We refer to functions f that re-arrange chromatin content from s_0 to s_1 as mass preserving (MP) mappings. Using the theory of optimal transport [72] we can thus establish a quantitative metric (Wasserstein distance) that compares the

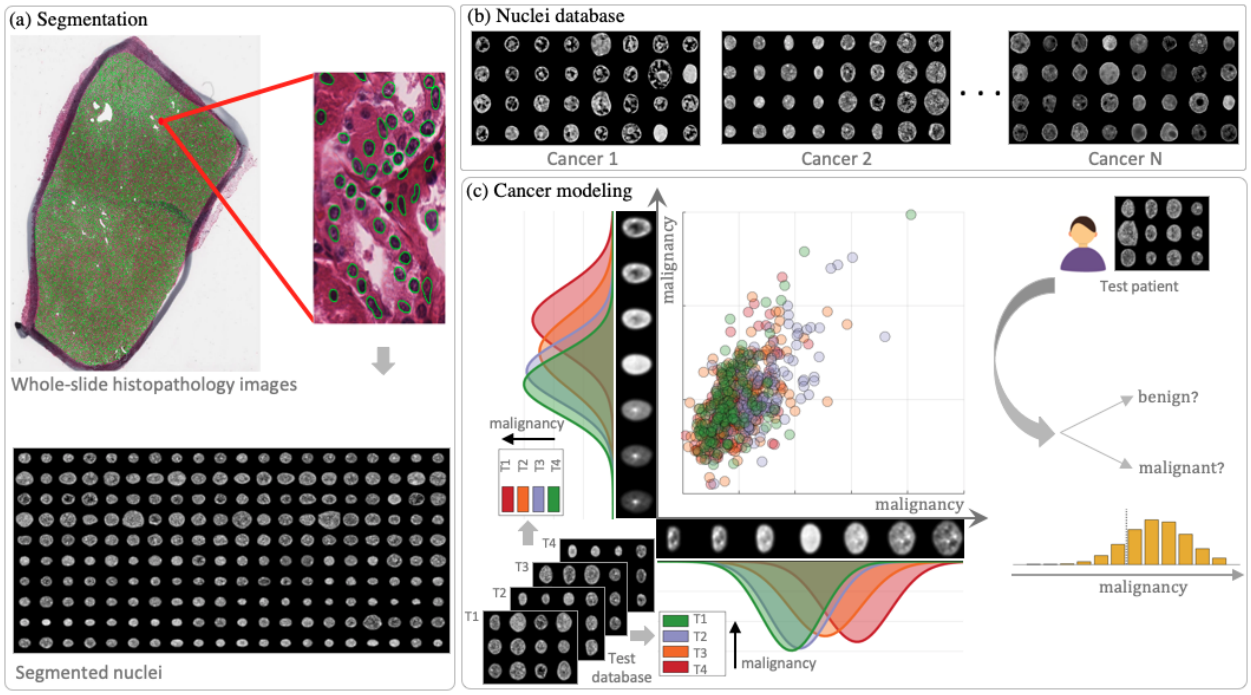

Figure 4.1: System diagram outlining the proposed cancer modeling approach. (a) Image segmentation techniques afford the ability to obtain a large-scale database of segmented nuclei from whole-slide histopathology images. (b) The proposed method takes segmented nuclei datasets obtained from various tissue types as inputs. (c) The proposed cancer modeling approach performs a joint regression in the transport space. The model can be used to visualize a specific feature, obtain malignancy potential rankings within a subset of tissue types, and classify patients, among other potential applications.

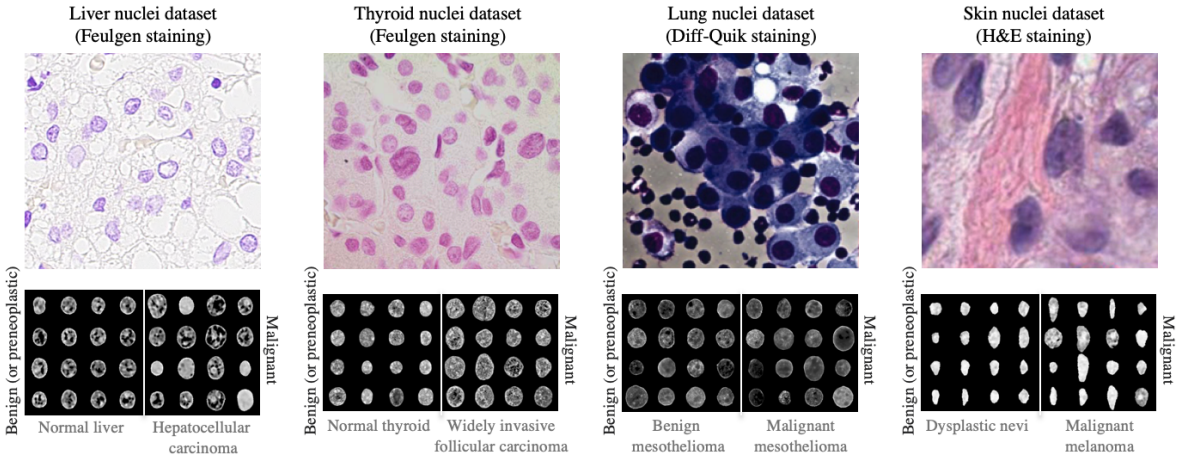

Figure 4.2: Sample nuclei from digital pathology images obtained from four tissue types: liver parenchyma, thyroid gland, lung mesothelium, and skin epithelium.

entirety of the normalized chromatin content between two nuclear images as the solution to the

following (continuous) optimization problem:

$$W_2^2(s_0, s_1) = \inf_{f \in MP} \int_{[0,1]^2} (f(y) - y)^2 s_0(y) dy. \tag{4.2}$$

The theory of optimal transport [72] allows us to interpret and re-write the optimization problem above in terms of fluid-dynamics formulation, where we seek for velocity vector field $v(x,t)$ that transports s_0 onto s_1 by incrementally "pushing" intensities according to the continuity equation[72] as follows:

$$\begin{aligned} W_2^2(s_0, s_1) &= \inf_{s,v} \int_0^1 \int_{[0,1]^2} |v(x,t)|^2 s(x,t) dx dt, \\ \text{s.t. } &\frac{\partial s}{\partial t} + \nabla \cdot (vs) = \rho, \end{aligned} \tag{4.3}$$

where $s(x,t)$ is the geodesic from s_0 to s_1. By solving the optimal transport continuity equation above, we can obtain the model for the rearrangement of the normalized intensity measurements from s_0 to s_1 as

$$s_1(x) = D_f(x) s_0(f(x)). \tag{4.4}$$

where, $f(x)$ denotes the mass-preserving optimal transport map, and $D_f(x)$ denotes the determinant of the Jacobian matrix of $f(x)$.

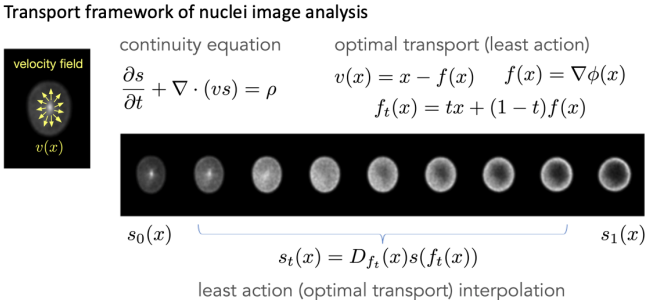

Figure 4.3: Nuclei image representation using optimal transport. An image s can be written in terms of a reference image s_0 through the use of a mapping function $f(x)$ (or equivalently, a velocity field $v(x)$). If the mapping function is chosen to be the gradient of a convex function (potential) ϕ then the transformation is also a metric (Wasserstein/optimal transport) between s_0 and the transported image s.

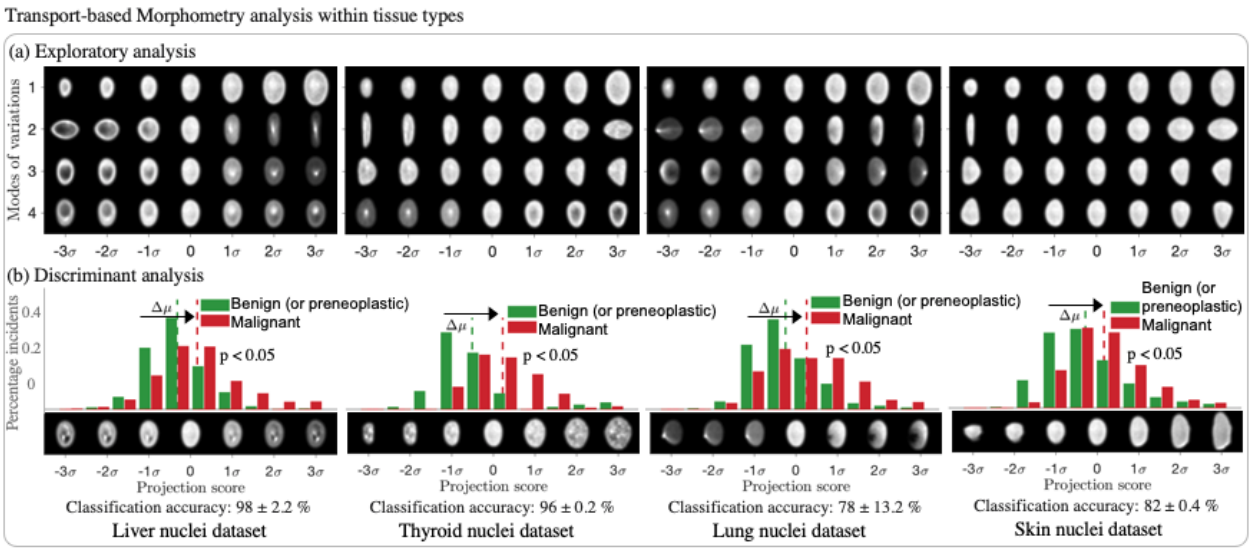

Figure 4.4: The TBM framework can model nuclear morphology within a specific tissue type accurately and efficiently. The exploratory analysis shows the main trends of nuclear structural variations in the datasets. The discriminant analysis shows that the histograms of projections of the malignant class in the test set are collectively localized towards the right of the projection axis (i.e., the malignant direction obtained from the training set) with statistically significant ($p < 0.05$) differences in means between the benign (or preneoplastic) and the malignant classes. The discriminant analysis also demonstrates high patient classification accuracy values when obtained in the discriminant feature space.

Now, take, for example, the task of modeling the chromatin structure change between two nuclear images $s_0(x)$ and $s_1(x)$ in Fig. 4.3 (leftmost and right most images). If we consider the image on the left (s_0) as fixed, we can represent the image on the right (or indeed any other image) by knowing s_0 (reference) as well as the velocity field $v(x) = x - f(x)$ that 'pushes' s_0 forward onto some other image. Thus, we can represent any image as well as the corresponding changes in nuclear structures within that image knowing the reference s_0 and the map $f(x)$, or equivalently, velocity $v(x)$ field (see Fig. 4.3).

The optimal transport-based approach has generated state of the art classification and estimation results for a variety of "segmented" signals/images including images of faces [104], cells [32], nuclei [5], digits [123], language characters [48], brain images [48], knee cartilage images [124], ECG, physiological signals [107, 108], and numerous other applications [1, 125]. Expanding on our previous work modeling nuclear morphological changes within a specific tissue type, this dissertation aims to combine transport-based image transforms, which we denote as transport-based

morphometry (TBM), with a set of new statistical regression methods, to synthesize and compare nuclear morphological changes between different tissue types. Our aim is to enable meaningful comparisons across a wide range of datasets, and to identify nuclear features that are shared by different cancer types. More details regarding the TBM methodology are described in Appendix 4.5.2. Next, we will highlight a few important aspects of TBM, which will help us understand its role in quantitative nuclear morphometry.

4.2.1 TBM provides standardized measurements

The TBM formulation provides us a physically meaningful metric (distance), which can be used to compare two nuclear images. We can use the Wasserstein metric (see equation (4.2)) to compare two nuclear images s_0 and s_1 and quantify the relative intensity changes between them as a representation of changes in chromatin structures. Note that, the metric W_2 can be expressed in terms of a well-defined unit: the unit of $normalized\ intensity \times m^2$. Thus, the TBM framework provides a standardized quantitative measurement of the distributions of chromatin structures where chromatin measurements are represented by the pixel intensities in nuclear images.

Because the proposed TBM approach enables us to measure the change of chromatin structures in terms of a well-defined unit, we can perform meaningful comparisons across a wide range of datasets, even when imaged using different protocols, resolutions, and staining patterns. Any nuclear morphological feature computed using the proposed approach can be expressed in terms of the same measurement unit of $normalized\ intensity \times m^2$. On the contrary, the other approaches, including feature engineering and end-to-end feature learning, do not usually provide well-defined units for the computed features, which makes it challenging to perform meaningful analyses in the joint feature space, compare information across datasets, and describe similarities or differences across datasets imaged with different pathology staining protocols.

4.2.2 TBM enhances interpretability

Unlike most deep learning and feature-based approaches, the TBM formulation allows us to visualize the morphological changes between two images. The TBM framework provides a geodesic (interpolation) between images that can improve understanding of related phenomena. Take, for example, the task of filling the gap (interpolating) between two nuclear chromatin measurements $s_0(x)$ and $s_1(x)$ (leftmost and rightmost images in Fig. 4.3). In addition to the transformation of the two images described by the function $f(x)$, we can obtain the Wasserstein geodesic between $s_0(x)$ and $s_1(x)$, described by the function $f_t(x) = tx + (1-t)f(x), 0 \leq t \leq 1$. This enables us to visualize the intermediate nuclei between $s_0(x)$ and $s_1(x)$ in the original image space (see Fig. 4.3). Thus, we can visualize the process of evolution from $s_0(x)$ to $s_1(x)$, which improves interpretability of the results and enhances our understanding of the underlying physical process.

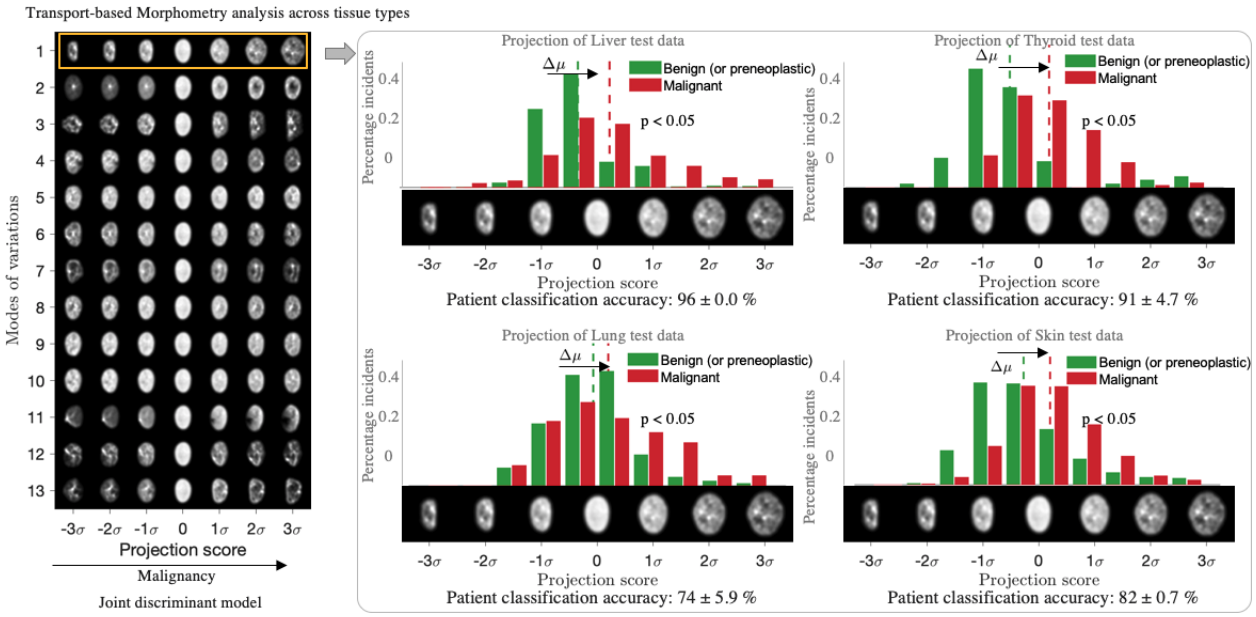

Figure 4.5: The proposed model identifies a set of nuclear morphological features of malignancy that are shared across cancer types (left panel). The projections of the test data in the malignant class of the four tissue types are collectively localized towards the right of the projection axis (i.e., the malignant direction obtained from the training set) with statistically significant ($p < 0.05$) differences in means between the benign (or preneoplastic) and the malignant classes (right panel). The patient classification performances are similar to the performances of individual tissue-specific models in Fig. 4.4, indicating high discriminating capacity of the learned features.

4.2.3 TBM models tissue-specific morphology

Our previous work applying TBM to tissue specimens can model nuclear morphology within a specific tissue type accurately and efficiently [5, 126, 127, 121, 128]. Fig. 4.4 summarizes the application of our previously described *tissue-specific* TBM model to each of the four tissue datasets we used in this study. Visual representations of the principal phenotype variability in each of the four tissue datasets (using the training set) were obtained using principal component analysis (PCA) in the transport space. [5] We present the main trends regarding size, shape, texture, and other nuclear structural variations in each dataset (see Fig. 4.4(a)). Using principal linear discriminant analysis (PLDA) in the transport space [5], we can visualize the principal nuclear morphological changes responsible for discriminating between the benign (or preneoplastic) and the malignant classes (see Fig. 4.4(b)). In the test dataset, the histograms of the malignant class, which are obtained as projections onto the direction of principal nuclear morphological change, are collectively located towards the right (i.e., the malignant direction obtained from the training set). Differences between the histograms of the benign (or preneoplastic) and the malignant classes are statistically significant ($p < 0.05$). The discriminant analysis also demonstrates high patient classification accuracy values when obtained in the discriminant feature space. Consistent with the findings of our previous papers [5, 126, 127, 121], the tissue-specific TBM model can accurately model nuclear morphology within a single tissue type in each of the aforementioned datasets. This is confirmed by its meaningful visual interpretation and effective discriminatory capability in the test dataset.

4.2.4 TBM for modeling shared cancer morphology

As discussed in the previous sections, the TBM framework has been demonstrated to perform well in numerous applications, including the modeling of tissue-specific nuclear morphological changes in cancer cells. We have explained that TBM can provide a physically meaningful standardized quantitative measurement metric (i.e., the Wasserstein metric) to compare nuclear structural changes in cancer. This can be used to better understand the underlying physical mechanism

that occurs during the evolution from a benign to a malignant cell. We hypothesized that the TBM framework has the potential to reliably synthesize and compare nuclear morphological information across different cancer tissue types from different datasets. In this dissertation, we present the results of an updated TBM framework utilizing optimal mass transport and a set of new statistical regression techniques to compare information across different datasets. More details regarding the methodology are described in Appendix 4.5.2.

4.3 Results

This section demonstrates that our proposed TBM approach can synthesize data across multiple tissue types and obtain nuclear morphological features of malignancy that are shared among different cancer types. These results validate our claim that the proposed method can provide meaningful quantitative comparisons across datasets, even when imaged using different protocols, resolutions, and staining patterns. We study the effectiveness of the proposed model by evaluating 1) the projections of held-out test data on the obtained model and 2) the patient classification accuracy assessed on the held-out test data projected on the obtained model. Finally, we show an example application where our model discovers information to rank malignancy (or histological grade) within the sub-types of other unseen cancer datasets.

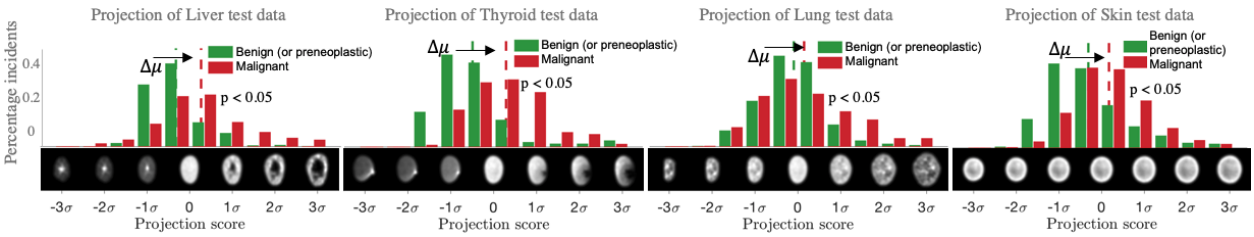

Figure 4.6: The proposed model under a modified experimental setup, where the model was trained using any three tissue types and tested using the fourth tissue type. This modified model correctly ranks the histological grade in the test tissue type using the nuclear morphological features learned from cancer tissues in the training set. The differences in projection means between the benign (or preneoplastic) and the malignant classes of the test set are statistically significant ($p < 0.05$).

4.3.1 Nuclear features shared across cancer types

The proposed TBM model predicts the existence of nuclear morphological features of malignancy that are shared across cancer types and our model can identify a set of shared features. Visual representations of the identified features are shown in the left panel of Fig. 4.5. Each of the identified features can be visualized as changes of nuclear images along a mode of variation, whereby changes from left to right indicate changes from the benign (or preneoplastic) to the malignant class, for each mode of variations (see Fig. 4.5). The main nuclear morphological changes described by the learned features can be visualized from the nuclear images of the above visualization figure. The proposed features were obtained from the training set comprising four tissue types (liver parenchyma, thyroid gland, lung mesothelium, and skin epithelium). Fig. 4.5 shows the most discriminant set of nuclear morphological alterations shared between tissue types that were obtained by the model as described in Appendix 4.5.2.

Projection of test data on the learned feature space

The histograms of the projections in the test set for each of the four tissue types (liver parenchyma, thyroid gland, lung mesothelium, and skin epithelium), on the proposed shared discriminant feature set are also shown in the right panel of Fig. 4.5. The horizontal axis represents the spread of projections in the unit of standard deviation. The representative image of the nuclear morphological feature corresponding to each histogram coordinate is shown below the horizontal axis. Each histogram bar indicates the percentage of nuclei in each class that closely resemble the nuclear morphological feature shown beneath that bar.

We observe that the projections of the test data in the malignant class are collectively located toward the right (i.e., the malignant direction obtained from the training set) of the projection axis compared with the benign (or preneoplastic) class (as indicated by the location of histogram means of these two classes). This trend of collective localization of the malignant class towards the right is consistent among each of the four tissue types tested, demonstrating the shared discriminatory capability of our learned feature model. Fig. 4.5 shows projections on the first learned

feature, however this observation can be seen in other derived shared morphological features also. The p-values of the differences of histogram-means between the benign (or preneoplastic) and the malignant classes (obtained by multivariate t-test) are less than 0.05, which indicates that the separation between the two classes is statistically significant.

We further evaluated the learned model's performance using a modified experimental setup. In this experiment, we trained the model using samples from three of the four tissue types comprising liver, thyroid, lung, and skin. The model was then applied to the fourth cancer type, to predict the relative histological grade, when defined as benign (or preneoplastic) versus malignant. We repeated this experiment for all four tissue combinations and report the ranking results in Fig. 4.6. The x-axis represents the spread of the projections (in the units of standard deviation) on the model trained on three cancer types. The representative image of the nuclear morphological feature corresponding to each histogram coordinate is shown below the horizontal axis. Each histogram bar indicates the percentage of nuclei from the fourth cancer type, in each class, that closely resemble the nuclear morphological feature from the trained cancer model. In all four examples of nuclear morphological features derived from cancers affecting three different tissue types and applied to a fourth tissue type, we observed collective localization of the projections of the test data in the malignant class towards the right of the projection axis (i.e., the malignant direction obtained from the training set) with statistically significant ($p < 0.05$) differences in means between the benign (or preneoplastic) and the malignant classes. This cross-validation model was observed to correctly rank the histological grade in the test tissue type using the nuclear morphological features learned from cancer tissues in the training set. It further highlights the discriminatory ability of our shared feature model.

Patient classification using the learned features

To evaluate the accuracy of the proposed model in ranking malignancy potential to estimate histological cancer grade, we used the learned nuclear morphological features shared between cancers affecting the four different tissue types to classify patients. We began by obtaining the

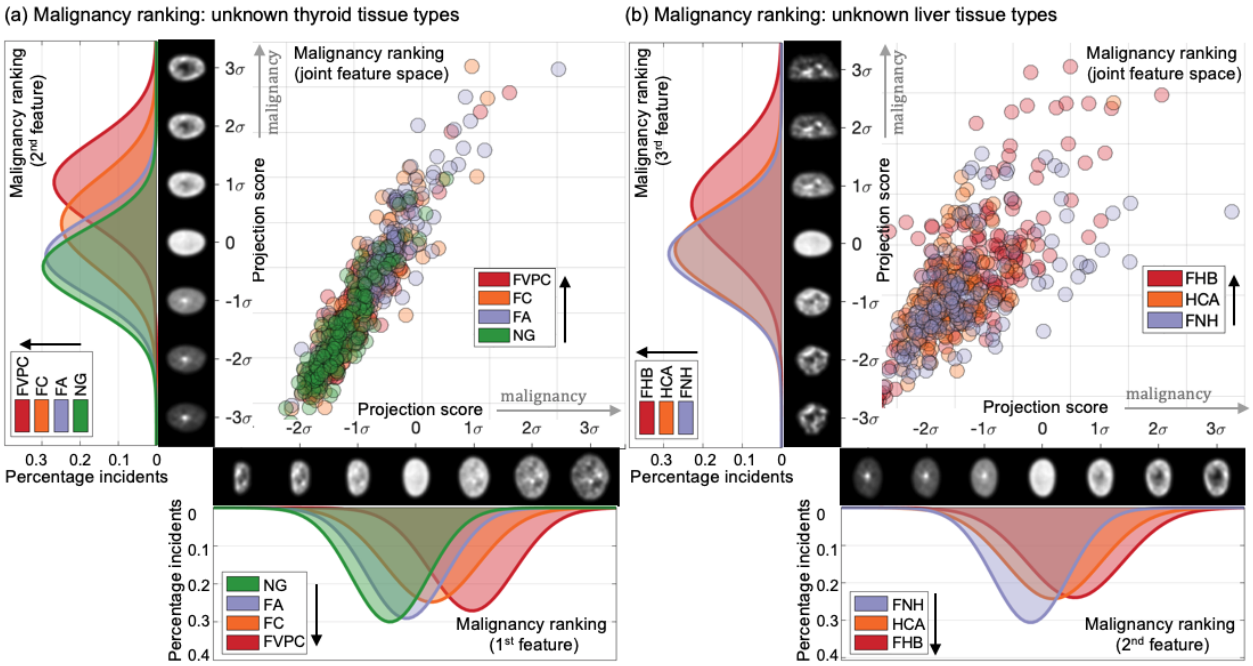

Figure 4.7: Application of the learned model in ranking the malignant potential within the subtypes of unseen cancer datasets from a particular organ: (a) malignancy ranking within the subtypes of thyroid tissue, (b) malignancy ranking within the subtypes of liver tissue. The rankings (from less malignant to more malignant) jointly predicted by the model are NG, FA, FC, FVPC and FNH, HCA, FHB for the thyroid and liver test tissue types, respectively.

histograms of projections of the nuclei for each patient in the training set on the shared discriminant morphological feature space. Next, we trained classifiers using these projections of nuclei and the corresponding histograms obtained from the training set. We obtained four sets of classifiers corresponding to the training set for each of the four tissue types. We also obtained the histograms of projections of the nuclei of the patients in the test set on the shared feature space (comprising all four tissue types). We used these test set patient histograms to test the performance of the trained classifiers in distinguishing patient samples from any of the four tissue types as benign (or preneoplastic) versus malignant. We evaluated several classifiers, including penalized discriminant analysis (PLDA), linear discriminant analysis (LDA), random forests (RF), logistic regression (LR), linear support vector machine (SVM-l), kernel support vector machine (SVM-k), and k-nearest neighbors (kNN). The best test accuracy values for the patient classification are provided in Fig. 4.5. It can be seen that the proposed model provides reasonably high patient classification test accuracy as compared with the chance accuracy (50%). The mean patient accuracy reached

as high as 96%, 91%, 74%, and 82.% for liver, thyroid, lung, and skin tissue types, respectively. These results are similar to the performances of the individual tissue-specific models presented in Fig. 4.4 for cancers affecting each of liver, thyroid, lung, and skin tissue types, respectively, which indicates high discriminating capacity of the shared cancer feature model when applied to heterogeneous tissue data. The detailed classification results are available in Appendix 4.5.3.

4.3.2 Application: Discovery of malignancy ranking within subtypes of unseen cancer datasets

The proposed shared cancer features explained in the previous sections have the potential to be used in many clinical and scientific applications, such as cancer screening for early diagnosis, prognostication, therapeutic development, biophysical studies of disease pathology, and large database analyses. As explained before, the learned features were obtained from four datasets in the training set. Here, we show an application where we utilize these learned features to rank the malignant potential within the subtypes of other unseen cancer datasets from a particular organ (e.g., thyroid and liver).

Fig. 4.7 shows the nuclear morphology-based histological grade ranking results among different cancer types in the thyroid and the liver. The x and y axes represent the spread of the projections (in the units of standard deviation) on a particular feature learned by the model. The images beneath and to the left of the x and y axes, respectively, represent the most discriminant nuclear morphological feature corresponding to each histogram coordinate. The corresponding Gaussian curves represent the mean and standard deviation of the projections of the test cancer types on the nuclear morphological features of the trained cancer model. The scatter plots show the projections of the different cancer types for each tissue (thyroid and liver) in the joint feature space. We show the ranking results obtained from individual features as well as the results obtained jointly on the five most discriminatory nuclear morphological features learned by the shared cancer model. The rankings (from less malignant to more malignant) jointly predicted by the model for the thyroid test tissue type are nodular goiter (NG), follicular adenoma (FA), follicular carcinoma (FC), and follicular variant of papillary carcinoma (FVPC). The rankings (from less malignant to more malignant)

jointly predicted by the model for the liver test tissue type is follicular nodular hyperplasia (FNH), hepatocellular adenoma (HCA), and fetal hepatoblastoma (FHB). These results demonstrate the potential for our model to rank or grade a spectrum of cancer types that vary in their malignant potential. They also highlight the out-of-distribution performance of our model.

4.4 Discussion and Conclusions

Improved understanding of the molecular mechanisms underpinning carcinogenesis has led to identification of biomarkers for risk assessment in cancer patients as we move further into the era of personalized medicine [129, 67]. Molecular biomarkers, derived from genomic and/or proteomics means primarily, are used clinically for diagnosis, prognosis, therapeutic interventions, and following cancer progression during treatment [67, 130]. In recent years, attention has turned towards identification of biomarkers applicable across a number of cancer types [131, 57]. Cross-cancer or universal cancer biomarkers may permit development of cost-effective and efficient cancer screening methods, elucidate common carcinogenesis pathways, and identify shared resistance and sensitivities to treatment [131, 57].

For a cancer biomarker to be clinically useful, it must address a specific stage in tumor development, reliably estimate risk and be actionable [132, 67]. As noted, nuclear morphological alterations are a feature commonly utilized by pathologists to grade tumors[27, 133]. By estimating the degree of deviation of the nuclear appearance from a normal cell, a histological grade is assigned and utilized to inform prognostic and therapeutic decisions [133]. Because nuclear morphological alterations affect all tumor cells, they represent a potentially useful biomarker for simultaneously evaluating multiple cancer types.

Computational studies using feature engineering and neural network-based methods attempting to model nuclear morphological alterations have suffered from numerous drawbacks including a limited knowledge of their internal workings which makes it difficult to safely implement them into a critical system, a requirement for large amounts of training data and a lack of robustness to adversarial information [28, 30, 31]. In addition, a reliable machine learning method has not

yet been found that can synthesize and summarize information across different cancers. This may be due to the lack of a quantitative metric that can be used to compare two nuclear images of any given cancer type. In this study, we present a TBM framework that uses a standardized quantitative metric (Wasserstein distance) to compare the entirety of the normalized chromatin content between two nuclear images. Our method, which preserves the information content of each image within a biologically meaningful, transport-based, representation, offers several advantages. First, it permits visualization of the change (or evolution) in nuclear structure between benign and malignant cells, enhancing understanding of the underlying biophysical process. Second, it detects and interprets persistent discriminating information between benign and malignant cells. Third, it can categorize and stratify patients or tissues by their histological grade in both known and unknown cancer types. Fourth, not only is it robust to variations in staining protocols and image resolutions, but it is also able to identify features shared by four different tissue types, enabling comparisons across a range of datasets.

In this dissertation, we presented visual exploratory analyses to highlight the common nuclear structural changes that were shared by cancers affecting the four tissue types. Our multivariate statistical analysis found a significant difference between the shared features discriminating benign or normal from malignant nuclei. This was consistent across all four tissue types. When we examined the discriminative ability of our shared feature model in classifying patients, we found it to estimate the tumor grade (malignancy ranking) with similar accuracy to the tissue-specific model. We further cross-validated our model's performance and found it to correctly estimate the tumor grade (malignancy ranking) in a tissue sample it was not previously trained upon, further highlighting the out of distribution performance of our shared feature model. Finally, we demonstrated our model's capacity to stratify unknown cancer subtypes acquired from a single tissue type that varied by their malignant potential. Our proposed method offers a novel approach to modeling the nuclear structure in cancer cells. We found it to accurately identify and measure the morphological changes that affected malignant cells, shared by the four different tissue types.

Several limitations to this work must be acknowledged. Our model was derived from a small

cohort of patient samples acquired from a limited number of centers and used digitized histological images from only four tissue types. In the absence of external validation, the broader generalizability of our results to a wide range of cancer types and to larger databases of patient samples remains unknown. Our analysis was solely based upon image features and we were unable to account for potential confounders including patient demographics, medical history information and treatment-related factors.

Our contributions in this dissertation are to (1) introduce a quantitative measurement metric that can be reliably used to discriminate nuclear morphological features of cancer cells, accounting for different tissue types, (2) demonstrate the potential for transport-based morphometry to overcome the limitations inherent to current techniques in digital pathology, and (3) present preliminary evidence that our transport-based morphometry method can make meaningful comparisons across a wide range of nuclei data. In combination with large datasets such as the human protein atlas and the cancer genome atlas, we believe that our proposed method has the potential to enable numerous clinical and scientific studies in fields such as population-based screening, development of personalized therapies, risk stratification, assessment of treatment response and understanding of carcinogenesis by, eventually, helping to elevate the potential role of nuclear morphometry as a universal cancer biomarker into a more quantitative science.

4.5 Appendix

4.5.1 Computational experiments

Experimental setup

The proposed model was built using a portion of the dataset for training and evaluated using the remaining portion as the test set. Two partitions were created using a 2-fold cross-validation method. Then the described computations were performed using these training and test sets. The p-values for differences in histograms between classes were obtained with a multivariate t-test.

Datasets

We utilized four cancer pathology datasets containing labeled nuclear microscopy images of liver, thyroid, lung, and skin cells. The thyroid and liver nuclei datasets were collected from the University of Pittsburgh Medical Center [134, 5, 126]. Cytology slides for lung tissue were obtained from Allegheny General Hospital and West Penn Hospital [127]. Hematoxylin and eosin (H&E) slides of skin tissue were retrieved from pathology archives [128] with Institutional Review Board approval. Each dataset comprises benign (or preneoplastic) and malignant populations, and class definitions are presented in Table 4.1.

Table 4.1: Details of the definitions used for each class of different tissue types

	Benign (or preneoplastic)	Malignant
Liver nuclei (Feulgen staining)	Normal liver	Hepatocellular carcinoma
Thyroid nuclei (Feulgen staining)	Normal thyroid	Widely invasive follicular carcinoma
Lung nuclei (Diff-Quik staining)	Benign mesothelioma	Malignant mesothelioma
Skin nuclei (Hematoxylin and Eosin staining)	Dysplastic nevi	Malignant melanoma

4.5.2 Methods

Preprocessing

To eliminate irrelevant variations in nuclei images, we normalized them before analysis following [135]. We translated the center of mass of each image to the center of view, aligned the

principal axis to a predetermined angle, and flipped the images for similar intensity weight distribution. We resized the data so that images from each tissue dataset have the same average area and pixel dimensions, and normalized the images such that the intensity values of all pixels sum to one.

Image transform: linear optimal transport

Our proposed image morphometry analysis generates unique representations for input images, enabling both visualization and quantitative analysis. The procedure employs the LOT distance from [73] and allows for an isometric embedding (LOT embedding [73]) of input image datasets onto the standard Euclidean space. Mathematical descriptions of these concepts are provided below:

The LOT distance: The LOT distance measures the optimal effort required to rearrange one structure to another and is constructed based on the tangent space approximation of the underlying Riemannian manifold representing the dataset geometry. Let $s = \sum_p a_p \delta_{\vec{x}_p}$ and $\varphi = \sum_q a_q^0 \delta_{\vec{y}_q}$ are the particle representations of a sample image and a reference structure, respectively. The LOT distance between s and φ is given by

$$d_{OT}^2(s, \varphi) = \min_{f \in \Pi(\mu, \mu_0)} \sum_p \sum_q \left|\vec{x}_p - \vec{y}_q\right|^2 f_{pq} \tag{4.5}$$

subject to $f_{pq} \geq 0$, $\sum_p f_{pq} = a_q^0$, and $\sum_q f_{pq} = a_p$. Here, f is the map corresponding to the optimal mass transport (mass corresponds to the image intesity in this context) between s and φ.

The LOT embedding: Using the optimal transport map from equation (4.5), the linear optimal transport (LOT) embedding of the sample image s is defined as follows:

$$\widehat{\mathbf{s}} = \left[\frac{1}{\sqrt{a_1^0}} \sum_p f_{p1} \vec{x}_p, \quad \frac{1}{\sqrt{a_2^0}} \sum_p f_{p2} \vec{x}_p, \quad \cdots \right]^T \tag{4.6}$$

Equation (4.6) was used to compute LOT embeddings for all sample images. The Euclidean average of all input training images was used to compute the reference structure φ. For more details, refer to [73, 5]. The procedure above generates linear representations of the data, significantly simplifying the data analysis.

Statistical modeling

Composite penalized linear discriminant analysis

We introduce a modified PLDA technique called composite penalized linear discriminant analysis (cPLDA) to derive a set of shared discriminant directions for all datasets. Let $\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j$ denotes the LOT embedding of the n-th sample image of the k-th class of the j-th dataset. The embeddings were first standardized by removing the mean and scaling to unit variance. The 'total scatter matrix' for the j-th dataset can be computed as

$$(\mathbf{S}_T)_j = \sum_k \sum_n \left(\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j - \left(\widehat{\underline{\mathbf{s}}}\right)_j\right)\left(\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j - \left(\widehat{\underline{\mathbf{s}}}\right)_j\right)^T$$

where $\left(\widehat{\underline{\mathbf{s}}}\right)_j = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{n=1}^{N_k}\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j$ denotes the average of the entire set of LOT embeddings of the j-th dataset. The 'within class scatter matrix' for the j-th cancer type is can be computed as

$$(\mathbf{S}_W)_j = \sum_k \sum_n \left(\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j - \left(\widehat{\underline{\mathbf{s}}}^{(k)}\right)_j\right)\left(\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j - \left(\widehat{\underline{\mathbf{s}}}^{(k)}\right)_j\right)^T$$

where $\left(\widehat{\underline{\mathbf{s}}}^{(k)}\right)_j = \frac{1}{N_k}\sum_{n=1}^{N_k}\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j$ denotes the average of the k-th class of the j-th dataset. The shared discriminant direction can be obtained by maximizing the following objective function:

$$\arg\max_{\mathbf{w}} \; Z(\mathbf{w}) = \frac{\mathbf{w}^T\left(\sum_j (\mathbf{S}_T)_j\right)\mathbf{w}}{\mathbf{w}^T\left(\sum_j (\mathbf{S}_W)_j + \alpha\mathbf{I}\right)\mathbf{w}}. \tag{4.7}$$

The optimization equation in equation (4.7) is equivalent to the following generalized eigen-

Table 4.2: Patient classification in the tissue-specific feature space

	Histogram means				Single nucleus classification (with maximum voting)				Complete histograms			
	LV	THY	LNG	SKN	LV	THY	LNG	SKN	LV	THY	LNG	SKN
LDA	87 ± 8.7	72 ± 0.9	59 ± 2.9	71 ± 3.6	74 ± 4.3	80 ± 1.3	74 ± 2.9	68 ± 0.4	96 ± 4.3	68 ± 5.1	69 ± 10.3	60 ± 5.0
PLDA	91 ± 4.3	68 ± 5.1	72 ± 1.5	75 ± 1.1	74 ± 4.3	80 ± 1.3	74 ± 2.9	68 ± 0.0	89 ± 2.2	72 ± 0.9	72 ± 7.4	81 ± 1.4
RF	91 ± 0.0	83 ± 3.6	75 ± 4.4	80 ± 0.7	91 ± 0.0	72 ± 0.9	71 ± 2.9	78 ± 1.4	93 ± 2.2	91 ± 0.4	74 ± 0.0	79 ± 1.4
LR	93 ± 2.2	78 ± 0.9	69 ± 1.5	76 ± 2.9	80 ± 2.2	81 ± 5.7	72 ± 1.5	68 ± 0.7	**98 ± 2.2**	70 ± 3.0	75 ± 7.4	81 ± 1.4
SVM-l	93 ± 2.2	66 ± 7.2	69 ± 4.4	77 ± 1.1	74 ± 0.0	24 ± 5.5	60 ± 10.3	35 ± 1.4	87 ± 4.3	67 ± 0.8	72 ± 7.4	80 ± 1.1
SVM-k	85 ± 2.2	76 ± 7.6	75 ± 4.4	**82 ± 0.4**	89 ± 2.2	81 ± 5.7	72 ± 4.4	79 ± 0.4	87 ± 4.3	**96 ± 0.2**	75 ± 10.3	81 ± 1.8
kNN	83 ± 4.3	80 ± 1.3	71 ± 8.8	82 ± 0.7	89 ± 2.2	72 ± 5.3	71 ± 2.9	79 ± 0.4	89 ± 2.2	55 ± 4.5	**78 ± 13.2**	79 ± 0.0

Table 4.3: Patient classification in the shared cancer feature space

	Histogram means				Single nucleus classification (with maximum voting)				Complete histograms			
	LV	THY	LNG	SKN	LV	THY	LNG	SKN	LV	THY	LNG	SKN
LDA	85 ± 2.2	76 ± 5.5	53 ± 5.9	75 ± 0.0	74 ± 0.0	79 ± 7.8	69 ± 4.4	66 ± 0.7	93 ± 2.2	80 ± 3.0	60 ± 1.5	58 ± 1.4
PLDA	85 ± 2.2	83 ± 3.6	57 ± 1.5	76 ± 0.0	74 ± 0.0	79 ± 7.8	69 ± 4.4	66 ± 0.7	91 ± 0.0	76 ± 7.6	65 ± 0.0	78 ± 2.2
RF	91 ± 4.3	80 ± 3.0	**74 ± 5.9**	81 ± 0.7	89 ± 2.2	74 ± 3.2	69 ± 1.5	80 ± 0.0	89 ± 6.5	**91 ± 4.7**	69 ± 10.3	78 ± 0.0
LR	87 ± 0.0	83 ± 3.6	65 ± 5.9	77 ± 1.1	80 ± 2.2	81 ± 5.7	69 ± 4.4	66 ± 0.0	91 ± 0.0	74 ± 1.1	71 ± 2.9	82 ± 1.1
SVM-l	**96 ± 0.0**	80 ± 1.3	68 ± 5.9	79 ± 0.4	80 ± 6.5	74 ± 3.2	60 ± 4.4	37 ± 0.4	89 ± 2.2	78 ± 5.3	71 ± 0.0	78 ± 2.5
SVM-k	83 ± 0.0	80 ± 1.3	72 ± 4.4	**82 ± 0.7**	87 ± 0.0	81 ± 5.7	72 ± 10.3	80 ± 0.0	93 ± 2.2	91 ± 9.1	69 ± 4.4	80 ± 1.1
kNN	87 ± 0.0	83 ± 3.6	69 ± 1.5	81 ± 0.7	87 ± 4.3	68 ± 5.1	65 ± 8.8	78 ± 0.0	89 ± 2.2	61 ± 11.4	72 ± 1.5	76 ± 1.1

decomposition problem:

$$((\sum_j (\mathbf{S}_W)_j + \alpha\mathbf{I})^{-1}(\sum_j (\mathbf{S}_T)_j))\mathbf{w} = \lambda\mathbf{w} \tag{4.8}$$

where, $\lambda = \max_{\mathbf{w}} Z(\mathbf{w})$. Let $\mathbf{w}^0$ is the solution of the equation (4.8) (also equation (4.7)), i.e., $\lambda = Z(\mathbf{w}^0)$. Then we removed the effect of the feature scaling from the obtained solution.

Hierarchical feature extraction

After obtaining the most discriminant direction $\mathbf{w}^0$, we removed the corresponding feature from the LOT embeddings of the datasets and continued the analysis with the filtered data to determine the other discriminant directions. Let $\widehat{\mathbf{s}}$ be the LOT embedding of a sample image s, and $\mathbf{b}_i$ be a basis vector spanning the ambient space. Next, we selected a basis set with $\mathbf{b}_1 = \mathbf{w}^0$ and defined the filtered representation of $\widehat{\mathbf{s}}$ as:

$$\widetilde{\mathbf{s}} = \sum_{i,i\neq 1} c_i\mathbf{b}_i; \quad \text{with, } c_i = < \mathbf{b}_i, \widehat{\mathbf{s}} >$$

We repeated the procedures above to obtain a set of shared discriminant features as follows:

$$W = \{\mathbf{w}^0_{(1)}, \mathbf{w}^0_{(2)}, \mathbf{w}^0_{(3)}, \cdots\} \tag{4.9}$$

The effectiveness of the computed discriminant feature-set is evaluated in the experimental section.

4.5.3 Patient classification

We began by projecting the nuclei of the patients in the training and test sets on the tissue-specific (Table 4.2) or the shared (Table 4.3) discriminant morphological feature space. Next, we trained the classifiers using three different descriptors: We used the means of histograms of projections of the nuclei, the single nucleus projections, or the complete histograms of projections of the training set to train different classifiers (see Tables 4.2 and 4.3). We obtained four sets of classifiers corresponding to four tissue types: liver (LV), thyroid (THY), lung (LNG), and skin (SKN). In the test phase, we used the same descriptors, i.e., the means of histograms of projections of the nuclei, the single nucleus projections, or the complete histograms of projections of the test set to predict the class of the patient. In the case of the single nucleus classification, we obtained the patient class prediction by applying the maximum voting procedure to the single nucleus class prediction results.

Chapter 5: Conclusion and Future directions

In this dissertation, we proposed a new computational framework for pattern analysis and recognition that can be applied in various disciplines of science and technology, including computer vision, biology, and health care. Our hypothesis is that data classes can be modeled as the instances of an unknown template (or templates) under the effect of unknown spatial deformations in a large subcategory of pattern analysis and recognition problems. We demonstrated that classification and modeling problems, involving data obtained under our proposed generative model, can be solved efficiently using a transport-based modeling approach in closed-form. Our approach is well-suited for modeling a wide range of processes in computer vision, biology, and medicine that involve some kind of movement or transport of mass or intensity of various entities, including pixels, tissue, molecules, proteins, and more. By mathematically modeling these processes using our proposed framework, we can gain deeper insights into their underlying mechanisms and potentially discover new avenues for intervention and treatment. The advantages of our modeling approach include better accuracy, generalizability, interpretability, data efficiency, and paths to discovering unknown processes. Furthermore, our methods can be implemented using a simple algorithm and are generally more computationally efficient than existing deep learning-based approaches. We also developed an accompanying Python software for easy implementation of our proposed framework. Overall, our study presents a promising opportunity to advance pattern analysis and recognition techniques in various scientific disciplines.

In Chapter 2, we introduced a novel end-to-end system for supervised image classification. This approach classifies images by computing the distance between the R-CDT transform of a given image and the linear subspaces estimated from the linear combination of the R-CDT transformed input training data. Our mathematical framework also enables invariances to a set of given image transformations. The proposed method is particularly relevant for image classification prob-

lems where image classes can be considered as an instance of a template observed under a set of spatial deformations. If these deformations are accurately modeled as a collection of smooth, one-to-one, and nonlinear transformations, then the image classes can be separated in the transform space via the properties outlined in [98]. These properties also allow for the approximation of image classes as convex subspaces in the R-CDT space, leading to a more suitable data model for the nearest subspace method. Consequently, the resulting classifier can achieve high accuracy, computational efficiency, and out-of-distribution robustness, as demonstrated in our experiments. Our proposed solution can benefit a wide range of image classification problems that can be formulated in a similar manner. For example, any classification problem where one image in a class can be obtained from another by a smooth rearrangement of pixel intensities, such as affine transformations, is a suitable fit for our generative model. More subtle examples include distortions resulting from the influence of a transparent medium in an optical communication channel [103], or morphological changes in MRI images due to the presence of a disease [3].

The proposed mathematical solution was shown to be highly effective in achieving accurate image classification results in a variety of real-world scenarios, both with low and high amounts of data. Furthermore, the method was demonstrated to significantly reduce computational costs associated with image classification tasks, while still maintaining high accuracy. One key advantage of this method is its mathematical coherence and simplicity, as it is non-iterative and does not require the tuning of hyper-parameters. The approach can also be implemented without the need for GPU support, although the method can benefit from parallelization on a GPU to further improve its computational efficiency. Additionally, the method was shown to be robust in challenging experimental scenarios, including the out-of-distribution setup. This is due to the fact that the method learns the underlying generative model of the image classes, allowing it to identify and classify images based on their underlying characteristics. Overall, this method provides a promising solution for a range of image classification problems, particularly those where images can be thought of as instances of a template observed under a set of spatial deformations.

We note that the method is well-suited for the problems where the data at hand conform to

the generative model stated in the dissertation. However, it is worth noting that the method is tailored towards modeling segmented images, making it less suitable for natural unsegmented images such as the CIFAR10 and imagenet datasets. Although it is possible to adapt the method for natural unsegmented images, it would require redefining the problem statement and generative model, which is left for future work. It is also important to note that the proposed model does not account for occlusions, introduction of other objects in the scene, or variations that cannot be modeled as a mass-preserving transformation on a set of templates. The results from the CIFAR10 dataset demonstrate that the proposed model lags far behind the standard deep learning classification methods in terms of classification accuracy. However, the approximations and assumptions made in the derivation of the spanning sets of shear and anisotropic scaling work reasonably well for practical purposes, as shown in the results.

There are many other adaptations of the method that can be explored in the future. For instance, the linear subspace method described above can be modified to utilize other assumptions regarding the set that best models each class, which may benefit certain classes of problems. Additionally, the method can be extended to be used on 3D images with the application of the 3D Radon transform, and the generative model can also be adapted for RGB images. Furthermore, the method can be adapted for unsupervised learning contexts such as subspace clustering. Overall, the proposed method has shown promising results and offers several potential directions for future research.

In Chapter 3, the transport-based frameworks for 1D (signals) and 2D (images) distributions were extended to N-dimensions and were utilized to analyze high-dimensional distributions. The extended transport-based embeddings (high-dimensional R-CDT and LOT) were employed to develop a framework for classifying high-dimensional distributions that could be used in numerous applications. The properties of the transport-based embeddings were also utilized to develop a simple and robust classification method with high accuracy. The study began with a generative model assumption for high-dimensional data, where data classes were modeled as the instances of an unknown template (or templates) high-dimensional distribution under the effect of unknown spatial deformations. A mathematical solution was then demonstrated, which showed better accuracy,

generalizability, interpretability, and data efficiency.

The proposed framework for high-dimensional distribution is applicable to problems where the data follows a given generative model, similar to its low-dimensional counterparts. Two approaches were attempted to solve high-dimensional classification problems, the high-dimensional R-CDT transform and the LOT transform. The R-CDT-based method involves obtaining a set of 1D projections of the high-dimensional distributions and applying the CDT transform. However, the choice of projection directions is an open problem that could affect performance. Currently, projections are obtained along the canonical axis directions, but the exploration of other projection directions may be necessary. The use of transport-based features may also benefit the analysis. For the R-CDT-based method, morphological numerical features such as area and solidity were used. The LOT-based method is more suitable for set-structured discrete data points, and to apply it to continuous high-dimensional distributions, the first step is to sample the continuous distributions. The LOT solution also requires solving a linear programming that adds to the computational cost. Finally, analyzing high-dimensional distributions is generally computationally intensive, and the computational complexity increases with dimensionality. Therefore, developing a more efficient and robust algorithm might be necessary, which we leave to future work.

In Chapter 4, we presented a series of computational analyses to investigate the common nuclear structural changes shared among cancers that affect four different tissue types. Our multivariate statistical analysis revealed significant differences between shared features that discriminate benign or normal nuclei from malignant ones, which was consistent across all tissue types. The discriminative ability of our shared feature model in classifying patients was evaluated and demonstrated similar accuracy in estimating the tumor grade (malignancy ranking) to the tissue-specific model. Cross-validation was further performed to assess the model's performance on an unseen tissue sample, highlighting its out-of-distribution performance. Our proposed method was also shown to have the ability to stratify unknown cancer subtypes acquired from a single tissue type based on their malignant potential. These findings suggest that our novel approach to modeling the nuclear structure in cancer cells can accurately identify and measure the morphological changes

that affect malignant cells and are shared among different tissue types.

Advancements in the understanding of the molecular mechanisms involved in carcinogenesis have led to the identification of biomarkers for risk assessment in cancer patients, particularly in the era of personalized medicine. Molecular biomarkers, primarily derived from genomics and proteomics, are utilized clinically for diagnosis, prognosis, therapeutic interventions, and monitoring cancer progression during treatment. Recently, attention has shifted towards the identification of biomarkers that are applicable across multiple cancer types. Universal cancer biomarkers may enable development of cost-effective and efficient cancer screening methods, identify shared resistance and sensitivities to treatment, and shed light on common carcinogenesis pathways. For a cancer biomarker to be clinically useful, it must address a specific stage in tumor development, reliably estimate risk, and be actionable. Nuclear morphological alterations are a commonly utilized feature by pathologists to grade tumors, as these alterations affect all tumor cells, making them a potentially useful biomarker for simultaneously evaluating multiple cancer types.

Our study introduces a transport-based morphometry (TBM) framework that utilizes the Wasserstein distance as a standardized quantitative metric to compare the normalized chromatin content of two nuclear images. Our method preserves the biologically meaningful information content of each image and offers several advantages. It allows for visualization of the change in nuclear structure between benign and malignant cells, detects persistent discriminating information, categorizes and stratifies patients or tissues by histological grade, and is robust to variations in staining protocols and image resolutions. Furthermore, it can identify shared features across four different tissue types, enabling comparisons across multiple datasets.

Several limitations of our work should be noted, including the small cohort of patient samples from a limited number of centers and the use of digitized histological images from only four tissue types. Our analysis was based solely on image features and did not account for potential confounders such as patient demographics or medical history. However, our work has introduced a quantitative measurement metric for reliably discriminating nuclear morphological features of cancer cells across different tissue types. Our transport-based morphometry method has the po-

tential to enable numerous clinical and scientific studies, including population-based screening, personalized therapies, risk stratification, treatment response assessment, and understanding of carcinogenesis. Our proposed method has the potential to elevate the role of nuclear morphometry as a universal cancer biomarker into a more quantitative science, particularly when combined with large datasets such as the human protein atlas and the cancer genome atlas.

References

[1] M. Nishikawa, H. Kanno, Y. Zhou, T.-H. Xiao, T. Suzuki, Y. Ibayashi, J. Harmon, S. Takizawa, K. Hiramatsu, N. Nitta, *et al.*, "Massive image-based single-cell profiling reveals high levels of circulating platelet aggregates in patients with covid-19," *Nature communications*, vol. 12, no. 1, pp. 1–12, 2021.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] S. Kundu, S. Kolouri, K. I. Erickson, A. F. Kramer, E. McAuley, and G. K. Rohde, "Discovery and visualization of structural biomarkers from mri using transport-based morphometry," *NeuroImage*, vol. 167, pp. 256–275, 2018.

[4] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," *Pattern recognition*, vol. 42, no. 6, pp. 1093–1103, 2009.

[5] S. Basu, S. Kolouri, and G. K. Rohde, "Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3448–3453, 2014.

[6] J. B. Schulz, J. Borkert, S. Wolf, T. Schmitz-Hübsch, M. Rakowicz, C. Mariotti, L. Schoels, D. Timmann, B. van de Warrenburg, A. Dürr, *et al.*, "Visualization, quantification and correlation of brain atrophy with clinical symptoms in spinocerebellar ataxia types 1, 3 and 6," *Neuroimage*, vol. 49, no. 1, pp. 158–168, 2010.

[7] A. Hadid, J. Heikkila, O. Silvén, and M Pietikainen, "Face and eye detection for person authentication in mobile phones," in *2007 First ACM/IEEE International Conference on Distributed Smart Cameras*, IEEE, 2007, pp. 101–108.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] R. F. Murphy, M. V. Boland, M. Velliste, *et al.*, "Towards a systematics for protein subcellular location: Quantitative description of protein localization patterns and automated analysis of fluorescence microscope images.," in *ISMB*, vol. 8, 2000, pp. 251–259.

[10] T. J. Keyes, P. Domizi, Y.-C. Lo, G. P. Nolan, and K. L. Davis, "A cancer biologist's primer on machine learning applications in high-dimensional cytometry," *Cytometry Part A*, vol. 97, no. 8, pp. 782–799, 2020.

[11] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, "Extracting a cellular hierarchy from high-dimensional cytometry data with spade," *Nature biotechnology*, vol. 29, no. 10, pp. 886–891, 2011.

[12] Y. Xu and U. Stilla, "Toward building and civil infrastructure reconstruction from point clouds: A review on data and key techniques," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2857–2885, 2021.

[13] Q. Wang, Y. Tan, and Z. Mei, "Computational methods of acquisition and processing of 3d point cloud data for construction applications," *Archives of computational methods in engineering*, vol. 27, pp. 479–499, 2020.

[14] L. Zhou, Y. Du, and J. Wu, "3d shape generation and completion through point-voxel diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5826–5835.

[15] A. Gupta, P. J. Harrison, H. Wieslander, N. Pielawski, K. Kartasalo, G. Partel, L. Solorzano, A. Suveer, A. H. Klemm, O. Spjuth, *et al.*, "Deep learning in image cytometry: A review," *Cytometry Part A*, vol. 95, no. 4, pp. 366–380, 2019.

[16] F. Pelon, B. Bourachot, Y. Kieffer, I. Magagna, F. Mermet-Meillon, I. Bonnet, A. Costa, A.-M. Givel, Y. Attieh, J. Barbazan, *et al.*, "Cancer-associated fibroblast heterogeneity in axillary lymph nodes drives metastases in breast cancer through complementary mechanisms," *Nature communications*, vol. 11, no. 1, pp. 1–20, 2020.

[17] K. Quintelier, A. Couckuyt, A. Emmaneel, J. Aerts, Y. Saeys, and S. Van Gassen, "Analyzing high-dimensional cytometry data using flowsom," *Nature protocols*, vol. 16, no. 8, pp. 3775–3801, 2021.

[18] S. G. Utz, P. See, W. Mildenberger, M. S. Thion, A. Silvin, M. Lutz, F. Ingelfinger, N. A. Rayan, I. Lelios, A. Buttgereit, *et al.*, "Early fate defines microglia and non-parenchymal brain macrophage development," *Cell*, vol. 181, no. 3, pp. 557–573, 2020.

[19] D. Zink, A. H. Fischer, and J. A. Nickerson, "Nuclear structure in cancer cells," *Nature reviews cancer*, vol. 4, no. 9, pp. 677–687, 2004.

[20] L. Beale, "Examination of sputum from a case of cancer of the pharynx and the adjacent parts," *Arch Med*, vol. 2, no. 44, pp. 1860–61, 1860.

[21] C. Uhler and G. Shivashankar, "Nuclear mechanopathology and cancer diagnosis," *Trends in cancer*, vol. 4, no. 4, pp. 320–331, 2018.

[22] E. Martinez-Ledesma, R. G. Verhaak, and V. Treviño, "Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm," *Scientific reports*, vol. 5, no. 1, pp. 1–14, 2015.

[23] A. Mazumder and G. Shivashankar, "Gold-nanoparticle-assisted laser perturbation of chromatin assembly reveals unusual aspects of nuclear architecture within living cells," *Biophysical journal*, vol. 93, no. 6, pp. 2209–2216, 2007.

[24] C. R. Pfeifer, J. Irianto, and D. E. Discher, "Nuclear mechanics and cancer cell migration," in *Cell Migrations: Causes and Functions*, Springer, 2019, pp. 117–130.

[25] C. Kaushal, S Bhat, D. Koundal, and A. Singla, "Recent trends in computer assisted diagnosis (cad) system for breast cancer diagnosis using histopathological images," *Irbm*, vol. 40, no. 4, pp. 211–227, 2019.

[26] R. W. Veltri and C. S. Christudass, "Nuclear morphometry, epigenetic changes, and clinical relevance in prostate cancer," *Cancer Biology and the Nuclear Envelope*, pp. 77–99, 2014.

[27] E. G. Fischer, "Nuclear morphology and the biology of cancer cells," *Acta cytologica*, vol. 64, no. 6, pp. 511–519, 2020.

[28] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE transactions on biomedical engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.

[29] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, IEEE, 2008, pp. 496–499.

[30] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature communications*, vol. 7, no. 1, pp. 1–10, 2016.

[31] X. Zhou and S. T. Wong, "Informatics challenges of high-throughput microscopy," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 63–72, 2006.

[32] M. Shifat-E-Rabbi, X. Yin, C. E. Fitzgerald, and G. K. Rohde, "Cell image classification: A comparative overview," *Cytometry Part A*, vol. 97, no. 4, pp. 347–362, 2020.

[33] M. V. Boland and R. F. Murphy, "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells," *Bioinformatics*, vol. 17, no. 12, pp. 1213–1223, 2001.

[34] M. Guillaud, D. Cox, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekirk, K. Adler-Storthz, N. Poulin, M. Follen, and C. MacAulay, "Quantitative histopathological analysis of cervical intra-epithelial neoplasia sections: Methodological issues," *Analytical Cellular Pathology*, vol. 26, no. 1-2, pp. 31–43, 2004.

[35] J. M. Prewitt and M. L. Mendelsohn, "The analysis of cell images," *Annals of the New York Academy of Sciences*, vol. 128, no. 3, pp. 1035–1053, 1966.

[36] R. Nosaka and K. Fukui, "Hep-2 cell classification using rotation invariant co-occurrence among local binary patterns," *Pattern Recognition*, vol. 47, no. 7, pp. 2428–2436, 2014.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[40] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[41] D. Wu, Y. Zhang, X. Jia, L. Tian, T. Li, L. Sui, D. Xie, and Y. Shan, "A high-performance cnn processor based on fpga for mobilenets," in *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, IEEE, 2019, pp. 136–143.

[42] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[43] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 10 990–10 997.

[44] J. H. Bappy and A. K. Roy-Chowdhury, "Cnn based region proposals for efficient object detection," in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 3658–3662.

[45] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *European conference on computer vision*, Springer, 2016, pp. 646–661.

[46] X. Zhang, X. Chen, L. Yao, C. Ge, and M. Dong, "Deep neural network hyperparameter optimization with orthogonal array tuning," in *International conference on neural information processing*, Springer, 2019, pp. 287–295.

[47] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020.

[48] M. Shifat-E-Rabbi, X. Yin, A. H. M. Rubaiyat, S. Li, S. Kolouri, A. Aldroubi, J. M. Nichols, and G. K. Rohde, "Radon cumulative distribution transform subspace modeling for image classification," *arXiv preprint arXiv:2004.03669*, 2020.

[49] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, "Wndcharm: Multi-purpose image classification using compound image transforms," *Pattern recognition letters*, vol. 29, no. 11, pp. 1684–1693, 2008.

[50] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *arXiv preprint arXiv:1805.12177*, 2018.

[51] X. Liang, D. Nguyen, and S. B. Jiang, "Generalizability issues with deep learning models in medicine and their potential solutions: Illustrated with cone-beam computed tomography (cbct) to computed tomography (ct) image conversion," *Machine Learning: Science and Technology*, vol. 2, no. 1, p. 015 007, 2020.

[52] G. V. Ponomarev, V. L. Arlazarov, M. S. Gelfand, and M. D. Kazanov, "Ana hep-2 cells image classification using number, size, shape and localization of targeted cell regions," *Pattern Recognition*, vol. 47, no. 7, pp. 2360–2366, 2014.

[53] Z. Gao, L. Wang, L. Zhou, and J. Zhang, "Hep-2 cell image classification with deep convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 21, no. 2, pp. 416–428, 2016.

[54] X. Qi, G. Zhao, J. Chen, and M. Pietikäinen, "Exploring illumination robust descriptors for human epithelial type 2 cell classification," *Pattern Recognition*, vol. 60, pp. 420–429, 2016.

[55] D. Strigl, K. Kofler, and S. Podlipnig, "Performance and scalability of gpu-based convolutional neural networks," in *2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing*, IEEE, 2010, pp. 317–324.

[56] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "On the limitation of convolutional neural networks in recognizing negative images," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2017, pp. 352–358.

[57] A. A. I. Sina, L. G. Carrascosa, Z. Liang, Y. S. Grewal, A. Wardiana, M. J. Shiddiky, R. A. Gardiner, H. Samaratunga, M. K. Gandhi, R. J. Scott, *et al.*, "Epigenetically repro-

grammed methylation landscape drives the dna self-assembly and serves as a universal cancer biomarker," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[58] D. Szüts, "A fresh look at somatic mutations in cancer," *Science*, vol. 376, no. 6591, pp. 351–352, 2022.

[59] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, e745–e750, 2021.

[60] J. Shen, C. J. Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, S. Y. Wong, P.-H. Fang, W.-K. Ming, *et al.*, "Artificial intelligence versus clinicians in disease diagnosis: Systematic review," *JMIR medical informatics*, vol. 7, no. 3, e10010, 2019.

[61] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of ai," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 088–30 095, 2020.

[62] A. Foote, A. Asif, N. Rajpoot, and F. Minhas, "Reet: Robustness evaluation and enhancement toolbox for computational pathology," *arXiv preprint arXiv:2201.12311*, 2022.

[63] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[64] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 142–151.

[65] L. Langnickel and J. Fluck, "We are not ready yet: Limitations of transfer learning for disease named entity recognition," *bioRxiv*, 2021.

[66] S. Soltan, H. Khan, and W. Hamza, "Limitations of knowledge distillation for zero-shot transfer learning," in *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, 2021, pp. 22–31.

[67] N. Goossens, S. Nakagawa, X. Sun, and Y. Hoshida, "Cancer biomarker discovery and validation," *Translational cancer research*, vol. 4, no. 3, p. 256, 2015.

[68] A. Kamb, S. Wee, and C. Lengauer, "Why is cancer drug discovery so difficult?" *Nature reviews Drug discovery*, vol. 6, no. 2, pp. 115–120, 2007.

[69] D. A. Ahlquist, "Universal cancer screening: Revolutionary, rational, and realizable," *NPJ precision oncology*, vol. 2, no. 1, pp. 1–5, 2018.

[70] G. Wolberg, "Image morphing: A survey," *The visual computer*, vol. 14, no. 8-9, pp. 360–372, 1998.

[71] S. Kolouri, S. R. Park, and G. K. Rohde, "The radon cumulative distribution transform and its application to image classification," *IEEE transactions on image processing*, vol. 25, no. 2, pp. 920–934, 2016.

[72] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE signal processing magazine*, vol. 34, no. 4, pp. 43–59, 2017.

[73] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, "A linear optimal transportation framework for quantifying and visualizing variations in sets of images," *International journal of computer vision*, vol. 101, no. 2, pp. 254–269, 2013.

[74] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[75] S. Kolouri, Y. Zou, and G. K. Rohde, "Sliced wasserstein kernels for probability distributions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5258–5267.

[76] S. R. Park, L. Cattell, J. M. Nichols, A. Watnik, T. Doster, and G. K. Rohde, "De-multiplexing vortex modes in optical communications using transport-based pattern recognition," *Optics express*, vol. 26, no. 4, pp. 4004–4022, 2018.

[77] F. Pontén, K. Jirström, and M. Uhlen, "The human protein atlas—a tool for pathology," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 216, no. 4, pp. 387–393, 2008.

[78] K. Tomczak, P. Czerwinska, and M. Wiznerowicz, "Review the cancer genome atlas (tcga): An immeasurable source of knowledge," *Contemporary Oncology*, vol. 2015, no. 1, pp. 68–77, 2015.

[79] S. R. Park, S. Kolouri, S. Kundu, and G. K. Rohde, "The cumulative distribution transform and linear pattern classification," *Applied and computational harmonic analysis*, vol. 45, no. 3, pp. 616–641, 2018.

[80] R. N. Bracewell and R. N. Bracewell, *The Fourier transform and its applications*. McGraw-Hill New York, 1986, vol. 31999.

[81] E. T. Quinto, "An introduction to x-ray tomography and radon transforms," in *Proceedings of symposia in Applied Mathematics*, vol. 63, 2006, p. 1.

[82] F. Natterer, *The mathematics of computerized tomography*. SIAM, 2001.

[83] Y. Brenier, "Polar factorization and monotone rearrangement of vector-valued functions," *Commun. Pure Appl. Math.*, vol. 44, no. 4, pp. 375–417, 1991.

[84] C. Villani, *Topics in Optimal Transportation*, 58. American Mathematical Soc., 2003.

[85] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[86] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[87] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Ieee, vol. 1, 2005, pp. 886–893.

[88] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, Ieee, vol. 2, 1999, pp. 1150–1157.

[89] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary, "Pywavelets: A python package for wavelet analysis," *Journal of Open Source Software*, vol. 4, no. 36, p. 1237, 2019.

[90] *Kaggle: Sign Language MNIST*, https://www.kaggle.com/datamunge/sign-language-mnist, Accessed: 2020-03-10.

[91] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Hoggles: Visualizing object detection features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1–8.

[92] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[93] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[94] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[95] F. C. Pampel, *Logistic regression: A primer*. Sage publications, 2020, vol. 132.

[96] A. H. M. Rubaiyat, K. M. Hallam, J. M. Nichols, M. N. Hutchinson, S. Li, and G. K. Rohde, "Parametric signal estimation using the cumulative distribution transform," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3312–3324, 2020.

[97] J. Nichols, T. Emerson, L Cattell, S Park, A Kanaev, F Bucholtz, A Watnik, T Doster, and G. Rohde, "Transport-based model for turbulence-corrupted imagery," *Applied optics*, vol. 57, no. 16, pp. 4524–4536, 2018.

[98] M. Shifat-E-Rabbi, X. Yin, A. H. M. Rubaiyat, S. Li, S. Kolouri, A. Aldroubi, J. M. Nichols, G. K. Rohde, *et al.*, "Radon cumulative distribution transform subspace modeling for image classification," *Journal of Mathematical Imaging and Vision*, vol. 63, no. 9, pp. 1185–1203, 2021.

[99] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[100] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Advances in neural information processing systems*, vol. 30, 2017.

[101] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[102] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[103] J. Nichols, T. Emerson, L Cattell, S Park, A Kanaev, F Bucholtz, A Watnik, T Doster, and G. Rohde, "Transport-based model for turbulence-corrupted imagery," *Applied optics*, vol. 57, no. 16, pp. 4524–4536, 2018.

[104] Y. Zhuang, S. Li, M. Shifat-E-Rabbi, X. Yin, A. H. M. Rubaiyat, G. K. Rohde, *et al.*, "Local sliced-wasserstein feature sets for illumination-invariant face recognition," *arXiv preprint arXiv:2202.10642*, 2022.

[105] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12 203–12 213, 2020.

[106] M. Shifat-E-Rabbi, Y. Zhuang, S. Li, A. H. M. Rubaiyat, X. Yin, and G. K. Rohde, "Invariance encoding in sliced-wasserstein space for image classification with limited training data," *Pattern Recognition*, vol. 137, p. 109 268, 2023.

[107] A. H. M. Rubaiyat, M. Shifat-E-Rabbi, Y. Zhuang, S. Li, and G. K. Rohde, "Nearest subspace search in the signed cumulative distribution transform space for 1d signal classifi-

cation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3508–3512.

[108] A. H. M. Rubaiyat, S. Li, X. Yin, M. Shifat-E-Rabbi, Y. Zhuang, and G. K. Rohde, "End-to-end signal classification in signed cumulative distribution transform space," *arXiv preprint arXiv:2205.00348*, 2022.

[109] A. Aldroubi, S. Li, and G. K. Rohde, "Partitioning signal classes using transport transforms for data analysis and machine learning," *Sampl. Theory Signal Process. Data Anal.*, vol. 19, no. 6, 2021.

[110] C. Moosmüller and A. Cloninger, "Linear optimal transport embedding: Provable wasserstein classification for certain rigid transformations and perturbations," *Information and Inference: A Journal of the IMA*, vol. 12, no. 1, pp. 363–389, 2023.

[111] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[112] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[113] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Fspool: Learning set representations with featurewise sort pooling," *arXiv preprint arXiv:1906.02795*, 2019.

[114] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python. the journal of machine learning research 12," 2011.

[115] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[116] Q. Wang, J. Xie, W. Zuo, L. Zhang, and P. Li, "Deep cnns meet global covariance pooling: Better representation and generalization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2582–2597, 2020.

[117] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 367–374.

[118] C. Garcia, "Point cloud mnist 2d," 2020.

[119] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[120] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[121] C. Liu, F. Shang, J. A. Ozolek, and G. K. Rohde, "Detecting and segmenting cell nuclei in two-dimensional microscopy images," *Journal of Pathology Informatics*, vol. 7, no. 1, p. 42, 2016.

[122] J. Mertz, *Introduction to optical microscopy*. Cambridge University Press, 2019.

[123] M. S. E. Rabbi, Y. Zhuang, S. Li, A. H. M. Rubaiyat, X. Yin, and G. K. Rohde, "Invariance encoding in sliced-wasserstein space for image classification with limited training data," *arXiv preprint arXiv:2201.02980*, 2022.

[124] S. Kundu, B. G. Ashinsky, M. Bouhrara, E. B. Dam, S. Demehri, M. Shifat-E-Rabbi, R. G. Spencer, K. L. Urish, and G. K. Rohde, "Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning," *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24 709–24 719, 2020.

[125] C. Zhang, M. Herbig, Y. Zhou, M. Nishikawa, M. Shifat-E-Rabbi, H. Kanno, R. Yang, Y. Ibayashi, T.-H. Xiao, G. K. Rohde, *et al.*, "Real-time intelligent classification of covid-19 and thrombosis via massive image-based analysis of platelet aggregates," *medRxiv*, 2022.

[126] J. A. Ozolek, A. B. Tosun, W. Wang, C. Chen, S. Kolouri, S. Basu, H. Huang, and G. K. Rohde, "Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning," *Medical image analysis*, vol. 18, no. 5, pp. 772–780, 2014.

[127] A. B. Tosun, O. Yergiyev, S. Kolouri, J. F. Silverman, and G. K. Rohde, "Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens," *Cytometry Part A*, vol. 87, no. 4, pp. 326–333, 2015.

[128] M. G. Hanna, C. Liu, G. K. Rohde, and R. Singh, "Predictive nuclear chromatin characteristics of melanoma and dysplastic nevi," *Journal of Pathology Informatics*, vol. 8, no. 1, p. 15, 2017.

[129] C. L. Sawyers, "The cancer biomarker problem," *Nature*, vol. 452, no. 7187, pp. 548–552, 2008.

[130] C. Oldenhuis, S. Oosting, J. Gietema, and E. De Vries, "Prognostic versus predictive value of biomarkers in oncology," *European journal of cancer*, vol. 44, no. 7, pp. 946–953, 2008.

[131] C. Arora, D. Kaur, and G. P. Raghava, "Universal and cross-cancer prognostic biomarkers for predicting survival risk of cancer patients from expression profile of apoptotic pathway genes," *Proteomics*, vol. 22, no. 3, e2000311, 2022.

[132] M. S. Tockman, P. K. Gupta, N. J. Pressman, and J. L. Mulshine, "Considerations in bringing a cancer biomarker to clinical application," *Cancer Research*, vol. 52, no. 9_Supplement, 2711s–2718s, 1992.

[133] A. H. Fischer, C. Zhao, Q. K. Li, K. S. Gustafson, I.-E. Eltoum, R. Tambouret, B. Benstein, L. C. Savaloja, and P. Kulesza, "The cytologic criteria of malignancy," *Journal of cellular biochemistry*, vol. 110, no. 4, pp. 795–811, 2010.

[134] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, and G. K. Rohde, "An optimal transportation approach for nuclear structure-based pathology," *IEEE transactions on medical imaging*, vol. 30, no. 3, pp. 621–631, 2010.

[135] G. K. Rohde, A. J. Ribeiro, K. N. Dahl, and R. F. Murphy, "Deformation-based nuclear morphometry: Capturing nuclear shape variation in HeLa cells," *Cytometry Part A*, vol. 73, no. 4, pp. 341–350, 2008.

Appendix A: Mathematical proofs and derivations

A.1 Proof of Property 1.1-A

Let $s(x)$ denote a normalized signal and let $\widehat{s}(x)$ be the CDT of $s(x)$. The CDT of $s_g = g's \circ g$ is given by

$$\widehat{s}_g = g^{-1} \circ \widehat{s}$$

Proof. Let r denote a reference signal. If $\widehat{s}$ and $\widehat{s}_g$ denote the CDTs of s and s_g, respectively, with respect to the reference r, we have that

$$\int_{-\infty}^{\widehat{s}(x)} s(u)du = \int_{-\infty}^{\widehat{s}_g(x)} s_g(u)du = \int_{-\infty}^{x} r(u)du$$

By substituting $s_g = g's \circ g$ we have

$$\int_{-\infty}^{\widehat{s}(x)} s(u)du = \int_{-\infty}^{\widehat{s}_g(x)} g'(u)s(g(u))du \tag{A.1.1}$$

By the change of variables theorem, we can replace $g(u) = v$, $g'(u)du = dv$ in equation (A.1.1):

$$\int_{-\infty}^{\widehat{s}(x)} s(u)du = \int_{-\infty}^{g\left(\widehat{s}_g(x)\right)} s(v)dv \tag{A.1.2}$$

From equation (A.1.2), we have that

$$g\left(\widehat{s}_g(x)\right) = \widehat{s}(x) \implies \widehat{s}_g(x) = g^{-1}\left(\widehat{s}(x)\right) \text{ or, } \widehat{s}_g = g^{-1} \circ \widehat{s}$$

□

A.2 Proof of Property 1.1-B

Recall that given two signals s and r, the Wasserstein metric $W_2(\cdot,\cdot)$ between them is defined in the following way:

$$W_2^2(s,r) = \int_{\Omega_r} (\widehat{s}(x) - x)^2 r(x) dx, \tag{A.2.1}$$

where $\widehat{s}$ is the CDT of s with respect to r.

Proof. Recall that an isometric embedding between two metric spaces is an injective mapping that preserve distances. Define the embedding by the correspondence $s \mapsto \widehat{s}$, it is left to show that

$$W_2^2(s_1,s_2) = \left\| (\widehat{s}_1 - \widehat{s}_2)\sqrt{r} \right\|^2_{L^2(\Omega_r)},$$

for all signals s_1, s_2. Let $f(y)$ be the CDT of s_2 with respect to s_1, then

$$W_2^2(s_2,s_1) = \int_{\Omega_{s_1}} (f(y) - y)^2 s_1(y) dy.$$

By the definition of CDT, $s_1 = f' s_2 \circ f$ and $r = \widehat{s}_1' s_1 \circ \widehat{s}_1$. Then by the composition property, $\widehat{s}_1 = f^{-1} \circ \widehat{s}_2$. Here again $\widehat{s}_1, \widehat{s}_2$ are CDT with respect to a fixed reference r. Let $y = \widehat{s}_1(x)$. Using the change of variables formula,

$$\begin{aligned} W_2^2(s_1,s_2) &= \int_{\Omega_r} (f(\widehat{s}_1(x) - \widehat{s}_1(x)) s_1(\widehat{s}_1(x)) \widehat{s}_1'(x) dx \\ &= \int_{\Omega_r} (\widehat{s}_2(x) - \widehat{s}_1(x))^2 r(x) dx \\ &= ||(\widehat{s}_2 - \widehat{s}_1)\sqrt{r}||^2_{L^2(\Omega_r)}. \end{aligned}$$

□

A.3 Proof of Property 1.3-A

Let $s(\mathbf{x})$ denote a normalized image and let $\widetilde{s}(t,\theta)$ and $\widehat{s}(t,\theta)$ are the Radon transform and the R-CDT transform of $s(x)$, respectively. The R-CDT of $s_{g^\theta} = \mathscr{R}^{-1}\left(\left(g^\theta\right)'\widetilde{s}\circ g^\theta\right)$ is given by

$$\widehat{s}_{g^\theta} = \left(g^\theta\right)^{-1}\circ\widehat{s}$$

Proof. Let r denote a reference image. Let $\widetilde{s}$ and $\widetilde{s}_{g^\theta}$ denote the Radon transforms of s and s_{g^θ}, respectively, and let $\widehat{s}$ and $\widehat{s}_{g^\theta}$ denote the CDTs of s and s_{g^θ}, respectively, with respect to the reference r. Then $\forall\theta\in[0,\pi]$, we have that

$$\int_{-\infty}^{\widehat{s}(t,\theta)}\widetilde{s}(u,\theta)du = \int_{-\infty}^{\widehat{s}_{g^\theta}(t,\theta)}\widetilde{s}_{g^\theta}(u,\theta)du = \int_{-\infty}^{t}\widetilde{r}(u,\theta)du$$

If we substitute $s_{g^\theta} = \mathscr{R}^{-1}\left(\left(g^\theta\right)'\widetilde{s}\circ g^\theta\right)$ or, $\widetilde{s}_{g^\theta} = \left(g^\theta\right)'\widetilde{s}\circ g^\theta$. Then $\forall\theta\in[0,\pi]$, we have

$$\int_{-\infty}^{\widehat{s}(t,\theta)}\widetilde{s}(u,\theta)du = \int_{-\infty}^{\widehat{s}_{g^\theta}(t,\theta)}\left(g^\theta\right)'(u)\widetilde{s}\left(g^\theta(u),\theta\right)du \tag{A.3.1}$$

By the change of variables theorem, we can replace $g^\theta(u) = v$, $\left(g^\theta\right)'(u)du = dv$ in equation (A.3.1):

$$\int_{-\infty}^{\widehat{s}(t,\theta)}\widetilde{s}(u,\theta)du = \int_{-\infty}^{g^\theta\left(\widehat{s}_{g^\theta}(t,\theta)\right)}\widetilde{s}(v,\theta)dv,\ \forall\theta\in[0,\pi] \tag{A.3.2}$$

From equation (A.3.2), we have that

$$\begin{aligned} & g^\theta\left(\widehat{s}_{g^\theta}(t,\theta)\right) = \widehat{s}(t,\theta) \\ \implies & \widehat{s}_{g^\theta}(t,\theta) = \left(g^\theta\right)^{-1}\left(\widehat{s}(t,\theta)\right) \text{ or, } \widehat{s}_{g^\theta} = \left(g^\theta\right)^{-1}\circ\widehat{s} \end{aligned}$$

□

A.4 Proof of Property 1.3-B

Recall that given two images s, r, using the correspondence in equation (6) the Sliced Wasserstein metric $SW_2(\cdot,\cdot)$ is defined as follows:

$$SW_2^2(s,r) = \int_{\Omega_{\tilde{r}}} (\widehat{s}(t,\theta) - t)^2 \tilde{r}(t,\theta) dt d\theta. \tag{A.4.1}$$

It can be shown that the above metric is well-defined [24], and in particular

$$SW_2^2(s_1,s_2) = \int_{\Omega_{\tilde{r}}} (\widehat{s}_1(t,\theta) - \widehat{s}_2(t,\theta))^2 \widetilde{r}(t,\theta) dt d\theta, \tag{A.4.2}$$

for all images s_1, s_2, the proof of which is essentially the same as in the CDT case in Appendix A.2.

Proof. Recall that an isometric embedding between two metric spaces is an injective mapping that preserve distances. Define the embedding by $s(\mathbf{x}) \mapsto \widehat{\mathbf{s}}(\mathbf{t},\theta)$ and the conclusion follows immediately from (A.4.2). □

Appendix B: Published works and other research contributions

Journal articles

Published

- Shifat-E-Rabbi M, Zhuang Y, Li S, Rubaiyat AH, Yin X, Rohde GK. Invariance encoding in sliced-Wasserstein space for image classification with limited training data. *Pattern Recognition*. 2022 Dec 15.

- Zhou Y, Nishikawa M, Kanno H, Xiao T, Suzuki T, Ibayashi Y, Harmon J, Takizawa S, Hiramatsu K, Nitta N, Kameyama R, Peterson W, Takiguchi J, Shifat-E-Rabbi M, Zhuang Y, Yin X, Rubaiyat AHM, Deng Y, Zhang H, Rohde GK, Iwasaki W, Yatomi Y, Goda K. Massive image-based single-cell profiling reveals high levels of circulating platelet aggregates in patients with COVID-19. *Nature Communications*. 2021 Dec 9;12(1):1-2.

- Zhang C, Herbig M, Zhou Y, Nishikawa M, Shifat-E-Rabbi M, Kanno H, Yang R, Ibayashi Y, Xiao TH, Rohde GK, Yatomi Y, Goda K. Real-time intelligent classification of COVID-19 and thrombosis via massive image-based analysis of platelet aggregates. *Cytometry Part A*.

- Shifat-E-Rabbi M, Yin X, Rubaiyat AHM, Li S, Kolouri S, Aldroubi A, Nichols JM, Rohde GK. Radon cumulative distribution transform subspace modeling for image classification. *Journal of Mathematical Imaging and Vision*. 2021 Aug 5:1-9.

- Kundu S, Ashinsky BG, Bouhrara M, Dam EB, Demehri S, Shifat-E-Rabbi M, Spencer RG, Urish KL, Rohde GK. Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning. *Proceedings of the National Academy of Sciences*. 2020 Sep 21.

- Shifat-E-Rabbi M, Yin X, Fitzgerald CE, Rohde GK. Cell Image Classification: A Comparative Overview. *Cytometry Part A*. 2020 Feb 10.

- Islam MS, Shifat-E-Rabbi M, Dobaie AM, Hasan MK. PREHEAT: Precision heart rate monitoring from intense motion artifact corrupted PPG signals using constrained RLS and wavelets. *Biomedical Signal Processing and Control*. 2017 Sep 1;38:212-23.

- Shifat-E-Rabbi M, Hasan MK. Speckle tracking and speckle content based composite strain imaging for solid and fluid filled lesions. *Ultrasonics*. 2017 Feb 1;74:124-39.

- Hasan MK, Shifat-E-Rabbi M, Lee SY. Blind deconvolution of ultrasound images using l1-norm-constrained block-based damped variable step-size multichannel LMS algorithm. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*. 2016 Jun 7;63(8):1116-30.

Submitted

- Shifat-E-Rabbi M, Ironside N, Ozolek JA, Singh R, Pantanowitz L, Rohde GK. Quantifying nuclear structures of digital pathology images across cancers using transport-based morphometry (Transport-based morphometry of nuclear structures of digital pathology images in cancer). submitted to *IEEE Journal of Biomedical Health and Informatics*.

- Rubaiyat AHM, Li S, Yin X, Shifat-E-Rabbi M, Zhuang Y, Rohde GK. End-to-End Signal Classification in Signed Cumulative Distribution Transform Space. submitted to *EEE Transactions on Pattern Analysis and Machine Intelligence*.

- Zhou Y, Nishikawa M, Kanno H, Xiao T, Suzuki T, Ibayashi Y, Harmon J, Takizawa S, Hiramatsu K, Nitta N, Kameyama R, Peterson W, Takiguchi J, Shifat-E-Rabbi M, Zhuang Y, Yin X, Rubaiyat AHM, Deng Y, Zhang H, Rohde GK, Iwasaki W, Yatomi Y, Goda K. The landscape of circulating platelet aggregates in COVID-19. *medRxiv*. 2021 Jan 1.

- Zhuang Y, Li S, Shifat-E-Rabbi M, Yin X, Rubaiyat AH, Rohde GK. Local Sliced-Wasserstein Feature Sets for Illumination-invariant Face Recognition. submitted to IEEE *Transactions on Image Processing*.

Ongoing works

- Shifat-E-Rabbi M, Rohde GK, et al. Linear optimal transport subspaces for point cloud classification.
- Nichols JM, Shifat-E-Rabbi M, Rohde GK, et al. Learning from spatial information in sports using Transport Based Morphometry.
- Pathan NS, Shifat-E-Rabbi M, Rohde GK, et al. Signed image classification using R-S-CDT.
- Ironside N, Shifat-E-Rabbi M, Rohde GK, et al. Automated prediction of hematoma growth in spontaneous intracerebral hemorrhage patients.
- Hassan MA, Shifat-E-Rabbi M, Rohde GK, et al. Eye tracking-based neurological disorder diagnosis using discrete R-CDT.

Conferences

- Rubaiyat AHM, Shifat-E-Rabbi M, Zhuang Y, Li S, Rohde GK. Nearest Subspace Search in The Signed Cumulative Distribution Transform Space for 1D Signal Classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*. 2022, Singapore.
- Shifat-E-Rabbi M, Rohde GK. Scientific tutorial: Cell Image Classification: An Overview of Methods with Software Examples. *CYTO, 2019*, Vancouver, Canada.
- (Abstract) Shifat-E-Rabbi M, Rubaiyat AM, Zhuang Y, Rohde GK. A sliced-Wasserstein distance-based approach for out-of-class-distribution detection.

- (Abstract) Zhang C, Herbig M, Zhou Y, Nishikawa M, Shifat-E-Rabbi M, Kanno H, Yang R, Ibayashi Y, Xiao TH, Rohde GK, Sato M, Kodera S, Daimon M, Yatomi Y, Goda K. Real-time intelligent classification of COVID-19 and thrombosis with label-free bright-field images of platelet aggregates. *JSAP* 2023.

Others

- Software developments: one of the contributors in PyTransKit , TBM package.
- Kundu S, Ashinsky BG, Bouhrara M, Dam EB, Demehri S, Shifat-E-Rabbi M, Spencer RG, Urish KL, Rohde GK. Reply to Roemer and Guermazi: Early biochemical changes on MRI can predict risk of symptomatic progression. *Proceedings of the National Academy of Sciences of the United States of America*. 2021 Mar 16;118(11):e2024679118.
- A Aldroubi, S Li, GK Rohde. Partitioning signal classes using transport transforms for data analysis and machine learning. 2020. Co-featured in Acknowledgement.
- M Shifat-E-Rabbi, X Yin, CE Fitzgerald, GK Rohde. Featured in a short highlight on the cover article for *Cytometry Part A*. Apr 2020 issue.
- MK Hasan, M Shifat-E-Rabbi, SY Lee. Cover article for *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency control*. Aug 2016 issue.