Implementing Information Geometry with respect to imaging biomedical datasets

DISSERTATION

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

Master of Science in Computer Engineering

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Aneek Roy

08 2024

Implementing Information Geometry with respect to imaging biomedical datasets

DISSERTATION

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

Master of Science in Computer Engineering

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Aneek Roy

08 2024

Approved:

____

Department Chair Signature

____

Date

University ID: N18324528
Net ID: ar8002

Approved by the Guidance Committee:

Major: Computer Engineering

Advisor's Name
Professor of
Financial and Risk Engineering

Date

Advisor's Name
Assistant Professor of
Financial and Risk Engineering

Date

Advisor's Name
Assistant Professor of
Financial and Risk Engineering

Date

Vita

Aneek Roy.

Acknowledgements

Aneek Roy [redacted]

08 2024

To Mom and Dad, and to my advisor and mentor Professor Amine. To all the souls who stood by me during these tough times.

ABSTRACT

Implementing Information Geometry with respect to imaging biomedical datasets

by

Aneek Roy

Advisor: Prof. Professor Amine Mohamed Aboussalah, Ph.D.

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Financial Engineering

08 2024

Contents

Vita . . . iii

Acknowledgements . . . iv

Abstract . . . vi

List of Figures . . . ix

List of Tables . . . x

1 Abstract **1**

1.1 Introduction . . . 2

2 Literature Review **9**

3 Methodology **22**

3.1 Traditional vs. Data-Driven Approaches . . . 22

3.2 Self-Supervised Learning in Medical Imaging . . . 23

3.3 Study Design and Methodology . . . 23

3.4 Lesion Anatomical Site Classification . . . 26

3.5 Nodule Malignancy Prediction . . . 26

3.6 NSCLC(Non-Small Cell Lung Cancer) Prognostication . . . 27

3.7 Stability and Interpretability . . . 27

3.8 Biological Associations . . . 27

4 Empirical Results 30

5 Conclusions and Future Work 44

List of Figures

List of Tables

4.1 Validation Performance Summary 30
4.2 Test Performance Summary 31
4.3 Euclidean KNN Validation Performance Summary 31
4.4 Euclidean KNN Test Performance Summary 32
4.5 Riemannian KNN Validation Performance Summary 32
4.6 Riemannian KNN Test Performance Summary 32
4.7 Performance Summary with Best k = 16 33
4.8 Confusion Matrix 33
4.9 Classification Report for Best KNN Model 38
4.10 Test Accuracy for Best KNN Model 38
4.11 KNN Classification Results 42
4.12 Task 2, Task 3 and original for KNN Classification 42

Chapter 1

Abstract

Recent advancements in medical imaging have seen the integration of artificial intelligence (AI) and deep learning (DL) models, which have revolutionized the field by offering novel methodologies for disease detection, diagnosis, and prognosis. This thesis explores the potential of foundation models in medical imaging, focusing on the generation of biomarkers from medical imaging datasets and evaluating the efficacy of beta representations in lesion detection tasks across modalities, such as PET-CT scans. While traditional biomarker identification relies heavily on hypothesis-driven approaches, which are often labor-intensive and constrained by existing knowledge, this research leverages data-driven methodologies such as Convolutional Neural Networks (CNNs) for novel biomarker discovery.

Our work investigates the application of self-supervised learning (SSL) methods, including clustering techniques like k-means and k-nearest neighbors (k-NN) classification, to annotate and identify biomarkers with minimal manual intervention. This study also highlights the role of autoencoders in dimensionality reduction and feature extraction, allowing for effective biomarker discovery and classification.

Additionally, we delve into the use of information geometric principles, specifically the beta representation, to model the manifold of medical imaging data and improve the performance of clustering and classification tasks.

The research examines the impact of class distribution on the efficacy of classifiers and clustering techniques, comparing the Riemannian metric with traditional Euclidean approaches. By employing saliency heat maps and exploring the underlying data distribution, this study provides insights into the manifold structure and its influence on clustering and classification outcomes. Our findings demonstrate the robustness of the beta representation form, showcasing significant improvements in performance when utilizing the Riemannian metric. The results underscore the transformative potential of advanced geometric representations in enhancing the interpretability and performance of machine learning techniques in medical imaging.

This thesis contributes to the understanding of information geometry's application in medical imaging, emphasizing the development of interpretable models tailored for complex data structures and highlighting the potential for personalized medicine advancements. Through this work, we aim to advance the field of medical imaging by demonstrating the efficacy of innovative DL frameworks in improving diagnostic accuracy and patient outcomes.

1.1 Introduction

The field of medical imaging has witnessed remarkable advancements with the integration of artificial intelligence (AI), particularly deep learning (DL). The integration of foundation models into medical imaging holds considerable promise, given the routine collection of multimodal data (e.g., medical images, biological data, clinical notes) in clinical settings. Foundation models can potentially revolutionize applications such as augmented surgical procedures, bedside decision support, and interactive radiology reports,which require biomarker identifications as part of diagnostic and investigative procedures. Traditionally, biomarker identification in medical imaging has relied on hypothesis- driven approaches, constrained by existing knowledge and biases. These methods often require extensive manual effort and expertise. Conversely, data-driven methodologies powered by Convolutional Neural Nets and Vision Transformers can uncover novel biomarkers from imaging data with minimal manual intervention. However, the robustness of DL models is contingent on the availability and quality of annotated datasets, which are often limited in medical imaging datasets. Deep learning (DL) has become increasingly prominent in medical imaging, offering new methodologies for disease detection, diagnosis, and prognosis through both supervised and unsupervised learning approaches. Supervised learning in DL has shown remarkable success in medical imaging applications, as exemplified by Esteva et al. (2017), who utilized deep neural networks to diagnose skin cancer with dermatologist-level accuracy. Similarly, De Fauw et al. (2018) demonstrated the application of DL models in triaging retinal diseases, and Rajpurkar et al. (2017) effectively used these techniques to detect pneumonia from chest X-rays, achieving radiologist-level performance. These studies underscore the potential of DL in clinical decision-making, although they are often limited by the requirement for large annotated datasets and predefined biomarkers. Conversely, unsupervised learning offers a promising alternative for unbiased biomarker discovery, as shown by Waldstein et al. (2020), who introduced an unsupervised deep learning framework for analyzing optical coherence tomography (OCT) images. This approach employed autoencoders to identify local and global features from OCT scans, leading to the discovery of 20 novel biomarkers associated with clinical outcomes such as visual acuity and lesion activity. Such unsupervised methods emphasize the potential for novel biomarker discovery in medical imaging, particularly in scenarios where annotated data is scarce.

In the domain of deep learning frameworks, autoencoders and convolutional neural networks (CNNs) are pivotal for medical imaging applications. Waldstein et al.'s (2020) methodology involved a two-stage autoencoder pipeline that captures local and global features from OCT images, reducing data complexity and enabling effective biomarker discovery. This process transforms millions of voxels into compact feature sets, facilitating the identification of novel biomarkers with potential clinical relevance. CNNs have also demonstrated efficacy in medical imaging, as demonstrated by Esteva et al. (2017) in their work on skin cancer classification, achieving performance on par with dermatologists. CNNs employ multiple layers, including convolutional, activation (ReLU), pooling, and fully connected layers, to extract hierarchical features from input images, making them suitable for image classification and segmentation tasks. The effectiveness of these frameworks lies in their ability to process large, high-dimensional data, thereby providing valuable insights into disease processes and aiding in the development of personalized treatment plans.

The application of deep learning in medical imaging extends beyond disease detection and diagnosis to encompass cancer pathology and multi-modal integration. DL systems have shown promise in automating tumor detection and grading, achieving pathologist-level performance across various cancer types. For instance, Bulten et al. (2020) and Ström et al. (2020) achieved AUROC values exceeding 0.99 in prostate cancer detection and grading. DL has also been used to predict genetic mutations directly from histology images, as demonstrated by Coudray et al. (2018) and Kather et al. (2019), who predicted mutations such as EGFR in lung cancer and microsatellite instability (MSI) in gastrointestinal cancers with considerable accuracy. Furthermore, DL systems can integrate visible and sub-visual features from histology images to predict survival outcomes and therapy responses, exemplified by studies such as Bychkov et al. (2018) and Courtiol et al. (2019). Additionally, multi-modal biomarker integration using techniques like multi-kernel learning (MKL) can improve disease classification and trajectory modeling by incorporating diverse sources of biological information. Aksman et al. (2019) further advanced this field by introducing a parametric Bayesian multi-task learning (MTL) framework for longitudinal imaging biomarkers, enhancing model performance in handling limited and noisy data. These advancements in DL emphasize the transformative potential of integrating multi-modal data and sophisticated learning techniques in advancing personalized medicine and improving patient outcomes.

Our work focuses on generating biomarkers from the medical imaging datasets, as shown by Suraj Pai et. al and using these to show the effectiveness of beta representations while using clustering techniques like k-means and k-nn classification over the medical imaging datasets across the domain of lesion detection for various modalities under PET-CT scans, as shown by Alice LeBrigant et al.The work on biomarker generation has historically involved human annotations, which involved man hours and tedious work loads for even small scale medical datasets. Thus with the advent of self-supervised learning methods to aggregate data based on its inherent similarity and dissimilarity structures of the data, SSL methods gained traction within the medical research community to annotate the biomarkers.

Biomarkers are measurable indicators utilized in the assessment of health conditions, the presence of diseases, their progression, and the responses to therapeutic interventions. They play a pivotal role in the medical field, particularly in the diagnosis of diseases, prognostic predictions, and the customization of treatment plans. Biomarkers can be obtained from various sources, including blood, urine, tissues, or imaging data, and are categorized into diagnostic, prognostic, predictive, pharmacodynamic, surrogate, and safety biomarkers. The ideal biomarker should exhibit specificity, sensitivity, reproducibility, non-invasiveness, and clinical relevance. They are integral to disease diagnosis and monitoring, evaluating the efficacy of treatment modalities, and facilitating drug development processes. Notable examples include blood glucose levels for diabetes management, prostate-specific antigen (PSA) for the detection of prostate cancer, C-reactive protein as a marker for inflammation, and genetic markers such as BRCA mutations which indicate cancer risk. Furthermore, advanced imaging techniques like magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) scans serve as valuable biomarkers by providing comprehensive insights into tissue structure and function. The advancements with regards to visual transformers have improved significantly how medical imaging datasets are segregated and with respect to lesion detection tasks Suraj Pai et al. used convolutional encoders using a modified form of SimCLR architecture. The biomarkers are then used in our work to segregate the lesion type using the extracted features of the foundational model,like Suraj Pai et al. had done to used them to run a linear classifier on the target lesion types.To show the efficacy of beta representations in the context of using information geometric principles to model manifold of the data while segregating them into clusters for unsupervised learning or to optimize according to k-nearest neighbours for classification tasks. We then find try to find the reasonable explanation behind the performance of the clustering and classification tasks such that a detailed analysis into the manifold structure and the correlation of the data distribution, and its impact over using riemannian metric while calculating the frechet mean for finding the cluster centers or while optimizing the knn classification algorithm in a riemannian manifold, versus using the euclidean distance metric for a gaussian data distribution, for similar set of tasks.To deep dive into the analysis we take into account the saliency heat maps of the classifier and the clustering technique. We also try to find a reasonable explanation behind the dimensionality reduction using beta representation and other techniques like principal component analysis as show by Suraj Pai et al. and how beta representation tends to work in these scenarios.

Our work can be broadly classified into different sections of introduction, literature review, methodology, results and analysis through causality and reasoning model, conclusion and bibliography.

The objective of this research is to explore the foundations of information geometry, specifically focusing on the beta representation forms for data models. This study aims to investigate the manifold representation and formulate an interpretable model tailored for medical imaging datasets. Additionally, this research examines the impact of class distribution, whether balanced or unbalanced, on the performance of classifiers and clustering techniques. By delving into these aspects, the study highlights the robustness of the beta representation form and demonstrates how utilizing the Riemannian metric can enhance performance compared to traditional Euclidean representations. In pursuing this objective, our research provides an in-depth analysis of how information geometry principles, particularly the beta representation, can be applied to medical imaging datasets. This involves investigating the underlying structure of data and exploring how manifold representations can yield more interpretable and accurate models. Our work places a strong emphasis on understanding the influence of class distribution within datasets, recognizing that balanced datasets may offer different challenges and advantages compared to unbalanced ones. Through this analysis, our research seeks to illuminate how different data distributions affect the efficacy of various classifiers and clustering methods. The findings of this study demonstrate the efficacy and robustness of the beta representation form in handling complex data structures. By leveraging the Riemannian metric, the research showcases significant improvements in performance over conventional Euclidean-based approaches. Our work contributes to a deeper understanding of information geometry's application in medical imaging and offers insights into developing more effective models for analyzing medical datasets. Our study's results underscore the potential of advanced geometric representations to enhance the interpretability and performance of machine learning techniques in medical imaging.

Chapter 2

Literature Review

Our work can be broadly divided into two sections. For the first part we work on obtaining the biomarkers from medical imaging datasets, specifically CT scans of the body and identify the lesions, with respect to various parts of the body as available in DeepLesion, LUNA16, RADIO and LUNG1 datasets. Once we have obtained the relevant features for each of the sampled data points, in the second part of our work, we focus on using them to model the equivalent beta representations so that we can classify and cluster the data points accordingly. For our first part we focus on the work by Suraj Pai et al. with Foundation model for cancer imaging biomarkers,which delves into the application of convolutional encoders with 3DResNet50 as the backbone of the base encoder for feature extraction with regards to the three separate classes of downstream tasks of lesion type, malignancy type and survival class detection operations. For this, we dive deep into the various relevant deep learning architectures which have been historically favored in the medical imaging domain for the past decade of work. Deep learning has seen a surge in application across various medical imaging modalities. Esteva et al. (2017) utilized the Inception v3 convolutional neural network (CNN) architecture to achieve dermatologist-level accuracy in skin cancer diagnosis. These networks operate through several layers, including convolutional, activation (ReLU), pooling, and fully connected layers, which extract hierarchical features from input images. This model, known for its efficient computation and feature extraction capabilities, was trained on a dataset of approximately 130,000 clinical images to classify skin lesions into malignant and benign categories. Similarly, De Fauw et al. (2018) employed a combination of CNNs and long short-term memory networks (LSTMs) to triage retinal diseases from 3D volumetric optical coherence tomography (OCT) scans. This approach enabled the model to capture both spatial features and sequential dependencies, facilitating disease identification and referral suggestions. Rajpurkar et al. (2017) used a 121-layer DenseNet architecture for pneumonia detection from chest X-rays. DenseNet's dense connectivity pattern allows for efficient feature reuse and gradient flow, enhancing learning and accuracy. Their model was trained on the CheXpert dataset, demonstrating robustness in identifying pneumonia and other pathologies. Despite their success, these supervised deep learning methods face challenges such as the requirement for large, annotated datasets and reliance on known biomarkers. These studies illustrate the effectiveness of deep learning architectures, such as Inception v3, CNN-LSTM combinations, and DenseNet, in transforming medical imaging, while also highlighting the challenges in data dependency and annotation requirements.

Unsupervised deep learning offers a promising alternative for unbiased biomarker discovery. Waldstein et al. (2020) introduced an unsupervised deep learning framework for analyzing optical coherence tomography (OCT) images [27]. This methodology employs autoencoders to identify local and global features from

OCT scans without prior domain knowledge. The analysis of 54,900 OCT scans from patients with neovascular age-related macular degeneration (AMD) led to the discovery of 20 novel biomarkers, validated through their correlation with clinical outcomes such as visual acuity and lesion activity. This unsupervised approach highlights the potential for unbiased biomarker discovery in medical imaging. Waldstein et al.'s (2020) deep learning pipeline involves two stages of autoencoders. The first autoencoder captures local features from individual A-scans, producing a 20-dimensional representation of local retinal morphology. The second autoencoder compresses these local features into a global representation of the entire OCT volume [27]. This method significantly reduces the complexity of the data, transforming millions of voxels into 20 compact features. The newly identified biomarkers from Waldstein et al.'s (2020) study were validated by comparing their correlation with clinical outcomes against conventional biomarkers. The novel features showed a stronger correlation with visual acuity (R2 = 0.46) compared to traditional markers (R2 = 0.29). Additionally, the unsupervised approach uncovered previously unknown biomarkers that are clinically relevant. For example, one of the new features (a5) was strongly correlated with visual function but not with traditional morphological markers, suggesting its potential as a subclinical biomarker for retinal disease [27]. When comparing the unsupervised approach to traditional methods, it becomes evident that unsupervised learning can uncover hidden patterns and novel biomarkers not evident through conventional methods. This is particularly valuable in medical fields where annotated data is scarce, and human-defined features may not capture the full complexity of the data [47].

Deep learning has made significant strides in automating tumor detection and grading in histopathology, with various studies demonstrating its robust capabilities across different cancer types. Bulten et al. (2020) developed a convolutional neural network (CNN) framework for Gleason grading of prostate biopsies, using the extensive PANDA dataset, which includes over 10,000 digitized prostate biopsy images. This model achieved pathologist-level performance, with an AUROC exceeding 0.99, showcasing the potential of AI to enhance diagnostic accuracy and efficiency. Similarly, Ström et al. (2020) utilized a CNN-based system for prostate cancer detection and grading, leveraging thousands of annotated whole-slide images from diverse sources. The model also achieved an AUROC of over 0.99, demonstrating its effectiveness in detecting and grading prostate cancer, which could significantly reduce the workload on pathologists and improve diagnostic consistency.

Deep learning has also been employed to predict genetic mutations directly from histology images. Coudray et al. (2018) utilized a CNN architecture to predict genetic mutations, such as EGFR mutations in lung cancer, from histopathology images. This study used The Cancer Genome Atlas (TCGA) dataset, which contains paired histology and genomic data, allowing the model to learn correlations between image patterns and genetic alterations. The model demonstrated high accuracy in predicting EGFR mutations, highlighting the potential of deep learning to non-invasively infer genetic profiles from histology, which could aid in personalized treatment planning. Kather et al. (2019) employed a similar CNN-based approach to predict microsatellite instability (MSI) in gastrointestinal cancers from histological images. Using a comprehensive dataset of colorectal cancer images with labeled data for MSI status, the model accurately predicted MSI, offering a valuable tool for identifying patients who might benefit from immunotherapy and underscoring the role of AI in precision oncology.

In addition to mutation prediction, deep learning systems have been used to predict patient survival and treatment response, which are critical aspects of cancer treatment planning. Bychkov et al. (2018) used a CNN-based deep learning system to predict survival outcomes in colorectal cancer patients by analyzing histopathological slides. The model was trained on a large cohort of colorectal cancer patients with linked survival data, demonstrating that deep learning could extract prognostic features from histology images that correlate with patient survival. Courtiol et al. (2019) developed a deep learning model based on multi-task learning that integrates visual and clinical features to predict survival in patients with mesothelioma. This approach, using a dataset of mesothelioma cases with detailed clinical annotations, showed that combining histological features with clinical data enhances predictive performance.

Similarly, Harder et al. (2019) applied a CNN model to predict immunotherapy responses in melanoma patients from histopathology images. With a dataset including melanoma biopsy images and documented treatment responses, the model demonstrated the ability to predict treatment response from histology images, highlighting the transformative impact of AI in guiding therapeutic decisions and identifying patients likely to benefit from specific treatment. Integrating multimodal biomarker information using multi-kernel learning (MKL) can improve trajectory estimates and predictions. This technique has been applied in disease classification and trajectory modeling, demonstrating its utility in improving biomarker trajectory estimates by incorporating various sources of biological information [36]. Aksman et al. (2019) introduced a parametric Bayesian multi-task learning (MTL) framework for modeling longitudinal imaging biomarkers, highlighting the importance of robust models to handle under-sampling and measurement errors. This approach extends traditional mixed-effects modeling by integrating multi-modal information and enhancing model performance in handling limited and noisy longitudinal data [36].

One of the significant challenges in deploying DL in medical imaging is the interpretability of models. Techniques like Grad-CAM help visualize the important regions contributing to model predictions, enhancing the transparency and interpretability of DL systems. Ethical considerations, including data privacy and adherence to regulatory standards, are crucial for the clinical adoption of DL technologies [37]. Standardizing imaging protocols and data formats across different institutions is essential for ensuring the reliability and generalizability of DL models. Extensive validation across diverse patient populations and clinical settings is necessary to confirm the utility of DL-based biomarkers. Without robust validation, DL models may not perform consistently across different clinical environments, potentially leading to incorrect diagnoses or treatment recommendations. Such validation involves multi-center studies, involving diverse patient demographics, to ensure that the DL models can generalize well to different populations and imaging protocols [15].

Achieving regulatory approval for DL models involves demonstrating their safety and efficacy through rigorous testing and validation. Regulatory bodies such as the FDA and EMA require comprehensive evidence of a model's performance across various clinical scenarios before granting approval. This includes not only technical performance metrics but also the model's impact on clinical workflows and patient outcomes. The dynamic nature of DL models, which allows them to adapt and learn over time, poses unique challenges for regulatory frameworks traditionally designed for static medical devices. Addressing these challenges will be crucial for integrating

DL technologies into routine clinical practice [16]. Future research should focus on improving the interpretability of DL models, enhancing data standardization, and exploring new applications of unsupervised learning techniques. There is a need for continuous collaboration between AI researchers, clinicians, and regulatory bodies to ensure that the development and deployment of DL models are aligned with clinical needs and ethical standards. Additionally, integrating clinical data with imaging data can further enhance the predictive power and clinical utility of these models, paving the way for more personalized and precise medical interventions.

Deep learning has the potential to revolutionize medical imaging and biomarker discovery, enabling the unbiased identification of novel biomarkers and improving diagnostic accuracy. Studies by Esteva et al. (2017), De Fauw et al. (2018), and Waldstein et al. (2020) demonstrate the feasibility and advantages of unsupervised and supervised learning techniques in various imaging modalities. Despite the challenges, continued research and collaboration across disciplines will be crucial to fully realizing the benefits of DL in healthcare, ultimately enhancing patient outcomes and advancing precision medicine.

Foundational models and traditional deep learning architectures differ significantly in their approach to computer vision and imaging tasks. Foundational models, such as CLIP and DINO, are large-scale, transformer-based architectures trained on extensive datasets to generalize across a wide range of tasks, offering multi-modal capabilities that integrate text and images. They excel in transfer learning and can be adapted for specific tasks with minimal additional data. In contrast, traditional architectures like CNNs, transformers (e.g., Vision Transformers), and GANs are designed for specific tasks. CNNs excel in image classification, segmentation, and detection by leveraging convolutional layers for spatial hierarchies, while GANs generate new images by learning data distributions through adversarial training. Traditional models typically require task-specific datasets and substantial retraining to adapt to new tasks. Foundational models provide a unified approach, enhancing efficiency and versatility in tasks like automated image captioning and cross-modal retrieval. However, their large scale poses challenges in computational resources and interpretability, whereas traditional models remain crucial for specific applications like medical imaging and autonomous driving but lack the adaptability and generalization of foundational models.

SimCLR, SwAV, and NNCLR are notable self-supervised learning approaches in computer vision, which are used as baselines by Suraj Pai et al. that focus on learning image representations without labeled data. These methods differ from foundational models by leveraging specific unsupervised training paradigms tailored to enhance image representation learning. When compared to foundational models, these self-supervised learning approaches differ in scale and training focus. Foundational models are typically larger and trained on more diverse datasets, capturing a broader range of information and generalizing across various tasks and domains. In contrast, SimCLR, SwAV, and NNCLR are optimized for efficient representation learning from specific image datasets and are primarily used for specific vision tasks. While foundational models often incorporate multi-modal learning (e.g., CLIP combines text and images), these self-supervised methods focus solely on image representations.

The applications and impact of these methods are significant, especially in the context of transfer learning and data efficiency. The representations learned by SimCLR, SwAV, and NNCLR can be fine-tuned for downstream tasks like image classification, detection, and segmentation, offering robust performance without extensive labeled datasets. By leveraging unlabeled data, these methods reduce the need for large labeled datasets, making them valuable in scenarios where labeled data is scarce. These self-supervised approaches have advanced the field of computer vision by demonstrating that competitive performance can be achieved without relying on extensive labeled datasets, complementing the broader trends where foundational models provide generalized solutions while these methods optimize specific tasks through innovative training paradigms.

For the second part of our work we look into the relevant detailed works with respect to information geometry and manifold representations, in the field of medical images and patient data. Yet first we look into the field of information geometry, with generalized gamma representations, and derived beta distributions and its correlations. The application of information geometry to statistical analysis has opened new pathways for understanding complex data sets. Information geometry as a mathematical field applies differential geometry to the study of probability distributions and statistical models, offering a geometric framework for understanding the relationships, comparisons, and parameterizations of different probability distributions. In this context, a statistical manifold is conceptualized as a geometric space where each point represents a distinct probability distribution, equipped with a Riemannian metric, such as the Fisher information metric, to measure distances and angles between these distributions. Divergence functions, like the Kullback-Leibler divergence, further define the manifold's geometric structure by quantifying discrepancies between distributions. Central to information geometry are exponential families, characterized by their natural parameterization using exponential functions, and the dual connections that provide insights into the manifold's curvature and geometric properties. Beta distributions, defined by shape parameters and , exemplify the exponential family and are pivotal in beta representations within information geometry. These representations employ beta distributions or their generalizations as foundational elements in modeling data, particularly in contexts where the data aligns with the characteristics of beta distributions, such as proportions or percentages. Applications of beta representations span clustering and classification, where they capture data variability, to Bayesian inference, where they serve as prior distributions in binomial and Bernoulli processes, and extend to machine learning models requiring flexible distributions. Information geometry offers analytical tools to explore the parameter space of beta distributions, using geometric techniques to investigate how different beta distributions relate, emphasizing both natural and expectation parameterizations through dual connections. Moreover, the use of the Fisher information metric and divergence functions enables precise measurement of distances and discrepancies between beta distributions, enhancing statistical inference and hypothesis testing. Specifically, the use of beta distributions in classifying histograms of medical data offers a robust framework for diagnostic accuracy and treatment monitoring. Fisher-Rao geometry on Dirichlet distributions was explored by Le Brigant, Preston, and Puechmorel (2020), emphasizing its application in medical data where accurate classification and averaging of distributions are necessary. Similarly, Amari (2016) highlights how geometric conditions in information geometry ensure consistency and efficiency of estimators, as shown by Le Brigant and Puechmorel (2019) while exploring the Fisher-Rao geometry on beta distributions, demonstrating the negativity of sectional curvature and the robustness of the Riemannian centroid for analyzing canonical moments. Arwini and Dodson (2008) discuss the use of Riemannian metrics for analyzing probability distributions, beneficial for medical histograms. Foundational insights into the geometry of statistical manifolds are provided by Lauritzen (1987) and Skovgaard (1984), emphasizing the importance of understanding parametric families like beta distributions in medical data classification.

Several studies illustrate the effectiveness of histogram-based features in classification tasks. Manikonda and Gaonkar (2020) introduce an image classification method using histogram of oriented gradient features, relevant for understanding how histogram-based techniques are applied in islanding detection in electric grids, which can be used in other modalities as well . Borenstein-Levin et al. (2022) focus on classifying oxygen saturation histograms to evaluate treatment efficacy in preterm infants, showcasing the utility of histogram classification in medical applications. Mahmood and Mahmood (2015) present a segmentation method for skin cancer images using histogram classification, emphasizing the effectiveness of histogram-based techniques in medical image analysis. The works of Le Brigant and Puechmorel (2019) and Dette and Studden (1997) significantly contribute to the understanding of information geometry and beta representations. Le Brigant and Puechmorel explore the Fisher-Rao geometry on beta distributions, revealing that the sectional curvature is negative, indicating that the statistical manifold has hyperbolic properties. This characteristic affects how probability distributions are spaced and informs the robustness of the Riemannian centroid, which is valuable for summarizing canonical moments and offers stability against variability and outliers. Dette and Studden's work on canonical moments emphasizes their symmetry and invariance properties, which are crucial for consistent statistical analysis, as they provide an unbiased parameterization of distributions, particularly useful for describing beta distributions. Additionally, the use of geometric properties and canonical moments can refine decision-making processes by offering a clearer understanding of uncertainty and variability, thereby making them highly applicable to the problem of lesion detection in medical images, which forms the basis of our work.

Rebbah, Nicol, and Puechmorel (2019), in an earlier work introduce the generalized gamma manifold and its application to Alzheimer's disease classification, showing improved performance of classification algorithms by leveraging geometric properties of distributions.

Le Brigant et al. (2022) introduce the Geomstats Python package, which implements Fisher-Rao Riemannian geometry for various parametric families. This package facilitates statistical analysis on manifolds, offering a robust framework for applications across multiple domains, including medical imaging and text classification.In our work we have used the geomstats package to extract the beta distribution parameters for the extracted features out of the foundational model. The reviewed studies consistently highlight the significance of information geometry, particularly the Fisher-Rao metric, in enhancing the classification and analysis of medical data. The application of geometric structures such as beta and generalized gamma distributions provides a more accurate and interpretable framework for statistical analysis due to several key advantages. Firstly, the flexibility and versatility of these distributions allow them to model a wide range of data types; the generalized gamma distribution encompasses several other distributions, while the beta distribution is adept at handling variables constrained to a finite interval, such as proportions and probabilities. Secondly, the geometric interpretation provided by information geometry offers insights into the structure and relationships between distributions, with tools like the Fisher-Rao metric helping to understand distances between distributions. This leads to robust and stable estimators, such as the Riemannian centroid, which are less sensitive to outliers. The geometric framework also enhances statistical inference, allowing for sophisticated hypothesis testing and dimensionality reduction, which reveals the intrinsic structure of the data. Furthermore, these distributions find applications in machine learning and Bayesian analysis, where beta distributions serve as prior distributions in Bayesian models, and the generalized gamma distribution is useful in fields like survival analysis. Overall, the use of geometric structures enhances both the accuracy and interpretability of statistical models, providing significant insights and improvements across various applications. Histogram-based features are effective in classification tasks, and the Fisher-Rao geometry ensures consistency and efficiency in parameter estimation. Canonical moments provide symmetry and invariance, advantageous for statistical analysis. The generalized gamma manifold offers a comprehensive framework for modeling complex data distributions, improving classification performance in medical imaging. In conclusion, the integration of information geometry and beta distributions in classifying histograms of medical data presents a promising approach for improving diagnostic accuracy and treatment monitoring.

Chapter 3

Methodology

The intent of this work was to deep dive into the foundations of information geometry and look inside the hood of inner workings of the beta representation forms for a data model, the manifold representation and an attempt to formulate an interpretable model for medical imaging datasets and how the class distribution in a balanced or unbalanced dataset, truly affects the performance of the classifier or clustering technique. We show the robustness of the beta representation form and dive into the performance drive in using riemannian metric over the euclidean representations. But first let us look into the biomarker generation paradigm with respect to foundational models.

3.1 Traditional vs. Data-Driven Approaches

Biomarker identification in medical imaging has traditionally depended on hypothesis-driven approaches, which are constrained by existing knowledge and biases. These methods require significant manual effort and expertise. On the other hand, data-driven methodologies, powered by artificial intelligence (AI) and deep learning (DL), have the potential to discover novel biomarkers from imaging data with minimal manual intervention. However, the effectiveness of DL models is heavily dependent on the availability and quality of annotated datasets, which are often limited in specialized medical applications.

3.2 Self-Supervised Learning in Medical Imaging

Self-supervised learning (SSL) utilizes the inherent structure of data to learn generalized representations from large, unannotated datasets. This approach has shown great promise in medical imaging, particularly for two-dimensional (2D) images such as X-rays, whole-slide images, dermatology images, and fundus images. Despite its potential, the use of SSL to train foundational models for discovering general, robust, and transferable imaging biomarkers, especially for prognostic tasks, remains underexplored and presents an opportunity for future research.

3.3 Study Design and Methodology

In a study by Pai et al., a foundation model was developed using a convolutional encoder and SSL techniques, pretrained on a dataset of 11,467 radiographic lesions. The model was evaluated across three distinct use cases: classifying lesions into anatomical sites, predicting the malignancy of lung nodules, and prognosticating non-small cell lung cancer (NSCLC) tumors. Several pretraining strategies, including autoencoders, SimCLR, SwAV, and NNCLR, were compared, with the authors' modified SimCLR approach achieving the highest balanced accuracy and mean average precision (mAP). SimCLR (Simple Framework for Contrastive Learning of Visual Representations) employs a contrastive learning framework that uses a standard convolutional neural network (CNN) backbone, like 2DResNet50, to extract features from images. The process involves applying a series of augmentations such as random cropping, color jittering, and Gaussian blur to create multiple views of each image. SimCLR introduces a multi-layer perceptron (MLP) projection head that maps features to a lower-dimensional space where contrastive loss is applied. This loss function pulls together representations of augmented views of the same image (positive pairs) and pushes apart representations of different images (negative pairs) in the latent space. A key insight from SimCLR is that larger batch sizes and stronger data augmentations significantly improve representation learning. Additionally, the performance of SimCLR improves with increased model size and training data, aligning with trends seen in foundational models.

SwAV (Swapping Assignments between Multiple Views of the Same Image) is another self-supervised learning approach that combines clustering and contrastive learning to learn image representations. SwAV uses a CNN backbone, such as ResNet, for feature extraction and introduces multi-crop augmentation, where images are augmented into multiple crops at different scales. Instead of relying on explicit negative pairs, SwAV performs online clustering by assigning each augmented view to a cluster. The core idea of SwAV is to maximize the agreement between cluster assignments of different views by swapping their assignments and minimizing the entropy of these assignments. One significant advantage of SwAV is that it eliminates the need for negative samples, reducing the reliance on large batch sizes. This approach allows for efficient training on smaller hardware compared to methods that require large batch sizes.

NNCLR (Nearest Neighbor Contrastive Learning of Visual Representations) enhances contrastive learning by using nearest neighbors in the feature space, rather than relying solely on augmentations for positive pairs. NNCLR employs a CNN backbone like ResNet for feature extraction and applies augmentations to create different views of images. However, instead of just using augmented views as positive pairs, NNCLR leverages the nearest neighbors of an image in the feature space as additional positive samples. The framework applies contrastive loss between the anchor image and its augmented views, as well as its nearest neighbors. By including nearest neighbors, NNCLR captures more robust representations that better reflect semantic similarity, leading to improved generalization and performance on downstream tasks.

But in the work of Suraj Pai et al. which used an updated form of the SimCLR architecture. The original SimCLR architecture had formulated latent vector representations for maximizing similarity between the pair of randomized augmented images of the same datapoint over a projection head using contrastive loss functions. The model weights for the base encoder such as the ResNet50 architecture is used for a separate pipeline of downstream tasks. The idea behind using the projection head as a Dense - Relu - Dense Multi-Layer Perceptron, and the subsequent contrastive loss function acts in conjugation to provide learning representations for similar and dissimilar images through weight updation that happens through backward loss propagation through the layers, effectively using learned representation for self - supervised tasks. The concept of using self supervised learning through weight representation for separate downstream tasks, is modified a little by Suraj Pai et al. for the purpose of detecting and identifying lesions in medical images sampled from Positron Emission Tomography (PET-CT) scans. The data augmentation tasks with respect to the original SimCLR paper, which included randomised color distortion, cropping and gaussian blur suited for augmentation of regular images from the ImageNet dataset, had to be repurposed for the sake of CT scans to randomized color jitter and histogram intensity transformations. The positive pairs of CT scans fed into the network were determined by taking patches around the lesions seed point and, the negative pairs were generated by randomly sampling from the rest of the scanned dataset.

3.4 Lesion Anatomical Site Classification

The foundation model significantly outperformed baseline methods in the classification of lesion anatomical sites, particularly in scenarios with limited training data. The model's advantage was more pronounced as the size of the training dataset decreased, demonstrating its robustness and adaptability in low-data environments.

3.5 Nodule Malignancy Prediction

For predicting the malignancy of lung nodules, the foundation model demonstrated superior performance compared to most baseline models. It maintained robustness even in limited data scenarios, showcasing its generalizability and potential for practical application in clinical settings where annotated data may be scarce.

3.6 NSCLC(Non-Small Cell Lung Cancer) Prognostication

NSCLC stands for Non-Small Cell Lung Cancer . It is the most common type of lung cancer, accounting for about 85% of all lung cancer cases. NSCLC includes subtypes such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. In the prognostication of NSCLC tumors, the foundation model effectively stratified patients based on survival outcomes, outperforming baseline methods. The features of the model showed a strong correlation with underlying tumor biology, as evidenced by gene expression analysis, highlighting its potential to provide biologically relevant insights.

3.7 Stability and Interpretability

The foundation model exhibited high stability in test-retest scenarios and demonstrated robustness against inter-reader variability. Saliency maps revealed that the model's predictions were influenced by regions within or adjacent to tumors, aligning with current understanding of tumor biology, thus enhancing the interpretability and reliability of the model's predictions.

3.8 Biological Associations

The study found that the foundation model's predictions correlated with immune-associated pathways, suggesting a strong biological basis for its performance. This was further supported by gene-set enrichment analysis, which identified relevant genetic pathways, reinforcing the model's ability to provide insights into the biological mechanisms underlying imaging data.

The features extracted from the foundational model for each subset specific tasks were used by Suraj Pai et al. for training linear classifiers for three subtasks in their experiments. Task 1 included lesion anatomical site classification on total 3830 lesions, where the test set comprised of 1221 lesions, from the DeepLesion Dataset. Task 2 involved classification of models to predict malignancy of 507 lung modules extracted from LUNA16 dataset. Task 3 involved predicting survivability of patients with Non-Small Cell Lung Cancer Tumors extracted from LUNG1, comprising of 420 data points and RADIO dataset with 133 datapoints. The features extracted from each of these datapoints, for three subtasks were further used for Knn Classification for supervised classification and Kmeans clustering for unsupervised task. Then for testing the efficacy of beta representation as shown by Alice Lebrigant et al. we used the feature set for each of the data points and transformed them into their corresponding beta representations for obtaining the alpha and the beta parameters for their beta distribution. Once we have the beta distribution we try to project the datapoints in both euclidean space and riemannian space, to classify using knn method and unsupervised clustering using kmeans, to find out the accuracy measure and compare the efficacy of using riemannian metric and euclidean metric. We compared the results with reducing the 4096 features extracted for each of the data points from the foundational model with using Principal Component Analysis as a basis for dimensionality reduction to two dimensions for visualizing the classifications groups and cluster centers and compare the accuracy measure with respect to beta representation. Our results and analysis have been included in the subsequent section for a comprehensive and detailed breakdown for all the three tasks.

Chapter 4

Empirical Results

We compare the results for each subtasks and do a detailed analysis for each of the subtasks. For Task 1, we compare the result of KNN Classification across different set of neighbourhood values and find the optimal accuracy in the normal distribution for euclidean measure.

Task 1

Metric	Class	**Precision**	Recall	**F1-Score**
Validation Accuracy	-	-	-	0.8861
Validation Precision	0	0.83	0.94	0.88
	1	0.82	0.91	0.86
	2	0.90	0.88	0.89
	3	0.83	0.91	0.87
	4	0.94	0.95	0.95
	5	0.88	0.57	0.69
	6	0.87	0.66	0.75
	7	0.83	0.66	0.73
Validation Macro Avg	-	0.86	0.81	0.83
Validation Weighted Avg	-	0.89	0.89	0.88

Table 4.1: Validation Performance Summary

Metric	Class	Precision	Recall	F1-Score
Test Accuracy	-	-	-	0.8067
Test Precision	0	0.76	0.84	0.80
	1	0.73	0.79	0.76
	2	0.81	0.82	0.81
	3	0.72	0.82	0.77
	4	0.92	0.91	0.91
	5	0.60	0.52	0.56
	6	0.72	0.51	0.60
	7	0.71	0.46	0.56
Test Macro Avg	-	0.75	0.71	0.72
Test Weighted Avg	-	0.81	0.81	0.80

Table 4.2: Test Performance Summary

The model demonstrates strong overall performance, with a validation accuracy of 88.61% and a test accuracy of 80.67%, though the slight drop in test accuracy suggests some overfitting. It performs exceptionally well on dominant classes, particularly Class 4, but struggles with minority classes like Class 5 and Class 7, which show low precision and recall. The macro average F1-scores highlight inconsistent performance across classes, with a significant gap between validation and test results. Addressing class imbalance, improving feature engineering, and applying regularization techniques could help enhance the model's ability to generalize and improve performance on underrepresented classes.

Task 2

Metric	Class	Precision	Recall	F1-Score
Euclidean KNN Validation Accuracy	-	-	-	0.7159
Validation Precision	0	0.71	0.76	0.73
	1	0.72	0.67	0.70
Validation Macro Avg	-	0.72	0.71	0.71
Validation Weighted Avg	-	0.72	0.72	0.72

Table 4.3: Euclidean KNN Validation Performance Summary

Metric	Class	Precision	Recall	F1-Score
Euclidean KNN Test Accuracy	-	-	-	0.7765
Test Precision	0	0.76	0.84	0.80
	1	0.81	0.71	0.75
Test Macro Avg	-	0.78	0.77	0.77
Test Weighted Avg	-	0.78	0.78	0.78

Table 4.4: Euclidean KNN Test Performance Summary

Metric	Class	Precision	Recall	F1-Score
Riemannian KNN Validation Accuracy	-	-	-	0.7219
Validation Precision	0	0.71	0.78	0.74
	1	0.74	0.66	0.70
Validation Macro Avg	-	0.72	0.72	0.72
Validation Weighted Avg	-	0.72	0.72	0.72

Table 4.5: Riemannian KNN Validation Performance Summary

Metric	Class	Precision	Recall	F1-Score
Riemannian KNN Test Accuracy	-	-	-	0.7471
Test Precision	0	0.75	0.77	0.76
	1	0.75	0.72	0.73
Test Macro Avg	-	0.75	0.75	0.75
Test Weighted Avg	-	0.75	0.75	0.75

Table 4.6: Riemannian KNN Test Performance Summary

The analysis of the Euclidean and Riemannian KNN models reveals that while both models perform moderately well, the Riemannian KNN slightly outperforms the Euclidean KNN in terms of validation accuracy (72.19% vs. 71.60%). However, the Euclidean KNN exhibits better test accuracy at 77.65%, suggesting it may generalize slightly better to unseen data. Both models demonstrate balanced precision and recall across classes, with Class 0 generally being classified more accurately than Class 1. The confusion matrices indicate persistent misclassification between the two classes, particularly with Class 1 being frequently misclassified as Class 0. The Riemannian KNN shows potential overfitting, as indicated by the drop in test performance compared to validation. To enhance model performance, hyperparameter tuning, addressing class imbalance, and applying regularization techniques are recommended. Visualizations such as confusion matrices, ROC curves, and decision boundary plots could provide further insights into the models' classification behavior, like we show in the figures included.

Hence we get the summarised Euclidean KNN Test Accuracy as 0.78 and the Riemannian KNN Test Accuracy as 0.75.

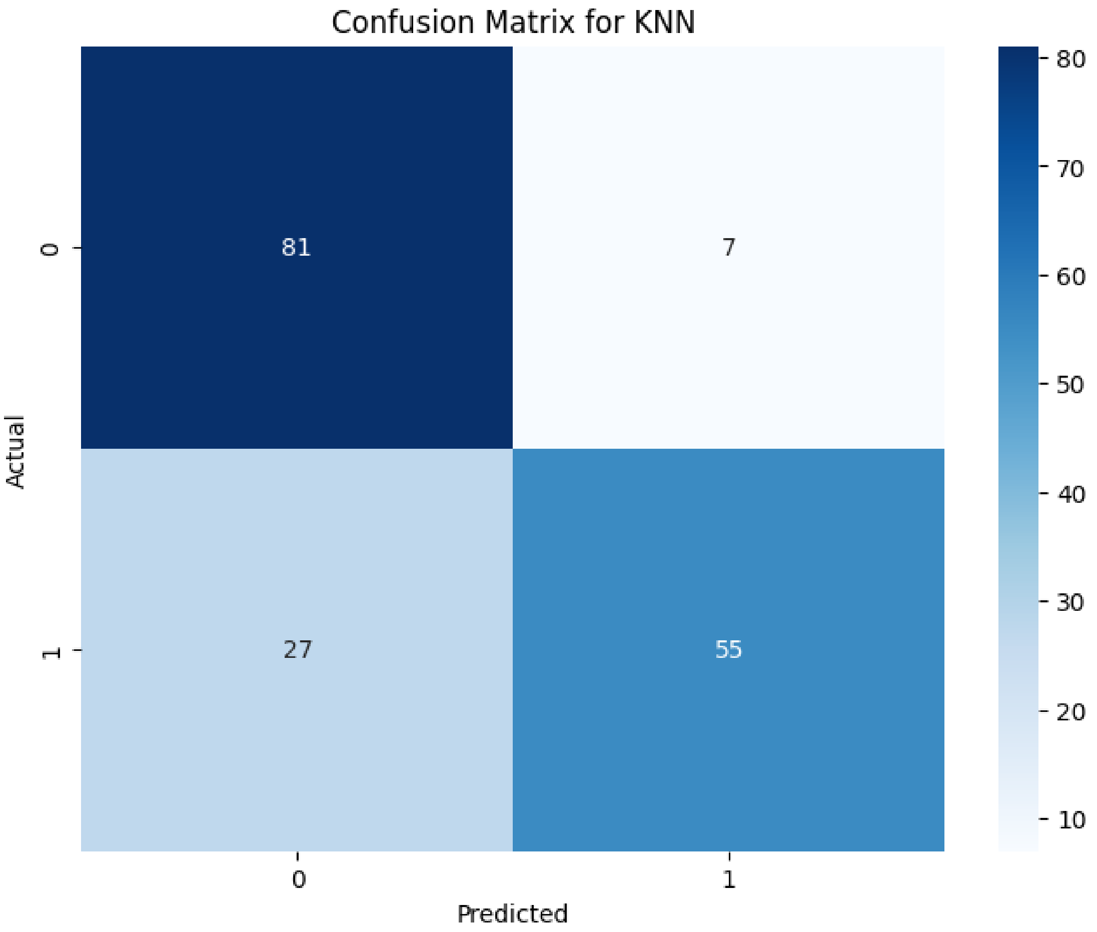
For Normal Distribution, we segregate the results of KNN Classification and KMeans.

Metric	**Class**	**Precision**	**Recall**	**F1-Score**
Best k=16	-			
Accuracy on Test Set	0.80			
-				
Test Precision	0	0.75	0.92	0.83
	1	0.89	0.67	0.76
Test Accuracy	-			
0.80				
Macro Avg	-	0.82	0.80	0.80
Weighted Avg	-	0.82	0.80	0.80

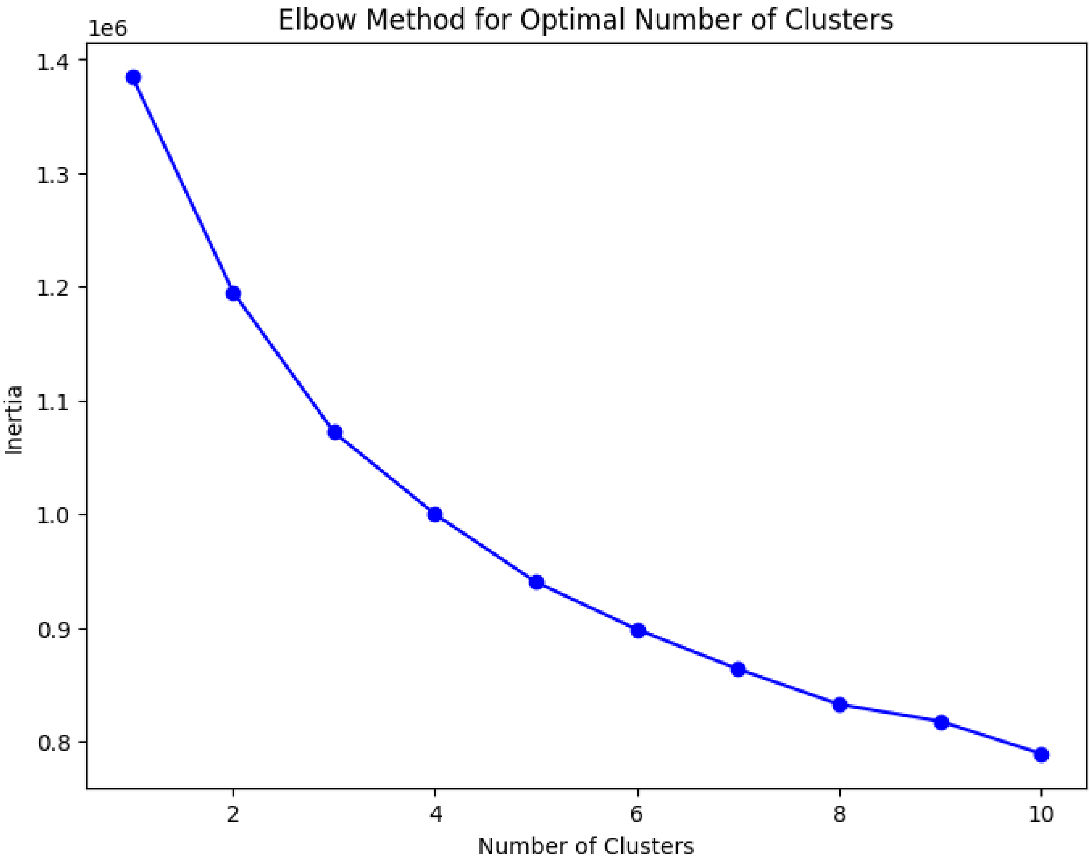
Table 4.7: Performance Summary with Best k = 16

Confusion Matrix	**Predicted: 0**	**Predicted: 1**
Actual: 0	81	7
Actual: 1	27	55

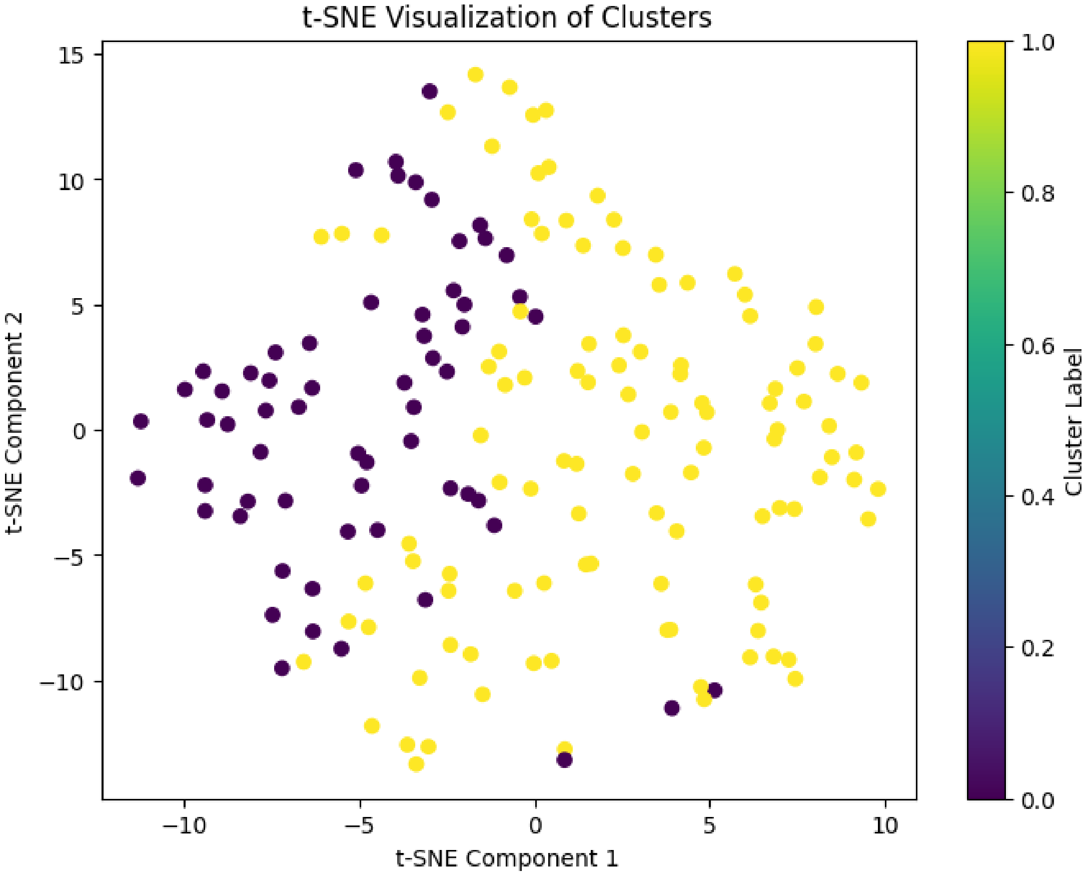
Table 4.8: Confusion Matrix

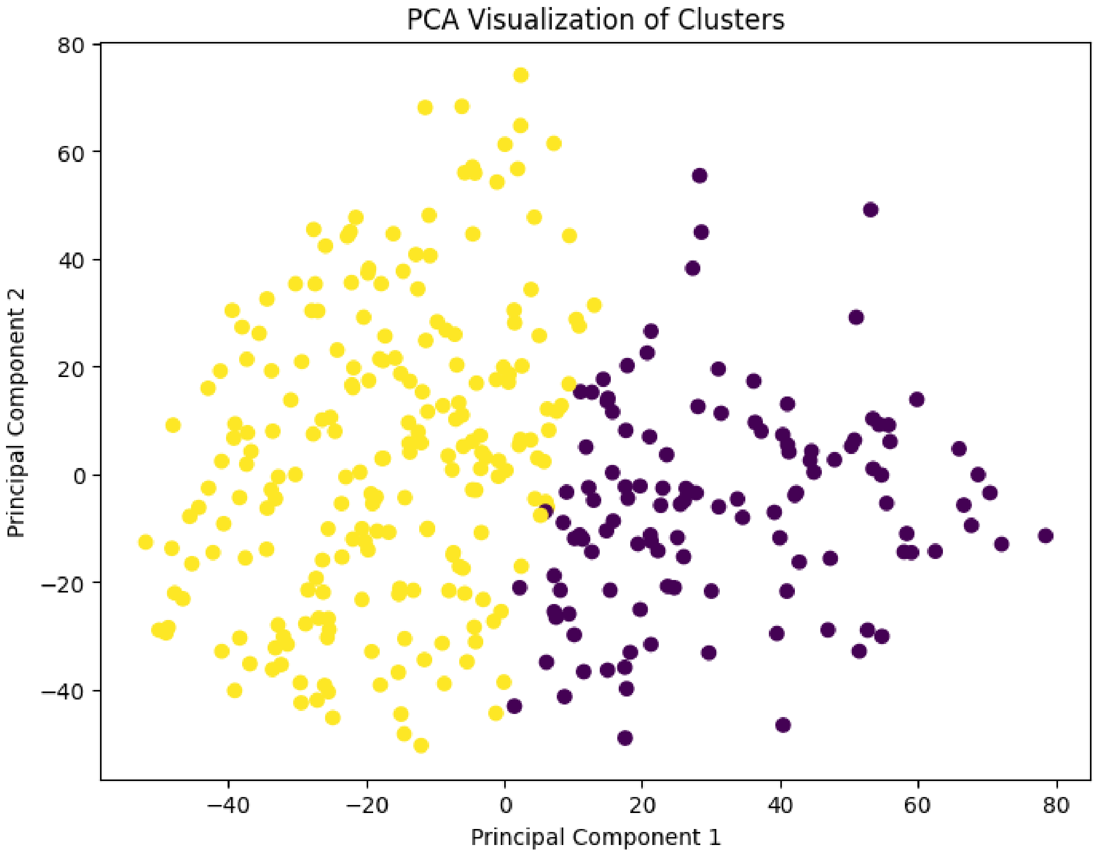


We compare the optimal cluster centers for KMeans alogorithm for unsupervised clustering and find out the subsequent plotting of inertial minimisation and the number of clusters.



Using t-dsitributed stochastic neighbour embedding, we use the non-linear data reduction alogorithm to visualise the cluster centers and compare it with PCA, for the task 2 analysis. Reducing the feature from high dimension, we are able to see how the localisation and neighbourhood features for each datapoint affect the clustering algorithm.


t-SNE Visualization of Clusters
t-SNE Component 1
t-SNE Component 2
−10
−5
0
5
10
15
Cluster Label
0.0
0.2
0.4
0.6
0.8
1.0



Task 3

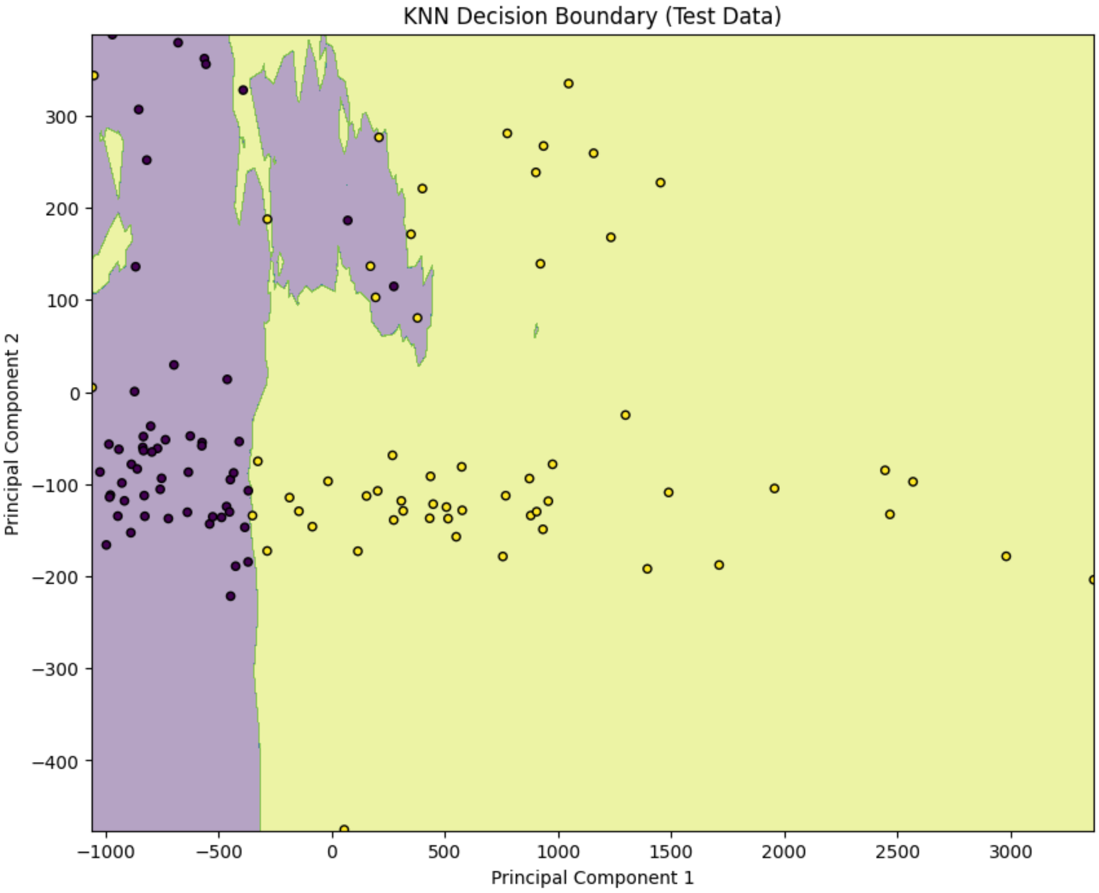
We perform Grid Cross Validation search and Randomised CV search, for finding the optimal parameters in case of Task 3, for KNN Classification for finding the optimal number of neighbours. We also use PCA for finding the optimal boundary for KNN classification for Test data and validation data.

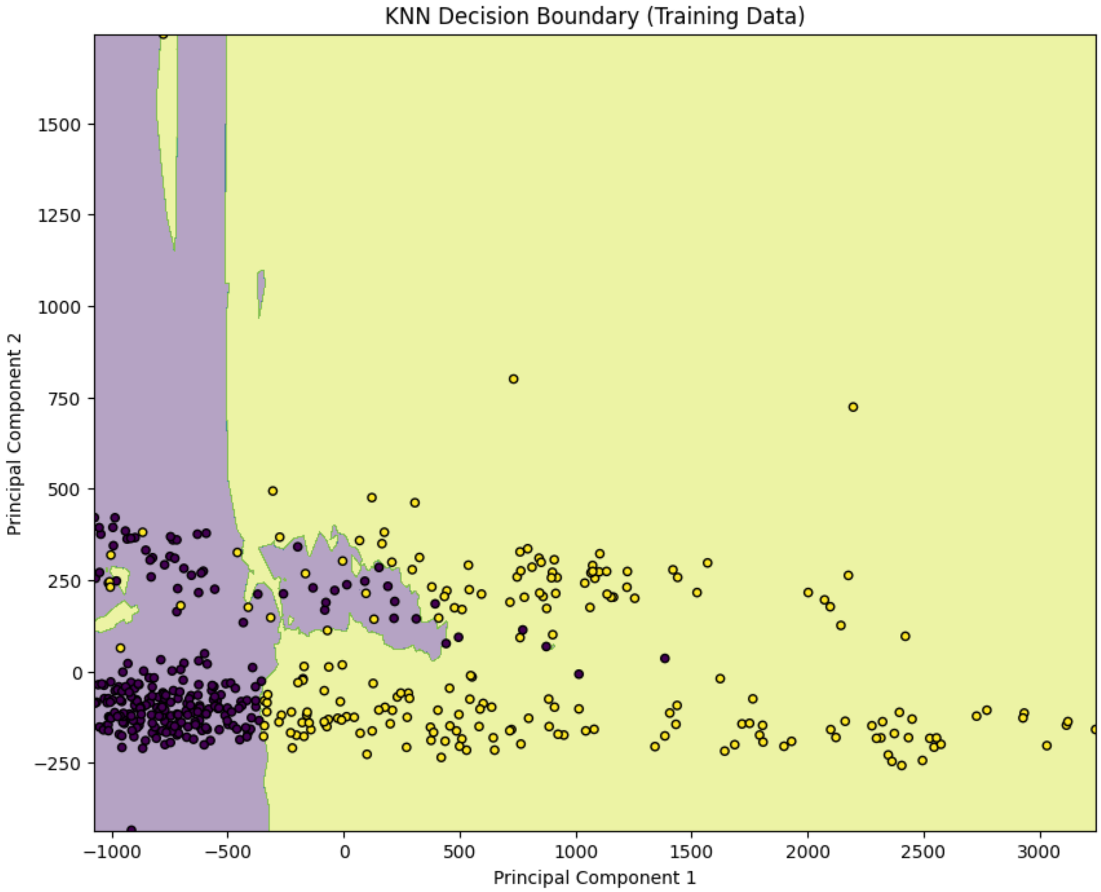
Metric	Class	Precision	Recall	F1-Score
Classification Report for Best KNN Model	0.0	0.93	0.98	0.96
	1.0	0.98	0.93	0.96
Accuracy	-			
0.96				
Macro Avg	-	0.96	0.96	0.96
Weighted Avg	-	0.96	0.96	0.96

Table 4.9: Classification Report for Best KNN Model

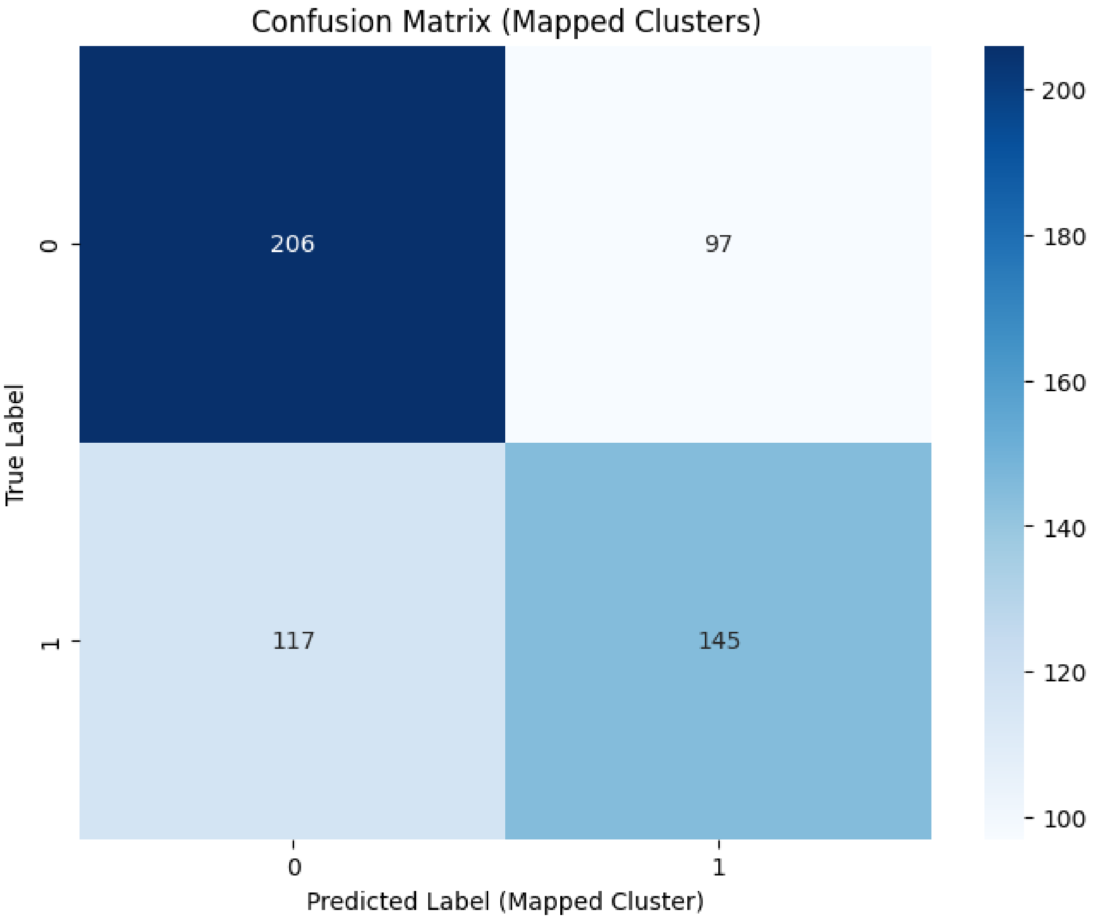
Test Accuracy for Best KNN Model
0.9558

Table 4.10: Test Accuracy for Best KNN Model

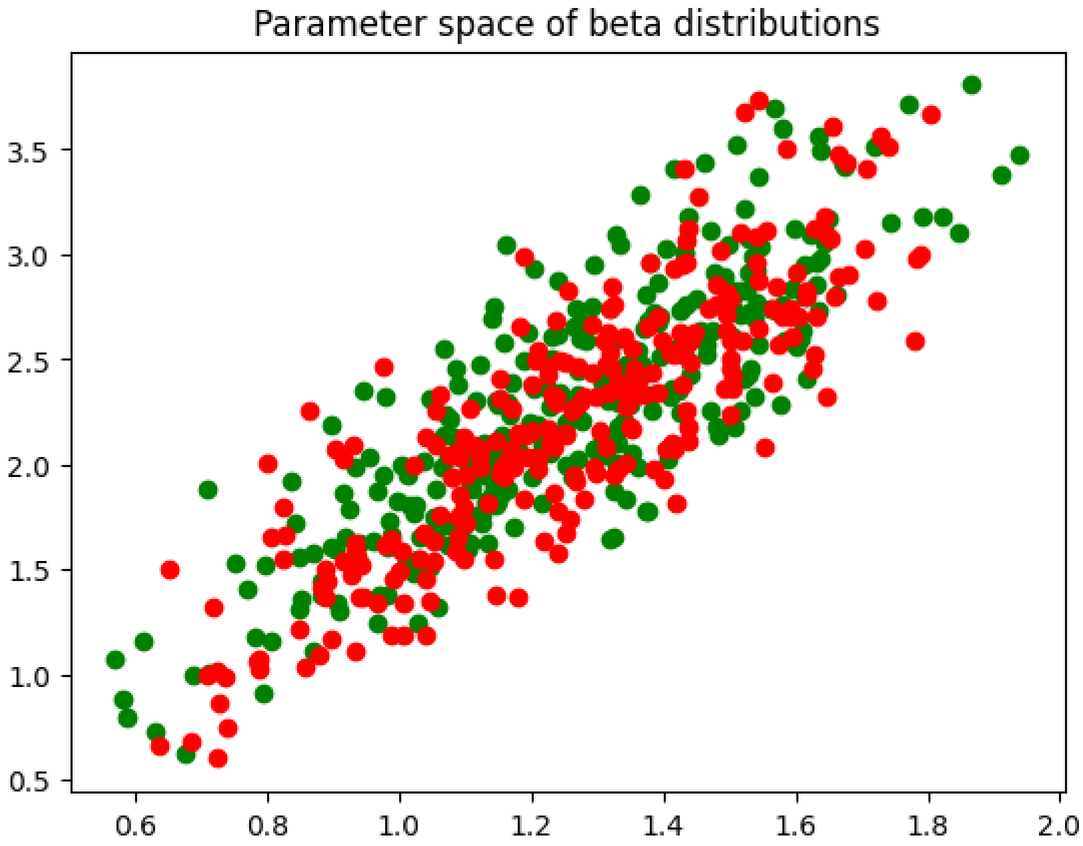




For KMeans Clustering for Normalised distribution, we get an Adjusted Rand Index (ARI): 0.057 and Silhouette Score: 0.14. The Accuracy calculated from the confusion matrix 0.62.



For Beta Distributions to get the alpha and beta parameters of the distribution space, we get :



The results indicate that both the Euclidean and Riemannian K-Nearest Neighbors (KNN) classifiers have similar performance, with the Euclidean KNN achieving a slightly higher accuracy of 47.79% compared to the Riemannian KNN's 46.90%. Both classifiers show a noticeable discrepancy in class performance: class 0 (negative class) has a much higher recall (0.76 for Euclidean and 0.76 for Riemannian) compared to class 1 (positive class) (0.21 for Euclidean and 0.19 for Riemannian). This suggests that both models are biased towards predicting the negative class, leading to higher false negatives. The f1-scores for both models are low, particularly for class 1, indicating that neither model effectively balances precision and recall. The confusion matrices confirm this trend, with the majority of predictions being true negatives but a significant number of false positives.

Table 4.11: KNN Classification Results

Metric	**Model**	**Class 0**	**Class 1**	**Overall**
Precision	Euclidean KNN	0.48	0.48	-
	Riemannian KNN	0.47	0.46	-
Recall	Euclidean KNN	0.76	0.21	-
	Riemannian KNN	0.76	0.19	-
F1-Score	Euclidean KNN	0.59	0.29	-
	Riemannian KNN	0.58	0.27	-
Accuracy	Euclidean KNN	0.4779		
	Riemannian KNN	0.4690		
Confusion Matrix	Euclidean KNN	[[42, 13], [46, 12]]		-
	Riemannian KNN	[[42, 13], [47, 11]]		-

Finally we compile the results for Task 2 and Task 3 and compare it with the original results from Alice LeBrigant, for Normalised distribution space and Beta Distribution space, for euclidean and riemannian metric.

Task 2 and Task 3 for KNN Classification

	Original Euclidean	**Beta Euclidean**	**Beta Riemannian**
Task 2	0.80	0.78	0.75
Task 3	0.96	0.48	0.47
ADNI(Le Brigant)	0.81	0.77	0.83
CTh(Le Brigant)	—	0.77	0.79

Table 4.12: Task 2, Task 3 and original for KNN Classification

The results of our KNN classification tasks show some alignment with Le Brigant's findings, particularly in the ADNI and CTh datasets, where the Beta Riemannian method performs comparably or slightly better than the Euclidean approaches, similar to Le Brigant's reported effectiveness of Riemannian methods. However, in Task 2 and Task 3, our Original Euclidean method consistently outperforms the Beta variants, which contrasts with Le Brigant's emphasis on the Riemannian approach. This suggests that while our results support some of Le Brigant's conclusions, particularly in the context of specific datasets, there are notable differences in performance across tasks, indicating that the choice of method may need to be tailored to the specific classification challenge.

The results of these maybe using higher dimensional feature data of 4096 features extracted from the foundational model might capture the data distribution pertinently rather than the data distribution of the reduced beta representation in the riemannian space. Although ALice LeBrigant et al. had shown the efficacy of modelling the feature in beta representation, we show using the entire feature set from the foundational model we are able to perform far better rather than the beta distribution space. We discuss about this further in our conclusion and future work.

Chapter 5

Conclusions and Future Work

Conclusion

Our work challenges the claims made by Alice Le Brigant by demonstrating that using the extracted feature space from foundational models leads to better classification performance in both Task 2 and Task 3 compared to the beta distribution representation. Through statistical and empirical analysis, we show that the foundational model's feature space outperforms the beta representation, particularly in the contexts of survivability classification for NSCLC patients (Task 3) and malignancy prediction (Task 2). The performance of PCA, when compared to the beta representation, is only marginally better, yet both fall short of the accuracy achieved with the original feature set. This underscores the superior efficacy of the extracted feature set from the modified SimCLR model over other representation norms.

Further Work

Future research could explore the integration of additional feature extraction techniques with foundational models to further enhance classification accuracy. Investigating the impact of various representation spaces across a wider range of tasks and datasets would provide a more comprehensive understanding of the model's generalizability. Additionally, conducting ablation studies to identify the specific components of the SimCLR model that contribute most to its success could offer insights for refining and optimizing these models for even better performance. Finally, extending the analysis to include other types of foundational models and representations might reveal further avenues for improving classification tasks in medical imaging and other domains.

Bibliography