ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

GENERALIZED MULTI-VIEW DATA PROLIFERATOR (GEM-VIP) FOR BOOSTING CLASSIFICATION

M.Sc. THESIS

Mustafa ÇELİK

Department of Computer Engineering

Computer Engineering Programme

AUGUST 2022

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

GENERALIZED MULTI-VIEW DATA PROLIFERATOR (GEM-VIP) FOR BOOSTING CLASSIFICATION

M.Sc. THESIS

Mustafa ÇELİK
(504131531)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Assist. Prof. Dr. Islem REKIK

AUGUST 2022

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GENELLEŞTİRİLMİŞ ÇOK BOYUTLU VERİ ÜRETİMİ İLE SINIFLANDIRMA HASSASLIĞININ YÜKSELTİLMESİ

YüKSEK LİSANS TEZİ

Mustafa ÇELİK
(504131531)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Assist. Prof. Dr. Islem REKIK

AĞUSTOS 2022

Mustafa ÇELİK, a M.Sc. student of ITU Graduate School student ID 504131531, successfully defended the thesis entitled "GENERALIZED MULTI-VIEW DATA PROLIFERATOR (GEM-VIP) FOR BOOSTING CLASSIFICATION", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assist. Prof. Dr. Islem REKIK** Istanbul Technical University	.....
Jury Members :	**Prof. Dr. Hatice KÖSE** Istanbul Technical University	.....
	Assist. Prof. Dr. Mustafa DAĞTEKİN Istanbul University - Cerrahpasa	.....

Date of Submission : 1 June 2022
Date of Defense : 8 August 2022

To my family,

FOREWORD

First and foremost, I would like to thank my advisors, Professor Islem REKIK for her guidance, effort and support. She has been a source of inspiration and motivation for me.

Professor Hatice KÖSE, Professor Tolga ENSARİ, Professor Yusuf YASLAN and Professor Mustafa DAĞTEKİN are also to be thanked for their useful recommendations and perceptive comments.

I would want to express my gratitude to all of my teachers for their help.

Finally, I would like to express my sincere appreciation and thanks to my wife, sister and parents for their understanding, patience and support.

August 2022 Mustafa ÇELİK

TABLE OF CONTENTS

Page

ABBREVIATIONS

MCI : Mild Cognitive Impairment
ML : Manifold Learning
NC : Normal Control
eMCI : early Mild Cognitive Impairment
dMRI : diffusion weighted MRI
lMCI : late Mild Cognitive Impairment
MRI : Magnetic Resonance Imaging
AD : Alzheimer's Disease
ROI : Regions of Interest
rsfMRI : resting state functional MRI
MBN : Morphological Brain Network
SMOTE : Synthetic Minority Oversampling Technique
MV-LEAP : Multi-View Learning Based Proliferator
SVM : Support Vector Machine
MVND : Multi-variate Normal Distribution
GEM-VIP : Generalized Multi-View Data Proliferator
VS-LEAP : View-Specific Learning Based Proliferator
CBT : Connectional Brain Template
GA : Genetic Algorithm
GAN : General Adversarial Network
PCA : Principal Component Analysis
WM : White Matter
GM : Gray Matter
TSE : Turbo Spin-Eco
LH : Left Hemisphere
RH : Right Hemisphere

SYMBOLS

n_v : View Number
n_c : Number of Initial Covariance Matrix
n_r : ROIs Number
n_f : Feature Number
n_s : Sample Number
n_{syn} : Synthetic Sample Number
itr_{ga} : Iteration Number of Genetic Algorithm
$\mathbf{x}_i$: Real Sample Vector $\in \mathbb{R}^{n_f}$ of i
$\tilde{\mathbf{x}}_i$: Synthetic Sample Vector $\in \mathbb{R}^{n_f}$ of i
$\mathbf{x^p}_i$: PCA applied real sample vector $\in \mathbb{R}^{n_f}$ of i
$\mathbf{X^s}_{(v_i)}$: data matrix view of i for subject $s \in \mathbb{R}^{n_s \times n_f}$
$\mathbf{B^s}_{(v_i)}$: brain network view of i for subject $s \in \mathbb{R}^{n_r \times n_r}$
$\mathscr{T}_s$ = : subject s's brain tensors with n_v frontal views $\in \mathbb{R}^{n_r \times n_r \times n_v}$
$\{\mathbf{B^s}_{(v_1)}, \mathbf{B^s}_{(v_2)}, \ldots, \mathbf{B^s}_{(v_q)}\}$
$\tilde{\mathscr{T}}$: representative tensor $\in \mathbb{R}^{n_r \times n_r \times n_v}$ of a population
μ_i : mean vector $\in \mathbb{R}^{n_f}$ of i-th feature in the given population
$\mathbf{V^s}_{ij}$: subject s's feature vector that is related to ROIs where i and $j \in \mathbb{R}^{n_v \times 1}$
$\mathbf{y}$: label vector $\in \mathbb{R}^{n_s}$
$\mathbf{C}_{xy}$: covariance matrix of x and y
$\mathbf{d}_{ij}$: the Euclidean distance between subject s and s
$c(att)$: cortical attribute measure assigned to vertex v

LIST OF TABLES

Page

LIST OF FIGURES

Page

GENERALIZED MULTI-VIEW DATA PROLIFERATOR (GEM-VIP) FOR BOOSTING CLASSIFICATION

SUMMARY

Multi-view network representation revealed multi-faced alterations of the brain as a complex interconnected system, particularly in mapping neurological disorders. Such rich data representation maps the relationship between different brain views which has the potential of boosting neurological diagnostic tasks. However, multi-view brain data is scarce and generally is collected in small sizes. Thus, such data type is broadly overlooked among researchers due to its *relatively* small size. Despite the existence of data proliferation techniques as a way to overcome data scarcity, to the best of our knowledge, multi-view data proliferation from a single sample has not been fully explored. Here, we propose to bridge this gap by proposing our *GEneralized Multi-VIew data Proliferator (GEM-VIP)*, a framework aiming to **proliferate synthetic multi-view brain samples from a single multi-view brain to boost multi-view brain data classification tasks.** For the given Connectional Brain Template (i.e., represents an approximation of brain graphs that captures the unique connection shared by a population's subjects), we set out the proliferate synthetic multi-view brain graphs using the inverse of multi-variate normal distribution (MVND). However, one needs two crucial components, which are the *mean* and the *covariance* of a given population. As such, *first*, our proposed GEM-VIP framework obtains a population-representative tensor (i.e., drawn from the prior CBT) which can be *mathematically* regarded as a *mean* of the population. Second, drawing inspiration from the genetic algorithm paradigm our proposed GEM-VIP learns the *covariance* matrix of the population using the given CBT. Lastly, it proliferates synthetic samples using the earlier obtained *representative tensor* and created *covariance* matrix of the population on the MVND equation. We evaluate our GEM-VIP against several comparison methods. The results show that our framework boosts the multi-view brain data classification accuracy of AD/ lMCI and eMCI/ normal control (NC) datasets. In short, our GEM-VIP method boosts the diagnoses of the neurological disorders.

GENELLEŞTİRİLMİŞ ÇOK BOYUTLU VERİ ÜRETİMİ İLE SINIFLANDIRMA HASSASLIĞININ YÜKSELTİLMESİ

ÖZET

Demans, dünyada sayısı en hızlı artan hastalıklardan biridir. Dünya sağlık örgütünün verilerine göre dünya genelinde 50 milyon demans hastası vardır. Her yıl bu sayıya 10 milyon yeni hasta eklenmekte ve 2050 yılına kadar bu sayının üç katına çıkması beklenmektedir. Demans toplumlar için ciddi bir ekonomik külfete sebep olmaktadır. Dünya genelinde, 2030 yılına kadar, yılda 2 trilyon ABD dolarını bulan bakım giderleri olacağı tahmin edilmektedir.

Neyse ki demans hastalığı, hafif bilişsel bozukluk (MCI) olarak adlandırılan ve erken tedavisi mümkün olan bir hastalığın ardından ortaya çıkmaktadır. Araştırmalar, hafif bilişsel bozukluğu olan hastaların, yalnızca yüzde 5 ile 10'unun ilk 10 yıl içerisinde demansa yakalandığını göstermektedir. Bu sebeple, MCI'nın erken tespiti ve tedavisi, demansın yukarda belirtilen etkilerinden korunmak için büyük önem taşımaktadır. Fakat, MCI tüm hastalar için ortak tek bir tedavi yönteminin olmadığı, nörolojik rahatsızlıklardan biridir. Yapılan son çalışmalarda, kişiye özel tedavi yöntemlerinin genel tedavi yöntemlerine kıyasla daha etkili olduğu beyan edilmektedir. Böylece tıpta, tüm hastaları kapsayan tek ve genel bir tedavi ile iyileştirme yaklaşımı yerine; alternatif bir yöntem olan kişiye özel tedavi yöntemleri önem kazanmaya başlamıştır.

Bu yöntemlerden biri de, örüntü tanıma da insan kapasitesinin çok üzerinde çalışan makine öğrenimi modellerinden faydalanmaktır. Bu bağlamda yapılan son araştırmalar, çok-görünümlü (multi-view) tıbbi görüntüleme verileri kullanılarak geliştirilen makine öğrenme modellerinin, biyobelirteç tespitlerinde etkili olduğunu göstermektedir. Bu bilgiler ışığında, hastaya özel tedavi yöntemlerini belirleme işlemleri önemli ölçüde kolaylaşmaktadır. Özetle çok-görünümlü veriler, içerisinde yer alan çeşitli öznitelik kümeleri sayesinde, hastalıkların teşhisinde önemli roller oynamaktadırlar.

Tıbbi görüntüleme kullanılarak teşhisi yapılan bir diğer hastalık da Alzaymırdır. Bu hastalığın beyindeki biyoişaretçileri, manyetik rezonans görüntüleme (MRI) yöntemiyle elde edilen beyin görüntüleri sayesinde keşfedilmiştir. Beyindeki iki anatomik ilgi bölgesi (ROI) arasındaki ilişki, bir beyin konektomu diye adlandırılmaktadır. Bu ilgi bölgeleri arasındaki ilişkiler de yapısal ve işlevsel beyin ağları olmak üzere 2 farklı şekilde gruplandırılır. Yapısal beyin ağları difüzyon ağırlıklı MRI (dMRI) yöntemi kullanılarak; işlevsel beyin ağları ise dinlenme-durumu işlevsel MRI (rsfMRI) yöntemiyle oluşturulmaktadır.

Ancak, bu tip beyin verileri genel olarak toplaması zor ve masraflıdır. Ayrıca hatalı olma ihtimalleri yüksek olduğundan, sonucunda elde edilen çıkarımların doğruluğu

olumsuz etkilenmektedir. Bu sorunu aşmak için, son yapılan çalışmalarda morfolojik beyin ağlarından (MBN) üretilen beyin konektomları tercih edilmeye başlanmış ve buradan elde edilen verilerin beyin hastalıklarının teşhisinde önemli bir iyileştirme yaptığı gözlemlenmiştir. Böylece, morfolojik beyin ağları Alzaymır hastalığının teşhisinde defakto bir yöntem olarak kullanılmaya başlanmıştır.

Fakat, Alzaymır'ın teşhisinde kullanılan çok-görünümlü bu yeni yöntem de (MBN), veri setinin toplanma aşamasında sorun çıkartabilmektedir. Bu sorunların başında dengesiz veri seti problemi gelmektedir. Bu durum, popülasyonda yer alan sınıflardan birindeki denek sayısının diğerine göre az olmasıdır ve sınıflandırma işleminde sorun oluşturmaktadır. Bu sorunun çözümü için çeşitli yöntemler kullanılmıştır. Bunlardan biri, sayısı çok olan sınıftaki deneklerin sayısının azaltılması ve az olan ile eşit hale getirilmesidir. Ancak deneklerin çıkarılmasından dolayı veri kaybı olacağından ve popülasyonun genel dağılımı etkileneceğinden dolayı her durumda kullanılması uygun değildir.

Başka bir çalışmada ise, fazla olan sınıfın denek sayısının azaltılması yerine, eksik olan sınıftaki verilerin sentetik olarak artırılması önerilmiştir. Temel formu SMOTE olarak bilinen bu yöntem ile, gerçek deneklerin en yakın komşuları arasındaki rastgele bir noktaya sentetik denek üretilmesi amaçlanmıştır. Bu sayede, sayısı az olan sınıftaki denek sayısı arttırılarak sınıflar arası denge sağlanmaya çalışılmıştır. Bu yöntem, uzun yıllar araştırmacılar tarafından kullanılmasına rağmen; çok boyutlu veriler söz konusu olduğunda yetersiz kalmıştır. Çok-görünümlü verilerde görünümler arası özniteliklerin aynı kapsamda olmaması, SMOTE kullanımında gürültülü sentetik veri üretilmesine sebep olmaktadır. Gürültü veriler üzerine eğitilen sınıflandırıcıların başarımı da negatif etkilenmektedir.

Bu çalışmalara alternatif olarak, araştırmamızda farklı görünümlerin öznitelikleri arasındaki ilişkilere odaklanılmıştır. Bu ilişkilerden, bir destek vektör makinesi sınıflayıcısının başarısını arttırmak için kullanılacak etkili bilgiler çıkarılabileceği varsayılmıştır. Bu kapsamda, bir sınıfın tamamını tek başına ifade edebilen bağlantılı beyin ağları (connectional brain networks) kullanılarak sentetik veri üretilmesini amaçlayan, Genelleştirilmiş Çok Boyutlu Veri Üretimi (GEM-VIP) yöntemi önerilmiştir. Bu yöntem aşağıda belirtilen üç önemli adımdan oluşmaktadır.

Bu yöntem, çok-değişkenli normal dağılımın (MVND) olasılık yoğunluk fonksiyonunun (PDF) tersi alınarak sentetik veri üretilmesini amaçlamaktadır. Bir çok-değişkenli normal dağılım, iki veya daha fazla tek-değişkenli normal dağılımın birleşimidir ve iki önemli parametreye sahiptir. Birincisi, çok-değişkenli normal dağılımı oluşturan her bir tek-değişkenli normal dağılımın popülasyon ortalamasının (mean) tutulduğu vektördür. İkinci parametre ise her bir tek-değişkenli normal dağılımın varyansının tutulduğu kovaryans matrisidir. Burada geçen tek-değişkenli normal dağılımlardan her biri çok-görünümlü verinin bir görünümüne karşılık gelmektedir. Özetlemek gerekirse, ortalama vektörü ve kovaryans matrisi bilinen bir sınıftan (AD veya MCI); yine o sınıfa ait çok-değişkenli normal dağılımın içerinde yer alan sentetik veri üretimi yapılabilmektedir.

İkinci adımda, çok-değişkenli normal dağılımda kullanılacak olan ortalama vektörünün elde edilmesi hedeflenmiştir. Burada her bir deneğin farklı görünümlerdeki öznitelikleri toplanmakta ve deneği ifade eden tek bir tensör oluşmaktadır. netNorm yöntemi kullanılarak da, popülasyondaki tensörlerin tamamını temsil eden merkezi tek

bir tensör üretilmektedir. Bu tensör, ilgili popülasyonun matematiksel bir ortalaması olarak kabul edilebilmektedir. Özetle, bu yöntem verilen bir popülasyondaki deneklerin en genel bilgilerini kullanarak, o popülasyonu tek bir tensör ile ifade eder ve ortalama (mean) olarak kullanılabilir.

Son adımda, çok-değişkenli normal dağılımda ihtiyaç duyulan kovaryans matris parametresinin üretilmesi hedeflenmiştir. Ancak, tıbbi verilerin sınırlı olmasından dolayı bu işlem sağlıklı yapılamamaktadır. Çünkü sınırlı veriden elde edilen kovaryans matrisi, sınıflandırma işleminde istenilen başarıyı sağlamamaktadır. Bu sorunun üstesinden gelmek için, basit ve etkili bir yöntem olan genetik algoritma yaklaşımı ile uygun kovaryans matrisi üretilmesi önerilmiştir. Bu yöntem, başlangıçta rastgele üretilen veya sınırlı sayıda denek üzerinden çıkartılan kovaryans matrisleri aday kovaryans matrisler olarak tanımlamaktadır. Bu tanımla matrislerlin her biri uygunluk fonksiyonuna sokulmakta ve uygunluk puanı hesaplanmaktadır. Bu puana göre sıralanan ve en yüksek puana sahip matrisler sonraki adım için ata kovaryans matrisler olarak seçilmektedir. Seçilen bu matrisler kendi aralarında eşlenmekte ve parametrelerin değer değişimi (crossover) yapılmaktadır. Ayrıca bu değişimden sonra rastgele seçilen parametreler, rastgele seçilen değerler ile değiştirilerekte genetik çeşitlilik sağlanmaktadır (mutation). Bu işlemler, uygunluk puanı sabitlenene kadar tekrarlanmakta ve sonunda optimum bir kovaryans matris elde edilmektedir.

Böylece, çok-değişkenli normal dağılımdan sentetik veri üretme aşamasında ihtiyaç duyulan ortalama tensörü ve kovaryans matris elde edilmiştir. Bu parametreler ve veri çeşitliliğini sağlayacak rastgele üretilmiş bir vektörün de yardımıyla çok-değişkenli normal dağılım denklemi tersten hesaplanarak sentetik veri üretilmesi sağlanmaktadır.

Önerilen yöntem ile sınıflandırması yapılacak popülasyonlara sentetik veri üretilmekte ve veri sayısı arttırılmaktadır. Sonrasında üretilen sentetik ve gerçek veriler ile SVM sınıflayıcısı eğitilmekte ve sınıflama sonuçlarında artış sağlandığı gözlemlenmektedir. Ayrıca farklı yöntemler ile üretilen veriler ile eğitilmiş SVM sınıflayıcıları ile önerilen yöntem kıyaslanmıştır. Farklı verisetleri üzerinde yapılan bu testlerde önerilen yöntemin başarısının diğer yöntemelere göre daha iyi sonuç verdiği gözlemlenmiştir.

Bu testlerin dışında, genetik algoritma yaklaşımı ile kovaryans matris üretme kısımları da değerlendirilmiştir. Burada üretilen kovaryans matrisinin uygunluğu, yine farklı yöntemler ile elde edilen diğer kovaryans matrisleri ile kıyaslanmıştır. Kıyaslama sonunda önerilen yöntemin daha iyi sonuç verdiği gözlemlenmiştir. Ayrıca, genetik algoritmanın çalışma süresi boyunca oluşturduğu çıktılar incelenmiş, zamanla algoritmanın optimum kovaryans matrise doğru evrildiği hesaplamalar ile tespit edilmiştir.

Bu tez çalışmasın, önerilen GEM-VIP yöntemi ile sentetik veri üretilerek çok-görünümlü beyin verilerinin sınıflandırma performansının arttırılması amaçlanmıştır. Veri üretimi aşamasında çok-değişkenli normal dağılım fonksiyonundan faydalanılmış ve bu fonksiyonun değişkenleri olan popülasyon ortalama ve kovaryans matrisi elde edilemeye çalışılmıştır. Bu aşamada netNorm ve genetik algoritma kullanılarak ilgili değişkenler üretilmiştir. Bu değişkenler ile sentetik veri üretilmiş ve sınıflandırma performansı artırılmıştır. Bu çalışma da 2 farklı veri seti kullanılmış, bu veri setlerinde yer alan sınıflar için ayrı ayrı bu yöntem uygulanıp sonuçlar gözlemlenmiştir. Gözlemler sonucunda, önerilen yöntemin her iki veri seti içinde sınıflandırma performansını pozitif yönte arttırdığı sonucuna varılmıştır.

1. INTRODUCTION

One of the fastest-growing illnesses within the world is Dementia. According to the world health organization [2], there are 50 million cases worldwide, 10 million new cases every year, and these numbers are expected to be tripled by the year 2050. Dementia is also estimated to have a serious economical impact with caregiving costs rising up to 2 trillion US dollars per year by 2030 [2]. Fortunately, dementia occurs following an earlier curable state of brain disorder called mild cognitive impairment (MCI). As such, studies have shown that only 5-10 % of the MCI patients experience a conversion to dementia in the first 10 years [3]. Hence, early detection and treatment of MCI is crucial to avoid the above-mentioned impacts of dementia. However, MCI is one of many neurological diseases where general therapy for all patients is unavailable [4]. In fact, recently, more and more studies proclaim that those general treatments are less and less efficient compared to the *personalized treatments* [5–8]. Hence, marking the end of the *one-size-fits-all* era in medicine and the beginning of the personalized treatments as a more efficient alternative for curing patients.

One way to explore in-depth the potential of personalized treatments is by leveraging machine learning models which surpasses the human capacity in pattern recognition [9]. In this context, recent studies suggested that the application of machine learning models on *multi-view medical imaging data* can enhance biomarker discoveries, therefore, leading to major refinements when proposing personalized treatments [10–12]. Basically, multi-view data includes various sets of features, each capturing a unique aspect of the data. Modeling the relationship between data views can perform better performance for classification in areas such as disease diagnosis in medical imaging [13–18].

1.1 Motivation

In the context of AD diagnosis, several studies have used brain images derived from magnetic resonance imaging (MRI) [19, 20] to explore AD brain biomakers. The

relation between pairs of anatomical ROIs can be represented by a brain connectome [20]. dMRI used for deriving structural networks and rsfMRI used for generating functional brain networks are *generally* used to measure the relationship between ROIs. However, this type of brain data is often time-consuming, costly to acquire and prone to noise [21]. To circumvent this issue, recent landmark studies [17, 18, 21–25] proposed to use of MBNs derived from the traditional T1-weighted MRI and reported a significant boost in brain disease classification results. Thereby, one can *safely* claim MBNs as the new defacto data type for AD diagnosis.

While such multi-view MBNs can be used for AD diagnosis, they might limit the generalizibility and accuracy of machine learning models where there is very limited samples to train on. Such problem is commonly solved by data generation/augmentation methods. For example in [26], in order to increase the number of samples the authors proposed to use Synthetic Minority Oversampling Technique (SMOTE). Specifically, SMOTE is based on a straightforward interpolation at random points laying between the k nearest neighbors of the ground truth samples. For many years, researchers have been characterizing SMOTE as rather an promising technique for data augmentation yet having few limitations [27]. For instance, when dealing with high-dimensional multi-view data it generates noisy synthetic samples because of predefined distances (e.g., Manhattan, Euclidean) usage. In order to address this limitation, [28] proposed MV-LEAP, a framework designed specifically to augment multi-view data given a few samples. However, MV-LEAP proliferates each brain network view independently thus disregarding the relationship between views.

1.2 Objectives and Scope

In this research, we are interested in the relationship between different views of the same dataset. We presume that this relationship may encode highly useful information which can potentially help a given classifier to achieve good generalizability from limited data. To do so, we adopt a different method and explore whether we can augment synthetic samples from only one ground-truth multi-view brain network. In such case, selecting/creating a representative and centered sample becomes very important. Obviously, generating synthetic samples from a single sample is only conceivable when we have relevant prior knowledge of the corresponding population

(e.g., AD). So, obtaining the centered and representative single sample that represents the whole samples of the population is one the key point. To solve this problem, it is proposed to use the newly developed paradigm of connectional brain template (CBT) [1,29] which represents the population of multi-view brain networks in a single connectivity matrix. It can be considered as a single *mean* sample that represents an *average* of the given population. With the help of such a centered and representative CBT, we propose GEneralized Multi-VIew data Proliferator (GEM-VIP) framework which aims to proliferate synthetic samples to compensate limited multi-view data Figure 1.1.

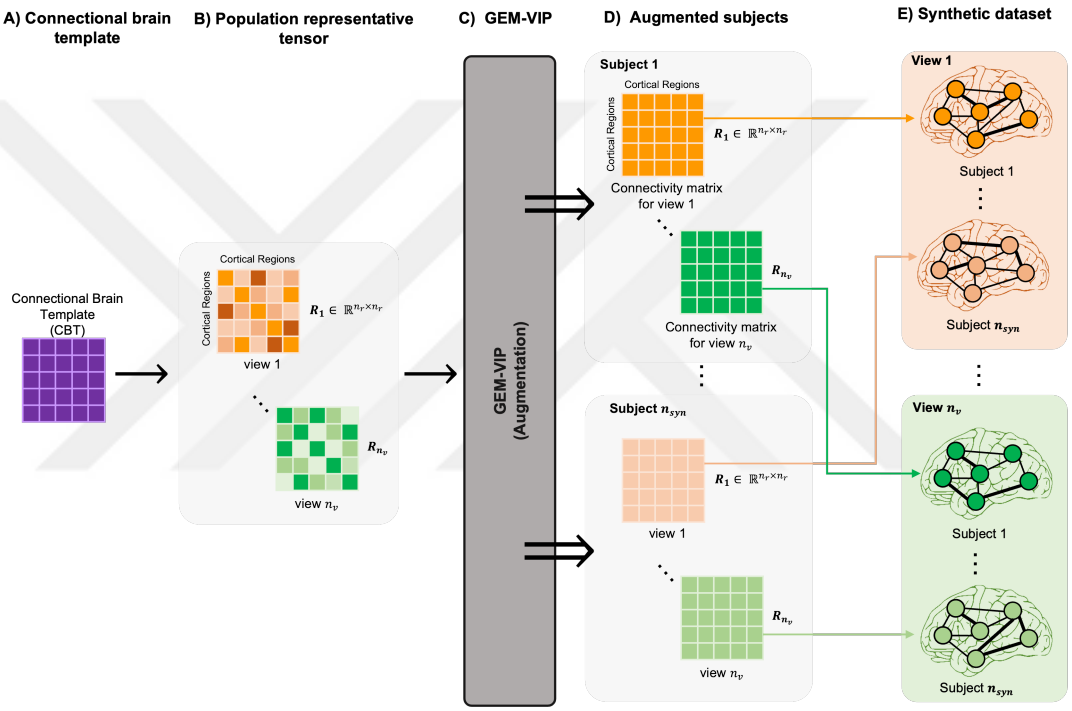

Figure 1.1 : A concept figure of GEM-VIP. (A) gets the CBT as an input. (B) creates the connectivity matrix (i.e., population-representative tensor) of the given population. (C) proliferates the augmented subjects (D) as output.

The proposed GEM-VIP method comprises three steps detailed as follows: *First,* unlike MV-LEAP [28], which proliferates each data view independently, we aim to model the relationship between views while augmenting the multi-view data. To do so, our framework relies on generating synthetic samples by using the inverse of the probability density function of multi-variate normal distribution (MVND) in which each variate represents a view of the multi-view data, and encodes the relationship between multi-variates (i.e., multi-views). Precisely, the MVND is a generalization of

two or more uni-variate (i.e., in our case each brain view) normal distributions, and has two parameters. *The first* is a vector encoding the mean of each uni-variate (i.e., view) normal distribution. The *second parameter* is the covariance matrix holding the variance of each uni-variate normal distribution. In other words, we require *mean* vector and *covariance* matrix of a population as a prior (e.g., AD or lMCI) in order to proliferate synthetic samples falling within in the prior multi-variate normal distribution of the population.

Second, with the given centered and representative CBT's of the population, we obtain a population-representative tensor. Despite being representative (i.e., and not the actual population-mean), this tensor can be *mathematically* considered as a given population's mean, and can be used as the first parameter of MVND equation. Precisely, the *population representative tensor* captures the most frequent cross-view features across each subjects' multi-view brain tensor [1]. As such, this tensor represents the overall most common *blue-print* for a given neurological disease.

Finally, we need to obtain the *covariance matrix* of the given population that will be used in MVND to be able to proliferate synthetic samples. However, this task is challenging since medical data is often limited, and we only have population CBTs. As a result, straightforwardly extracting covariance matrix from limited data may not provide a meaningful matrix to represent the whole population. To overcome this issue, we propose to use genetic algorithm [30], which is an effective method for resolving an issue for which there is limited information. Initially, genetic algorithm (GA) creates random initial candidates covariance matrices, and generates synthetic samples for each candidate matrices. Then, measures the fitness of each candidate matrices based on cross-distance between generated samples and the population CBT. After that, GA selects best ones for following iterations (i.e., generations). It repeats same process till converges to the optimum candidate covariance matrix.

After obtaining the above-mentioned *representative tensor* and *covariance matrix* of the given populations, a random vector is generated to provide diversity for data proliferation task. Finally, synthetic samples are proliferated by applying obtained parameters over MVND.

1.3 Contributions

A set of novel approaches is proposed in this thesis solving the multi-view data proliferation from a single sample and boosting the multi-view brain data classification result of the given datasets.

The main contributions of our method can be summarized as follows:

1. We propose to use the inverse of the probability density function of multi-variate normal distribution for generation of synthetic samples.
2. To the best of our knowledge, this work is the first to solve the problem of multi-view data proliferation from a single representative template (i.e., CBT).
3. We propose to use a prior CBT to obtain the population-representative tensor which can be regarded as a *mean* of population.
4. We propose a novel method that learns a covariance matrix between different views of a given population using genetic algorithm.
5. Our method is a generic approach that can be applied to augment any multi-view brain data.
6. With the proposed method, we improve the accuracy results of the classifiers helping the diagnoses of the neurological disorders.

1.4 Organization of the Thesis

The thesis is structured as follows.

Chapter 1 presents the introduction, motivation, objectives and scopes of the thesis. **Chapter 2** has provided the background of the research. It explains the multi-view data augmentation , data augmentation problems and their solutions, high dimensional multi-view data and the genetic algorithm. **Chapter 3** explains the details of our GEM-VIP method which is based on the multi-variate normal distribution equation. The steps of obtaining the mathematically regarded mean of the given populations by using population-representative tensor is given in **Chapter 4**. **In Chapter 5**, the genetic algorithm approach that is used to obtain the covariance matrix between different views of a population is introduced. This is followed by the data proliferation, and the classification of the samples (i.e., synthetic and ground-truth) in **Chapter 6**.

Chapter 7 provides the details of the dataset that are used in the experiments. Also, the steps of the experiment, evaluation and comparison methods are explained in this chapter. The insights of the experiments, limitations and the possible suggestions for future works are mentioned in **Chapter 8**. **Chapter 9** summarizes the outcomes of the research and draws conclusions from them. Furthermore, the research's limitations are discussed, as well as suggestions for future works.

2. LITERATURE SURVEY

In this chapter, we first give an overview of multi-view data augmentation. Next, we provide an overview for data proliferation techniques called SMOTE and MV-LEAP. At the end, we explain the genetic algorithm approach that we use to create covariance matrix of the given populations.

2.1 Multi-view Data Augmentation

Data augmentation is a set of methods for creating synthetic data samples from ground-truth data to expand the number of samples of a given population. The more representative and centered the augmented data, the more successful it will be in areas of application (e.g., classification).

A simple and effective traditional data augmentation technique is making small changes (e.g., flipping, rotating, cropping) to real samples [31]. Nevertheless, flipping or rotating a brain network matrix completely changes the order of ROIs.

Another type of data augmentation technique is interpolation-based oversampling techniques [26, 32–34]. Since they have many variations and are widely-used in the researches, they may be unable to handle high-dimensional brain networks, leading to important information loss. Moreover, such techniques fail to generalize to graph-based data living in a non-Euclidean multi-view space. To circumvent this issue, [35] proposed a GAN-based approaches to proliferate synthetic samples from a given single-view dataset called BrainNetGAN.

Unlike BrainNetGAN, which proliferates synthetic samples from a single-view data, [28] proposed MV-LEAP a framework specifically developed to augment multi-view data. Besides, MV-LEAP proliferates each data view separately (i.e., ignoring the relationship between the views). Furthermore, it appears to be less effective when the aim is to proliferate synthetic samples from only a few ground-truth samples.

Alternatively, [36] proposed one representative-shot learning based on learning from only one sample. Even though this method achieved good performance on brain disorder classification, the data is still limited. A large amount of data is required for statistical analysis and machine learning applications. Since this method has no data augmentation, it limits the leveraging the proven efficiency of machine learning applications in brain disorder classification. To address this limitation, we propose GEM-VIP framework to generate synthetic multi-view samples from a single sample while preserving the relationship between the views. In the first step, we us a population-driven CBT as a single representative template (i.e., mean of a population). Next, we apply genetic approach on randomly generated candidate covariance matrices using prior given CBTs as a fitness function to converge the optimum covariance matrix of the population. Lastly, we proliferate synthetic samples using the prior obtained mean and covariance matrix.

2.2 Imbalanced Datasets, SMOTE and MV-LEAP

Classification is a very crucial method for pattern recognition, especially in disease diagnostic problems. Many learning methods, like as decision trees and support vector machines, have been widely developed and used. Although this approaches are very effective, they may suffer from imbalanced class distribution [37]. The distribution is described as being unbalanced because certain classes have much more samples than others. Since uncommon events occur infrequently, collecting samples for some classes is difficult. In such cases, samples from small classes are misclassified more frequently than those from more common classes [38]. For example in a diagnosis of Alzheimer's disease, the AD case is often rare compared to normal subjects. It is a crucial problem because the classification goal is to detect the disease not the normal subjects.

Several approach to enhance a classifier's performance under the data imbalance condition have developed. Depending on the strategy used, they can be divided into three types: those who generate synthetic samples for the minority class, those who remove samples from the majority class, and those who do both. [39] proposed using a method called EasyEnsemble. Essentially, this strategy involves removing a certain number of samples and decreasing the size of the majority class, so balancing the

distribution across all classes. Removing samples, on the other hand, may change the distribution of samples, resulting in information loss [37]. Another work [26] proposed using Synthetic Minority Oversampling Technique (SMOTE) to avoid information losses. As its name implies in order to balance the training set, this method proliferates synthetic samples that belong to the minority class.

To get an understanding of this framework, consider SMOTE to be a framework based on simple interpolation at random points located between the k nearest neighbors of the ground truth samples. SMOTE has been described by academicians for many years as a particularly exciting approach for oversampling with few constraints [27]. As such, the machine learning community has seen an overflow of SMOTE variations. To mention a few, [32] proposed SMOTEBoost a framework that adds a "re-weighting" procedure to the original SMOTE algorithm. In other words, SMOTEBoost changes the weights of the misclassified subjects by generating synthetic samples which compensate for the skewed distribution, instead of updating the weights of the misclassified samples (i.e, using only boosting procedure). In their paper [33], the authors proposed borderline SMOTE a technique aiming to over-sample only the minority borderline samples (i.e., samples that are close to majority class but labeled as a minority class). In another paper, [34] proposed CURE-SMOTE as an attempt to generate synthetic examples by clustering the minority class, removing the noises and outliers, and applying SMOTE to the remaining samples. Although the above-mentioned methods displayed a significant boost of the classification results for imbalanced data, these methods still fail to accurately pick the nearest neighbors. This is because of predefined distances (e.g., Manhattan, Euclidean) which generates noisy synthetic samples when dealing with high-dimensional multi-view data.

[28] suggested MV-LEAP, a framework intended particularly to deal with multi-view data representation given multi-view data. To overcome this limitation, they propose the View-Specific LEArning-based Proliferator (VS-LEAP), that generates single-view training samples based on learning the similarities within the minority class samples [28]. They determine the nearest neighbors using the sample-to-sample similarity matrix created by multi-kernel ML [40] then generate the synthetic samples based on specified neighbors. MV-LEAP, on the other hand, generates each data view independently, ignoring the relation between views. Instead, we are interested

in the relationship between different views of the same dataset in this research. This relationship, we believe, may carry highly relevant information that might enhance a Support Vector Machine (SVM) classifier in achieving effective generalization for limited data.

2.3 Genetic algorithm

Genetic Algorithm is a programming method based on the logic of evolutionary theory. The Genetic Algorithm is a problem-solving approach that aims to produce the plausible solution possible [41]. They are the most effective method for resolving an issue for which there is limited information. Due to the its generality, GA performs well in any search space. It leverages the concepts of selection and evolution o find the plausible solution for the given problem. The GA is activated by a collection of possible solutions to the issue, as well as a measure called a fitness function that allows each solution to be scored numerically. These candidates might be chosen at random or could be tried-and-true solutions that the GA is searching for. Then, GA evaluates each candidate based on their fitness score. Candidates with higher score have a better chance of being chosen than candidates with lower fitness score. These suitable candidates are preserved and given the opportunity to reproduce. Then, candidates are paired and mated. The crossover points can be selected randomly, and the offspring solutions are generated. In order to provide genetic diversity for next generation, the mutation part applies a random changes on candidate solutions. These newly created candidate solutions creates a new pool of candidate solutions. Those candidates which has worse fitness score are deleted. The one that has better fitness score are picked and kept for the next generation, and the process repeats itself. Each iteration, it is expected that the population's average fitness will improve. A plausible solution to the problem can be obtained after hundreds of thousands of iterations [30].

3. MULTI-VARIATE NORMAL DISTRIBUTION AND GEM-VIP

This chapter describes the MVND, and the method that is used to proliferate synthetic samples for the given populations (e.g., AD, NC, eMCI, lMCI).

3.1 Multi-variate Normal Distribution Equation

A random vector $\mathbf{x} = \{X_1, X_2, ..., X_n\}$ is a multivariate normal distribution if any linear combination of its vectors $\{X_1, X_2, ..., X_n\}$ follows a Gaussian distribution. In other words,

$$\mathbf{x} = a_1 X_1 + a_2 X_2 + ... + a_n X_n \tag{3.1}$$

has a normal distribution for any set of constants $a_1, a_2, ..., a_n$.

Similarly, linear transformation of a set of independent standard normal random vector can be viewed as multivariate normal distribution. To put it another way, if $\mathbf{z}$ is random vector of standard random variables, then there exists a matrix $\mathbf{B}$ and vector μ such that:

$$\mathbf{x} = B\mathbf{z} + \mu \tag{3.2}$$

if $\mathbf{x}$ is multivariate normal, it also has a *mean vector* μ such that:

$$\mu = (\mathbb{E}(X_1), \mathbb{E}(X_2), ..., \mathbb{E}(X_n)) \tag{3.3}$$

where $\mathbb{E}(X)$ is the expected value (mean) of X. $\mathbf{x}$ also has a *covariance matrix* Σ such that:

$$\Sigma_{i,j} = Cov(X_i, X_j) \tag{3.4}$$

where $Cov(X_i, X_j)$ is the covariance of X_i and X_j. Moreover, we can define $\mathbf{x}$ by μ and Σ. Therefore, it is convenient to write:

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma) \tag{3.5}$$

The probability density function of a multivariate normal distribution $\mathbf{x}$ is given by:

$$p(\mathbf{x}; \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right)}{\sqrt{(2\pi)^k |\Sigma|}}, \tag{3.6}$$

where $exp(\mathbf{x}) = \mathbf{e}^{\mathbf{x}}$.

3.2 Generalized Multi-view Data Proliferator (GEM-VIP)

For the given CBT of a population (e.g., AD) and the definition of MVND equation (3.6) which has *three parameters* to initialize, we aim to proliferate synthetic samples belonging to the same multi-variate normal distribution for a given population. A CBT is a fused form of a population-representative tensor (i.e., mosaic representation) that extracts the most frequent cross-view feature vectors across subjects. So, for the given CBT there exists a population representative tensor $\tilde{\mathcal{T}} \in \mathbb{R}^{n_r \times n_r \times n_v}$. We aim to use this tensor as the *mean parameter* (μ_r) of the given MVND. Moreover, since we have no prior information about *covariance matrix* (Σ) of the population, we use genetic algorithm approach to obtain it as the second parameter of the MVND. Ideally, a genetic algorithm allows us to reliably solve searching problems [30]. After obtaining the μ_r and Σ parameters of the MVND equation and generating an initial *random vector* ($\mathbf{x}$) which are the three required parameter to initiate MVND equation, we proliferate multi-view synthetic samples for the given population Figure A.1.

Capitalizing on the above-mentioned definitions, we propose GEM-VIP, a framework aiming to proliferate synthetic samples. To do so, we aim to obtain both the *representative tensor (μ_r)* and *covariance matrix (Σ)* of the target population. First, we propose to use a population-representative tensor as the population mean. Second, we use genetic algorithm technique to derive a plausible *covariance matrix (Σ)* for the population. After collecting the key variables and a randomly generated vector, we have all the necessary parameters of (3.6) to be able to proliferate synthetic samples.

4. USING POPULATION-REPRESENTATIVE TENSOR AS THE MEAN (μ_r) OF A POPULATION

In this section, given a CBT of a population (e.g., AD, NC, eMCI or lMCI), we aim to obtain the μ_r of the population distribution. Figure 4.1 illustrates the key steps for obtaining the μ_r of the population.

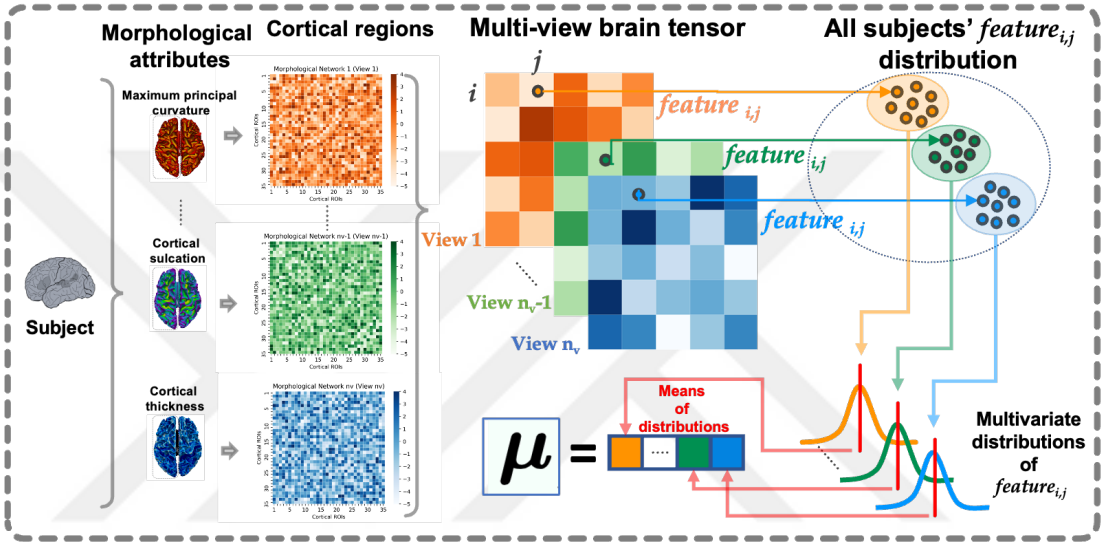

Figure 4.1 : The key step of obtaining the representative tensor (μ_r) of the population is described. We obtain the μ_r which extracts the most frequent cross-view feature vector within each subjects' multi-view brain tensor, and can be used as a mathematically regarded representative mean of population.

We denote by $\mathbf{V}_{ij}$ vector, the connectivity between an ROI i and ROI j across the views (i.e., feature), and each subject s has its own feature $\mathbf{V^s}_{ij}$. The proliferator requires the representative $\mu_{\mathbf{ij}}$ vector of each feature $\mathbf{V}_{ij}$ in the population distribution in order to generate the synthetic samples.

As such, we propose to use the given population-representative tensor $\tilde{\mathcal{T}}$ [1] that extracts the most frequent cross-view feature vectors within the subjects in a selective manner as the mean of population. In [1], for each subject s, a multi-view tensor, where each view represents a distinct brain attribute, has been associated. Using the Euclidean distance across all feature vector of subjects, the vector that satisfies the

minimum mean distance to all other subjects' feature vector has been picked as a population-specific feature vector (i.e., *representative vector mathematically regarded as mean*) . Hence, [1] builds a tensor that holds all the population-specific feature vectors $\tilde{\mathcal{T}}$.

4.1 Morphological Brain Networks

A single-view MBN is defined as a graph where nodes representing ROIs and edges representing the similarity between the ROIs by using cortical attributes such as maximum curvature, sulcal depth Figure 4.2.

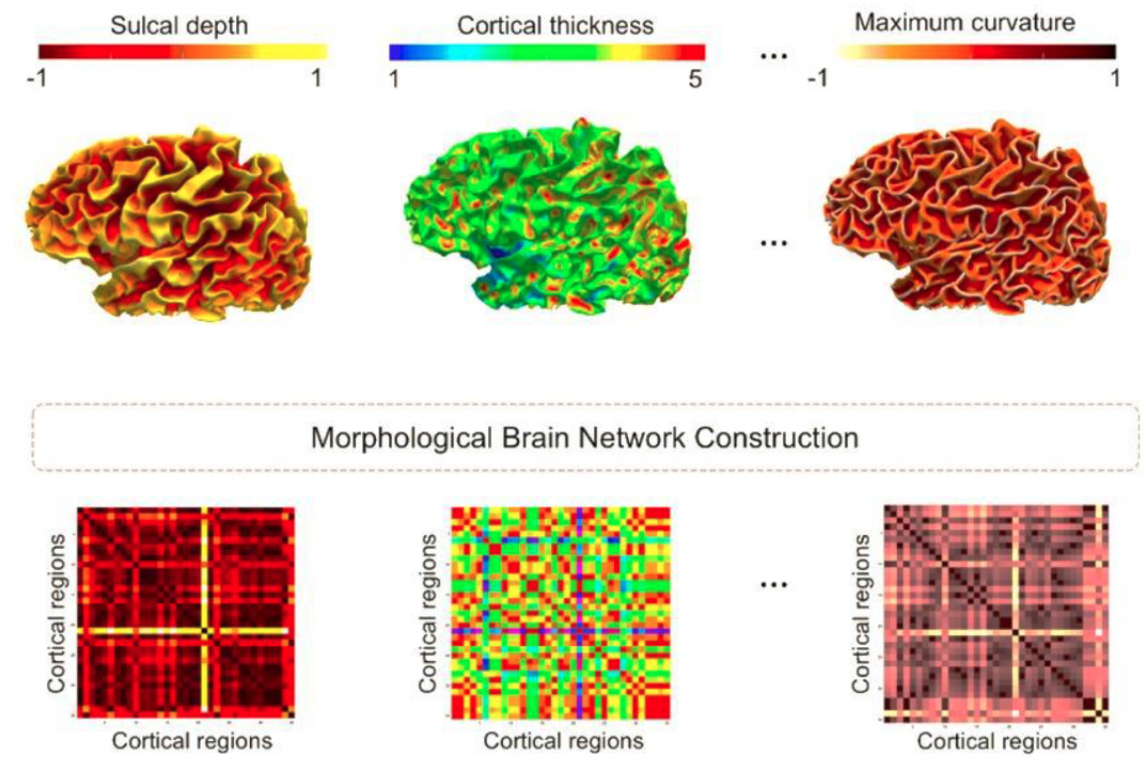

Figure 4.2 : MBN construction for a representative subject utilizing different views of the cortical surface [1].

Where n_r represents the total ROI numbers, each network is represented by a matrix $\mathbf{X}$ in $\mathbb{R}^{n_r \times n_r}$. Next, by calculating the absolute difference between each ROIs i and j, we specify each element of $\mathbf{X}(i, j)$, and generate a *connectivity matrix* that represents the similarity between different ROIs for the related cortical attribute (i.e., sulcal depth, cortical thickness, maximum principal curvature, average curvature). Also, applying the same method on different cortical attributes, we proliferate views of multiple

network for each subject. For a given population where each subject s is expressed with n_v MBNs, $\mathbf{X_k^S}$ represents the k^{th} network view (i.e., connectivity matrix). Then, it can be denoted that each subject s as a tensor $\mathscr{T}^s \in \mathbb{R}^{n_r \times n_r \times n_v}$ by combining each views. After that, for each ROI pairs i and j, we can construct a cross-view feature vector $\mathbf{V}_{ij}^s$ where $1 \leq i, j \leq n_r$. This yields:

$$\mathbf{V}_{ij}^s(k) = \mathscr{T}^s(i,j,k) \ , \ \forall\, 1 \leq k \leq n_v \ , \ \forall\, 1 \leq i, j \leq n_r \tag{4.1}$$

$\mathbf{V}_{ij}^s$ and $\mathbf{V}_{ij}^{s'}$ represent the cross-feature vector of subjects s and s'. For ROIs i and j, the Euclidean distance among subject s and s' is computed as follows:

$$\mathbf{d}_{ij} = \sqrt{\sum_{k=1}^{n_v} \left(\mathbf{V}_{ij}^s(k) - \mathbf{V}_{ij}^{s'}(k) \right)^2} \tag{4.2}$$

By the help of equation (4.2), we can calculate the overall distance from feature vector $\mathbf{V}_{ij}^s$ for subject s to whole individuals in the population. So, it is convenient to write:

$$\mathbf{D}_{ij}(s) = \sum_{s'=1}^{n} \sqrt{\sum_{k=1}^{n_v} \left(\mathbf{V}_{ij}^s(k) - \mathbf{V}_{ij}^{s'}(k) \right)^2} \tag{4.3}$$

Then, the subject-specific feature vector, having the smallest distance to other subjects $min(\mathbf{D}_{ij}(s))$ where $1 < s < n$, can be seen as the most representative and centered cross-view feature vector across the population. The vector can be used as a mathematically regarded $mean(\mu)$ vector of the population for given ROIs i and j. Applying equation (4.3) on the given connectivity matrices of a population, we can obtain the *mean* vector (i.e., has smallest distance to other vectors) of each ROIs i and j, where $1 < i, j < n_r$. Then, we can create the population-representative tensor $\mathscr{T}$ holding the mean vectors of each feature in the population Figure 4.3.

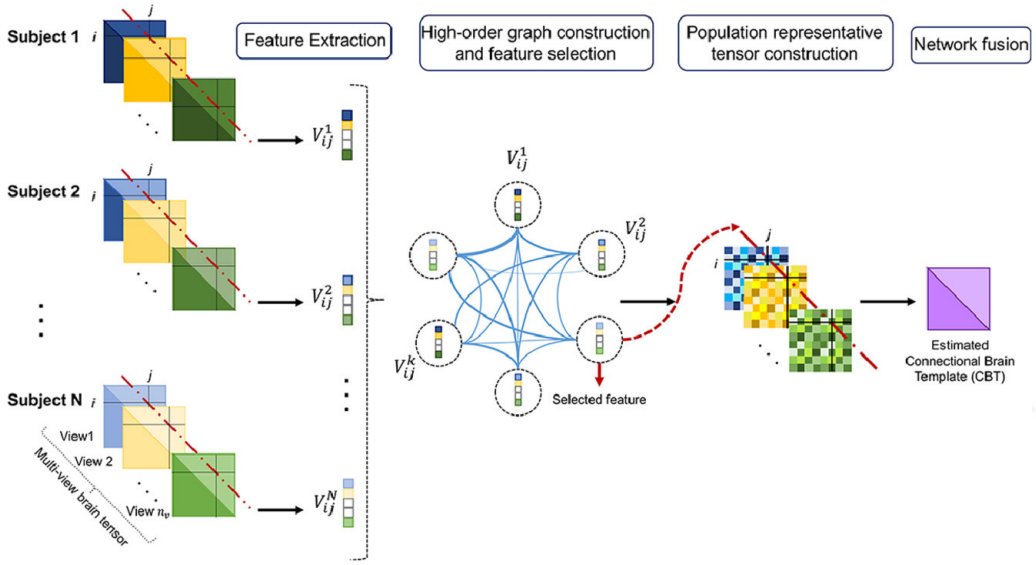

Figure 4.3 : Overview of connectional brain template (CBT) and population representative tensor extraction from multi-view brain networks [1].

5. OBTAINING PLAUSIBLE COVARIANCE MATRIX BETWEEN DIFFERENT VIEWS OF A POPULATION

This section describes the details of the genetic algorithm that is used to generate covariance matrices of the given populations. To proliferate synthetic samples, for the given CBTs of a population (e.g., AD, NC, eMCI or lMCI), equation (3.6) needs the covariance matrix Σ of the population. In other words, we aim to generate the covariance value of each feature in the population. However, we only have the dimension of the covariance matrix. Therefore, due to its efficient multi-variable searching scheme, we propose to leverage the genetic algorithm (GA) approach [30] to build the optimal covariance matrix Σ.

5.1 Genetic Algorithm

In genetic approach Figure 5.1, we propose to use, as a *fitness function*, the cross distance (5.1) between the CBT and the synthetic samples generated with the help of the candidate covariance matrix (i.e., randomly generated initial matrix to trigger GA). We use this distance to evaluate the suitability of the current candidate covariance matrix where we iterate our GA until it reaches the optimum candidate covariance matrix. In the next paragraph, we will detail each step of the GA that allows us to obtain the desired covariance matrix.

5.1.1 Creation of initial population

The process of genetic algorithm starts with a set of subjects in a population, in our case candidate covariance matrices are the subjects of the population. These candidate matrices are generated randomly to initiate the genetic algorithm.

5.1.2 Fitness function

The fitness function determines the how fit a candidate matrix. A fitness score is calculated for each candidate matrix in the population. We propose to use, as a *fitness*

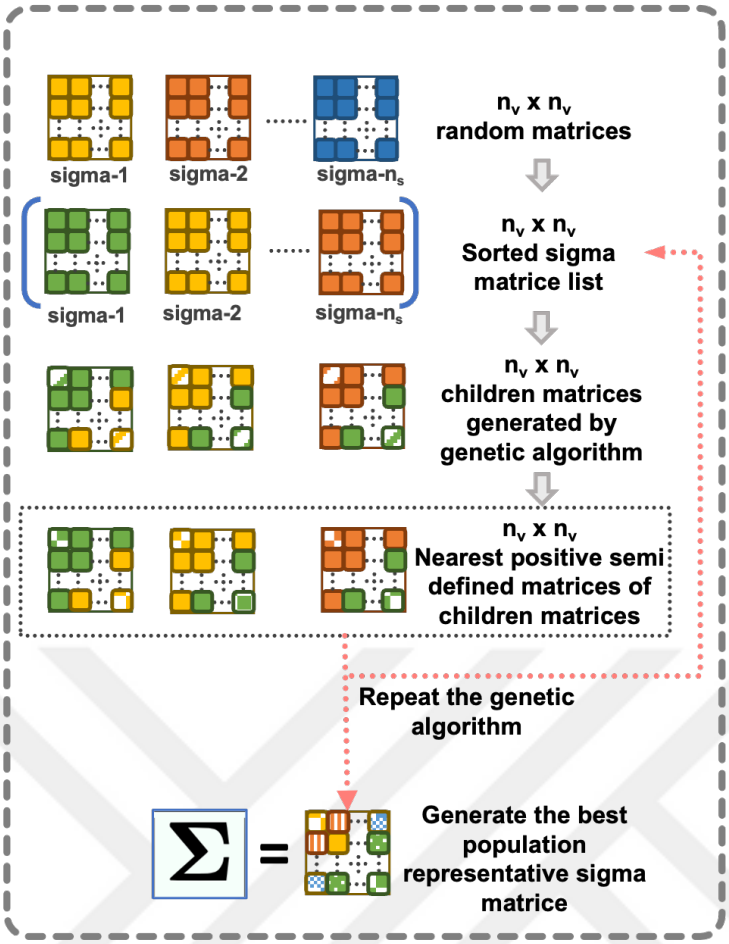

Figure 5.1 : Overview of genetic algorithm that is used to proliferate covariance matrix.

function, the cross distance (5.1) between the ground truth samples and the synthetic samples generated with the help of the candidate covariance matrix (i.e., randomly generated initial matrix to trigger GA). We use this distance to evaluate the suitability of the current candidate covariance matrix where we iterate our GA until it reaches the optimum candidate covariance matrix.

$$d_F(A,B) = \sqrt{\Sigma_i \Sigma_j \left| a_{ij} - b_{ij} \right|^2} \tag{5.1}$$

5.1.3 Selection of best candidate matrices

Selection part is used to select the fittest candidate covariance matrix for the next step of the genetic algorithms. Based on the fitness score of the candidate covariance matrix, we select n_c of them as parent covariance matrix for the next steps.

5.1.4 Crossovering the candidate matrices

This part is one of the most significant step of the genetic algorithm. It provides genetic diversity in the population Figure 5.2. The parent covariance matrices are paired, and mated. The crossover points are randomly selected, and children matrices are generated by exchanging the values of the selected points.

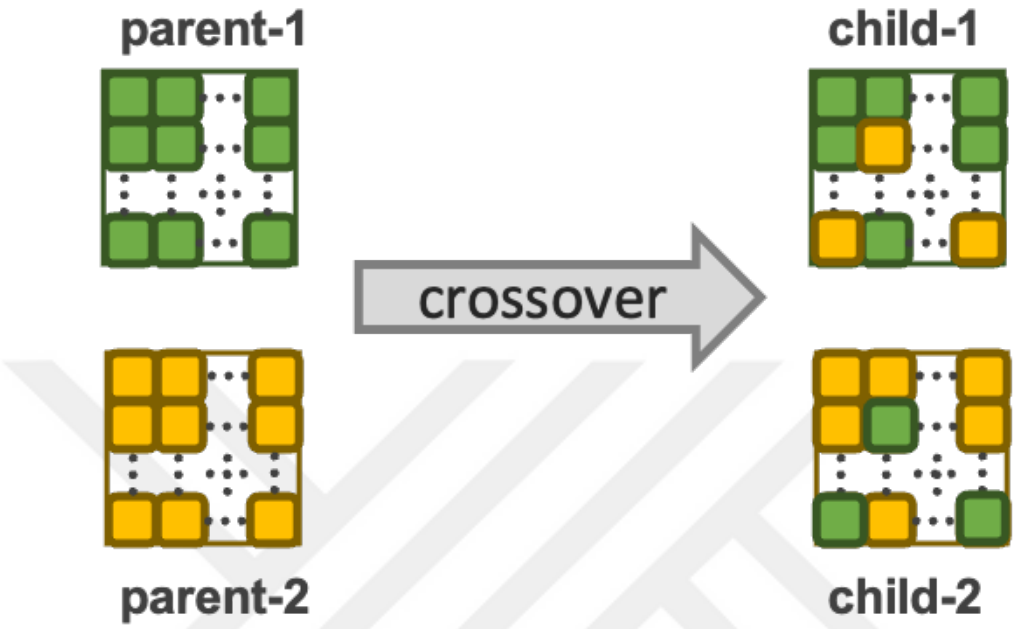

Figure 5.2 : Overview of parent matrices' crossover.

5.1.5 Children matrices mutation

This step is used to prevent premature convergence to local optima and provide genetic diversity in the population Figure 5.3. Randomly selected points of the cross-overed children matrices are changed by a random value that is between the range[1] of covariance matrix.

5.1.6 Converging the nearest positive semi-definite covariance matrix

The equation (3.6) that will generate the synthetic samples for the given multivariate normal distribution needs covariance matrix of the population (i.e., output of genetic algorithm). However, the covariance matrix must be positive symmetric semi-definite matrix [42]. In order to project the generated matrices to their nearest positive symmetric semi-definite matrix, we applied [43]'s method. This method both

[1] A random value between the minimum and maximum value within the covariance matrix.

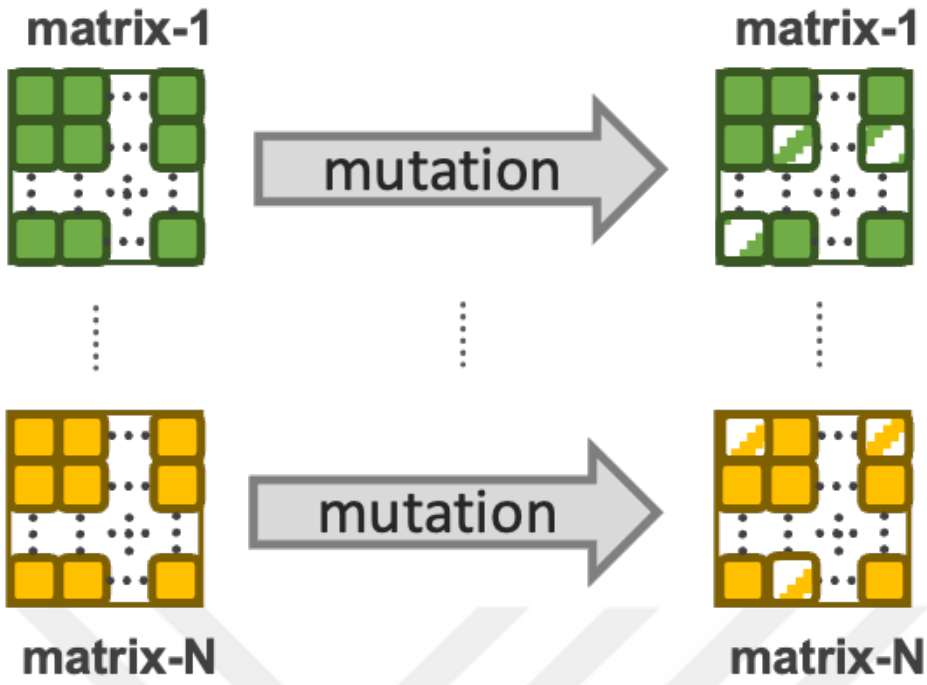

Figure 5.3 : Overview of covariance matrix mutation.

computes the nearest positive semi-definite matrix and provides a genetic diversity for the children matrices.

5.2 End-to-end Covariance Matrix Generation

To generate a plausible covariance matrix for the given population (e.g., AD), we use equation (5.1) as a fitness function to evaluate the suitability of the newly generated children matrices. The genetic algorithm's steps are as follows:

1. *Generate initial matrices to trigger GA.* The initial matrices can be generated by randomly or can be extracted from the given population (e.g., AD).

2. *Selecting the promising parent matrices.* By using the initial covariance matrices and the earlier obtained μ_r on the equation (3.6), synthetic samples are generated for corresponding covariance matrices (i.e., initial matrices). Then, the covariance matrices can be sorted by comparing the cross-distance between synthetic and ground-truth samples (i.e., from each view of the each synthetic samples to the each view of the each ground-truth samples). The best covariance matrices (i.e.,

that generates the synthetic samples closest to ground-truth samples) are selected as parent matrices for the next steps of GA.

3. *Crossovering the parent matrices to generate children matrices.* All subjects (i.e., parent covariance matrices) in the population are paired with each other. The elements paired-parent covariance matrices are swapped with each other to produce new generation children matrices. This operation is applied to each paired parents Figure 5.2.

4. *Applying genetic mutation to children matrices.* Genetic mutation is essential for maintaining genetic diversity. The diversity is provided by either randomly increasing or decreasing the values of randomly selected elements in children covariance matrices Figure 5.3.

5. *Converging the nearest positive semi-definite covariance matrix.* The children covariance matrices must be positive symmetric semi-definite to be applied on equation (3.6). However, the positive symmetric semi-definiteness of the children matrices is broken when applying the crossover and mutation. The [43]'s method projects the children matrices to their nearest positive symmetric semi-definite matrices. This conversion promotes genetic variety while also satisfying the equation's (3.6) criterion.

6. *Preparing the matrices for the next GA iteration.* At this point, the population contains both parent and newly produced children matrices which are assigned as initial matrices for the genetic algorithm's next iteration. The genetic algorithm repeats the procedure until the optimum covariance matrix with the smallest cross-distance[2]to ground-truth samples is found.

[2] When the same distance repeats itself 10 times in a row.

6. PROPOSED GEM-VIP DATA PROLIFERATION AND CLASSIFICATION

In this chapter, the data proliferation phase of the proposed GEM-VIP method, and the classifier model that is trained on the ground-truth and synthetic samples (i.e., generated by GEM-VIP) are introduced.

Let μ_r be a *mathematically regarded mean* (i.e., earlier obtained population-representative tensor) of the given population, Σ be a *covariance matrix* that is generated by genetic algorithm for the corresponding population, and let $\mathbf{x}$ be a *randomly generated vector*. Figure A.1 – E shows the samples' generation using the multivariate normal distribution equation (3.6) which takes as input the earlier obtained μ_r, its corresponding *covariance matrix* (Σ) and a random vector ($\mathbf{x}$) $\in \mathbb{R}^{1\times n_v}$ where n_v is number of views.

Applying GEM-VIP method on the given populations (e.g., AD/lMCI) individually, we generate synthetic samples for each population. The classifier model is trained both the ground-truth subjects and synthetic samples to evaluate the effect of generated samples on classification accuracy. Then, we flatten and concatenate all samples' features (i.e., connectivity matrices) into a single vector for classification. However, due to its size, the obtained vector may suffer from the curse of dimensionality. Hence, we propose to reduce the dimensionality for each sample i vector $\mathbf{x_i}$. To accomplish so, we propose leveraging the widely used Principal Component Analysis (PCA) to extract essential features in order to ensure a straightforward reproducibility of our method [44]. Finally, we classify the reduced 2-dimensional *PCA vector* $\mathbf{x_i^p}$ (i.e., the output of the PCA function) using a simple yet efficient linear SVM classifier to distinguish the subjects (e.g., AD/lMCI or eMCI/NC).

7. EXPERIMENTS AND RESULTS

The methods for evaluating the proposed GEM-VIP model, as well as the dataset used in the research, are described in this chapter. It also includes the results of the methods' evaluation.

7.1 Datasets

We evaluated GEM-VIP on two different datasets as detailed in Table 7.1. For each subject, the cortical hemispheres are reconstructed from structural T1-w MRI by using FreeSurfer analysis suite [45]. As described in [46], the latter procedure includes motion correction, skull stripping, topology correction, averaging of two T1-w images, intensity normalization and segmentation of subcortical white matter (WM) and gray matter (GM) volumetric structures to identify GM/WM and GM/cerebrospinal fluid boundaries. Then, using Desikan-Killiany atlas [47] each cortical hemisphere is divided into 35 cortical regions. We select four cortical attributes which are: maximum principal curvature, cortical thickness, sulcal depth and average curvature. Each subject's T1-w MRI is obtained using a Siemens head-only 3T scanner (Allegra, Siemens Medical System, Erlangen, Germany) with a circularly polarized head coil, by using a turbo spin-echo (TSE) sequences: TR = 7380 nos TE = 119 mss with a Flip Angle = 150°, and resolution = 1.25 X 1.25 X 1.95mm3 [48], 70 transverse slices were acquired. We used T1-weighted images acquired from a 1.2mm isotropic resolution. With a correlation range of 0.75 for the estimated cortical thickness of the right medial prefrontal cortex to 0.99 for the estimated intracranial volume, test-retest reliability was determined [49].

Table 7.1 : AD/lMCI and eMCI/NC data distribution.

	Dataset 1		Dataset 2	
	AD	lMCI	eMCI	NC
Subject Number	41	36	18	41
Mean Age	75.27	72.54	70.4	74.1

For each cortical attribute of each dataset, the intensity of the connectivity between two ROIs i and j is calculated as follows:

$$\bar{c}_i = \frac{1}{\#\{att \in R_i\}} \sum_{att \in R_i} c(att), \tag{7.1}$$

where $\#\{att \in R_i\}$ denotes the connectivity number att belonging to ROI i while $c(att)$ refers the measure of cortical attribute given to vertex v. Each element $\mathbf{C}_c(i,j)$ in the cortical attribute c is calculated as the absolute difference value within the average cortical attribute in ROIs i and j: $|\bar{c}_i - \bar{c}_j|$ [18,21,22]. Given n_r cortical regions in each hemisphere where the size of each view $n_r \times n_r$. We note that, as two ROIs i and j become morphologically similar, their morphological connectivity $\mathbf{C}_c(i,j)$ tends to 0.

7.1.1 Dataset-1: AD/lMCI

We used T1-weighted brain MRIs of 77 subjects (36 lMCI and 41 AD) from ADNI GO dataset as the first dataset [50]. We select 21 AD and 21 lMCI subjects as ground-truth training samples, and 15 AD and 15 lMCI as ground-truth testing samples. We applied our method only on the left-hemisphere of the AD/lMCI dataset.

7.1.2 Dataset-2: eMCI/NC

Public ADNI GO dataset [51] which has 18 eMCI and 41 NC subjects, is used. We select 12 eMCI and 12 NC as ground-truth training samples, and 6 eMCI and 6 NC as ground-truth testing samples. We applied our method on both left and right-hemisphere of the eMCI/NC dataset.

7.2 Method Parameters

For genetic algorithm parameters, we switch randomly chosen indexes of parent matrices as a crossover. To provide genetic mutation, we stochastically increase or decrease the randomly selected matrix element by randomly selected value between [-1, 1]. For the generation of children matrices, we choose the best first seven covariance matrices as parent matrices. We paired all of the first seven people to each other, and each paired-parents produced 2 children. At the end of each iteration, there are 42 children and 7 parent matrices which are the initial matrices for the next generation. The stopping condition of the genetic algorithm is determined as finding

same results (i.e., same covariance matrix as global/minimum optima) 10 times in a row.

The nearest neighbors number is set to 5 for SMOTE.

N ground-truth and $3N$ synthetic samples are used to train the classifiers. The AD/lMCI dataset has a N value of 21, while the eMCI/NC dataset has a N value of 12.

7.3 Performance Measure

We conducted 3-fold cross-validation to evaluate the classification model's performance. The proposed method generates synthetic samples by using ground-truth samples. After that, a linear SVM classifier is trained by ground-truth and synthetic samples. The SVM model is evaluated by the testing samples. We test the proposed model by using the accuracy metric. The model's overall performance is measured by its accuracy, that is calculated as the number of properly identified samples divided by the number of whole samples:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{7.2}$$

where TP and TN are, respectively, the number of positive and negative samples that are correctly classified. FP and FN represent the number of positive and negative samples that are incorrectly classified.

7.4 Experiments

7.4.1 Experiment 1: left-hemisphere of eMCI/NC dataset

We choose 12 eMCI as ground-truth training samples and 6 eMCI as ground-truth testing samples based on the connectivity matrices of the left hemisphere of the eMCI population.

First, we derive the eMCI population's mathematically regarded population representative mean μ_r tensor. Then, at random, we create 49 covariance matrices and extract the eMCI population's original covariance matrix (i.e., in total 50 initial covariance matrices).

We generate 36 synthetic samples for each initial covariance matrix using previously obtained mu_r on equation (3.6). The fitness function (5.1) calculates the cross-distance between synthetic and ground-truth samples, and finds the fitness score of each initial covariance matrices. The fitness scores of the initial covariance matrices are then sorted, and the best seven are chosen as parent covariance matrices. All seven parents are paired with one another, yielding a total of 21 pairs. The crossover and mutation processes are used, and two children matrices are generated for each couple (details are discussed in chapter five). The positive symmetric semi-definiteness must be met by a total of 42 newly produced children covariance matrices. To do so, we compute each children matrix's nearest positive symmetric semi-definite matrices (i.e., chapter-five details computation). Finally, there are 42 children and 7 parent matrices that will be used in the next iteration of the genetic algorithm. We repeat the genetic process until the best 7 matrices do not change for 10 consecutive iterations.

After genetic process, the obtained μ_r tensor, Σ matrix and a random vector are used to proliferate 36 (i.e., 3 times the ground-truth samples) synthetic samples for the eMCI population. The processes outlined above are followed exactly for the left-hemisphere of the NC dataset, and 36 synthetic samples are proliferated for it. At the end of the process, the number of samples increases to 48 for each of the eMCI and NC populations.

We train the classifier with extended data to see how created samples affect classification accuracy. For classification, we first flatten and concatenate all samples' features (i.e., connectivity matrices) into a single vector. However, because of its length, the obtained vector may be subject to the curse of dimensionality. As a result, the dimensionality of each sample i vector $\mathbf{x_i}$ should be reduced. To accomplish so, we used PCA to extract essential features. Next, we use a simple yet efficient linear SVM classifier to differentiate the subjects (i.e., eMCI/NC) from the reduced 2-dimensional *PCA vector* $\mathbf{x_i^p}$. Finally, we utilize baseline and other approaches to evaluate the classifier's accuracy.

7.4.2 Experiment 2: right-hemisphere of eMCI/NC dataset

We chose 12 eMCI and 12 NC as ground-truth training samples and 6 eMCI and 6 NC as ground-truth testing samples for the given connectivity matrices of the

right-hemisphere of the eMCI/NC dataset. All the operations in the first experiment are done exactly the same for this data set.

7.4.3 Experiment 3: left-hemisphere of AD/lMCI dataset

For the given connectivity matrices of the right-hemisphere of the AD/lMCI dataset, we selected 21 AD and 21 lMCI as ground-truth training samples and 15 AD and 15 lMCI as ground-truth testing samples. As in the first experiments, we repeat the same operations exactly in this one.

7.5 Evaluation and Comparison Methods

We evaluate our GEM-VIP using four different perspectives.

1. We compare the accuracy of the model (i.e., trained on our proposed GEM-VIP's synthetic samples) against three different models as detailed in Table 7.2. These models are also trained with synthetic samples that are generated by other methods (i.e., SMOTE and MV-LEAP).

Table 7.2 : SVM-classifiers trained on using different synthetic samples of different proliferators.

	Dataset		**Proliferation Method**			**Feature Extraction**	**Classifier**
	Ground Truth	**Synthetic Samples**	**SMOTE** chawla2002	**MV-LEAP** graa2019	**GEM-VIP**	**PCA**	**SVM**
Baseline-classifier	+	-	-	-	-	+	+
SMOTE-inspired classifier	+	+	+	-	-	+	+
MV-LEAP-inspired classifier (built-in PCA)	+	+	-	+	-	+	+
Ours	+	+	-	-	+	+	+

2. We aim to compare the centeredness, representativeness, and dispersion of the synthetic samples (i.e., proliferated by GEM-VIP) with the ground-truth by using *t-Distributed Stochastic Neighbor Embedding (t-SNE).*

3. We examine the evolution of candidate covariance matrices for each iteration of the GA by calculating the cross-distance between ground-truth and synthetic samples (i.e., generated by GEM-VIP) to show how candidates matrices converge to a reasonable covariance matrix.

4. We aim to evaluate the covariance matrix (i.e., generated by GA) against three different covariance matrices by comparing the models that are trained on synthetic samples (i.e., generated by using these covariance matrices) shown in Table 7.3

Table 7.3 : SVM-classifiers trained on using different synthetic samples of different covariance matrices.

	Ground Truth & Synthetic Dataset Generated by Covariance matrix				**Feature Extraction**	**Classifier**
	Ground Truth (n)	Random Covariance Matrix (3n)	Population Covariance Matrix (3n)	GEM-VIP Covariance Matrix (3n)	PCA	SVM
Baseline-classifier	+	-	-	-	+	+
Random-classifier	+	+	-	-	+	+
Population-classifier	+	-	+	-	+	+
Ours	+	-	-	+	+	+

7.5.1 Comparison with other proliferators

On Figure 7.1, we compared classification models trained on ground-truth and synthetic samples generated by different proliferators to classify AD/lMCI (i.e., only left-hemisphere) and eMCI/NC (i.e., both left/right-hemisphere).

1. **Baseline-classifier** is an SVM classifier trained on using only PCA applied ground-truth samples

2. **SMOTE-inspired-classifier** is an SVM classifier trained on using PCA applied ground-truth and synthetic samples which are generated by SMOTE.

3. **MV-LEAP-inspired-classifier** is an SVM classifier trained on using both ground-truth and synthetic samples which are the outputs of MV-LEAP. Note that the MV-LEAP framework has built-in PCA step.

4. **Ours** is an SVM classifier trained on using PCA applied ground-truth and synthetic samples which are proliferated by proposed GEM-VIP.

The classifiers' overall accuracy is used to evaluate their performance. Ground-truth samples are divided into training and testing. Table 7.4 shows the final results. When compared to the above-mentioned methods, classifiers trained on synthetic samples (i.e., generated using our method) achieved better outcomes in both the left and right-hemispheres.

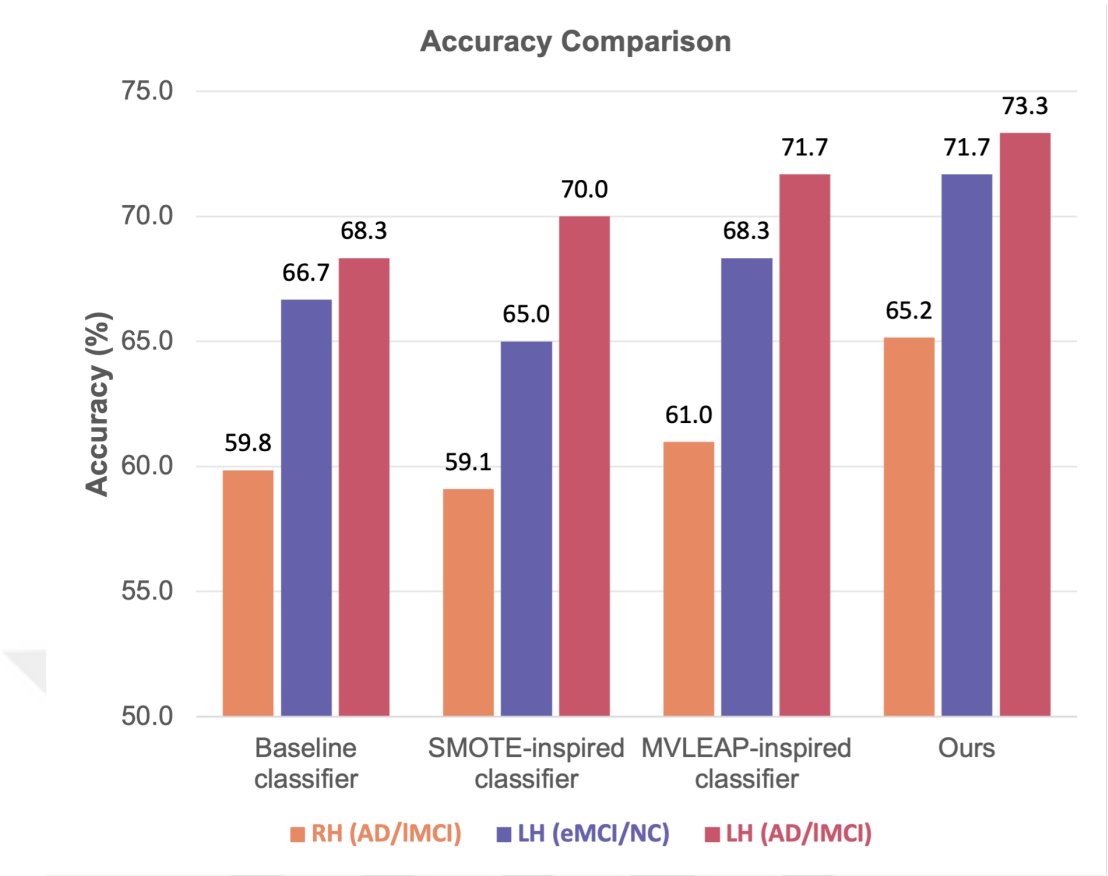

Figure 7.1 : Accuracy comparison of the eMCI/NC and AD/lMCI classification models trained on using ground-truth and synthetic dataset for both hemispheres.

Table 7.4 : Accuracy (%) results of baseline, SMOTE-inspired, MV-LEAP-inspired and our methods. LH: left hemisphere. RH: right hemisphere.

	LH eMCI/NC(%)	RH eMCI/NC(%)	LH AD/lMCI(%)
Baseline-classifier	68.333	66.667	59.800
SMOTE-inspired-classifier	70.000	65.000	59.090
MV-LEAP-inspired-classifier	71.667	68.333	60.974
Ours	**73.333**	**71.667**	**65.150**

7.5.2 Observing centeredness, representativeness, and dispersion over t-SNE graph

t-SNE is an unsupervised technique used for reducing high-dimensional data [52]. Particularly, t-SNE is a non-linear that has ability to reveal population stratification [53]. As illustrated in Figure 7.2 (i.e., for synthetic samples with ground-truths), the synthetic eMCI and NC samples scatter separately in two-dimensional visualization. Moreover, for both hemispheres, synthetic samples are spread among the ground-truths

of their respective class. Same *distribution-behavior* is also observed for the left-hemisphere of AD/lMCI dataset.

7.5.3 Genetic algorithm evaluation of the covariance matrix

We measure the performance of each candidate covariance matrix iteration to analyze the GA's evolution. Each GA iteration generates a set of possible covariance matrices. The lowest cross-distance between ground-truths and synthetic samples (i.e., generated by corresponding candidate matrix) is then calculated using equation (5.1), for each covariance matrix, and the one scoring the smallest cross distance is chosen. We can compare the computed cross-distances (i.e., between ground-truth and synthetic samples) to observe the evolution of the covariance matrix in Figure 7.3. The better the covariance matrix quality, the smaller the cross-distance in both hemispheres of eMCI/NC and the left hemisphere of AD/lMCI datasets. Following empirical tuning, we decided to run our GA for 150 iterations. The eMCI/NC dataset's covariance matrices converge to a desirable value around the 140th iteration, while AD/lMCI converges around the 100th iteration.

7.5.4 Comparison of covariance matrices

To test the centeredness and representativeness of the covariance matrix generated by GA , we use two additional covariance matrices. One of them is produced at random (i.e., random covariance matrix). The second (i.e., population covariance matrix) is calculated using the population's ground-truth samples. The population covariance matrix, in other words, reflects the ground-truth samples. Table 7.3 shows the details of the classifiers that are trained with ground-truth and synthetic samples generated with the aid of these covariance matrices.

1. **Baseline-classifier.** For n samples drawn from the ground-truth population, we apply PCA to reduce its feature dimensionality. Then, we train a baseline SVM classifier.

2. **Random-classifier.** Here, we train an SVM classifier that is using reduced features by PCA of n ground-truth **and** $3n$ synthetic samples proliferated using a *randomly* generated *covariance matrix*.

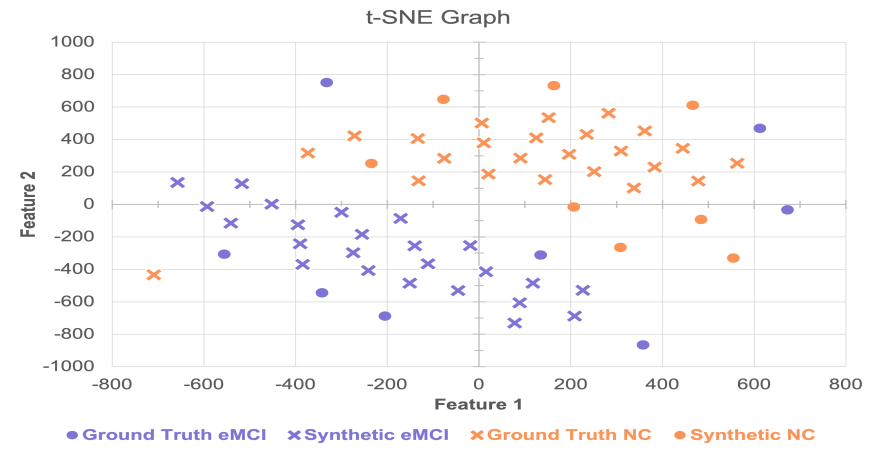

Figure 7.2a : Left-Hemisphere of eMCI/NC Dataset

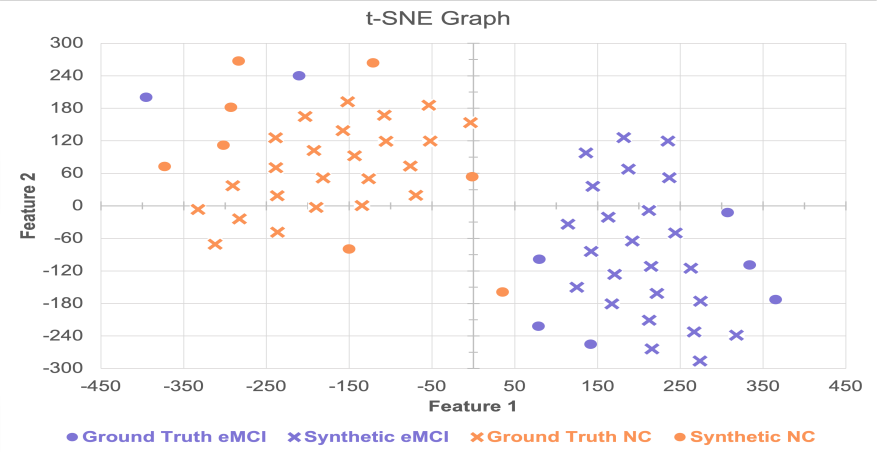

Figure 7.2b : Right-Hemisphere of eMCI/NC Dataset

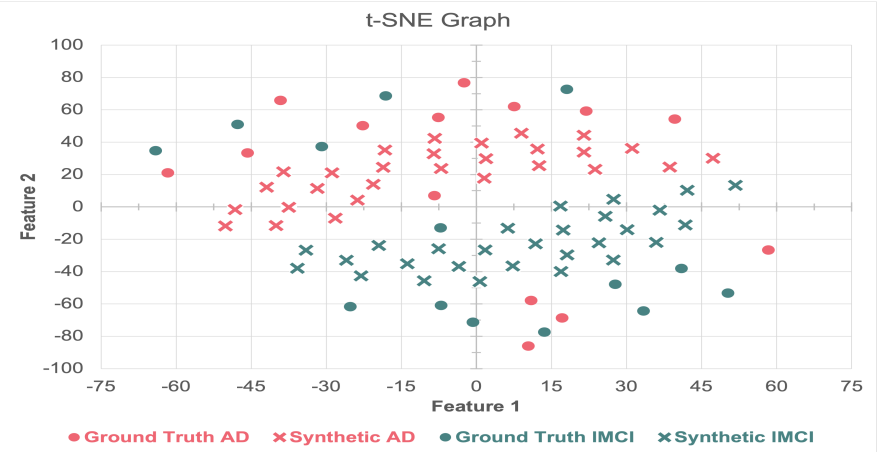

Figure 7.2c : Left-Hemisphere of AD/lMCI Dataset

Figure 7.2 : t-SNE graph shows the dispersion of the synthetic and ground-truth samples. Graph (a) and (b) shows the t-SNE of left and right hemisphere of eMCI/NC dataset. Graph (c) shows the t-SNE of the left-hemisphere of AD/lMCI dataset. Synthetic samples of each class are close to their own ground-truth samples and separated from other samples.

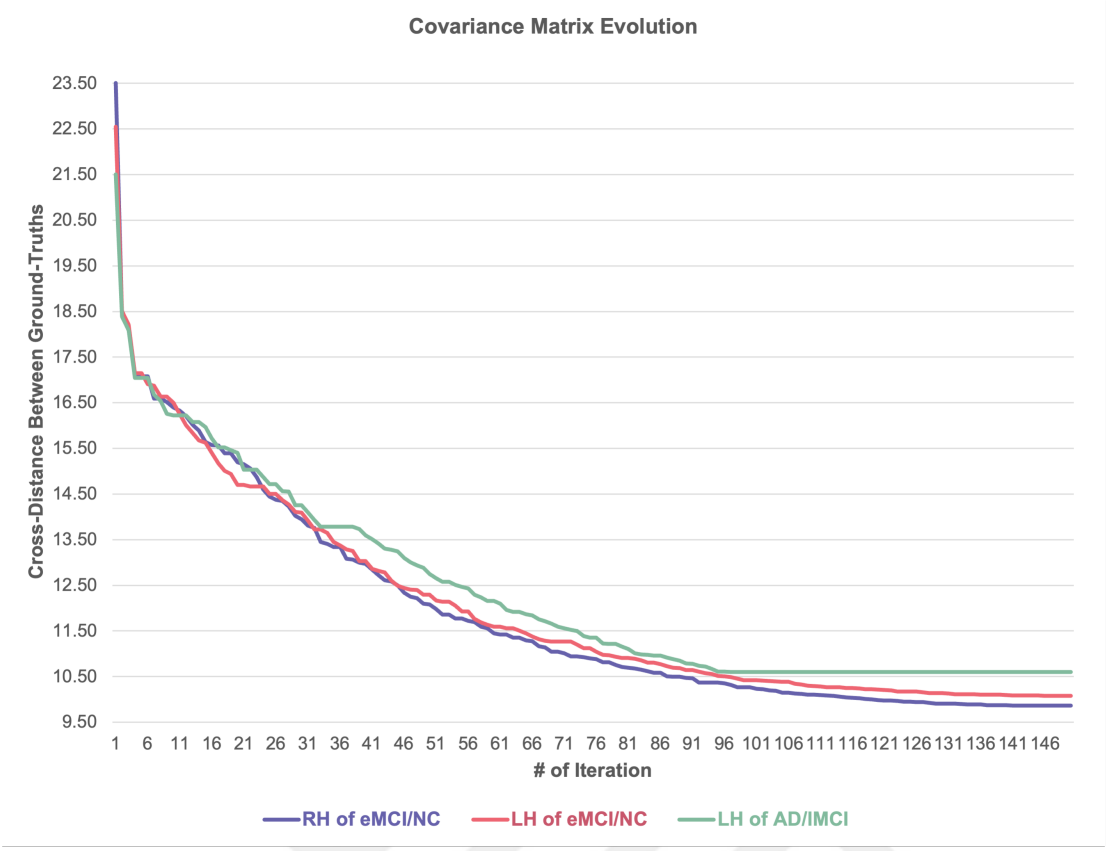

Figure 7.3 : Graph of cross-distance between ground-truth and synthetic samples that are proliferated by GEM-VIP's covariance matrix which is generated by each iteration of the genetic algorithm.

3. **Population-classifier.** An SVM classifier is trained using PCA reduced features of n ground-truth **and** $3n$ synthetic samples proliferated by leveraging the *population covariance matrix* (i.e., derived from the population).

4. **Ours.** We train an SVM classifier on using PCA on n ground-truth **and** $3n$ synthetic samples generated by utilizing the *GEM-VIP covariance matrix* which is obtained by applying GEM-VIP's GA approaches.

Figure 7.4 shows the accuracy of each classifier. Overall, our proposed method's covariance matrix yielded not only the best outcome but also outperformed the ground-truth driven *population covariance matrix*.

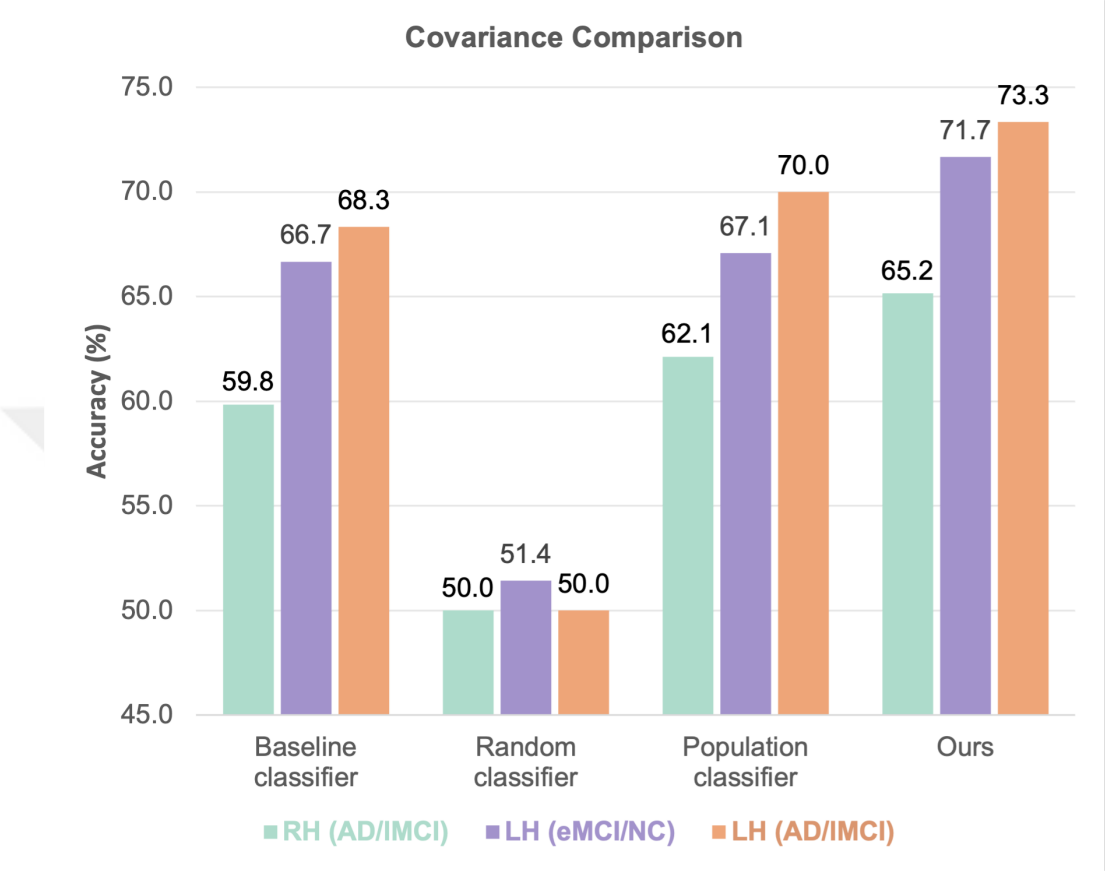

Figure 7.4 : Accuracy comparisons of the eMCI/NC and AD/lMCI classifiers trained on four different dataset. *First* classifiers are trained on only n ground-truth samples without any synthetic samples. *Second* ones are trained on both n ground-truth and $3n$ synthetic samples that is proliferated by using *random covariance matrix*. *Third* classifiers are trained on both n ground-truth and $3n$ synthetic samples that is proliferated by using *population covariance matrix*. *Fourth* classifiers are trained on both n ground-truth and $3n$ synthetic samples that is proliferated by proposed *GEM-VIP covariance matrix*. Although the synthetic samples that is generated by population covariance matrix have high accuracy than baseline model, proposed method achieved the highest accuracy. On the other hand, the synthetic samples that are generated by random covariance matrix decrease the accuracy of the classifiers.

8. INSIGHTS, LIMITATIONS AND FUTURE WORKS

We focused on insights, limitations, and future work in this chapter. To begin, we discussed the insights of our study. Then, we laid out the limitations and possible solutions to these limitations as future works.

8.1 Insights

In four distinct perspectives, we showed the GEM-VIP method's insights. To evaluate the influence of our GEM-VIP method on classification accuracy, we first analyzed our classifier that was trained using GEM-VIP's synthetic samples with other classifiers. Second, we demonstrated how the genetic algorithm evolved while converging on a reasonable covariance matrix. Third, we used the t-SNE distribution to show the dispersion of synthetic samples. Finally, we compared our covariance matrix by other covariance matrices using classifiers (i.e., that are trained on synthetic samples that are generated by these covariance matrices).

8.1.1 Insights into baseline method and data proliferators

As shown in Table 7.4, SMOTE-inspired classifier improved the accuracy for LH of eMCI/NC dataset. However, it decreased the accuracy for RH of eMCI/NC and LH of AD/lMCI dataset. This result can be explained by the fact that SMOTE uses predefined distances (e.g., Euclidean Distance), to find the nearest neighbour samples to generate synthetic sample. Therefore, it can cause noisy synthetic samples harming the overall classification accuracy. MV-LEAP-inspired classifier increases the accuracy for all datasets. This observation can be justified by the fact that, unlike SMOTE, MV-LEAP uses similarity matrix, generated by multi-kernel manifold learning, to find the nearest neighbours. However, MV-LEAP proliferates each data view independently and loses the information between the views. Yet, our proposed GEM-VIP proliferates synthetic samples using the information amid the views, and achieved the highest accuracy for all datasets.

8.1.2 Insights into genetic algorithm

To find the plausible covariance matrix of the population we use the cross-distance (i.e., as fitness function) between CBT and synthetic samples that is proliferated by using the covariance matrix which is generated after each iteration of the genetic algorithm. After each iteration we proliferate synthetic samples and measures the distance between the CBT. Considering the result of each iteration of the genetic algorithm in Table 7.3, the cross-distance is decreasing and GA converges the plausible covariance matrix.

8.1.3 Insights into dispersion of synthetic samples generated by GEM-VIP

To analyze the centeredness and representativeness of the synthetic samples, we use t-SNE distribution of the synthetic samples and ground-truths. The results shown in Figure 7.2 can emphasize that the dispersion of synthetic samples are well-distributed thanks to the ability to be clustered within their own classes and separated from other classes.

8.1.4 Insights into classifiers trained by different covariance matrices' synthetics samples

From Table 8.1 and Figure 7.4, we can make the following observations. *First*, randomly generated covariance matrix contribute to a lower classification accuracy. Such observation can justified by the fact that synthetics samples may not representative for their population because the classification accuracy with limited data has higher accuracy. *Second*, population covariance matrix increased the accuracy of each dataset, and it can be explained by the fact that population derived covariance matrix represents the population and generates well-dispersed synthetic samples. *Last*, GEM-VIP covariance matrix achieved the best results for each datasets, because genetic algorithm converges to an *optimal* covariance matrix of the population, even it is better than population derived matrices.

8.2 Limitations and Future Works

Although we have good results, GEM-VIP has a few limitations.

Table 8.1 : Accuracy (%) results of baseline, random, population and our classifier. LH: left-hemisphere. RH: right-hemisphere. GT: ground-truth. N: number of samples.

	LH eMCI/NC(%)	RH eMCI/NC(%)	LH AD/lMCI(%)
Baseline-classifier	68.33	66.67	59.84
Random-classifier	50.00	52.42	50.00
Population-classifier	70.00	67.08	62.12
Ours	**73.33**	**71.67**	**65.15**

$$n_f = \frac{(n_r - 1)n_r}{2}, \tag{8.1}$$

1. *First*, multi-view connectomic data has 35 ROIs (n_r) and 595 (n_f) features for each view, and GEM-VIP obtains a representative vector for each features. However, we obtain only one covariance matrix that represents all features because the execution time for the genetic algorithm (GA) is computationally expensive. In future works, we aim to obtain more dedicated covariance matrix for each feature individually.We assume that a better covariance matrix can contribute to a more centered and representative synthetic features.

2. Above-mentioned *second limitation* is the execution time of GA to obtain the covariance matrix. While executing GA, a high performance (i.e., faster) method can be developed. Additionally, the most time consuming step (i.e., conversion of a simple matrix to it's nearest symmetric positive semi-definite matrix) of the genetic algorithm would be changed by a faster method.

3. *Third*, mutation and crossover steps of genetic algorithm would be changed by other methods to increase genetic diversity for converging better covariance matrix.

9. CONCLUSIONS

In this research, we proposed GEM-VIP method for boosting classification performance of multi-view connectomic data by generating synthetic samples for given classes. It compromises three major steps. *First*, we propose to use the population-representative tensor as the mathematically regarded mean (μ_r) of the population [1]. *Second step*, we aim to obtain plausible covariance matrix between different views of the population. *Third step* generates synthetic samples by using the earlier obtained μ_r and covariance matrix in first and second steps based on multivariate normal distribution. Overall, our method boosted the classification result for LH of AD/lMCI and LH/RH of eMCI/NC datasets by proliferating synthetic samples. In our future work, we aim to obtain more dedicated covariance matrix for each features instead of using one covariance matrix for all features. Moreover, we can generate better covariance matrix by increase the genetic diversity of GA, and increase the performance by implementing faster conversion of nearest symmetric positive semi-definite matrix.

REFERENCES

[1] **Dhifallah, S.**, **Rekik, I.**, **Initiative, A.D.N.** *et al.* (2020). Estimation of connectional brain templates using selective multi-view network normalization, *Medical Image Analysis*, *59*, 101567.

[2] **Organization, W.H.** *et al.* (2019). Risk reduction of cognitive decline and dementia: WHO guidelines.

[3] **Mitchell, A.J. and Shiri-Feshki, M.** (2009). Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies, *Acta Psychiatrica Scandinavica*, *119*(4), 252–265.

[4] **Higdon, R.**, **Earl, R.K.**, **Stanberry, L.**, **Hudac, C.M.**, **Montague, E.**, **Stewart, E.**, **Janko, I.**, **Choiniere, J.**, **Broomall, W.**, **Kolker, N.** *et al.* (2015). The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders, *Omics: a journal of integrative biology*, *19*(4), 197–208.

[5] **Atan, O.**, **Jordon, J. and van der Schaar, M.** (2018). Deep-Treat: Learning Optimal Personalized Treatments From Observational Data Using Neural Networks., *AAAI*, pp.2071–2078.

[6] **Owens, C. and Irwin, M.** (2012). Neuroblastoma: the impact of biology and cooperation leading to personalized treatments, *Critical reviews in clinical laboratory sciences*, *49*(3), 85–115.

[7] **Arnedos, M.**, **Soria, J.C.**, **Andre, F. and Tursz, T.** (2014). Personalized treatments of cancer patients: a reality in daily practice, a costly dream or a shared vision of the future from the oncology community?, *Cancer treatment reviews*, *40*(10), 1192–1198.

[8] **Rossi, G.**, **Pelosi, G.**, **Graziano, P.**, **Barbareschi, M. and Papotti, M.** (2009). A reevaluation of the clinical significance of histological subtyping of non—small-cell lung carcinoma: diagnostic algorithms in the era of personalized treatments, *International journal of surgical pathology*, *17*(3), 206–218.

[9] **Kononenko, I.** (2001). Machine learning for medical diagnosis: history, state of the art and perspective, *Artificial Intelligence in medicine*, *23*(1), 89–109.

[10] **Zhao, J.**, **Xie, X.**, **Xu, X. and Sun, S.** (2017). Multi-view learning overview: Recent progress and new challenges, *Information Fusion*, *38*, 43–54.

[11] **Li, Y.**, **Yang, M. and Zhang, Z.M.** (2018). A Survey of Multi-View Representation Learning, *IEEE Transactions on Knowledge and Data Engineering*.

[12] **Sun, S.** (2013). A survey of multi-view machine learning, *Neural Computing and Applications*, *23*(7-8), 2031–2038.

[13] **Wu, F.**, **Jing, X.Y.**, **You, X.**, **Yue, D.**, **Hu, R. and Yang, J.Y.** (2016). Multi-view low-rank dictionary learning for image classification, *Pattern Recognition*, *50*, 143–154.

[14] **Luo, Y.**, **Liu, T.**, **Tao, D. and Xu, C.** (2015). Multiview matrix completion for multilabel image classification, *IEEE Transactions on Image Processing*, *24*(8), 2355–2368.

[15] **Yang, M.**, **Deng, C. and Nie, F.** (2019). Adaptive-Weighting discriminative regression for multi-view classification, *Pattern Recognition*, *88*, 236–245.

[16] **Soussia, M. and Rekik, I.** (2017). High-order connectomic manifold learning for autistic brain state identification, *International Workshop on Connectomics in Neuroimaging*, 51–59.

[17] **Soussia, M. and Rekik, I.** (2018). Unsupervised Manifold Learning Using High-Order Morphological Brain Networks Derived From T1-w MRI for Autism Diagnosis, *Frontiers in Neuroinformatics*, *12*.

[18] **Dhifallah, S.**, **Rekik, I.**, **Initiative, A.D.N.** *et al.* (2019). Clustering-based multi-view network fusion for estimating brain network atlases of healthy and disordered populations, *Journal of neuroscience methods*, *311*, 426–435.

[19] **Baron-Cohen, S.**, **Ring, H.A.**, **Wheelwright, S.**, **Bullmore, E.T.**, **Brammer, M.J.**, **Simmons, A. and Williams, S.C.** (1999). Social intelligence in the normal and autistic brain: an fMRI study, *European journal of neuroscience*, *11*(6), 1891–1898.

[20] **Sporns, O.** (2013). Structure and function of complex brain networks, *Dialogues in clinical neuroscience*, *15*(3), 247.

[21] **Lisowska, A.**, **Rekik, I. and disease neuroimaging initiative (ADNI), T.A.** (2018). Joint pairing and structured mapping of convolutional brain morphological multiplexes for early dementia diagnosis, *Brain connectivity*, *9*(1), 22–36.

[22] **Mahjoub, I.**, **Mahjoub, M.A. and Rekik, I.** (2018). Brain multiplexes reveal morphological connectional biomarkers fingerprinting late brain dementia states, *Scientific reports*, *8*(1), 4103.

[23] **Lisowska, A.**, **Rekik, I.**, **Initiative, A.D.N.** *et al.* (2017). Pairing-based ensemble classifier learning using convolutional brain multiplexes and multi-view brain networks for early dementia diagnosis, *International Workshop on Connectomics in Neuroimaging*, Springer, pp.42–50.

[24] **Raeper, R.**, **Lisowska, A. and Rekik, I.** (2018). Cooperative Correlational and Discriminative Ensemble Classifier Learning for Early Dementia

Diagnosis Using Morphological Brain Multiplexes, *IEEE Access*, *6*, 43830–43839.

[25] **Nebli, A. and Rekik, I.** (2019). Gender differences in cortical morphological networks, *Brain Imaging and Behavior*.

[26] **Chawla, N.V.**, **Bowyer, K.W.**, **Hall, L.O. and Kegelmeyer, W.P.** (2002). SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, *16*, 321–357.

[27] **Sáez, J.A.**, **Luengo, J.**, **Stefanowski, J. and Herrera, F.** (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, *Information Sciences*, *291*, 184–203.

[28] **Graa, O. and Rekik, I.** (2019). Multi-view learning-based data proliferator for boosting classification using highly imbalanced classes, *Journal of neuroscience methods*, *327*, 108344.

[29] **Rekik, I.**, **Li, G.**, **Lin, W. and Shen, D.** (2017). Estimation of brain network atlases using diffusive-shrinking graphs: application to developing brains, *International conference on information processing in medical imaging*, Springer, pp.385–397.

[30] **Thengade, A. and Dondal, R.** (2012). Genetic algorithm-survey paper, *MPGI National Multi Conference*, Citeseer, pp.7–8.

[31] **Shijie, J.**, **Ping, W.**, **Peiyi, J. and Siping, H.** (2017). Research on data augmentation for image classification based on convolution neural networks, *2017 Chinese automation congress (CAC)*, IEEE, pp.4165–4170.

[32] **Chawla, N.V.**, **Lazarevic, A.**, **Hall, L.O. and Bowyer, K.W.** (2003). SMOTEBoost: Improving prediction of the minority class in boosting, *European conference on principles of data mining and knowledge discovery*, Springer, pp.107–119.

[33] **Han, H.**, **Wang, W.Y. and Mao, B.H.** (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *International conference on intelligent computing*, Springer, pp.878–887.

[34] **Ma, L. and Fan, S.** (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests, *BMC bioinformatics*, *18*(1), 169.

[35] **Li, C.**, **Wei, Y.**, **Chen, X. and Schönlieb, C.B.**, (2021). BrainNetGAN: Data augmentation of brain connectivity using generative adversarial network for dementia classification, Deep Generative Models, and Data Augmentation, Labelling, and Imperfections, Springer, pp.103–111.

[36] **Guvercin, U.**, **Gharsallaoui, M.A. and Rekik, I.** (2021). One Representative-Shot Learning Using a Population-Driven Template

with Application to Brain Connectivity Classification and Evolution Prediction, *International Workshop on PRedictive Intelligence In MEdicine*, Springer, pp.25–36.

[37] **Chawla, N.V.**, **Japkowicz, N. and Kotcz, A.** (2004). Special issue on learning from imbalanced data sets, *ACM Sigkdd Explorations Newsletter*, *6*(1), 1–6.

[38] **Sun, Y.**, **Wong, A.K. and Kamel, M.S.** (2009). Classification of imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719.

[39] **Liu, X.Y.**, **Wu, J. and Zhou, Z.H.** (2009). Exploratory undersampling for class-imbalance learning, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550.

[40] **Wang, B.**, **Ramazzotti, D.**, **De Sano, L.**, **Zhu, J.**, **Pierson, E. and Batzoglou, S.** (2017). SIMLR: a tool for large-scale single-cell analysis by multi-kernel learning, *bioRxiv*, 118901.

[41] **Mitchell, M.** (1996). An introduction to genetic algorithms mit press, *Cambridge, Massachusetts. London, England*, *1996*.

[42] **Johnson, R. and Wichern, D.** The multivariate normal distribution, *Applied Multivariate Statistical Analysis. Englewood Cliffs, NJ: Prentice-Hall Inc*, 150–173.

[43] **Higham, N.J.** (1988). Computing a nearest symmetric positive semidefinite matrix, *Linear algebra and its applications*, *103*, 103–118.

[44] **Abdi, H. and Williams, L.J.** (2010). Principal component analysis, *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

[45] **Fischl, B.** (2012). FreeSurfer, *Neuroimage*, *62*(2), 774–781.

[46] **Dale, A.M.**, **Fischl, B. and Sereno, M.I.** (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction, *Neuroimage*, *9*(2), 179–194.

[47] **Desikan, R.S.**, **Ségonne, F.**, **Fischl, B.**, **Quinn, B.T.**, **Dickerson, B.C.**, **Blacker, D.**, **Buckner, R.L.**, **Dale, A.M.**, **Maguire, R.P.**, **Hyman, B.T.** *et al.* (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, *Neuroimage*, *31*(3), 968–980.

[48] **Gillmore, R.**, **Stuart, S.**, **Kirkwood, A.**, **Hameeduddin, A.**, **Woodward, N.**, **Burroughs, A.K. and Meyer, T.** (2011). EASL and mRECIST responses are independent prognostic factors for survival in hepatocellular cancer patients treated with transarterial embolization, *Journal of hepatology*, *55*(6), 1309–1316.

[49] **Holmes, A.J.**, **Hollinshead, M.O.**, **O'keefe, T.M.**, **Petrov, V.I.**, **Fariello, G.R.**, **Wald, L.L.**, **Fischl, B.**, **Rosen, B.R.**, **Mair, R.W.**, **Roffman, J.L.** *et al.* (2015). Brain Genomics Superstruct Project initial data release with

structural, functional, and behavioral measures, *Scientific data*, *2*(1), 1–16.

[50] **Mueller, S.G.**, **Weiner, M.W.**, **Thal, L.J.**, **Petersen, R.C.**, **Jack, C.**, **Jagust, W.**, **Trojanowski, J.Q.**, **Toga, A.W. and Beckett, L.** (2005). The Alzheimer's disease neuroimaging initiative, *Neuroimaging Clinics*, *15*(4), 869–877.

[51] **Mueller, S.**, **Weiner, M.**, **Thal, L.**, **Petersen, R.**, **Jack, C.**, **Jagust, W.**, **Trojanowski, J.**, **Toga, A. and Beckett, L.** (2005). Neuroimaging Clin, *North Am*, *15*(4), 869–877.

[52] **Van der Maaten, L. and Hinton, G.** (2008). Visualizing data using t-SNE., *Journal of machine learning research*, *9*(11).

[53] **Li, W.**, **Cerise, J.E.**, **Yang, Y. and Han, H.** (2017). Application of t-SNE to human genetic data, *Journal of bioinformatics and computational biology*, *15*(04), 1750017.

APPENDICES

APPENDIX A.1 : The main figure that describes the each steps of the proposed GEM-VIP method.

APPENDIX A.1

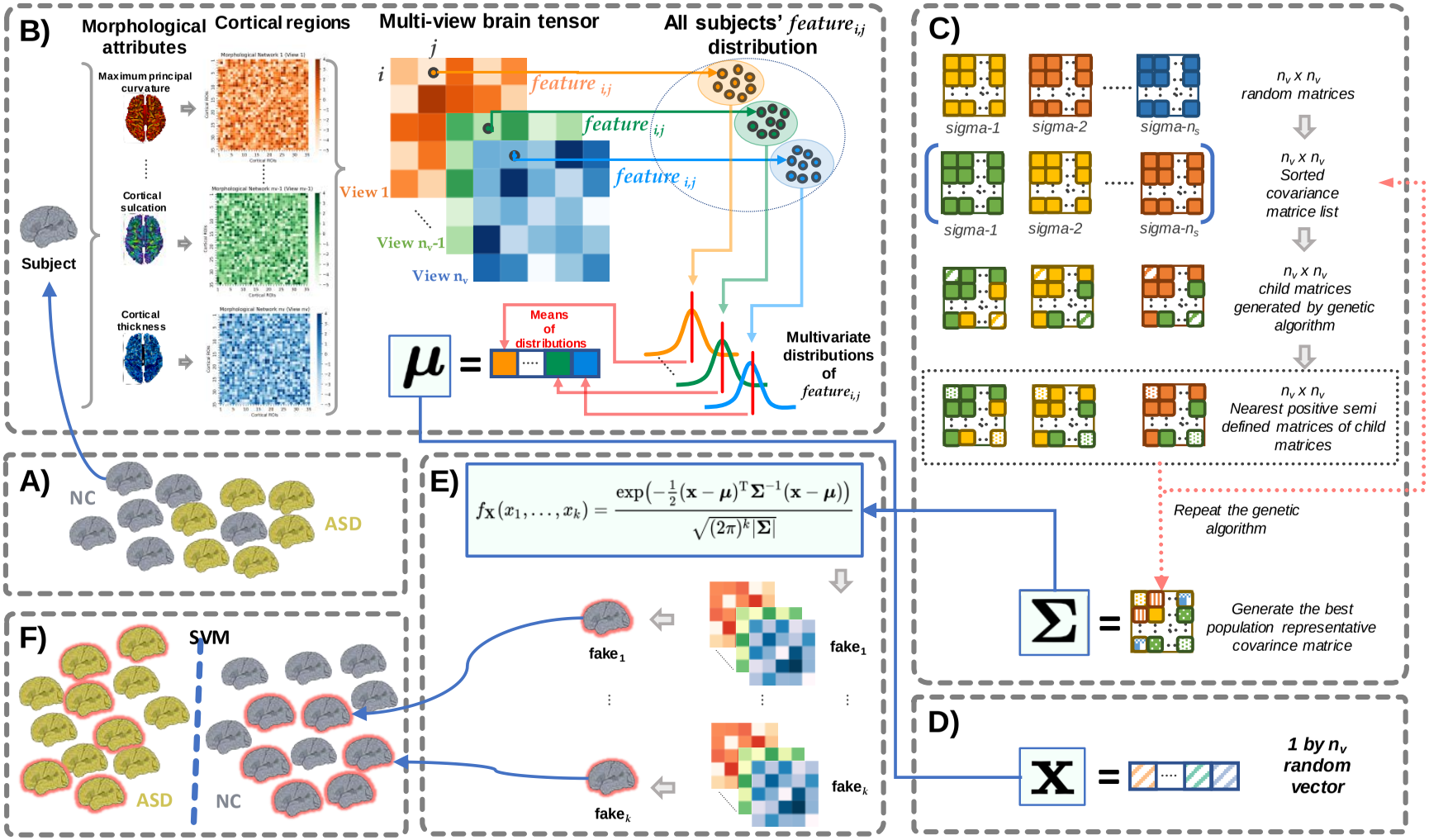

Figure A.1 : Illustration of the proposed GEM-VIP method.

(A) shows the ground truth subjects of the given populations. (B) describes the key step of obtaining the representative tensor (μ_r) of the population. We obtain the μ_r which captures the most frequent cross-view feature vector across each subjects' multi-view brain tensor, and can be used as a mathematically regarded representative mean of population. (C) represents steps of the genetic algorithm which obtains the covariance matrix of the population. Firstly, it generates random matrices as initial covariance matrices. Then, sorts the matrices by using the *fitness function*, and selects the best matrices as parent matrices. Next, children matrices are generated by applying mutation and crossover. After that, the nearest positive semi-definite matrices of each children matrices are calculated. Next, these semi-definite matrices are used as parent matrices for the next generation (i.e., iteration) of the genetic algorithm. Same steps repeat until the fitness function converges a plausible value. (D) a random vector is generated in order to provide diversity for data proliferation. (E) The obtained μ_r, covariance matrices, and the random vector are used to proliferate synthetic samples. (F) Proliferated synthetic samples are added to the ground-truth samples and used as training samples for an SVM classifier in order to boost the classification performance.

CURRICULUM VITAE

Name Surname : Mustafa ÇELİK

EDUCATION :

- **B.Sc.** : 2013, Fatih University, Faculty of Engineering, Department of Computer Engineering
- **B.Sc.** : 2013, Fatih University, Faculty of Engineering, Department of Electrical and Electronics Engineering

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Celik M., Rekik I.** 2022. Data Augmentation of Morphological Brain Network Using Multi-variate Normal Distribution. *International Conference on Computer Science and Engineering,* September 14–16, 2022 Diyarbakır, Turkey.