Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Metrik Öğrenmesi Kullanarak Çeşitli Kanser Dokularına Ait Mikro Dizi Gen Verilerinin Sınıflandırılması

Fırat İSMAİLOĞLU [a,*]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Sivas Cumhuriyet Üniversitesi, Sivas, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: fismailoglu@cumhuriyet.edu.tr*
DOI: 10.29130/dubited.886353

ÖZ

Kanserli dokuların heterojen doğası gereği birçok kanserin alt türü vardır, ve bu alt türler tespit edilmedikçe kanser tedavisi hedefi bulamaz. Mikrodizi gen teknolojisi ve veri teknolojisinin gelişmesiyle beraber, son yıllarda kanserli dokulara ait mikro dizi gen ifadesi verilerini kullanarak makine öğrenmesi yardımıyla kanserlerin alt türünü tespit etmek yaygınlaşmıştır. Fakat burada asıl problem, veri setinde her bir gene bir özniteliğin karşılık gelmesi, bu yüzden yüksek boyut probleminin ortaya çıkmasıdır. Bu çalışmada üç farklı metrik öğrenmesi metodu (LMNN, ITML ve NCA) ayrı ayrı kullanılarak çeşitli kanser türlerine ait mikro dizi gen veri setleri boyutu azaltılmış uzaylara transfer edilmiştir. Bu sayede, PCA gibi klasik boyut azaltma yöntemlerinden farklı olarak boyutu azaltılmış uzayda, aynı sınıfa (kanser alt türüne) ait örnekleri birbirine yaklaştırılırken, farklı sınıflara ait örnekleri birbirinden uzaklaştırılmıştır. t-SNE metodu yardımıyla azaltılmış boyutlu uzaylar görüntülenerek sınıfların birbirinden ayrıştığı teyit edilmiştir. İlaveten, bu yeni uzaylarda sınıflama algoritmalarının daha performanslı çalıştığını göstermek amacıyla, *k*-NN, en yakın merkez ve LVQ gibi örnek temelli (instance-based) sınıflama algoritmaları çalıştırılmış ve bu algoritmaların kanser türlerini tespit etmede orjinal uzaydaki performanslarına göre yaklaşık %30'a kadar performanslarının arttığı gözlemlenmiştir.

Anahtar Kelimeler: *Kanser Sınıflandırma, Metrik Öğrenmesi, Mikro Dizi Gen Verisi, Örnek Temelli Sınıflama*

Classifying Microarray Gene Data of Various Cancerous Tissues Using Metric Learning

ABSTRACT

Due to the heterogeneous structure of cancerous tissues, they have several subclasses. Unless the subclasses are detected, the cancer treatment cannot be carried out accurately. With the advent of microarray gene technology and data science technology, employing machine learning methods that use the microarray gene expression data of the cancerous tissues for classifying the cancer subclasses has gained an increasing popularity. However, as there exists one feature for each gene, the issue of the curse of dimensionality arises. In the present study, the microarray gene expression data of various cancer types were transferred to some dimensionality reduced spaces by the means of three metric learning methods: LMNN, ITML and NCA. As a result, the instances of the same classes come closer in the reduced space; while those from different classes locate far from each other, which is different from the conventional dimensionality reduction methods, such as PCA, do. To verify this, dimensionality reduced spaces created by the t-SNE method were monitored. Additionally, to show that the classification algorithms perform better in such new spaces, instance-based classifiers, e.g. *k*-NN, the nearest mean classifier and the LVQ, were built and then it was observed that the performances of the classifiers increased up to 30% in comparison with their performances in the original space.

Keywords: *Cancer Classification,Metric Learning, Microarray Gene Expressions, Instance-based Classification*

Geliş: 24/02/2021, Düzeltme: 22/05/2021, Kabul: 28/05/2021

I. GİRİŞ

Dünyada en fazla ölüme yol açan hastalıklar içerisinde ilk sıralarda yer alan kanserin tedavisinde, kanser tanısı almış kişiye en uygun tedavi programının uygulanabilmesi için tespit edilen kanserin alt türünü de doğru olarak tespit edebilmek gerekir [1]. Çünkü kanserli dokular genelde oldukça heterojen yapılara sahiptirler; bu yüzden her bir doku türünün tedaviye vereceği cevap birbirinden farklı olmaktadır. Öte yandan, klasik patolojik testlere ek olarak kanserli hücrelere karşılık gelen gen ifadelerine ait verilerin yapay zeka/makine öğrenmesi metotları ile sınıflandırılmasının önemi veri depolama, veri işleme ve biyoteknoloji alanındaki gelişmelere paralel olarak giderek artmaktadır [1,2]. Hatta bu yöntem, subjektif görülerden uzak, otomatik ve standart olması bakımından konvansiyonel kanser sınıflandırması yöntemlerine göre daha sık kullanılmaktadır [2].

Tek bir mikro dizi deneyinde binlerce hatta on binlerce gen aynı anda analiz edildiğinden, mikro dizi gen verileri çok yüksek boyutludur. Makine öğrenmesi bakımından bu, şu problemleri ortaya çıkarır: yüksek boyutlu uzaylarda örnekler birbirinden uzaklaşır, uzaklık-yakınlık kavramları bulanıklaşır; böylece sınıflandırma algoritmalarının başarıları ciddi anlamda azalmaya başlar [3]. Literatürde mikro dizi gen verilerinin boyutunu veri madenciliği/makine öğrenmesi alanındaki alışılagelmiş özellik seçimi yönetemleri ile azaltma çalışmaları mevcuttur [3,4,5]. Fakat bu çalışmalar genelde ya filtreleme temelli olup, her defasında yalnızca bir adet geni (özelliği) diğer özelliklerden bağımısız olarak ele alır; ya da sarmalama (wrapping) temelli olup özellik kümelerini bir arada ele alarak bu kümeler üzerine kurulan sınıflayıcıların performansından etkilenir. Buda, özellik yani gen seçiminin kullanılan sınıflama algoritmasına bağlı olmasına neden olur [4].

Bir makine öğrenmesi dalı olan "*metrik öğrenmesi*" eldeki veriye özel olarak birbirine anlamsal olarak benzer olan veri örneklerini bir araya getirirken, farklı örnekleri birbirinden uzakta tutmaya çalışan uzaklık metriği öğrenmeye çalışır [6]. Öğrenilen bu metrik, hem boyut azaltma (dimensionality reduction) anlamında, hem de direkt bir sınıflama ya da gruplama algoritması içerisinde kullanarak faydalanılabilir; böylece bu algoritmaların performansı gözle görülür şekilde artmaya başlar.

Bu çalışmada farklı kanser türlerine ait mikro dizi gen verilerinin boyutunun klasik filtreleme ve sarmalama özellik seçimi yöntemlerinden farklı olarak metrik öğrenmesi metotları yardımıyla azaltılması, indirgenmiş boyutlu uzaylarda makine öğrenmesi sınıflama algoritmaları kullanarak ilgilenilen kanserlerin alt türlerini tahmin edilmesi çalışılmıştır. Burada metrik öğrenmesi ile elde edilen temel kazanç, boyutu indirgenmiş uzayda aynı sınıfa (aynı kanser alt türüne) ait örneklerin birbirine yakın olması; öte yandan farklı sınıflara ait örneklerin birbirine uzak olmasıdır. Bu, klasik özellik seçimi yöntemlerinin sağlayamayacağı bir avantajdır. Ayrıca sarmalama yönteminin aksine, boyut azaltımı için metrik öğrenmesinin kullanılması, boyut azaltımının herhangi bir sınflayıcıya bağlı olmasını da engeller. Böylece boyutun azaldığı ve sınıfların birbirinden ayrıldığı uzayda, makine öğrenmesi uygulayıcısı tercih ettiği herhangi bir sınıflama algoritmasını kullanabilir. Yinede bu çalışmada sınıfların birbirinden ayrı olmasına ihtiyaç duyan *k*-en yakın komşu, LVQ gibi örnek temelli (instance-based) sınıflama algoritmaları, metrik öğrenmesiyle boyutu azaltılmış uzayda kullanılmış, bu algoritmaların sahip olduğu başarıların ciddi oranda (%30'a kadar) arttığı gözlenmiştir.

Makalenin geri kalanında, önce kanser mikro dizi gen verilerinin sınıflandırılması ve bu verilerin boyutlarının azaltılması ile ilgili benzer çalışmaları içeren literatür taraması bölümü verilmiştir. Bölüm 3'te metrik öğrenmesi ile ilgili genel bilgilerden ve bu çalışmada kullanılan metrik öğrenmesi metotlarından bahsedilmiştir. Bölüm 4'te ise üzerinde çalışılan mikro dizi kanser gen veri setlerinin bilgileri ve ilgilenilen metrik öğrenmesi metotlarının bu veri setleri üzerindeki uygulamalarının sonuçları verilmiştir. Bölüm 5 ile çalışma özetelenerek ortaya konulan sonuçlar sunulmuştur.

II. LİTERATÜR ÇALIŞMASI

Ekim 1990'da başlayıp, Nisan 2003'te sonlanan İnsan Genom Projesi (Human Genome Project) insan ırkına ait bütün genetik izleri ortaya çıkarmaya çalışmış, bu anlamda yaklaşık 21,000 genden oluşan

insan genomunu elde etmiştir. Böylece insan genomuna ait tüm DNA dizileri açığa çıkmıştır. Buna paralel olarak, gelişen mikro dizi teknolojisiyle mikroskopik (çapı 250 mikrondan az) DNA spotları cam ya da silikon bir çip üzerine sistematik bir şekilde dizilerek DNA mikro dizi oluşturulabilmiş, bu sayede hem binlerce genin ifadesi ölçülebilmiş, hemde bu genlerin birbirleriyle etkileşimini anlayabilmek mümkün hale gelmiştir [7]. Bu şekilde ortaya çıkan mikro dizi gen verileri makine öğrenmesi dahil birçok farklı disiplin tarafından çalışılmaya başlanmıştır [7,8].

Çeşitli kanser dokularına ait mikro dizi gen verilerini makine öğrenmesi/veri madenciliği alanlarına ait gruplama (clustering) algoritmalarıyla anlamakla ve sınıflama (classification) algoritmalarıyla kanser alt türlerini tahmin etmekle ilgili çalışmalar İnsan Genom Projesi ile birlikte başlamış ve günümüze kadar kesintisiz olarak gelmiştir [8,4]. Dahası, son yıllardaki veri depolama ve veri işleme teknolojilerindeki ilerleme ve bollaşma sayesinde bu çalışmalar her zamankinden daha popüler hale gelmiştir.

Ulusal ve uluslararası birçok çalışmada, farklı kanser türleri, makine öğrenmesinin sınıflama algoritmalarının bu kanserli dokulara ait mikro dizi gen verileri ile eğitilmesiyle sınıflandırılmıştır. Örneğin Kılıçarslan ve ark. [8] Destek Vektör Makineleri ve *k*-en yakın komşu algoritmalarını prostat kanserine ait mikro dizi gen verileri ile eğiterek prostat kanserini sınıflamaya çalışmış; Haznedar ve ark., [9] ise Ağ Tabanlı Bulanık Mantık Çıkarım Sistemi (ANFIS) ve Genetik Algoritma (GA) algoritmalarını birleştirerek oluşturdukları hibrit algoritma ile karaciğer kanserinin alt türlerini sınıflandırmaya çalışmıştır. Uluslararası çapta ise öne çıkan çalışmalardan biri Dwivedi [2] 'nin yapay sinir ağlarını lösemili dokulardan alınmış mikro dizi gen verileri üzerinde eğiterek löseminin iki türü olan akut lenfoblastik lösemi ile akut miyeloid lösemiyi efektif olarak ayrıştırmasıdır. Yine çok yakın bir zamanda Morais-Rodrigues ve ark. [10], meme kanseri hastalarından toplanmış mikro dizi gen verileri ile lojistik regresyon sınıflyacısını eğitmiş, bu sayede %80'den fazla başarı ile meme kanserinin alt türlerini tahmin edebilmiştir.

Giriş bölümünde de değinildiği gibi, mikro dizi gen verileri kanserle ilişkili ya da ilişkisiz binlerce gene ait bilgiyi tutarlar. Bu, makine öğrenmesi bakımından binlerce özellik (öznitelik-feature) anlamına gelir; bu da kanserli doku örneklerine karşılık gelen özellik vektörlerinin boyutlarının binlerle ifade edilmesine neden olur. Aşırı uyum (overfitting) ile birlikte makine öğrenmesinin en önemli sorunlarından biri olan yüksek boyut problemi ortaya çıkar. Bu nedenle, mikro dizi gen verilerini kullanarak kanser sınıflandırması çalışmaları kadar, literatürde bu verilerin boyutunu azaltma (dimensionality reduction) ile ilgili çalışmalar da aktif bir çalışma alanıdır. Örneğin Yıldız ve ark. [11], Fisher korelasyon analizi, t-Skor ve Welch'in t-istastiği özellik filtreleme yöntemlerini ayrı ayrı kullanarak, her bir yöntem ile 100 adet gen seçmiş, daha sonra bu genleri birleştirerek oluşturduğu özellik vektörleri ile bir SVM sınıflayıcısını eğiterek meme kanserini sınıflandırmaya çalışmıştır. Bununla birlikte sarmalama (wrapping) yöntemi ile de kanser ile ilgili genlerin alt kümelerinin seçimi yaygındır [12]. Sarmalama yöntemi özellikleri (genleri) birer birer değil alt kümeler halinde birlikte ele aldığından özellikler (genler) arasındaki etkileşimi de dikkate almış olur. Bu haliyle filtreleme yöntemlerine göre daha gelişmiş sayılabilirken; özelliklerin bütün alt kümelerini incelemesi bakımından hesaplama anlamında pahalı, hemde ele aldığı alt küme üzerinde bir sınıflandırıcı eğiterek bu alt kümenin işe yararlılığını test ettiğinden, ortaya çıkan sonuç kullanılan sınıflandırma algoritmasına bağımlı olur. Yinede hem filtreleme hemde sarmalama yöntemleri kanserle ilişkili genlerin seçimi için aktif olarak kullanılmaya devam etmektedir. İlgili okuyuculara kanser sınıflandırma amacı için gen seçimi çalışmalarını derleyen [3] ve [4] kaynakları tavsiye edilir.

III. METRİK ÖĞRENMESİ

Metrik öğrenmesi, ya da uzaklık metrik öğrenmesi, makine öğrenmesi alanında verilen bir göreve özel olarak veriden bir uzaklık metriği öğrenmeyi amaçlar [5, 15]. Öğrenilen bu metrik, anlamsal olarak yakın örnekler arası uzaklığı az, anlamsal olarak uzak olan örnekler arası uzaklığı fazla ölçer. Burada bahsedilen anlamsal olarak yakınlık kavramı verilen göreve göre değişkenlik gösterir. Örneğin bir

sınıflandırma probleminde aynı sınıfa ait örnekler anlamsal olarak yakın olarak düşünülür; şu halde bir sınıflandırma problemi için öğrenilen bir metrik aynı sınıflara ait örnekler arası uzaklığı az, farklı sınıflara ait örnekler arası uzaklıkları fazla olarak ölçmesi beklenir.

Yukarıda verilen örnek, metrik öğrenmesinin denetimli (supervised) öğrenmedeki kullanımına bir örnektir. Metrik öğrenmesi aynı zamanda denetimsiz (unsupervised) öğrenmede de kendine sıkça kullanım alanları bulur [13]. Örneğin bir gruplama (clustering) probleminde, belirli örnek çiftlerinin aynı grup içerisinde olması ya da olmaması bilgisi ön bilgi olarak verilebilir. Bu durumda, gruplama algoritmasında kullanılacak metriğin, beraber olması şartı koşulan örnekleri aynı gruba düşürebilmesi için bu çiftler arasındaki uzaklığı minimum seviyede ölçerken, farklı gruplara düşmesi istenilen örnek çiftleri arası uzaklığı maksimum seviyede hesaplar. Böylece, metrik öğrenmesi yalnızca sınıf (etiket) bilgisini değil aynı zamanda hangi örnek çiftlerinin bir arada bulunması ya da bulunmaması şeklinde verilen zayıf bir denetim bilgisini dahi öğrenebilir. Bu, metrik öğrenmesinin temel avantajıdır [13,14].

Hem denetimli hemde denetimsiz makine öğrenmesi problemlerinde kullanılabildiğinden, metrik öğrenmesinin kullanım alanı oldukça geniştir. Bu alanlar içerisindeki en önemli kullanım alanlarından biri bilgisayarlı görme (computer vision)'dir [6]. Özellikle görsel sınıflandırma (image classification), yüz tanıma (face recognition), kişilerin poz tahmini (human pose estimation) gibi problemlerde metrik öğrenmesi başarılı bir şekilde kullanılmaktadır [15]. Ayrıca yapısal ve yapısal olmayan metinlerin analizinde ve sınıflandırılmasında metrik öğrenmesi gruplama ve sınıflandırma görevleri için başarılı sonuçlar vermektedir [15]. Bu iki kullanım alanına ek olarak, son yıllarda mikro dizi teknolojisinin gösterdiği gelişme ile birlikte bir dokuya ait binlerce hatta onbinlerce genin DNA ifadesinin ölçümlerinin kolaylaşmıştır. Aynı zamanda bulut bilgisayar teknolojilerindeki ilerlemeye paralel olarak DNA ölçümlerinin saklanması ve işlenmesi çok daha mümkün hale gelmiştir ve böylece metrik öğrenmesi doku örneklerinin içerdiği hastalıklara göre sınıflandırılmasında önemli rol oynamaya başlamıştır [6]. Bu çalışma da bu yönde olup, farklı kanser türlerinin alt türlerini metrik öğrenmesinin mikro dizi gen ifadelerine uygulanmasıyla sınıflandırmaya çalışmaktadır.

Metrik öğrenmesinin iki temel amacı vardır: *i)* örnek temelli (instance based) ya da hafıza temelli (memory based) olarak adlandırılan, test örneklerini eğitim setindeki örneklerle kıyaslayarak modelleme yapan makine öğrenmesi algoritmalarına (örneğin k-en yakın komşu, anomali algoritmaları) eklemlenerek bu algoritmaların çok daha başarılı olmalarını sağlamak, *ii)* doğrusal diskriminant analizi (linear discriminant analysis) gibi denetimli bir şekilde boyut azaltımı (dimensionality reduction) yapmak [14]. Bu çalışmada metrik öğrenmesi mikro dizi gen verileri üzerinde her iki amaç için kullanılmıştır.

Kulis [6] metrik öğrenmesini informal olarak şu şekilde formülize etmiştir. Metrik öğrenmesi, $\boldsymbol{x}_i$ ve $\boldsymbol{x}_j$ örnekleri için verilen bir uzaklık $u(\boldsymbol{x}_i, \boldsymbol{x}_j)$ fonksiyonunu (örneğin öklid uzaklığını) ve bu örnekler için verilen benzerlik ve benzemezlik bilgilerini kullanarak yeni bir uzaklık fonksiyonu tanımlar. Öyleki bu yeni uzaklık fonksiyonun, orjinale göre benzer çiftleri daha yakın, ayrı çiftleri ise daha uzak ölçmesi gerekmektedir. Daha formal olarak, $\boldsymbol{x}_i$ ve $\boldsymbol{x}_j$ elementleri reel sayılar olan d boyutlu birer vektör olmak üzere (yani $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$ iken) metrik öğrenmesi bir $M \in \mathbb{R}^{d \times d}$ pozitif yarı tanımlı (öz değerleri negatif olmayan) matrisini öğrenmeyi amaçlar öyleki, burada M aşağıdaki şekilde tanımlanan Mahalanobis uzaklığının parametresidir:

$$u_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T M (\boldsymbol{x}_i - \boldsymbol{x}_j)} \quad (1)$$

Eğer M matrisi birim matris olarak alınırsa, (1)'de tanımlanan Mahalanobis uzaklığı sıkça kullanılan öklid uzaklığına dönüşür. Ayrıca $r \leq d$ pozitif tam sayısı M matrisinin rankı olmak üzere, M matrisini $M = P^T P$ $(P \in \mathbb{R}^{r \times d})$ şeklinde ayrıştırarak (1)'de tanımlanan Mahalanobis uzaklığını, aşağıdaki şekilde de tanımlanabilir:

$$u_P(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(P\boldsymbol{x}_i - P\boldsymbol{x}_J)^T (P\boldsymbol{x}_i - P\boldsymbol{x}_J)} \quad (2)$$

(2)'de dikkat edilmesi gereken nokta orjinal $\mathbb{R}^d$ uzayında yer alan tüm $\boldsymbol{x} \in \mathbb{R}^d$ örneklerinin $P\boldsymbol{x} \in \mathbb{R}^r$ boyutu azaltılmış uzayına transfer edilmesi, ve bu uzayda klasik öklid uzaklığının hesaplanmaya çalışılmasıdır. Böylece metrik öğrenmesi problemi, orjinal uzayda yer alan örnekleri daha küçük boyutlu uzaylara transfer etme problemine dönüşmüş olur. Bu şekilde hem boyut azaltımı sağlanmış, hemde transfer edilen uzayda birbirine yakın (uzak) olması gerek örnekler daha yakın (uzak) olmuş olur. Ayrıca şu not edilmelidir ki bir $\boldsymbol{x} \in \mathbb{R}^d$ örneğini, $P\boldsymbol{x} \in \mathbb{R}^r$ 'e dönüştürmek lineer bir dönüşümdür. Daha genel olması bakımından, bu çalışmada lineer dönüşümleri dikkate alınmaktadır. Bundan başka, örnekleri önce lineer olmayan (nonlinear) bir dönüşümle dönüştürüp ardından bir lineer öğrenme uygulayan metrik öğrenme çalışmaları da mevcuttur [15].

Aşağıda metrik öğrenme alanında en sık kullanılan üç (lineer) metot kısaca açıklanmıştır. Bu üç metodun motivasyonları farklı olsa da, üçünün ortak noktası (1)'deki M matrisini hesaplamalarıdır. Ayrıca bu üç metot ilgilenilen mikro dizi gen veri setleri üzerine uygulanmış ve deneyler bölümünde sonuçları paylaşılmıştır.

A. BAZI ÖNEMLİ METRİK ÖĞRENMESİ METOTLARI

A. 1. Large-Margin Nearest Neighbors (LMNN)

Geniş Marjinli En Yakın Komşu (Large-Margin Nearest Neighbors (LMNN) [16]) metodu, metrik öğrenmesinde en sık kullanılan metotlardan biridir, ve birçok varyantı bulunur [6]. Özünde LMNN'nin amacı çok açıktır: aynı sınıfa (etikete) ait örnekleri birbirine yakın hale getirmek; farklı sınıflara ait örnekleri birbirinden uzaklaştırmak. LMNN, bu amaçla aşağıdaki amaç fonksiyonunu tanımlar ve bu fonksiyonu minimize edecek M matrisini hesaplamaya çalışır:

$$min_{M \succcurlyeq 0} \sum_{(i,j) \in B} u_M(\boldsymbol{x}_i, \boldsymbol{x}_j) + \lambda \sum_{(i,j,k) \in R} [1 + u_M(\boldsymbol{x}_i, \boldsymbol{x}_j) - u_M(\boldsymbol{x}_i, \boldsymbol{x}_k)]_+ \quad (3)$$

Burada B kümesi aynı sınıfta yer alan örnek çiftlerinin kümesini, R kümesinin elemanı olan (i, j, k) üçlüsü $\boldsymbol{x}_i$ ve $\boldsymbol{x}_j$ örneklerinin aynı sınıfta olduğunu, $\boldsymbol{x}_k$ örneğinin $\boldsymbol{x}_i$ ve $\boldsymbol{x}_j$'nin sınıfından farklı bir sınıfta olduğunu; ve λ kullanıcı tarafından girilen regülarizasyon parametresini göstermektedir.

Ayrıca burada $[\ ]_+$ ile gösterilen terim menteşe kaybı (hinge loss) olarak adlandırılan bir kayıp türüdür; $[z]_+ = \max\{0,z\}$ olarak tanımlanır. Bu amaç fonksiyonu konvekstir; ve artık değişkenler (slack variables) yardımıyla bir yarı tanımlı programlama (semidefinite programming) problemine dönüştürülerek lokal minimum noktalarından etkilenmeyecek şekilde çözülür [16].

A. 2. Neighborhood Component Analysis (NCA)

Komşuluk Bileşen Analizinde (Neighborhood Component Analysis (NCA)) [17] amaç, eğitim setindeki her örneğin kendi sınıfındaki diğer örneklerle komşu olabilme olasılığını artırmaktır. Bu anlamda, aynı sınıfta olduğu varsayılan $\boldsymbol{x}_i$ ve $\boldsymbol{x}_j$ örneklerinin birbirlerine komşu olma olasılığı:

$$p_{i,j} = \frac{exp\left(-u_M(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)}{\sum_{k \neq i} exp(-u_M(\boldsymbol{x}_i, \boldsymbol{x}_k))} \quad (4)$$

C_i, i ile aynı sınıfta olan örneklerin kümesini göstermek üzere, NCA'nın maximize etmeye çalıştığı amaç fonksiyonu aşağıdaki gibidir:

$$max_{M \succcurlyeq 0} \sum_{i} \sum_{j \in C_i, j \neq i} p_{i,j} \quad (5)$$

Dikkat edilirse (5) herhangi bir regülarizasyon terimi içermez, ve (3) 'ün aksine konveks değildir. Bu yüzden (5) fonksiyonu, M matrisi $M = P^T P$ şeklinde ayrıştırılıp P ye bağlı olarak dereceli azalma (gradient descent) yardımıyla çözülür [17].

A. 3. Information Theoretic Metric Learning (ITML)

Enformasyon teoretik metrik öğrenmesi (Information Theoretic Metric Learning (ITML)) [18] aşağıda gösterilen amaç fonksiyonunu minimize etmeyi amaçlar:

$$\begin{aligned} &max_{M \succcurlyeq 0} \ iz\,(M) - \log\det(M), \\ &\text{öyleki}, u_M(\boldsymbol{x}_i, \boldsymbol{x}_j) < s, \ \ (\boldsymbol{x}_i, \boldsymbol{x}_j) \in B \\ &u_M(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq t, \ \ (\boldsymbol{x}_i, \boldsymbol{x}_j) \in F \end{aligned} \quad (6)$$

Burada s kullanıcı tarafından belirlenen yeterince küçük bir sayı; t ise yine kullanıcı tarafından belirlenen yeterince büyük bir reel sayıdır. İlaveten, F kümesi birbirinden uzakta olması istenilen örnek çiftlerinin kümesi (bu çiftler farklı sınıflara ait örneklerden oluşabilir); B kümesi birbirine yakın olması istenilen örnek çiftlerinin kümesidir. $iz(M)$, M matrisinin izi (trace); $\log\det(M)$, ise M matrisinin determinantının logaritmasıdır.

ITML'nin amaç fonksiyonu, LMNN'de olduğu gibi artık değişkenler (slack variables) yardımıyla bir sınırsız (unconstrained) optimizasyon problemi haline getirilir ve Bregman projeksiyonuyla çözülür [18].

Görülüyor ki özünde üç metrik öğrenmesinin de motivasyonu aynıdır: beraber olması düşünülen örnekler çiftlerini beraber, uzak olması düşünülen örnek çiftlerini uzak hale getirmek. Bu metriklerin ayrıştıkları nokta tanımladıkları amaç fonksiyonunun yapısı ve bu fonksiyonların çözülme biçimidir. Doğal olarak bu fonksiyonların çözülebilmesi için gereken çalışma zamanı da metrikten metriğe değişkenlik göstermektedir.

IV. UYGULAMA

Bu bölümde, bir önceki bölümde anlatılan metrik öğrenmesi metotları çeşitli türden kanser dokularından alınan örneklere ait mikro dizi gen verileri üzerine uygulanmıştır. Bu veri setlerine ait istatistikler Tablo 1'de verilmiştir. Buna göre çalışılan kanser türleri, meme kanseri, Çocukluk Çağı Beyin Tümörü (ÇÇBT), lösemi, lenfoma ve Küçük Yuvarlak Mavi Hücreli Tümör (KYMHT)'dir. Ayrıca alt tür sayısı, kanserin kaç farklı alt türü olduğunu (makine öğrenmesi dilinde sınıf sayısını) vermektedir.

Tablo 1. *Çalışmada kullanılan mikro dizi gen veri setleri*

Veri Seti	Kanser Türü	Örnek Say.	Gen Say.	Alt Tür Sayısı
Gravier [19]	Meme Kanseri	168	2905	2
Sorlie [20]	Meme Kanseri	85	456	5
Pomeroy [21]	ÇÇBT	60	7128	2
Golub [22]	Lösemi	72	7129	2
Shipp [23]	Lenfoma	77	6817	2
Khan [24]	KYMHT	63	2308	4

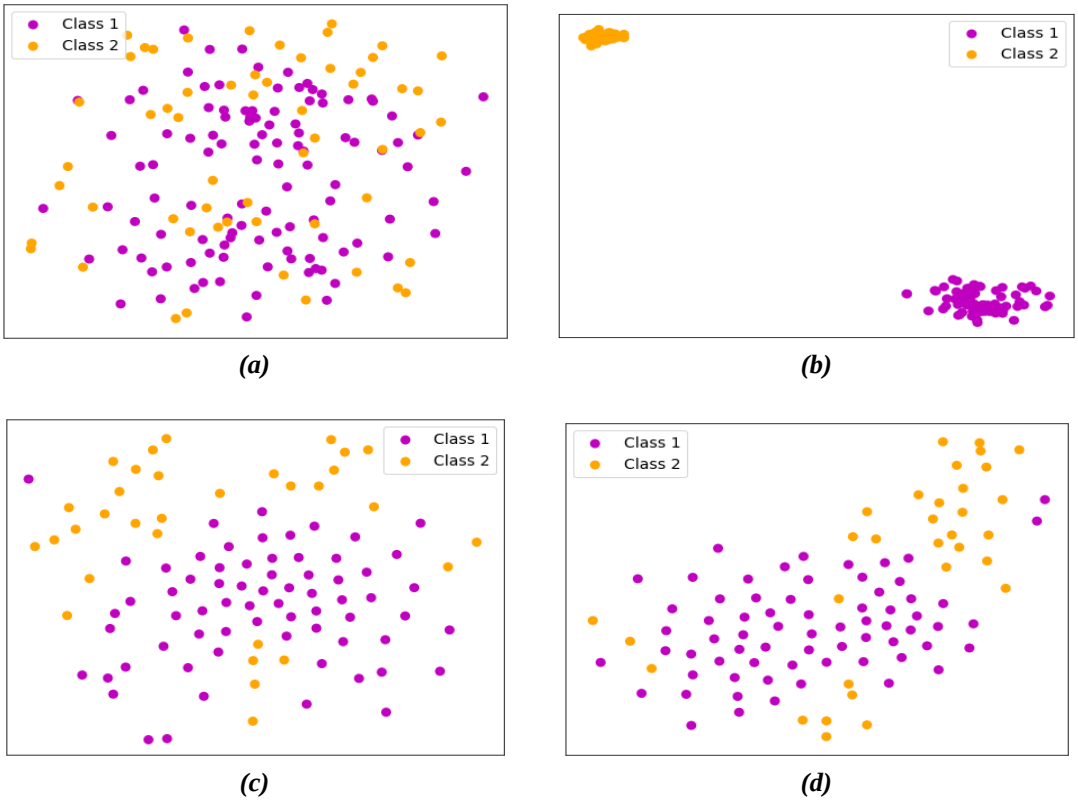
Hatırlanırsa, metrik öğrenmesinin denetimli boyut azaltma ve örnek temelli (instance-based) sınıflandırma algoritmalarına ön hazırlık oluşturma gibi iki önemli amacı olduğunu belirtilmişti. Bu

çalışmada, Tablo 1'de detayları verilen kanser veri setleri kullanılarak metrik öğrenmesinin bu iki amacı test edilmiştir.

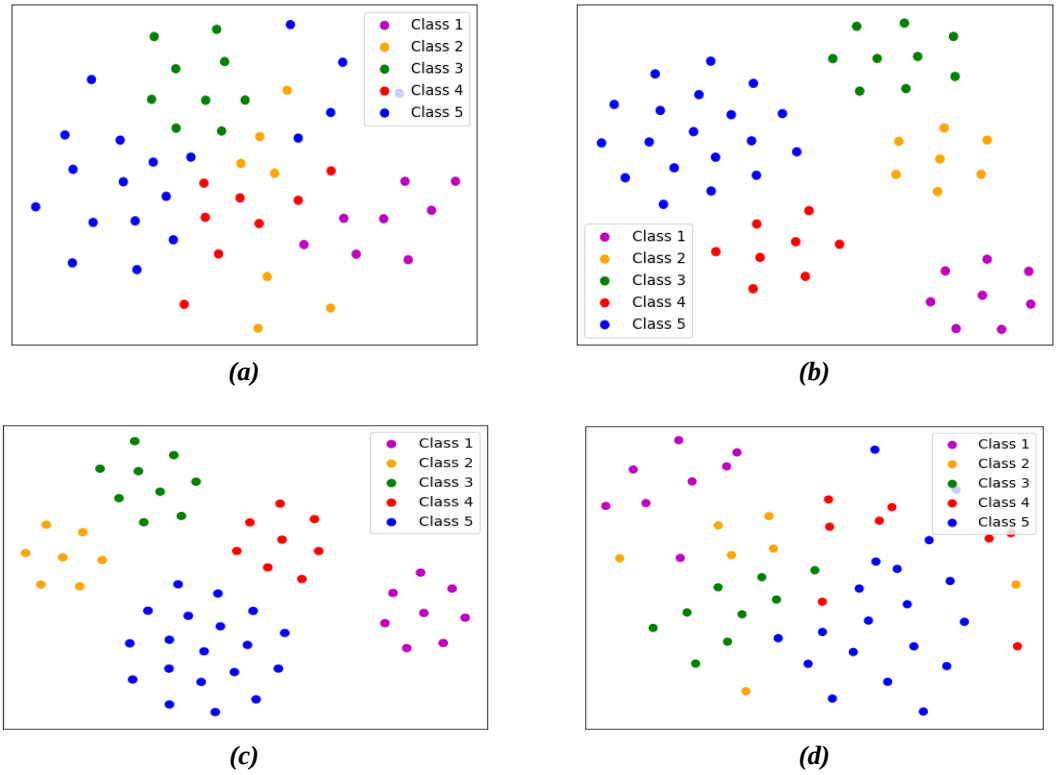
A. METRİK ÖĞRENMESİ İLE ÖZELLİK UZAYLARININ BOYUTUNU İNDİRGEME

Tablo 1'de görüldüğü gibi ilgilenilen kanser veri setleri farklı sayılarda gen içerdiğinden farklı boyutlara sahiptirler. Öyleki bu boyutlar 456 – 7129 aralığındadır. Eğer tüm veri setlerinin boyutu tek bir ortak boyuta indirgenirse, her veri setinin bozulma oranları farklı farklı olur, sınıflar birinde ayrışırken diğerinde ayrışmayabilir. Yinede, tüm veri setlerinde ortak bir boyut indirgeme prosedürü kullanmak adına bu çalışmada her veri setinin boyutu kullanılan metrik öğrenemsi metotlarıyla %30 azaltılmıştır. Bununla birlikte %20 - %80 bandında farklı indirgenme yüzdeleri de denenmiş; neredeyse tüm veri setleri için sınıflayıcıların sınıflandırma başarıları korunurken, en yüksek indirgenme yüzdesi %30 olarak tespit edilmiştir.

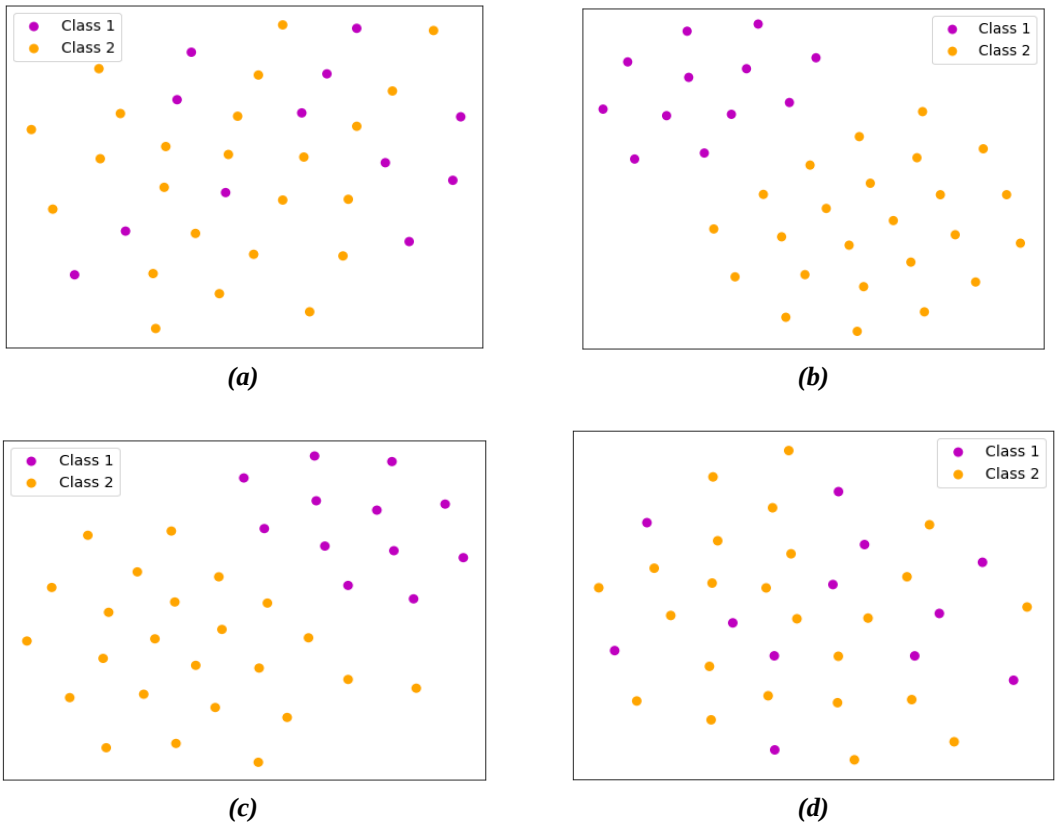
1, 2, 3, 4, 5 ve 6 nolu şekiller metrik öğrenmesinin azaltılmış boyutlu uzaylarda sınıf içi varyansı azaltabildiğini, öte yandan sınıflar arası varyansı artırabildiğini göstermektedir. Burada, kullanılan metrik öğrenmesi sayesinde boyutu indirgenmiş uzayda sınıfların birbirinden ayrıştığını görebilmek için bu uzayların görüntüleri t-distributed stochastic neighbor embedding (t-SNE) [25] metodu ile verilmiştir. Kısaca bahsetmek gerikirse t-SNE, yüksek boyutlu verileri iki ya da üç boyutlu uzaylara indirgeyerek görselleştirebilmeyi sağlayan, temellerini stokastik komşu gömme metodundann alan popüler bir istatiksel metotdur. t-SNE'yi bu denli popüler yapan özelliği yüksek boyutlu orjinal uzaydaki uzaklık-yakınlık ilişkilerini düşük boyutlu uzayda da çoğunlukla koruyabilme başarısıdır.



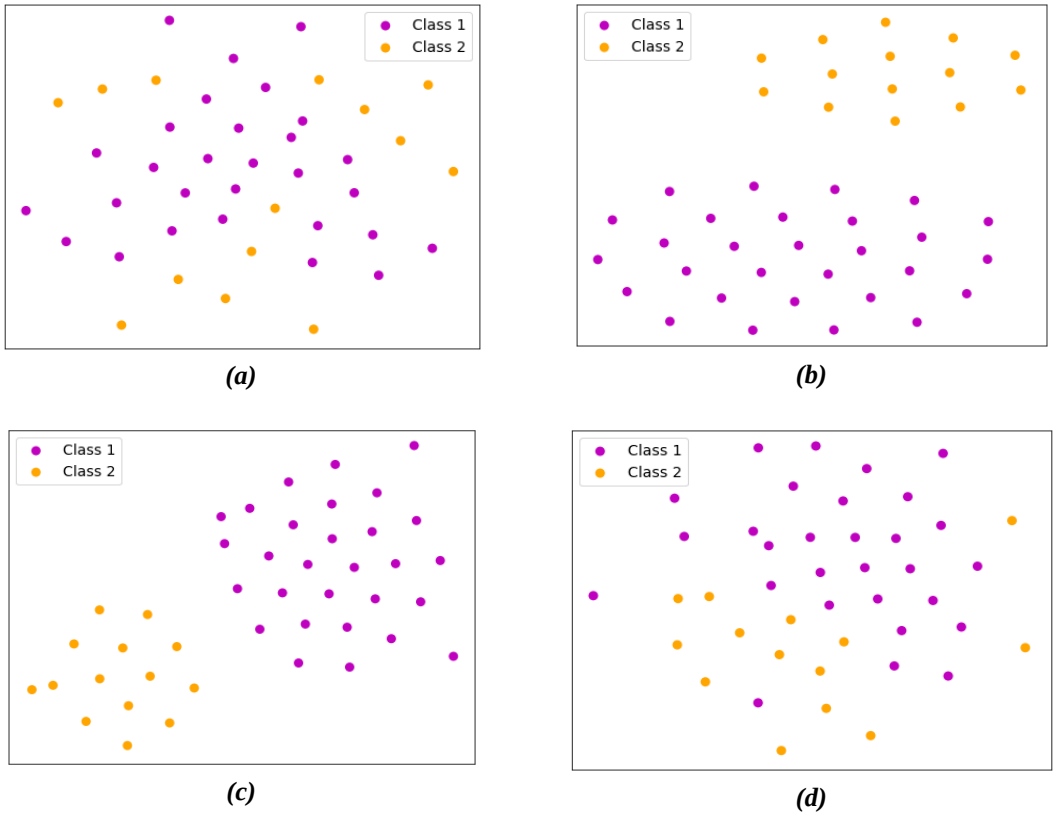
Şekil 1. *Gravier [19] Meme Kanseri Veri Seti* ***(a)*** *Orjinal Uzay* ***(b)*** *LMNN ile oluşturulmuş uzay* ***(c)*** *ITML ile oluşturulmuş uzay* ***(d)*** *NCA ile oluşturulmuş uzay.*



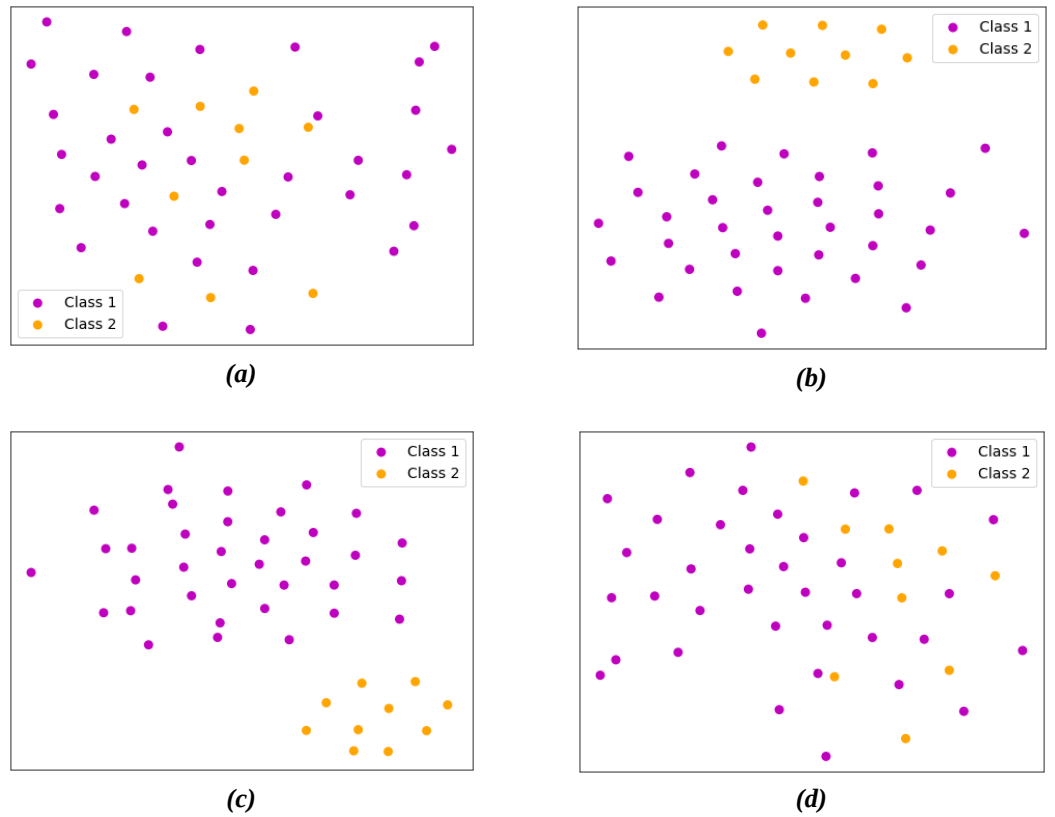
Şekil 2. *Sorlie [20] Meme Kanseri Veri Seti* ***(a)*** *Orjinal Uza* ***(b)*** *LMNN ile oluşturulmuş uzay* ***(c)*** *ITML ile oluşturulmuş uzay* ***(d)*** *NCA ile oluşturulmuş uzay.*



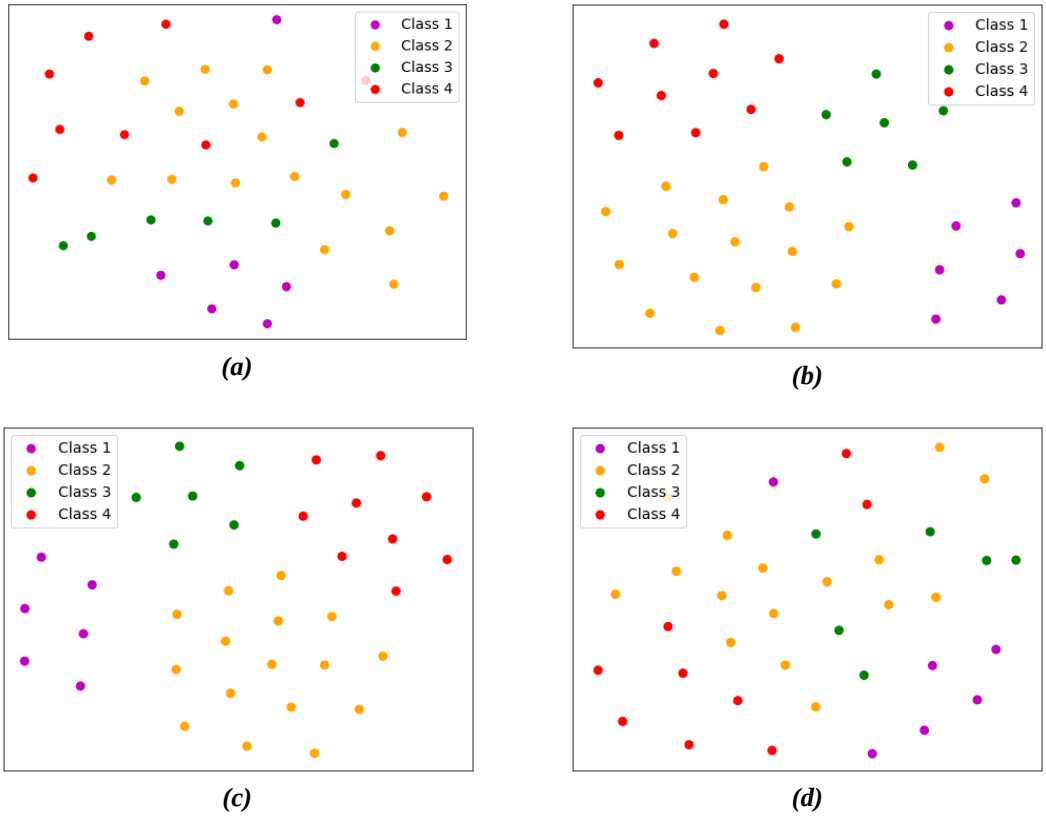
Şekil 3. *Pomeroy [21] ÇÇHT Veri Seti* ***(a)*** *Orjinal Uzay* ***(b)*** *LMNN ile oluşturulmuş uzay* ***(c)*** *ITML ile oluşturulmuş uzay* ***(d)*** *NCA ile oluşturulmuş uzay.*



Şekil 4. *Golub [22] Lösemi Veri Seti* ***(a)*** *Orjinal Uzay* ***(b)*** *LMNN ile oluşturulmuş uzay* ***(c)*** *ITML ile oluşturulmuş uzay* ***(d)*** *NCA ile oluşturulmuş uzay.*



Şekil 5. *Shipp [23] Lenfoma Veri Seti* ***(a)*** *Orjinal Uzay* ***(b)*** *LMNN ile oluşturulmuş uzay* ***(c)*** *ITML ile oluşturulmuş uzay* ***(d)*** *NCA ile oluşturulmuş uzay.*



Şekil 6. *Khan [24] KYMHT Veri Seti* ***(a)*** *Orjinal Uzay* ***(b)*** *LMNN ile oluşturulmuş uzay* ***(c)*** *ITML ile oluşturulmuş uzay* ***(d)*** *NCA ile oluşturulmuş uzay.*

t-SNE [25] yardımıyla görüntülenen uzaylarda görülüyor ki metrik öğrenmesi ile oluşmuş uzaylarda sınıflar birbirinden lineer olarak ayrılabilir haldedir. Bu, incelenen altı veri seti için de geçerlidir. Böylece örnek temelli sınıflandırma algoritmalarının bu şekilde oluşmuş uzaylarda orjinal uzaya göre çok daha başarılı olması beklenir. Bu, bir sonraki bölümde test edilmiştir.

İncelenen üç metrik öğrenmesini karşılaştırılırsa, LMNN'nin sınıfları birbirinden ayrıştırmada en başarılı metot olduğu yukarıda verilen şekillerden gözlemlenebilir. Gerçekten, incelenen 6 veri seti için LMNN sınıfları ayırmada başarılı olmuştur. Bu anlamda LMNN'yi ITML metodu takip etmiş, Gravier [19] veri seti hariç diğer veri setlerinde sınıfları ayrıştırabilmiştir. En son olarak NCA'nın da Golub [22] ve Shipp [23] veri setlerinde ayrı sınıflara ait örnekleri birbirinden uzaklaştırmada başarılı olduğu söylenebilir.

B. İNDİRGENMİŞ BOYUTLU UZAYLARDA ÖRNEK TEMELLİ SINIFLANDIRMA

Uygulamanın ikinci kısmında, bahsedilen metrik öğrenmesi metotlarını kullanarak oluşturulan indirgenmiş boyutlu uzaylarda örnek temelli (instance-based) sınıflandırma yapılmıştır. Yapılan sınıflandırmaların başarısı orjinal uzayda yapılan sınıflandırmalar ile kıyaslanmıştır. Bunun için örnek temelli üç sınıflandırma algoritması kullanılmıştır: $k-$en yakın komşu olarak bilinen $k-$Nearest Neighbor (kNN) , En Yakın Merkez Sınıflandırıcı (nearest centroid classifier) (EYMS) ve vektör niceleme olarak da çevrilen Learning Vector Quantization (LVQ) [26]. Bu sınıflandırma algoritmalarının ilgilenilen gen veri setleri üzerine olan sonuçları verilmeden önce, B.1. başlığında bu algoritmalar kısaca anlatılmıştır.

B.1 Kullanılan Sınıflandırma Algoritmaları

$k-$en yakın komşu sınıflayıcısı (k-NN), her bir test örneğini kendine en yakın k adet eğitim örneğinde en sık görülen sınıfa eşler. Formal olarak $\mathcal{Y} = \{1, \dots, n\}$ kümesi, $\boldsymbol{x} \in \mathbb{R}^d$ örneklerinin ait oldukları sınıfları göstermek üzere, eğitim setinde olmayan bir $\boldsymbol{x}_i$ test örneği (7)'ye göre sınıflandırılır:

$$argmax_{y_j \in y} \quad \frac{1}{k} \sum_{x_i \in \mathcal{A}} I(y_j = y) \tag{7}$$

Burada $\mathcal{A}$ kümesi, test örneği olan $\boldsymbol{x}_i$ örneğine öklid uzaklığı kullanılarak bulunan en yakın k adet komşunun kümesi; I ise aldığı değer doğru olunca 1'e; değilse 0'a giden indikatör fonksiyonudur.

EYMS olarak kısaltılan en yakın merkez sınıflayıcısı, test örneğini en yakınındaki merkezin ait olduğu sınıfa eşler. Genel olarak sınıfların merkezi, her özelliğin (boyutun) her sınıfta aldığı değerlerin ortalaması alınarak tahmin edilir. Formal olarak, $\mu_1, \dots, \mu_n \in \mathbb{R}^d$ sınıf merkezleri olmak üzere, EYMS (8)'e göre, bir $\boldsymbol{x}_i$ test örneğini sınıflandırır:

$$argmin_{y_j \in y} \quad u\left(\boldsymbol{x}_i, \mu_{y_j}\right) \tag{8}$$

Burada u öklid uzaklığı gibi bir uzaklık fonksiyonudur.

LVQ algoritması EYMS'deki gibi her sınıf için özelliklerin her sınıfta aldığı değerlerin ortalaması alınarak bir başlangıç merkezi oluşturur, daha sonra aşağıda detayları verilen algoritma ile bu merkezlerin tahminlerini iyileştirir. Kullanıcı tarafından önceden belirlenen iterasyon sayısı kadar (örneğin 1000) LVQ eğitim setindeki bütün örnekleri ele alır. Ele aldığı her bir eğitim örneği için bu örneğe en yakın sınıf merkezini bulur. Bu merkez, kendi sınıfının merkezi ise (9)'daki kurala göre bu merkezi kendine doğru çeker:

$$\mu_{y_j} := \mu_{y_j} + \text{öğ}renme_{oranı} \cdot \left(\mu_{y_j} - \boldsymbol{x}_j\right) \tag{9}$$

μ_{y_j} burada $\boldsymbol{x}_j$ eğitim örneğine karşılık gelen sınıfın merkezidir. Farklı olarak, ilgilendiğimiz eğitim örneğine en yakın sınıfın merkezi kendi sınıfının merkezi değilse bu merkez, (10)'daki gibi eğitim örneğinden uzaklaştırılır:

$$\mu_{y_s} := \mu_{y_s} - \text{öğ}renme_{oranı} \cdot \left(\mu_{y_s} - \boldsymbol{x}_j\right) \tag{10}$$

μ_{y_s}, $\boldsymbol{x}_j$ eğitim örneğine en yakın sınıf merkezini göstermektedir. Yukarıda (9) ve (10) da görülen denklemlerde $\text{öğ}renme_{oranı}$ öğrenme oranı (learning rate) dir. Bu, her iterasyonda merkeze yaklaşmanın ya da merkezden uzaklaşmanın miktarını belirler. Kullanıcı tarafından karar verilen bu parametre için genel yaklaşım her iterasyonda miktarını azaltmaktır [26]. T toplam iterasyon sayısını ve α başlangıç öğrenme oranını (genelde 0.3 alınır) göstermek üzere, t. iterasyondaki öğrenme oranı (11)'deki gibi hesaplanır:

$$\text{öğ}renme_{oranı} = \alpha \cdot \left(1 - \left(\frac{t}{T}\right)\right) \tag{11}$$

B.2 Sınıflandırma Sonuçları

Yukarıda bahsedilen örnek temelli sınıflandırma algoritmalarının, Tablo 1'de detayları verilen çeşitli kanser türlerine ait mikro dizi gen veri setlerinin boyutlarının LMNN, ITML ve NCA metrik öğrenmesi metotlarıyla indirgenmiş uzaylarda uygulanmasıyla elde edilen kesinlik (accuracy) sonuçları aşağıda paylaşılmıştır. Şunu not etmekte fayda vardır ki: bu sınıflandırma algoritmaları orjinal halleriyle, öklid uzaklığı kullanarak, boyutu indirgenmiş uzaylarda kullanılmıştır. Bu şekilde elde edilen sonuçlar, metrik öğrenmesi uygulanmamış orjinal uzaylarda bahsi geçen algoritmaların elde ettiği sonuçlarla kıyaslanmıştır.

Paylaşılan sonuçlar 10-katlı çapraz doğrulama (ten-fold cross validation) iterasyonlarından alınan sonuçların ortalamasıdır. Ek olarak, her veri setinde her sınıflandırma algoritması için her metrik

öğrenmesi metotunu, algoritmanin orjinal uzaydaki kullanımıyla (yani tablolardaki ilk sütun) ile kıyaslanmştır. Bu karşılaştırmalarda, tek taraflı t-testine göre istatiksel olarak anlamlı bulunan farkları daha kalın olarak yazılarak istatiksel anlamlılık belirtilmiştir.

Tablo 2. *Çalışmada kullanılan mikro dizi gen veri setlerinin sınıflandırma kesinlik (accuracy)sonuçları*

Veri Seti	Sınıflandırma Alg.	Orj. Uzay	LMNN	ITML	NCA
Gravier [19] (Meme K.)	*k*-NN	0.666	**0.832**	**0.736**	**0.791**
	EYMS	0.743	0.743	0.745	0.698
	LVQ	0.704	**0.761**	**0.722**	**0.723**
Sorlie [20] (Meme K.)	*k*-NN	0.749	**0.863**	**0.872**	**0.801**
	EYMS	0.873	**0.881**	0.877	0.916
	LVQ	0.941	0.939	0.945	0.946
Pomeroy [21] (ÇÇHT)	*k*-NN	0.633	**0.676**	0.619	0.626
	EYMS	0.771	0.736	0.645	0.771
	LVQ	0.581	**0.676**	0.594	**0.691**
Golub [22] (Lösemi)	*k*-NN	0.854	**0.953**	**0.953**	**0.897**
	EYMS	0.893	**0.944**	**0.944**	**0.910**
	LVQ	0.865	**0.939**	**0.939**	0.862
Shipp [23] (Lenfoma)	*k*-NN	0.941	**0.994**	**0.994**	0.941
	EYMS	0.841	**0.947**	**0.946**	0.836
	LVQ	0.754	**0.992**	**0.992**	0.754
Khan [23] (KYMHT)	*k*-NN	0.654	**0.973**	**0.976**	**0.688**
	EYMS	0.831	**0.973**	**0.977**	0.799
	LVQ	0.771	**0.985**	**0.977**	0.771

İlk bakışta Tablo 2 şunu ortaya koymaktadır ki farklı kanser türlerine ait mikro dizi gen veri setlerinde bir ön işlem olarak metrik öğrenmesi uygulamak kanserlerin alt türlerini belirlemek anlamında bir gelişme sağlamaktadır. O kadar ki, bu gelişme yaklaşık olarak %30'u bulabilmektedir. Bu sonuçlar Şekil 1-6 da görülen t-SNE yardımıyla görüntülenen boyutu indirgenmiş uzaylardaki görüntülerle uyuşmaktadır. Gerçekten, orjinal uzaylara göre metrik öğrenmesi yardımıyla oluşturulan uzaylarda sınıfların birbirinden ayrıldığı apaçık görülmektedir; bu da bu uzaylarda daha başarılı sınıflandırma yapabilmeyi sağlamaktadır. Bu nedenle Tablo 2 'de iyileştirilmiş sınıflandırma sonuçları görülmektedir.

Veri setleri ayrı ayrı ele alındığında metrik öğrenmesinin en fazla faydayı Khan [23] veri setinde elde ettiği görülmektedir. Orjinal uzayda uygulanan *k*-NN sınıflandırıcısının buradaki ortalama başarısı yaklaşık 0.65 iken, LMNN metrik öğrenmesi ile oluşturulmuş uzayda bu başarı 0.97'ye çıkmaktadır. Yine, bu uzayda EYMS sınıflayıcısın başarısı yaklaşık %14 artmıştır, LVQ algoritmasında ise %20 civarinda bir ilerleme kaydedilmiştir. Sorlie [20] veri setinde *k*-NN sınıflayıcısı LMNN ile oluşturulmuş uzayda yaklaşık %12 daha başarılı iken, Shipp [23] veri setinde bu ilerleme %5 civarındadır.

İlgilenilen sınıflandırma algoritmaları *k*-NN, En Yakın Merkez Sınıflayıcısı (EYMS) ve LVQ algoritmaları kendi içinde kıyaslandığında, EYMS'nin az bir farkla da olsa diğer iki algoritmadan daha başarılı olduğu görülmektedir. Gerçekten, EYMS altı kanser veri setinin üçünde en yüksek kesinliği yakalamıştır. Aynı zamanda *k*-NN ve LVQ algoritmalarına göre hesaplama maliyeti çok daha az olması bakımından EYMS'nin mikro dizi kanser veri setleri için uygun bir sınıflayıcı olduğu söylenebilir. Hatta bu başarılı ve görece maliyetsiz sınıflayıcının performansı metrik öğrenmesi metotları ile %18'e kadar artırılabilir. Benzer şekilde Lenfoma ve KYMHT veri setleri dikkate alındığında LVQ algoritmasının başarısı da metrik öğrenmesi sayesinde %24'e kadar artmıştır.

Tablo 2'nin gösterdiği bir başka gerçeklik örnek temelli sınıflayıcıların kanser alt türünü tahmin etmedeki performanslarının farklı kanser türleri için farklı olduğudur. Örneğin *k*-NN'nin performansı %65 - %94 bandındadır. *k*-NN'nin en düşük sınıflandırma performansını elde ettiği Khan [23] veri setinin orjinal uzaydaki görüntüsü Şekil 6a'da görülmektedir. Bu şekilden basitçe görülmektedir ki,

neredeyse her bir eğitim örneğinin en yakın komşuları farklı sınıflara ait örneklerden oluşmaktadır; bu da k-NN'algoritmasının yanılmasına neden olmaktadır. Öte yandan k-NN'nin en yüksek performansa eriştiği Shipp [23] veri setinde lenfomanın birinci alt türü baskındır, çok daha fazla görülmektedir. Gerçekten, veri setinin yaklaşık %75'i birinci alt türe, geri kalan %25 ise ikinci alt türe aittir. Üstelik Şekil 5a'dan görülmektedir ki aynı sınıfa ait örnekler birbirine yakındır. Bu nedenle sınıflandırma algoritmaları genel olarak bu veri setinde başarılıdır.

EYMS'nin kanser alt türünü tahmin etme başarısı ise k-NN'ye göre daha dar bir aralıkta değişkenlik gösterir; yaklaşık %74 - %89 bandında değişir. %74 sınıflandırma başarısının elde edildiği Gravier [19] veri setinde Khan [23] veri setinde olduğu gibi farklı sınıflara ait örnekler birbirleriyle iç içe geçmiştir (bakınız Şekil 1a). k-NN gibi bir lineer sınıflandırıcı olan EYMS'nin sınıfların bu şekilde iç içe olduğu veri setlerinde görece başarız sonuçlar alması bu yüzden şaşırtıcı değildir. Bir başka lineer sınıflandırıcı olan LVQ algoritması için de durum aynıdır. Farklı sınıflara ait örneklerin lineer bir doğru ile birbirinden ayrılamadığı uzaylarda sınıflandırma başarısı düşük kalır. Gravier [19] veri setindeki sınıf tahminin yeterince başarılı olmaması bu şekilde açıklanabilir. Özel olarak LVQ sınıflayıcısı Pomeroy [21] veri setinde en düşük sınıflandırma başarısına ulaşmıştır (%58). Şekil 5a'ya bakıldığında aynı sınıfa ait örneklerin uzay içersinde dağılımının fazla olduğu görülmektedir. Bu nedenle her bir sınıfın uzay içersindeki merkezini belirlemek güçtür. Bu'da LVQ'nun sınıf merkezlerini belirlemede güçsüz kalmasına neden olmaktadır.

Son olarak, metrik öğrenme metotları algoritmaların sınıflandırma başarısını artırma anlamında kıyaslandığında, en yüksek performansı LMNN'nin gösterdiği görülmektedir. İlgilenilen her veri setinde neredeyse her algoritma için veri setine uygulanması başarıyı artırmıştır; ve bu başarılar istatiksel olarak anlamlıdır. Bu sonuç 1-6 nolu şekillerdeki görüntülere paraleldir. Gerçekten, LMNN her veri setinde aynı sınıftan örnekleri bir araya getirirken farklı sınıflara ait örnekleri birbirinden uzaklaştırabilmiştir. Bu anlamda, LMNN'yi takip eden metot ITML olmuştur. ITML de altı veri setinin dördünde sınıflandırma kesinliğini istatiksel olarak anlamlı bir şekilde artırmıştır. NCA'nin performansı ise LMNN ve ITML'ye göre biraz daha geride kalmıştır. Yinede meme kanseri veri setlerinde k-NN algoritmasının başarısını %14'e kadar artırabilmiştir.

Uygulanan üç metrik öğrenmesi metodu içerisinde LMNN'nin sınıfları birbirinden ayırabilmesinde dolayısıyla en yüksek sınıflandırma yüzdesi sağlamasında altta yatan neden LMNN'nin minimize etmeye çalıştığı amaç fonksiyonun (3)'ün konveks olması dolayısıyla analitik çözümünün olmasıdır. Öte yandan ne NCA'nın nede ITML'nin amaç fonksiyonları, sırasıyla (5) ve (6), konveks değildir; ve dereceli azalma (gradient descent) ve Bregman projeksiyonu yardımıyla yaklaşık bir çözüm bulunur. Bu ise pratikte sınıfların birbirinden ayrılmasını her zaman garanti etmez.

IV. SONUÇ

Kanserli dokuların heterojen yapısı nedeniyle kanser kendi içinde çok büyük farklılıklar gösterir; bu farklılık kanserlerin alt türlerinin ortaya çıkmasına neden olur. Bu yüzden kanser tedavisinde alt türün tespit edilmesi elzemdir. Alt türün tespitinde konvansiyonel klinik yöntemlere ek olarak, kanserli dokulara ait mikro dizi gen verilerinin makine öğrenmesi/veri madenciliği yöntemleriyle tahmin edilmesinin son yıllardaki veri depolama ve veri işleme teknolojilerinin ilerlemesi ve bollaşması sayesinde popülerliği giderek artmaktadır.

Makine öğrenmesi algoritmalarının kanserli dokulara ait mikro dizi gen verileri ile eğitilerek kanserin alt türünün tespit edilmesi çalışmalarında karşılaşılan temel güçlük, doku örneklerinin ilgili-ilgisiz binlerce hatta onbinlerce genden oluşmasından dolayı, çalışılan veri setinin çok büyük boyutlu olması problemidir. Bu sorunla baş etmek için bu çalışmada, yüksek boyutlu uzaylarda yer alan veri setleri *metrik öğrenmesi* ile çok daha küçük boyutlu uzaylara transfer edilmiştir; üstelik bu uzaylarda aynı sınıfa (kanser alt türüne) ait veri örnekleri bir araya getirilmiş, farklı sınıflara ait veri örnekleri birbirinden uzaklaştırılabilmiştir. Bu görev için üç farklı metrik öğrenmesi (LMNN, ITML ve NCA), altı farklı

kanser türüne ait mikro dizi gen verisi üzerine uygulanmıştır. Uygulama sonucunda kanser alt türlerinin boyutu azaltılmış uzaylarda birbirinden ayrıldığı t-SNE yardımıyla görüntülenmiştir. Ek olarak, bu uzaylarda k-NN, en yakın merkez ve LVQ örnek temelli (instance-based) sınıflama algoritmaları çalıştırlmış, orjinal (boyutu azaltılmamış) uzaya göre bu sınıflayıcıların kesinlik (accuracy) anlamında performanslarının yaklaşık olarak %30'a arttığı gözlemlenmiştir.

V. KAYNAKLAR

[1] H. Salem, H. G. Attiya and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Applied Soft Computing*, vol. 50, pp. 124–134, 2017.

[2] A. K. Dwivedi, "Artificial neural network model for effective cancer classification using microarray gene expression data," *Neural Computing And Applications*, vol. 29, no. 12, pp. 1545–1554, 2018.

[3] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017.

[4] N. Almugren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*, vol. 7, pp. 78533–78548, 2019.

[5] Z. M. Hira and D.F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances In Bioinformatics*, vol. 1, no. 198363, 2015.

[6] B. Kulis, "Metric learning: A survey," *Foundations and trends in machine learning*, vol. 5, no. 4, pp. 287–364, 2012.

[7] S. B. Cho and H. H. Won, "Machine learning in DNA microarray analysis for cancer classification," in *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics*, Adelaide, Australia, 2003, vol. 19, pp. 189–198.

[8] S. Kılıçarslan, K. Adem ve O. Cömert, "Parçacık sürü optimizasyonu kullanılarak boyutu azaltılmış mikrodizi verileri üzerinde makine öğrenmesi yöntemleri ile prostat kanseri teşhisi," *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, c. 7, s. 1, ss. 769–777, 2019.

[9] B. Haznedar, M. T. Arslan ve A. Kalınlı, "Karaciğer mikroarray kanser verisinin sınıflandırılması için genetik algoritma kullanarak ANFIS'in eğitilmesi," *Sakarya University Journal of Science*, c. 21, s. 1, ss. 54–62, 2017.

[10] F. Morais-Rodrigues, R. Silv́erio-Machado, R. B. Kato and D. L. N. Rodrigues, "Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression," *Gene*, vol. *726*, pp. 144–168, 2020.

[11] O. Yıldız, M. Tez, H. Ş. Bilge, M.A.Akcayol ve İ. Güler, "Meme kanseri sınıflandırması için gen seçimi," *IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, İstanbul, Türkiye, 2012, ss. 18–20.

[12] R. Ruiz, J. C. Riquelme and J. S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognition*, vol. 39, no. 12, pp. 2383–2392, 2006.

[13] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, "Constrained k-means clustering with background knowledge," in *Proceedings of the 18th International Conference on Machine Learning*, Florida, USA, 2001, vol. 1, pp. 577–584.

[14] W. De Vazelhes, C. J. Carey, Y. Tang, N. Vauquier and A. Bellet, "Metric-learn: metric learning algorithms in Python," *Journal of Machine Learning Research*, vol. 21, no. 138, pp. 1–6, 2020.

[15] F. Wang and J. Sun, "Survey on distance metric learning and dimensionality reduction in data mining," *Data Mining and Knowledge Discovery*, vol. 29, no. 2, pp. 534–564, 2015.

[16] K. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. 2, 2009.

[17] J. Goldberger, S. Roweis, G. Hinton and R. Salakhutdinov, "Neighbourhood components analysis," *Advances in Neural Information Processing Systems*, vol. 17, pp. 513–520, 2004.

[18]. J. V. Davis, B. Kulis, B. P. Jain, S. Sra and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th International Conference on Machine Learning*, New York, USA, 2007, pp. 209–216.

[19] E. Gravier, G. Pierron, A. Vincent-Salomon, A. Gruel, N. Raynal, V. Savignoni and A. Fourquet, "A prognostic DNA signature for T1T2 nodenegative breast cancer patients," *Genes, Chromosomes and Cancer*, vol. 49, no. 12, pp. 1125–1134, 2010.

[20] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen and A. L. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *National Academy of Sciences*, vol. 98, no. 19, pp. 0869–10874, 2001.

[21] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, and M. E. McLaughlin, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[22] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov and E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[23] M.A. Shipp, K.N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok and R.C. Aguiar, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[24] J. Khan, J. S. Wei, M. Ringner, L. H.Saal, M. Ladanyi, F. Westermann and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol.7, no. 6, pp. 673–679, 2001.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no.11, 2008.

[26] T. Kohonen, *Self-Organizing Maps*, 1st ed., Berlin, Germany: Springer, 1995, pp. 245–26.