Data Surgery Task: ‘Soru.pdf’ Dosyasında Yazılmış Maddelere Yanıtlar:

1. Veri seti olarak https://data.mendeley.com/datasets/wmy84gzngw/1 linkinden temin ettiğim Breast Cancer veri setini kullandım. Veri seti, benign (100)/malign (150) olarak kategorize edilmiştir. Burada tarafımdan ön işleme olarak sadece resize ve scaling kullanılmıştır.
2. Burada ise bir sistem tasarımı istenmiştir. Bu kısımda maalesef kullanıcı etkileşimli bir tasarım veremiyorum. Ancak model yaratıldıktan sonra inference kodunu yazmak için tek gereken train ederken uygulanan ön işlemlerin kullanıcı girdisine uygunlanması gerektiğidir. Yani model önce kaydedilir (burada joblib kullanılabilir). Daha sonra ise kullanıcıdan resim alınır ve trainde uygulanan ön işlemeden geçirilir. Model load edilerek predict ile basitçe tahmin edilebilir. Yazdığım train kodu gittedir. Diğer uygulama (konsol, önyüz vs.) çalışmaları mevcut değildir.
3. Bulmuş olduğum sınıflandırma raporu Şekil 1. üzerinde gösterilmiştir. Bu aynı zamanda confusion matrisin de yorumudur.

	precision	recall	f1-score	support
0	0.95	0.90	0.92	20
1	0.95	0.98	0.97	43
accuracy			0.95	63
macro avg	0.95	0.94	0.94	63
weighted avg	0.95	0.95	0.95	63

Şekil 1. Sınıflandırma Raporu

Yukarıdaki sonuçlara bakılırsa sınıflandırma başarısı iyidir denilebilir (0: benign, 1: malignant). Ancak sette sadece iki sınıf bulunmaktadır bu işi biraz da kolay kılmaktadır. Sistemin train/test verisi dışında farklı kullanıcılara ait resimlerle denenmesi sistemin başarısını görmek açısından en gerçekçi yol olacaktır. Şekil 1. de genel accuracy yerine precision yada recall daha kritik olabilir. Riske göre değiştirilmiş katsayılı f1 score formülleri kullanılabilir. Burada genel olarak skorlar iyi gözükmektedir.

Kendi yorumumla ‘kanser değilsiniz’ demenin daha riskli olduğunu görüyorum. Burada yanlış kişiye ‘kanser değilsiniz’ denilirse sonuçları diğerlerinden daha kötü olabilecektir. Bu da önem derecesi olarak 1 etiketindeki precision kısmına denk gelmektedir.

Sonuçlar genel anlamda iki şekilde iyileştirilebilir. Bunlar iyi bir ön işleme ve iyi bir parametre araştırmasıdır. Aslında veri ön işlemesi ve algoritma parametre araştırması birbirini etkileyen unsurlardır. İdeal algoritmanın, ideal parametreler ile ideal ön işlenmiş veriler üzerinde çalışabilmesi gerçekten çok zorlu bir konudur. Bu, denenip görülmesi gereken bir durumdur. Donanım olanakları zaman kazanmak açısından çok önemlidir.

4. Bu soruda kısıt konusu verinin tamamen ham olması ve etiketlemenin muğlak yapılmış olmasıdır. Malign setleri indirdiğimde sadece xml dosyaları vardı. Tüm seti indirdiğimde ise

xml dosyaarında malign benign ayrımı yoktu sadece tirad ayrımı vardı. Hem ciddi bir ön işleme hem de sınıflandırma için ayrıyeten programlamatik işler gerekliydi. Bu sebeple zamanım oldukça dar olduğundan modeli sunabilmek için daha oturmuş bir görüntü seti kullandım.

Ben model eğitimi için CNN kullandım. Sebebi ise CNN'in en başta görüntü işleme amaçlı olarak literatüre kazandırılmış olmasıdır. Başarılı olacağını düşündüğüm için kullandım. Buna ilaveten elde hazır başka bir model olsaydı transfer learning de uygulanabilirdi. Ayrıca CNN algoritmasının çeşitli gelişmiş versiyonları da bulunmaktadır.

5. Model benign/malign teşhisini yanlış yapıyorsa, confusion matristen hangi tiradların karıştığı görülmelidir. O noktalarda belirgin şekilde birbiri ile karışan tiradlar varsa o noktada ek ön işleme ya da veri seti zenginleştirme ile çözüm bulunabilir.