

DATASURGERY Yazılım ve Danışmanlık Ltd. Şirketi 1.Aşama Değerlendirme Soruları cevapları:

- 1.) Soruda vermiş olduğunuz <http://cimalab.intec.co/applications/thyroid/index.php> linkindeki data setini kullandım. İnternet üzerinde <https://www.cancerimagingarchive.net/> sayfası üzerinde tiroit kanseri ile ilgili bazı görüntüler bulmama rağmen bunların hangilerinin benign hangilerinin manign oldukları açıklanmadığından dolayı bu site içerisindeki görüntüleri kullanamadım. Google görseller üzerinde ve makalelerde kullanılmış bazı görüntüler olmasına rağmen yukarıda açıkladığım sebepten dolayı sadece sizin referans gösterdiğiniz veri tabanını kullanmayı tercih ettim.

Cimala'a ait olan görüntüler web sayfasının ön yüzü incelendiğinde ana hat olarak ikiye ayrılmış durumda. Bunlar benign ve manign. Yani iyi huylu ve kötü huylu. Bu iki ana başlık altında tirads skorlarına göre alt başlıklara ayrılmış durumda. Benigns altında 2 ve 3 tirad skorlu görüntüler ve Manigns altında 4a, 4b, 4c ve 5 tirad skorlu görüntüler bulunmaktadır.

Toplu data setini indirip xml dosyaları incelendiğinde bu data seti içerisinde daha fazla görüntü olduğu görülmekte. Data seti içerisindeki görüntü isimleri ile site önyüzünde bulunan görüntü isimleri eşleşmemekte. Bu yüzden sınıflamak için xml dosyaları içerisindeki tirads skorları kullanıldı. Python ile yazılan kod ile xml dosyalarının içerisindeki tirads skorları ile tekrar sınıflama yapıldığında 2, 3, 4a, 4b, 4c, 5 tirad skorları hariç tirad skoru olmayan None tirad skorlu görüntülerde bulundu.

Xml dosyası içerisinde tirads özniteliği dışında number, age, sex, composition, echogenicity, margins, calcifications gibi başka öznitelikler bulunmakta. Fakat uygulamada sadece tirads özelliği görüntüleri sınıflandırmak için kullanılmıştır.

- 2.) Thyroid_diagnosis_using_image_processing.ipynb
- 3.) Dört farklı etmeni en iyi şekilde bir araya getirerek en iyi sonucu bulmaya çalıştım. Bunlar sırayla eğitim için verilen girdi sayısı; verilerin train, test ve validation şeklinde hangi oranla bölüneceği; model oluşturma ve model hiperparametrelerin uygun bir şekilde ayarlanmasıdır.

- İlk etmen için data setine bakıldığında manign değerlerin benign değerlerden fazla olduğu görülmektedir.

İlk olarak None değerler hariç bütün değerler ile eğitim denendi. %90 gibi bir doğruluk oranı çıkmasına rağmen bütün tahminler manign olarak tahmin edildi. Çünkü test için 10 tane benign değer varsa eğer buna karşın 60 tane manign değer olmakta. Manign değerlerin hepsini tahmin etse bile %90 gibi yüksek bir doğruluk oranı çıkmakta. Burada veri sayısı eşit olmadığı için doğruluk oranı model iyi çalışıyormuş gibi gösterebilir.

İkinci olarak eğitime eşit miktarda girdi sağlanırsa eğer model1 için doğruluk oranı %75 gibi bir seviyeye gerilemekte fakat benign değerlerde daha fazla doğru tahmin yapmakta. Modellerde her zaman tahmin değerini manign olarak tahmin etme gibi bir eğilim var. Eğer

test verilerinde 20 benign 20 manign değer varsa benign değerlerinin 10 tanesini doğru tahmin edip kalan 30 değeri manign olarak tahmin etmekte.

- İkinci etmen olarak girdi verilerinin train, test ve val olarak hangi oranda ayrılacağı.

İlk olarak verileri %70 %15 %15 olarak üçe ayrıldı. Bu ayrılma sonucunda test ve val değerleri için her sınıftan sadece 4 görüntü kaldı. Model eğitiminden sonra model test verileri ile test edildiğinde %87 gibi bir sonuç çıkmakta. Fakat test verileri çok az olduğu için bu sonuç yüksek olsa da her zaman çok güvenilir olmayabilir.

İkinci olarak verileri %70 %25 %25 olarak üçe ayrılması. Toplamı bütünden fazla olmasına rağmen istenilen şekilde metot bölme yapabilmekte. Test ve val girdi sayıları arttırıldığında train için kalan girdi sayısı da otomatik olarak düşmekte. Bunun sonucunda modelin doğruluk oranı %75 gibi bir orana düşmekte. Fakat genel olarak daha güvenilir bir oran. Model2’de %50 ML algoritmalarında ise sırasıyla %72 ve %67 gibi bir doğruluk oranı olmaktadır.

- Üçüncü olarak modelin oluşturulması veya standart bir modelin uygulanması. Bu veriler için iki CNN ve iki Makine Öğrenmesi modeli kullanıldı. İlk CNN modeli sigmoid aktivasyon fonksiyonu ile ikili(binary) sınıflandırma yapan bir model. İkinci CNN modeli softmax ile regresyona dayalı tahmin yapan VGG16 standart modeli. Üçüncü ve dördüncü model ise Support Vector Machine ve Random Forest olarak bilinen makine öğrenmesi modelleri. En yüksek oranı % 75 gibi bir oranla birinci model sağladı. İkinci model özellikle az verilerde %50 gibi bir çıktı sağladı ki bu sadece benign değerleri bildiği için. Makine öğrenmesi modelleri ise sırayla %72ve %67 gibi bir oranları yakalayabildiler.
- Dördüncü etken ise model parametreleri. Hangi optimizelerin kullanılması, number_of_epochs , batch size, steps_per_epoch sayılarının birbirlerini dengeleyecek şekilde verilmesi. Son yazdığım 3 parametreyi doğru şekilde verebilmek için image_count= number_of_epochs*batch size*steps_per_epoch formülünden yararlandım.

4.) Beni kısıtlayan en büyük engel kesinlikle veri sayısının orantısız ve az olmasıydı. Her iki sınıfında eşit olabilmesi için her iki sınıftan 60-60 eleman aldım. Doğru sonuç için en az her sınıftan 300 girdi verisi olması gerekir. İkinci olarak daha fazla zamanım olsaydı daha fazla araştırma yapabilirdim. Yeni bir modelle birlikte Transfer Learning kullanılabilirdi. Örnek vermek gerekirse Microsoft’un ResNet veya MobilNet gibi daha önce birçok farklı görüntüyle eğitilmiş bir modelin ağırlıkları amacımız doğrultusunda bizim tahmin etmesini istediğimiz görüntülere yönlendirilirse eğer daha az eforla daha iyi sonuçlar elde edilebilir.

5.) Eğer belli şekilleri olan kedi köpek gibi görüntülerin sınıflanması istenseydi doğruluk oranını arttırmak için üçüncü soruda verdiğim cevapları uygulayabilir ve ek olarak daha çok veriyle eğitim yapıp doğruluk oranını arttırabilirdim. Fakat bu durumda tirads skorlarının tam olarak resimde bize neyi gösterdiğini tıbbi açıdan açıklamasının ne olduğunu için uzmanı bir doktordan öğrenip çıkarabileceğim yeni özellikler ile yeni bir eğitim yapmayı denerdim.