TİROİD NODÜLLERİNİN YAPAY ÖĞRENME MODELİ İLE SINIFLANDIRILMASI

ÖZET

Çalışma kapsamında , son derece popüler derin öğrenme algoritmalarından biri olan evrişimli sinir ağları ile tiroid nodüllerinin iyi huylu ve kötü huylu olarak sınıflandırılması yapılmıştır.

AYDIN YENİL

Elektronik ve Haberleşme Mühendisi

Github Adresi : https://github.com/aydinynl

DATASURGERY

1. Hangi veri setini kullandınız, ilgili veri seti için öznitelikler nelerdir, açıklayınız.

Tiroid içerisinde oluşan nodüller ,1' den 6 ' ye kadar TIRADS skorlaması yapılarak sınıflandırılır. Bu skorları hastalığın doğasını baz alarak sınıflandırdığımızda ; 1-3 arası "iyi huylu", "4a-4b-4c-5-6" arası skorlar ise "kötü huylu" olarak iki sınıfta ayrılır.

Bu çalışma kapsamında da yukarıda bahsedildiği gibi tiroid nodüllerini TIRADS skorlarına göre sınıflandırıp ilgili skor içerindeki görüntüler için hasta bilgisi ve uzman notları barından bir veri seti, **TDID (Thyroid Digital Image Database)** veri seti kullanmıştır. Veri seti içerisinde 99 adet vakaya ait tiroid bezi ultrason görüntüsü ve bu görüntülerde toplamda 102 adet tiroid nodülü bulunmaktadır. Görüntüler 0(siyah) 'dan 255(siyah) 'e piksel değerleri alan grayscale görüntülerdir.

Veri setini avantajı bakımından ele aldığımızda; tiroid görüntülerindeki iyi huylu ya da kötü huylu nodüllerin yeri klinik doktoru tarafından belirtilmiştir. Aksi takdirde ultrason görüntüsündeki ilgili nodül bölgeleri , kompleks görüntü işleme veya makine öğrenmesi algoritmaları ile segmente edilmelidir. Burada amaç ; ilgisiz alanı çıkartarak analiz/sınıflandırma performansını arttırmaktadır. Bu işlem model eğitime sokulmadan yapılması gereken bir ön hazırlılardan biridir. Bu çalışma kapsamında da veriler aşağıdaki şekilde işlenmiştir.
Literatürlere bakıldığında aşağıdaki işlemleri doğrudan tüm ultrasonlu bölge üzerinde yapan çalışmalarda görülmüştür.

Veri setinin dezavantajı ise çeşitlilik ve hacim olarak küçük olmasıdır. Büyük veriyi , büyük veri yapan 2V kuralı (volume , variability) bu veri setinde maalesef eksiktir. Veri setindeki görüntülerin sayısı , çeşitliliği , kalitesi arttıkça eğitim başarısı da uygun network mimarisi altında artacaktır. Genelleme kabiliyeti de bir o kadar yüksek olacaktır.

2. İstediğiniz yapay öğrenme modelini kullanarak oluşturacağınız bir sistem tasarlayın ve kaynak kodunuzu githuba yükleyin. Bu aşamada performans metriklerinizi (accuracy, f-score…) listeleyin. Görselleştirme için hazır toollardan yararlanabilirsiniz.

(Python kodları github adresimdedir.)

Bu alanda literatür taraması yapıldığında farklı algoritmalar ile tiroid nodüllerinin sınıflandırılması / tespiti yapıldığı gözlemlenmiştir. Bunlar arasında ;

-Temel sinyal ve görüntü işleme algoritmaları (Fast Fourier Transform ile)

-SVM , Neural Network , Decision Tree , Naıve Bayes gibi sınıflandırıcı Makine Öğrenimi algoritmaları

-Deep Neural Network , Conv. Neural Network , Recurrent neural network , Cascade Conv. Neural Network gibi Derin Öğrenme algoritmaları mevcuttur.

Genel olarak yapılan çalışmalarda uygun veri seti eşliğinde(bir sınıf için 5k ve üzeri data) CNN ve hatta Cascade CNN ile geliştirilen modellerin daha iyi genellemeler yapabildiği , daha yüksek performans sergilediği gözlemlenmiştir.

Bu çalışma kapsamında da Convolutional Neural Network kullanılarak model geliştirilmiştir. Sinir ağı ve eğitim algoritmasına ait hiper parametreler literatür taraması ve tekrar tekrar yapılan eğitim sonuçları baz alınarak güncellenmiş , son aşamada aşağıdaki hiper parametlerin kullanılmasına karar kılınmıştır.

HİPER PARAMETRELER	DEĞER	ÖZELLİK
FC1 NÖRON	256	Birinci tam bağlı katmandaki nöron sayısıdır
DROPOUT	0.25	FC katmanına ait nöronların yarısını random olarak inaktif yapar.
AKTİVASYON FONKSİYONU	ReLU,Softmax	Evrişim ve FC katmanında relu, çıkış katmanında ise softmax kullanılır.
EVRİŞİM FİLTRE BOYUTU	3x3 and 5x5	Evrişim katmanında özellik çıkarıcı her filtrenin boyutudur
ORTAKLAMA FİLTRE BOYUTU	2x2	Ortaklama katmanında kullanılan filtre boyutudur.
OPTİMİZASYON LOSS FONKSİYONU	ADAM CATEGORİCAL CROSS ENTROPY	Hata oranını hesaplamak ve ardından düşürmek için loss ve optimizasyon fonk.
ÖĞRENME ARALIĞI	0.0001	Eğitim sırasında ağ ağırlıklarının ne kadar hızlı güncelleneceğini belirler.

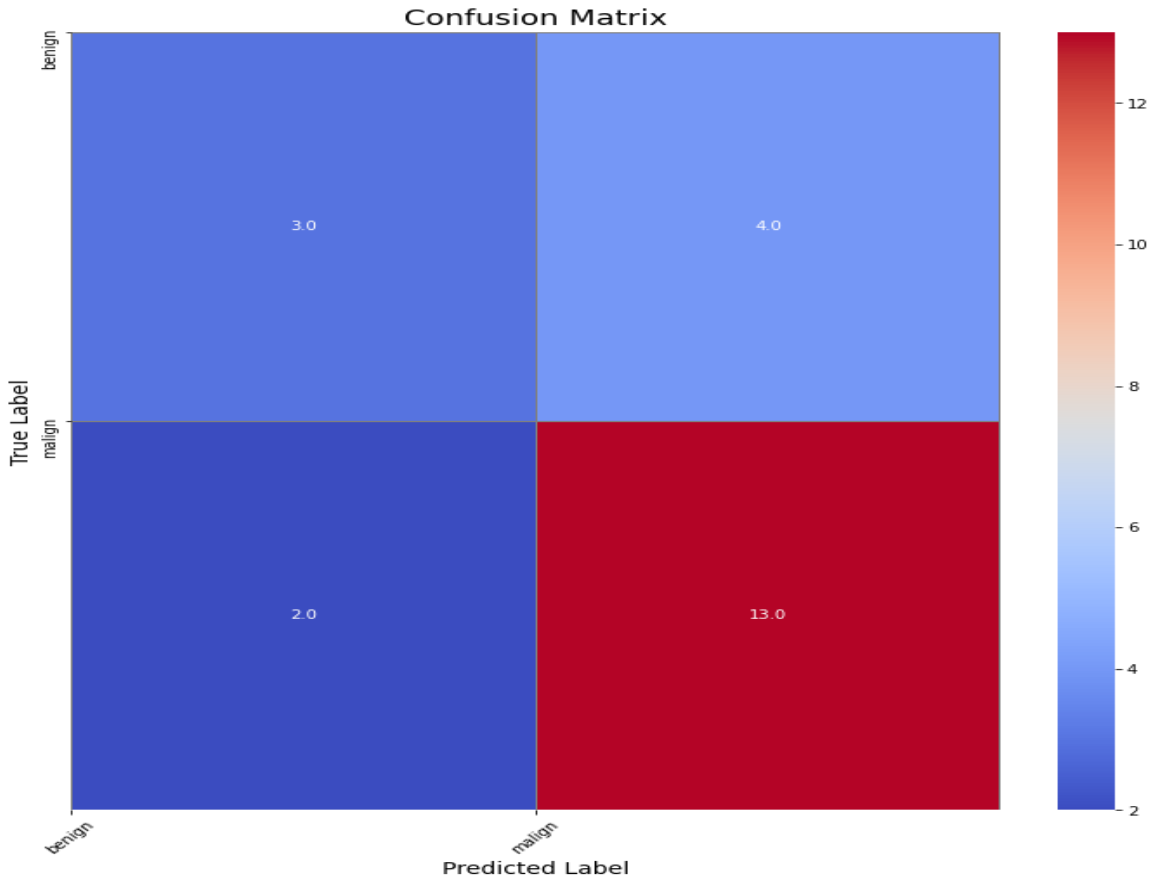
Modele bağlı hiper parametler ile ağırlıkların doğru bir şekilde güncellenerek , lossun azaltılıp model başarısının arttırılması sağlanır. Model başarısı için mimarinin yanında veri seti de önemli unsurlardan biridir. Veri seti çeşitli ve hacimli olmalıdır. Yeterli veri yok ise GAN, veri sentezleme gibi işlemler veri setine uygulanır. Çalışmada kullanılan veri seti , bir CNN modeli eğitimi için çok az örnek içerdiğinden dolayı bu çalışmada veri sentezleme methodu kullanılmıştır. Mevcut örnekler üzerinde parlaklık, rotasyon gibi işlemler yeni örnekler üretilmiştir.

Oluşturulan bir modelin performansı hakkında Accuracy-Valid. Eğrisi , ROC eğrisi ,Confusion Matrisi gibi yapılar bize bilgi verir.

-Confusion Matrix

Veri seti , 80 adet traning(53-malign V 27-benign) ve 22 adet test(Test verileri random atandı.Az veri olduğu için cross-validation yapılmadı.)görsellerinden oluşmaktadır. Mevcut veri az olduğu için test seti , validation seti olarakta kullanılmıştır. Aşağıdaki matris ise eğitim sonunda test setlerindeki benign ve malign tiroid nodüllerinin değerlendirilmesidir. Matrise bakıldığında kötü huylu nodüllerin tespiti için performans yüksek iken iyi huylu nodüller için bunu söylemek mümkün değildir. 2 adet kötü huylu nodül , iyi huylu olarak tespit edilmiştir. Aynı şekilde 4 adet iyi huylu nodül , kötü huylu olarak tespit edilmiştir.

	precision	recall	f1-score	support	
benign	0.60	0.43	0.50	7	
malign	0.76	0.87	0.81	15	
accuracy			0.73		= (13+3)/(13+3+2+4)
macro avg	0.68	0.65	0.66	22	
weighted avg	0.71	0.73	0.71	22	



3. Bulduğunuz sonucu raporlayarak, yorumlayın. Daha iyi bir model sonucu elde etmek için neler yapılabilir?

Matris incelendiğinde yukarıda bahseldiği gibi iyi huylu nodülleri saptamada çok başarılı değildir. CNN modeli eğitimindeki başarısızlıklar temelde iki sebepten kaynaklanır. Birincisi veri setinin çeşitlilik ve hacim bakımından eksikliği , ikincisi ise model mimarisi ve hiperparametrelerin doğru bir şekilde(veri setine uygun bir şekilde) belirlenmeyişidir.

Matrise bakıldığında iyi huylu nodüller için daha cidli problem varmış gibi gözükmektedir. Ancak traning veri setinin çok az olmasından kaynaklı , model kötü huylu nodüller için iyi bir sınıflandırma yapar gibi gözükse de , farklı nodüller ile karşılaştığında iyi bir sınıflandırma yapamaz. Yani her ne kadar accuracy olarak %73 ‘lerde olsada genelleme kabiliyeti düşük bir modeldir. (Loss grafiği de bunu desteklemektedir.)

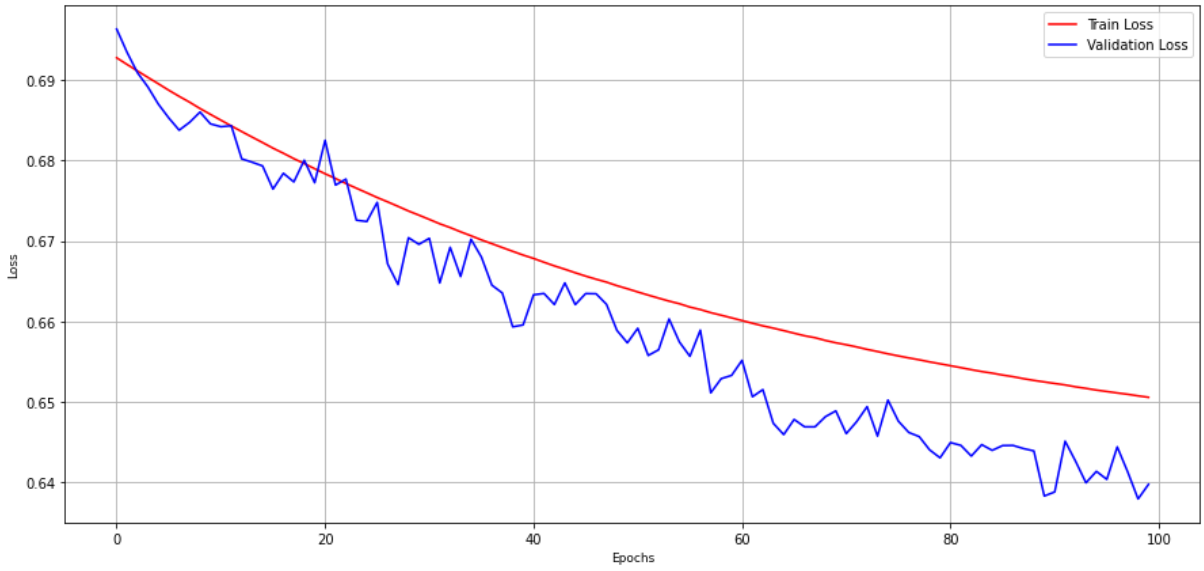
Bir de modelin başarısızlığının iyi huylu nodüllerden kaynaklı olması bir diğer senaryayo göre daha iyi gözükebilir. Kısacası , ‘’Kötü huylu nodülü olan hastaya iyi huylu tespiti koymak mı? Yoksa iyi huylu nodülü olan hastaya kötü huylu tespiti koymak mı ? ‘’ daha cidli sorunlara yol açabilir , diye düşünebiliriz.

Peki CNN modelimizi nasıl iyileştirebiliriz?(Ya veri seti ya da model / model mimarisi ile oynanır.)

*Hem iyi huylu hem de köyü huylu nodüller için daha çok örnek barındıran , çeşitlilik bakımından zengin veri setleri ile eğitim yapabiliriz. Data setlerini manuel arttırmak için veri sentezinin yanı sıra GAN yapıları da kullanabiliriz. Aynı zamanda veri setleri içerisinde yer alan traning ve test örneklerini de çapraz-validation ile dağıtmak daha doğru olur.

*Model – veri seti arasında dinamik bir ilişki vardır. Eğer veri seti büyük ve kompleks değil ise aşırı parametreye sahip, karmaşık sinir ağı mimarisinden uzak durmakta fayda vardır. Bu çalışmadaki model , veriye göre daha kompleks bir yapıdadır.

Ayrıca eğitim sonunda traning ve validation loss grafiği de çıkartılmıştır. Grafiğe bakıldığında over-fitting(aşırı öğrenme) tarzı bir durum gözlemlenmektedir. Yani model , matrisi yorumlar iken de dediğim gibi genellemeden uzak , veriye bağımlı bir modeldir.



4. Bu soruda iyi bir sonuç için sizi kısıtlayan aşamalar nelerdir, farklı bir model kullanmak isteseniz hangi model/leri kullanırdınız sebebi ile açıklayınız?

Literatür taraması yapıldığında da CNN ve Cascade CNN ile geliştirilen sınıflandırma modellerinde diğerlerine göre daha yüksek başarı elde edildiği gözlemlenmiştir. Ancak burada belirleyeci noktalardan biride elimizdeki veri setidir. CNN ile iyi bir sınıflandırma modeli geliştirmek istiyorsanız veri setiniz hacimli ve çeşitli olmak zorundadır. Bu çalışmada geliştirilen model ile istenilen başarının sağlanamamasının başlıca sebebi veri seti örneklerinin miktar ve çeşit bakımında çok yetersiz olmasıdır. İyi bir sınıflandırıcı modelde , bir sınıfa ait 5k ve üzeri örnekler ile genelleme yapabilen , fiziksel ortamlara taşınacak-operasyonlar da kullanılacak bir model geliştirilebilir. Veri seti az olduğu için modelim de parametre sayısını az tutmaya çalıştığım zaman model öğrenimini sağlayamadı.(Under-fit) Biraz kompleks bir model kurmak zorunda kaldım. Bu seferde az da olsa bir over-fit durumu gözlemledim.

Peki aynı veri seti için tekrar yapay öğrenme modeli geliştirsem , hangi modeli kullanırım ?

Sınıflandırma amaçlı kullanılan birçok makine öğrenme algoritmaları mevcuttur. Bunlardan en bilinenleri ; yapay sinir ağları , karar ağarları , random forest ve destek vektör makineleridir. Literatür taraması yapıldığında çalışmamızdaki veri seti ile eşdeğer şekilde karar ağaçları ve random forest algoritması ile yapılan modeller ile %78 doğruluk , yapay

sinir ağları ile yapılan modeller de ise %74 oranlarında doğruluk elde edilmiştir. CNN ile elde edilen doğruluk oranı bu modellerin performanslarına yakındır. Belki de parametrelerde ve ağ mimarisinde biraz daha değişim yaparak bu başarı oranları elde edilebilir.

Aynı zamanda doğrudan sinir ağları ile yapılan derin modeller CNN ile geliştirilen modellerden daha çok parametreye sahiptir ve eğitim süreci daha uzun ve zahmetlidir. Tam da bu sebep aslında evrişimli ağların çıkış noktasıdır.

Bir diğer sınıflandırıcı algoritma ise destek vektör makineleridir. Literatürlere bakıldığında çok küçük veri setleri ile bile destek vektör makineleri ile yapılan modellerin güzel sonuçlar verdiği gözlemlenmiştir. %85 - %95 oranında başarı oranlarına sahiptir. Destek vektörlerinin yukarıda bahsi geçen diğer makine algoritmalarına göre daha iyi sonuçlar vermesinin sebeplerini ;

-Tüm veri sınıflarından farklı fonksiyonları destekleyen , lineer olmayan modellerdir.

-Örnek dışı genelleme sağlayabilirler ve optimallik problemleri convextir.

şeklinde sıralayabiliriz.

Kısacası ; en baştan bu veri seti ile tekrar yapay öğrenme modeli oluşturuyor olsam , küçük veri seti ile bile yüksek başarı oranı sahip olabilen , CNN algoritmasına göre daha az kompleks olan Destek Vektör Makineleri algoritmasını seçerdim.

5. Tirads skoru kullanarak yapacağınız bir sınıflandırma modelinde, modelinizin sonuçları ile, etiketli benign/malign teşhisi arasında uyumsuzluklar varsa, bu uyumsuzlukları nasıl karşılaştırırsınız, çözüm için öneriniz ne olur?

Tirads numaralarına göre tiroid nodülleri 1-2-3-4a-4b-4c-5-6 olarak sınıflandırılır.Yani sınıf sayısı bakımından categorical bir sınıflandırma diyebilir. Diğer taraftan nodülü doğasına göre sınıflandırdığımızda iyi huylu ve kötü hutlu olarak 2 sınıfımız olur. Buradaki sınıflandırma ise binary sınıflandırma diyebiliriz.

Yukarıdaki olayı örnekleyecek olursak ; elimizde aslında Tirad3 sınıfıda ait bir görüntü olsun. Bu görüntü doğasına göre oluşturulan model tarafından iyi huylu olarak sınıflandırılsın. Ancak skora göre oluşturulan model tarafından Tırad2 olarak bir çıktı versin. Tırad3 benign bir nodüldür. Doğası gereği yapılan sınıflandırma için doğru ancak skora göre sınıflandırma da yanlış tespit olur.

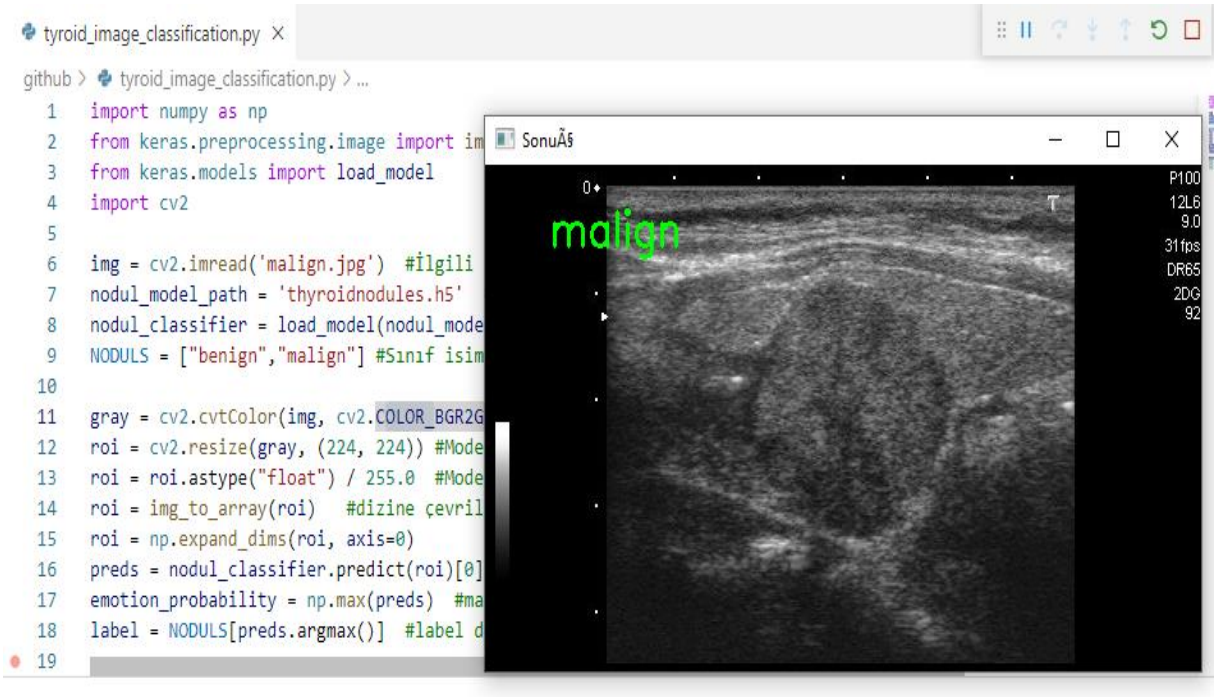
Bu tarz bir hatayı veri seti ile ilişkilendirebiliriz. Örneğin elimizde 1000 tane image içeren veri seti olsun. Bunların 500 tanesi iyi huylu ve 500 tanesi kötü huylu sınıfına ait olsun. Bir sınıf için 500-1000 arası örnek CNN modeli için konuşacak olursak orta decerede bir başarı yakalayabilir.

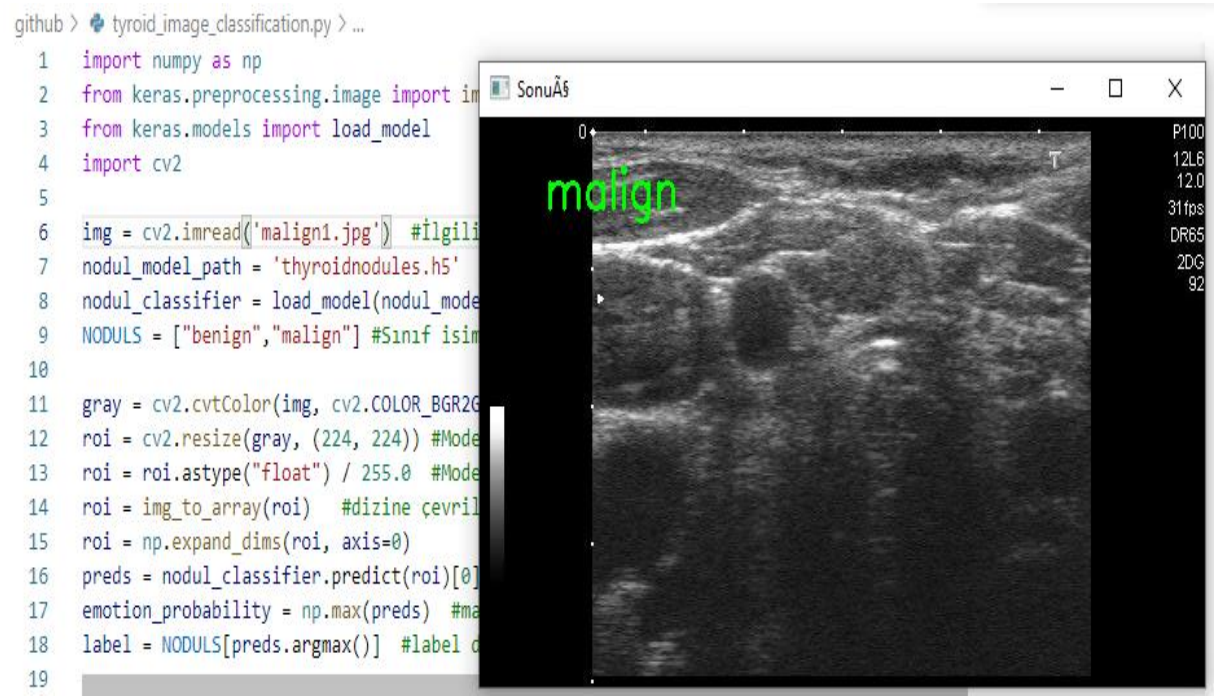
Diğer taraftan Tirads skorlarına göre bir veri dağılımı gerçekleştirdiğimizde 6 sınıf olucaktır. Her sınıf için ise 100-200 arası örnek dağıldığını düşünürsek , bu veri sayıları ile CNN modelimizin başarısının yüksek olacağını bekleyemeyiz.

>>Kısaca veri setinin yelpazelenmesi sonucunda sınıflar için verinin yetersizleşmesi bu duruma sebep olabilecek sebeplerden biridir. Veri sayıları arttırılmalıdır.

>>Tirads skorlarına göre görüntüler sınıflandırıldığında , Tirads2 ve Tirads3 nodülleri birbirine benzeyebilir. Ardışık nodüllerdir. Biri diğerinin daha ilerlemişi gibi düşenebiliriz. Ya da Tirads4a-4b-4c nodülleri arasındaki benzerlikler gibi. Model tarafından ardışık Tiradslardaki resimlerin ayırt edici , ona ait olan özellikleri daha rahat öğrenilebilmesi için görüntü işleme ve makine öğrenimi algoritmaları ile kompleks işlemler yapılabilir.

Aşağıda eğitilen modelin ağırlık dosyası kullanılarak yapılan nodül sınıf tespitine ait ekran görüntüleri yer almaktadır. Verilen her tiroid görüntüsünü doğru sınıflandırmasa bile ortalama olarak traing ve validation verileri üzerinde nodül ayrımı yapabilmektedir. Test amaçlı dışarıdan farklı görüntüler verildiğinde başarı oranı düşmektedir. Uygulama kısmında da görebilildiğimiz üzere genelleme kabiliyeti yüksek olmayan bir modeldir.





Kariyer hedefim olmasına rağmen uzun zamandır, bu konulara uzak kalmıştım. Umarım ilginizi çekebilecek bir çalışma ortaya koymuşumdur. Daha önce mailde belirttiğim gibi aktif olarak başka bir yerde çalıştığım içinde çarşamba günü itibariyle projeye bakmaya fırsat bulabildim. Pazar gününü beklemeden , biter bitmez yolluyorum. Tekrardan anlayışınız için çok teşekkür ederim. İyi çalışmalar dilerim.