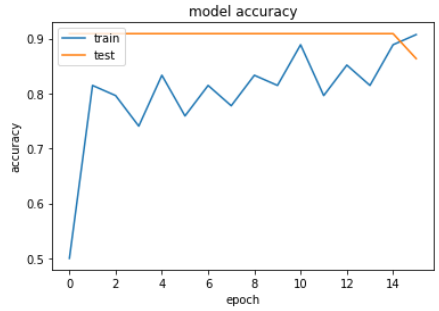
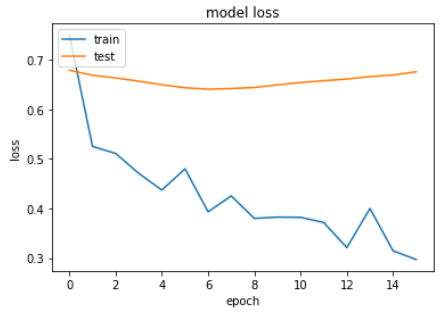
DATASURGERY CASE STUDY

1.Aşama Değerlendirme Cevapları

1. http://cimalab.intec.co/applications/thyroid/index.php adresindeki gruplandırılmış (bening, maling) verisetini kullandım, bu verisetinin öznitelikleri: Number, Age, Sex, Composition, Echogenicity, Margins, Calcifications, Tirads





3. Modelin eğitim metrikleri olması gerektiği gibi görünüyor accuracy yeni epochlarda artış trendi gösteriyor ve loss azalıyor ancak teste baktığımızda testin loss oranı yüksek kalıyor accuracy si ise 12. Epoch dan sonra düşmeye başlamış bunu engellemek için test accuracy fazla yükselmeden modeli train etmeyi durduracak bir callback kullandım.

Daha iyi bir model sonucu elde etmek için daha büyük bir veri seti kullanılabilir. Eksik öznitelikler tamamlanabilir ve cinsiyet, ekojenisite, kalsifikasyon normalize edilebilir. Test veri seti büyütülebilir karşılaştığım durumda test veri seti küçüktü bu da test metriklerinin doğru sonuçlar göstermemesine yol açıyor.

Genel olarak internetten veri seti dışında bulduğum ultrason resimlerini yükleyince yüksek oranla tiroit tipini doğru tahmin ediyor.

4. Bu soruda beni kısıtlayan en büyük etken verilerin az olması bu zaman zaman verileri train ederken overfitting yaşamama sebep oldu. Model görünürde düzgün çalışsa da accuracy ve loss değerleri pek de iyi görünmüyor. Hatta epoch sayısını 50 ye çıkarttığımızda son 2 veya 3 epoch da accuracy 1 ve validation accuracy 0 oluyor. Bunu engellemek için accuracy %95 üzerinde olmasın diye bir callback fonksiyonu oluşturdum.

Ayrıca verilerin ve resimlerin aynı anda işlenmesi gerekmekte ancak daha önce böyle bir deneyimim olmadı. İlk defa görsel bir veri seti ile çalıştım.

Başka bir model kullanmak yerine sadece görselleri işlemek yerine verileri de analiz edip onları da bu klasifikasyon çalışmasına dahil ederdim. Özniteliklere chi-square test uygulanıp hangilerinin etkili olduğu bulunup random forest gibi bir karar ağacı modeli kullanılıp klasifikasyon yapılabilir, bu sayede görüntü işleme haricinde tiroit tipi belirleme için alternatif bir araç olur. Modeli değiştirmek gerekirse eğer XGBoost kullanılabilir. Hem random forest benzeri bir klasifikasyon metodu oluğu için hem de boosting kullandığı için accuracy değerleri de yükselebilir.

5. Bu veri setinde göze çarpan ilk durum veri sayılarındaki eşitsizlik. Bu da sonucun false positive veya false negative olmasına yol açabilir. Bu durumda resimleri eşit sayıda sağlamak gerekebilir. Bunun için az resim olan gruba aynı resimlerin döndürülmüş veya reflekte edilmiş hallerini ekleyebiliriz ve sayılarını eşitleyebiliriz.