KSÜ Mühendislik Bilimleri Dergisi, 28(2), 2025
Araştırma Makalesi

KSU J Eng Sci, 28(2), 2025
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 07.01.2025
Kabul Tarihi : 18.04.2025

Received Date : 07.01.2025
Accepted Date : 18.04.2025

AÇIKLANABİLİR YAPAY ZEKÂ YÖNTEMLERİYLE MR GÖRÜNTÜLERİNDEN BEYİN TÜMÖRÜ TESPİTİ

BRAIN TUMOR DETECTION FROM MRI IMAGES WITH EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS

Muhammet Doğukan İLİ[1*] (ORCID: 0009-0006-1518-0720)
Fatih ÖZYURT[1] (ORCID: 0000-0002-8154-6691)

[1] Fırat Üniversitesi, Yazılım Mühendisliği Bölümü, Elazığ, Türkiye

*Sorumlu Yazar / Corresponding Author: Muhammet Doğukan İLİ, dogukanili25@gmail.com

ÖZET

Bu çalışma açıklanabilir yapay zeka yöntemleri kullanılarak MR görüntülerinden beyin tümörlerinin tespit edilmesini amaçlamaktadır. GradCAM, LIME ve Shapley görselleştirme yöntemleri CNN modellerine entegre edilerek, dört farklı beyin durumu (No Tumor, Glioma, Meningioma, Pituitary) sınıflandırılmıştır. GradCAM yöntemi modelin genel odaklanma alanlarını belirlerken, LIME modelin kararlarını detaylandırmış, Shapley ise modelin genel performansını ve eksikliklerini ortaya koymuştur. Çalışmada bu yöntemlerin birlikte kullanılması, model performansının artırılması için önemli bir yol gösterici olarak sunulmuştur.

Anahtar Kelimeler: Açıklanabilir yapay zeka, derin öğrenme, makine öğrenmesi, GradCAM, beyin tümörü

ABSTRACT

In this study, the aim is to detect brain tumors from MR images using explainable artificial intelligence methods. GradCAM, LIME, and Shapley visualization methods were utilized as part of CNN models in the study. The classification in the model developed during the study was examined under four groups: No Tumor, Glioma, Meningioma, and Pituitary. As a result of the study, GradCAM proved effective in identifying the general focus areas of the model, LIME provided a detailed explanation of the model's decisions, and Shapley revealed the overall performance and shortcomings of the model. The combined use of these techniques enables the provision of more data or the implementation of necessary improvements to ensure the model works more reliably and effectively.

Keywords: Explainable artifical intelligence, deep learning, machine learning, GradCAM, brain tumor

GİRİŞ

Günümüzde yapay zeka teknolojisinin gelişmesiyle birlikte günlük yaşamdan ekonomik faaliyetlere kadar birçok konuda başarının elde edildiği bilinmektedir. Bu başarı ile yapay zeka teknolojilerinin otomotiv sektöründen tıp alanına kadar geniş bir yelpazede aktif olarak kullanılmaya başlandığı söylenebilmektedir (Pannu, 2015). Sağlık alanında yapay zeka, ilaçların keşfedilmesi, hasta takibinin yapılabilmesi, hastalıkların tanılarının konulabilmesi, risk yönetimi gibi çeşitli uygulama alanlarında kullanılmaktadır. Tıbbi görüntüleme de yapay zeka teknolojilerinin aktif olarak kullanılmasının sebebi klinik açıdan etkinlik ve verimliliğin arttırılabilmesidir (Reddy, 2018). Özellikle günümüzde beyin tümörlerinin tespit edilmesi ve sınıflandırılabilmesi açısından yapay zeka teknolojilerinin önemli başarılara imza attığı bilinmektedir. Bu sebeple sağlık alanında özellikle önemli hastalıkların tespit edilmesinde bu teknolojilerin sıklıkla kullanılmaya başlandığı söylenebilmektedir (Manne ve Kantheti, 2021).

Makine öğrenmesi, tıbbi görüntüleme alanında önemli bir potansiyele sahiptir ve MR görüntülerinden beyin tümörlerinin tespit edilmesi amacıyla yaygın olarak kullanılmaktadır (Ellah vd., 2019). Beyin tümörlerini tespit

ToCite: İLİ, M. D. & ÖZYURT, F., (2025). AÇIKLANABİLİR YAPAY ZEKÂ YÖNTEMLERİYLE MR GÖRÜNTÜLERİNDEN BEYİN TÜMÖRÜ TESPİTİ. *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 28(2), 1092-1109.

etmek amacıyla geniş bir MR görüntüleme veri setine ihtiyaç duyulmaktadır. Bu veri setleri genel olarak sağlıklı beyin MR görüntüleri, çeşitli türlerde tümör içeren görüntüler ve klinik etiketleme ile sınıflandırmalar içermektedir (Abdusalomov vd., 2023). Makine öğrenmesi modellerinden doğru sonuçların alınabilmesi için ön işleme gerekmektedir. Bu süreçte farklı cihazlardan kaynaklanan parlaklıklar ve kontrast farkları düzeltilmektedir. Görüntülerin boyutlarının standart hale getirilmesi gerekmektedir. MR görüntülerindeki parazitlerin filtrelenmesi ve tümörlerin manuel olarak işaretlenmesi, segmentasyon modeller için temel gerekliliklerdir (Rahman, 2019). Farklı türlerdeki makine öğrenmesi modelleri MR görüntülerinden beyin tümörünün tespit edilmesinde kullanılmaktadır. Örneğin, derin öğrenme modellerinden CNN'ler görüntü özelliklerini çıkarmada ve sınıflandırmada yaygın olarak kullanılmaktadır (Aamir vd., 2022).

Açıklanabilir yapay zeka, herhangi bir modelin karar verme süreçlerinin daha anlaşılır ve şeffaf olmasını amaçlamaktadır (Pillai, 2024). Görüntü işleme algoritmaları ile evrişimli sinir ağları (CNN) gibi farklı derin öğrenme modellerinde kullanılan açıklanabilir yöntemler, ilgili modelin herhangi bir görüntü üzerinde karar verme işlemini nasıl yaptığının görselleştirmek ve yorumlama amacıyla kullanılmaktadır (Singh vd., 2020). CNN, özellikle görüntü işleme, nesne tanıma ve çeşitli görsel veriler ile ilgili görevlerde kullanılan yapay sinir ağı türüdür. Bu ağlar, biyolojik sinir sisteminin işleyişinden esinlenmekte ve yapılarında bulunan evrişim katmanları aracılığıyla veriden özellik çıkarımı yapabilme becerisine sahiptirler (Kriegeskorte, 2015). CNN'ler genel olarak bilgisayarla görü, doğal dil işleme, tıp ve otonom araçlar başta olmak üzere farklı alanlarda kullanılabilmektedir (Turay ve Vladimirova, 2022). Günümüzde en sık kullanılan yöntemler ise LIME, GradCAM ve Shapley'dir. Yapılan literatür taramaları sonucunda beyin tümörü teşhisinde açıklanabilir yapay zeka teknolojilerinin kullanılarak yüksek başarıların elde edildiği bildirilebilmektedir (Aslan, 2024; Marmolejo ve Kose, 2024; Orman, 2021; Khan vd., 2020).

Yapılan çalışma kapsamında açıklanabilir yapay zeka teknikleriyle MR görüntülerinden beyin tümörünün tespit edilmesi amaçlanmıştır.

Literatür Araştırması

Baran (2024) çalışmasında belirli nöropsikolojik rahatsızlıkların (anksiyete, şizofreni, otizm spektrum bozukluğu, depresyon ve demans) yapay zeka temelli sınıflandırılması ve performanslarının incelenmesi amaçlanmıştır. Çalışmada makine öğrenmesi modelleri olarak Rastgele Orman, k-En Yakın Komşu, XGBoost, LightGBM ve Destek Vektör Makineleri kullanılmıştır. Çalışma sonucunda Destek Vektör'ün %97 doğruluk oranıyla en fazla doğruluk oranının olduğu ortaya konulmuştur. Karakaya (2024) çalışmasında meme kanseri tahmininde makine öğrenmesi algoritmaları ile AUTOML'nin incelenmesi amaçlanmıştır. Çalışma da Lojistik Regresyon, Karar Ağacı, KNN, Naive Bayes, Destek Vektör Makinesi, Rassal Orman, Stokastik Gradyan İniş, Adaboost, XGBoost, LightGBM, Yapay Sinir Ağları kullanılmıştır. Çalışma sonucunda AutoML yöntemlerinin makine öğrenmesi algoritmalarının tamamından daha yüksek doğruluk ve F1 skoruna sahip olduğu ortaya konulmuştur. Gülle vd. (2024) çalışmasında derin öğrenme yöntemleri kullanılarak böbrek hastalıklarının tespiti ve çoklu sınıflandırılması amaçlanmıştır. Çalışmada Classic CNN, ANN, ALEXNET, VGG16, VGG19 ağları ve Poly CNN modelleri kullanılmıştır. Çalışma sonucunda Poly CNN'in %99,94 oranıyla en yüksek doğruluğa sahip olduğu ortaya konulmuştur. Nancy ve Sathyarajasekaran (2024) çalışmasında beyin tümörü segmentasyonu gibi karmaşık görevlerin açıklanabilir yapay zeka modellerinin yorumlanabilirliğinin ve etkinliğinin incelenmesi amaçlanmıştır. Çalışmada MACE, Gradient Shap, LIME, GradCAM ve Guided GradCAM yöntemleri kullanılmıştır. Çalışma sonucunda beyin tümörü görüntülerinin etkili ve yorumlanabilirlik açısından en iyi performans gösteren modelin GradCAM olduğu ortaya konulmuştur. Amin vd. (2024) çalışmasında beyin tümörü teşhisinde GradCAM, LIME ve Shapley yöntemlerinin kullanılabilirlik performanslarının incelenmesini amaçlamıştır. Çalışma sonucunda her üç yönteminde doğrulama doğruluğunun %98 olduğu ve beyin tümörlerinde kullanılabilir olduğu ortaya konulmuştur. Aslan (2024) çalışmasında MR görüntüleri ile beyin tümörlerinin tespit edilmesi amaçlanmıştır. Çalışmada Evrişimli Sinir Ağı, uzun Kısa Süreli Bellek kullanılmıştır. Çalışma sonucunda çalışmada önerilen LSTM-ESA modelinin standart ESA modelinden daha iyi performans gösterdiği ve beyin tümörleri %98,1 doğruluk oranıyla tespit ettiği ortaya konulmuştur.

MATERYAL VE METOT

Yapılan çalışmada açıklanabilir yapay zeka yöntemleri kullanılarak MR görüntülerinden beyin tümörlerinin tespit edilmesi amaçlanmıştır. Bu bölümde çalışmada kullanılan materyal ve metotlar kapsamlıca verilmiştir.

Materyal

Çalışmada beyin tümörü sınıflandırmak için MR görüntüleri içeren veri seti kullanılmıştır. Bu veri seti dört farklı beyin durumunu içermektedir; NoTumor, Glioma, Meningioma ve Pituitary'dir. Mevcut görüntüler, etiketlenmiş olarak sınıflandırılmış ve modelin eğitimi, doğrulaması ve testi için ayrılmıştır.

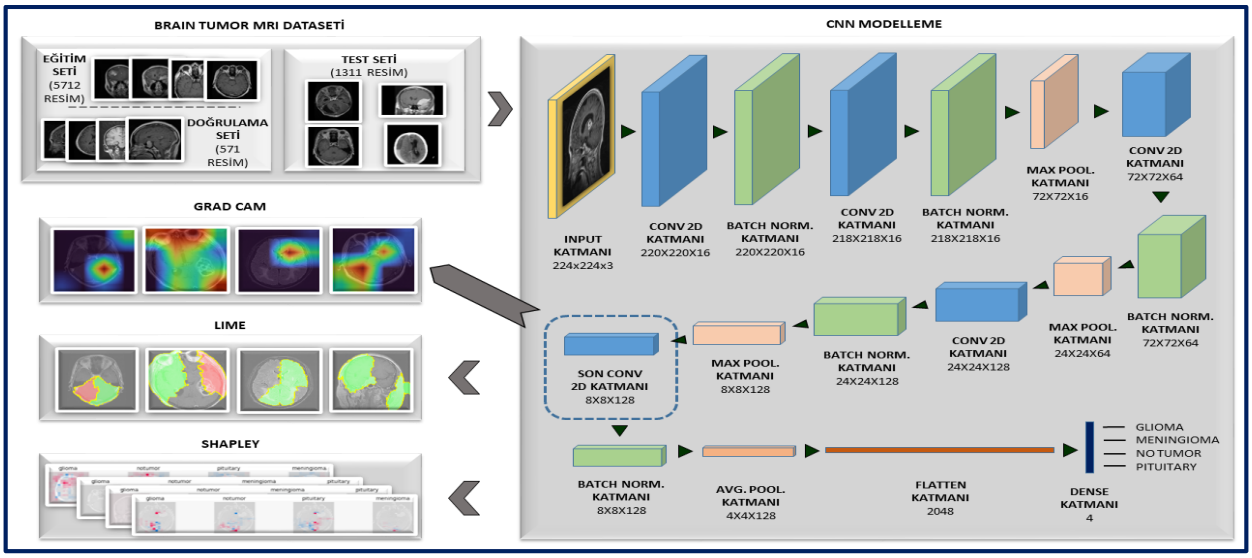
Çalışmanın genel akışı; veri setinin hazırlanması, ön işleme adımlarının uygulanması, derin öğrenme modeli ile sınıflandırmanın gerçekleştirilmesi ve açıklanabilir yapay zeka teknikleri ile model kararlarının görselleştirilmesi aşamalarından oluşmaktadır. İlk olarak, çalışmada kullanılacak MR görüntüleri içeren veri seti elde edilerek etiketlenmiş dört farklı beyin durumu (No Tumor, Glioma, Meningioma, Pituitary) için sınıflandırma yapılabilecek şekilde organize edilmiştir. Çalışmada kullanılan veri setindeki dört sınıftan (Glioma, Meningioma, Notumor ve Pituitary) üçer adet ve toplamda 12 adet resim seçilerek elde edilen sonuçların karşılaştırılabilmesi için görselleştirilmiştir. Daha sonra, veri seti üzerinde çeşitli ön işleme adımları uygulanarak görüntülerin boyutlandırılması, normalizasyonu ve veri artırma teknikleri kullanılmıştır. Ardından, CNN tabanlı bir derin öğrenme modeli eğitilmiş ve test edilmiştir. Modelin sınıflandırma kararlarının açıklanabilirliğini artırmak amacıyla açıklanabilir yapay zeka yöntemlerinden olan Grad-CAM, LIME ve Shapley değerleri kullanılmıştır. Bu yöntemler, modelin hangi bölgelerden yola çıkarak karar verdiğini görselleştirerek, tahmin sürecinin anlaşılmasını sağlamıştır. Grad-CAM yöntemi, derin öğrenme modelinin görüntü üzerinde odaklandığı bölgeleri ısı haritası şeklinde gösterirken, LIME yöntemi giriş görüntüsünü değiştirerek modelin verdiği kararın nedenlerini analiz etmiştir. Shapley değerleri ise modelin her bir özniteliğe olan bağımlılığını ölçerek sınıflandırma sürecindeki önemli özellikleri belirlemiştir. Modellemede kullanılan resimler giriş boyutunda 224x224x3 olarak modelde kullanılmıştır. Modelde sırasıyla Input Layer'ı takip eden birden fazla Convolution Layer, Batch Normalization, Max Pooling, Average Pooling ve Dense layerlardan oluşan katmanlarla model mimarisi kurulmuştur. Modelde aktivasyon fonksiyonu olarak "ReLu", Dense Layer için "Softmax" seçilmiş, Optimizer parametresi olarak "Adam" kullanılmış ve learning rate 0.0005 olarak belirlenmiştir. Loss function olarak "Categorical Cross Entropy", metrikler ise "Categorical Accuracy" olarak atanmıştır. Model eğitiminde 15 epoch ile eğitim gerçekleştirilmiştir.

Çalışmada kullanılan veri seti açık kaynak olarak tutulan Kaggle platformundan alınmıştır (Kaggle, 2025). Glioma (Eğitim Seti:1321, Test Seti:300), Meningioma (Eğitim Seti:1339, Test Seti:306), Notumor (Eğitim Seti:1594, Test Seti:405) ve Pituitary (Eğitim Seti:1457, Test Seti:300) sınıflarına görseller bulunan veri seti çalışma kapsamında eğitim ve test olmak üzere iki parçaya ayrılmıştır. Eğitim setinde toplam 5712, Test setinde 1311 resim bulunmaktadır. Eğitim setinin %10'u (571 resim) model eğitiminde kullanılmak üzere doğrulama seti olarak ayrılmış ve kalan 5141 adet resim eğitim setinde kullanılmıştır. İmgeler modele girmeden önce ön işlem (görüntü boyutlandırma, normalizasyon, veri artırma, veri kümesinin ayrılması) uygulanmıştır. Görselleştirmelerde tutarlılık sağlanması göz önünde bulundurularak her sınıftan (NoTumor, Glioma, Meningioma, Pituitary) üçer tane olacak şekilde test setinden resimler seçilmiştir. Bu resimler her görselleştirme modelinde karşılaştırılabilmesi için sabit olarak tutulmuştur. Çalışmada ele alınan sınıflamada No Tumor; sağlıklı beyin yapısını yani tümör bulunmayan beynin MR görüntüsünü, Glioma; beyin dokusunda gelişen tümörleri yani kötü huylu ve hızlı yayılabilen tümörleri ifade etmektedir. Meningioma; beyin zarlarında gelişen iyi huylu ancak büyüyerek baskı oluşturabilen tümörleri ve Pituitary ise hipofiz bezi tümörlerini ifade etmektedir.

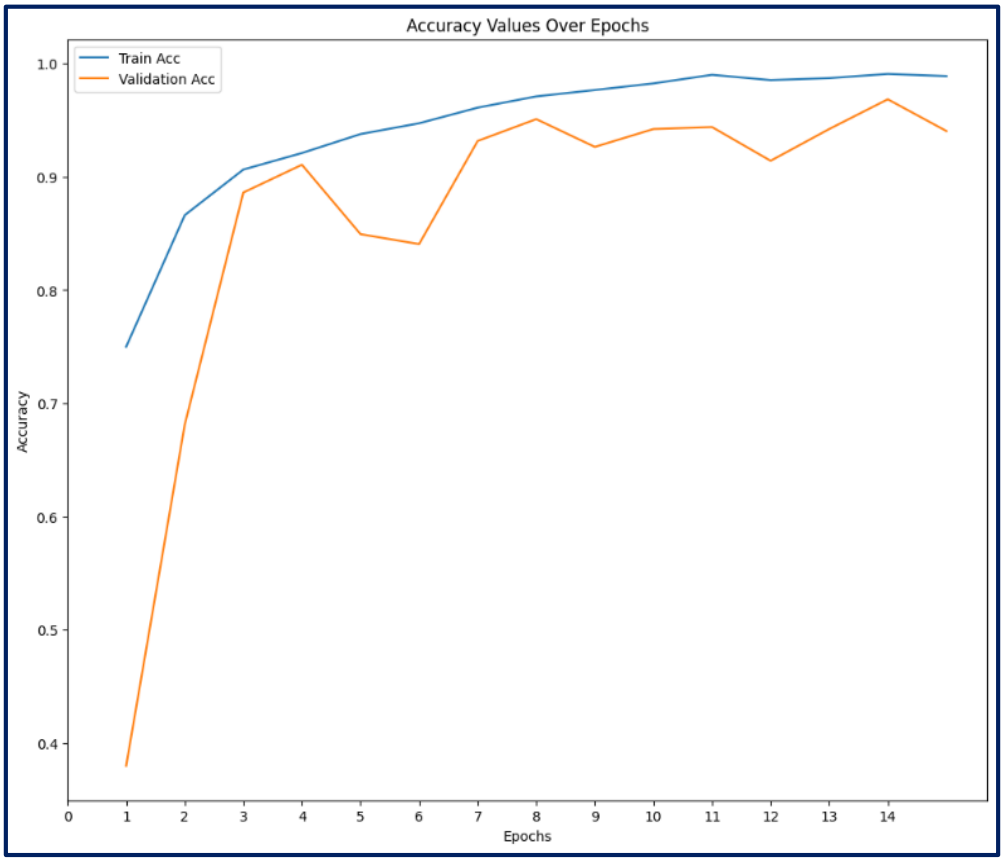
Metot

Bu bölümde çalışmada kullanılan CNN modelin blok diyagramı, epoch doğruluk grafiği, karmaşıklık matrisi ve açıklanabilir yapay zeka araçları olan GradCAM, Shapley ve LIME yöntemleri açıklanarak resimlerin görselleştirme işlemleri yani modelin açıklanma işlemi gerçekleştirilmiştir.
Şekil 1'de eğitimi yapılan CNN modelinin blok diyagramı görülebilmektedir.

Şekil 1'den de görüldüğü gibi diyagramda veri seti CNN modeline girer ve modelin her katmanından sıra ile geçer. Katmanlarda işlem uygulanan resimler dört sınıftan ağırlığı en fazla sınıf olarak tahmin çıktısı alınır. Ayrıca modellemeden sonra açıklanabilir yapay zeka tekniklerinden olan GradCAM, Lime ve Shapley ile görselleştirmeler modellenmiştir. Bu tekniklerden GradCAM doğrudan "Son Convolution Katmanı"ndaki ağırlıklara göre görselleştirmeler sağlarken, Lime ve Shapley tekniklerinde modelin tamamından yararlanılarak görselleştirmeler ortaya konulmuştur. Eğitim 15 epoch ile gerçekleşmiştir. Şekil 2'de eğitimi gerçekleştirilen CNN modeline ait Epoch-Doğruluk grafiği görülebilmektedir.



Şekil 1. Eğitimi Yapılan CNN Modelinin Blok Diyagramı



Şekil 2. Epoch Doğruluk Grafiği

Şekil 2'den de görüldüğü gibi epoch ilerledikçe genel olarak modelin doğruluk eğrilerinin de artış gösterdiği tespit edilmiştir. İlk epoch eğitim seti doğruluğu %75,00 ile başlamış ve on dördüncü epochta %99,09'a ulaşmıştır. Doğrulama setinde ise ilk epoch doğruluğu %38,00 ile başlamış ve on dördüncü epochta %96,85'e ulaşmıştır. Tablo 1'de eğitilen modelin ortalama sonuçlarının final metrikleri verilmiştir.

Tablo 1. Model Metrikleri

Final Model Metrikleri:	
Doğruluk (Accuracy):	0.9443
Kesinlik (Precision):	0.9430
Duyarlılık (Recall):	0.9401
F1 Skoru:	0.9413

Tablo 1'den de görüldüğü gibi eğitilen modelin doğruluğunun %94,43 olduğu, kesinliğinin %94,30 olduğu, duyarlılığının %94,0, olduğu ve F1 skorunun %94,13 olduğu tespit edilmiştir. Şekil 4'de CNN modelinin eğitimi sonucunda elde edilen karmaşıklık matrisi verilmiştir. Karmaşıklık matrisinin sol dikey ekseni gerçek sınıfları, alt yatay eksen ise tahmin edilenleri temsil etmektedir.

Açıklanabilir Yapay Zeka ve Araçları

Açıklanabilir yapay zeka, yapay zeka modellerinin iş işleyişlerini, tahminlerini ve kararlarını insanlar için anlaşılır yapmayı amaçlayan bir yaklaşıma sahiptir (Angelov vd., 2021). Geleneksel yapay zeka modelleri özellikle derin öğrenme gibi karmaşık olanlar genel olarak kara kutu olarak nitelendirilmektedir. Bunun sebebi, nasıl çalıştıklarının açıklanmasında güçlük çekilmesidir. Açıklanabilir yapay zeka ise modellerin şeffaf, güvenilir ve yorumlanabilir olması açısından daha avantajlı olmaktadır (Hassija vd., 2024). Görüntü işleme algoritmalarıyla evrişimli sinir ağları (ESA) gibi çeşitli derin öğrenme modellerinde, açıklanabilir teknikleri, modelin bir görüntü üzerindeki kararını nasıl verdiğini görselleştirmek ve yorumlamak amacıyla kullanılmaktadır. Görüntü işleme ve model açıklanabilirliği açısından en sık kullanılan yöntemler olan GradCAM, Shapley ve LIME detaylıca açıklanmıştır.

GradCAM Yöntemi

GradCAM, derin öğrenme modellerinin özellikle de görüntü işleme amacıyla kullanılan CNN'lerin kararlarının açıklanması için kullanılan bir tekniktir. Herhangi bir modelin tahmin yaparken giriş görüntüsündeki hangi bölgelere odaklanıldığının görselleştirilmesi amacıyla kullanılmaktadır (Selvaraju vd., 2020). GradCAM, CNN modelinin son evrişim katmanındaki aktivasyon haritasını ve gradyanlarını kullanmakta ve modelin tahmin yaparken hangi bölgeleri dikkate aldığını gösteren sıcaklık haritası (heatmap) oluşturmaktadır. Bu haritalar modelin karar alma sürecini daha anlaşılır kılmaktadır (Chattopadhay vd., 2018). GradCAM yönteminde görselleştirme işlemlerinde belirli adımlar izlenmektedir. İlk olarak modelin tahmini yapılmaktadır. Bu aşamada giriş olarak bir görüntü verilmekte ve modelin bir sınıf tahmini yapması sağlanmaktadır. İkinci aşamada hedef sınıfın gradyanları hesaplanmaktadır. Yani modelin tahmin edilen sınıfına yönelik kayıp fonksiyonları hesaplanır. Ayrıca kayıp fonksiyonlarına göre son evrişim katmanındaki aktivasyonların gradyanları da hesaplanmaktadır. Özellik haritası ağırlıklarının hesaplanması adımında; gradyanların ortalaması alınarak her bir özellik haritasının ağırlıkları hesaplanmaktadır. Sıcaklık haritasının oluşturulması aşamasında; ağırlıklar ve aktivasyon haritaları birleştirilmekte ve sınıf aktivasyon haritası oluşturulmaktadır. Son olarak haritanın görselleştirilmesi aşamasında da sınıf aktivasyon haritası, giriş görüntüsüyle birleştirilmekte ve sıcaklık haritası oluşturulmaktadır. Bu harita ile modelin hangi bölgelerde yoğunlaştığı gösterilmektedir (Selvaraju vd., 2020). GradCAM, CNN'de çalışan modellerde kullanılmakta ve genel olarak son evrişim katmanına odaklanmaktadır. Ayrıca derinlikli bilgileri analiz etmektedir. Bu yöntemle görüntülerdeki belirli alanların, modelin belirli bir sınıfı tahmin etmesinde ne denli etkili olduğu görselleştirilmektedir. GradCAM yöntemi, CNN modellerine kolay bir şekilde entegre edilebilmekte ve kod kütüphaneleriyle etkin bir şekilde uygulanabilmektedir. GradCAM yönteminin çeşitli avantajları vardır. Bunlar; karar verme sürecinin açıklanması sebebiyle güvenilir bir uygulama alanı sunması, modellerin değiştirilmeden uygulanabilirlik sağlaması ve görüntülerdeki birden fazla sınıflar için farklı sıcaklık haritalarını oluşturmasıdır (Kumar vd., 2023). GradCAM yönteminin önem ağırlığı formülü Denklem 1'de görülebilmektedir.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

Denklem 1'de görülen "α_k^c"; hedef sınıf c'nin k nöronu önem ağırlığını sresimlemektedir. "y^c"; c'nin softmax katmanından önceki skorunu, "A_{ij}^k"; i, j konumundaki özellik haritasındaki aktivasyonunu k nöronu için sresimlemektedir. "Z" ise normalleştirme sabitidir. Ayrıca yöntemde kullanılan bir diğer fonksiyon olan ReLu fonksiyonunun formülü de Denklem 2'de görülebilmektedir (Selvaraju vd., 2020).

$$L_{GradCam}^C = \text{ReLU}\left(\sum_k \alpha_k^C A^k\right) \quad (2)$$

Denklem 2'deki "$L_{GradCam}^C$"; c sınıfındaki ayırt edici yerelleşme haritasını sresimlemektedir (Selvaraju vd., 2020).

Çalışma kapsamında görselleştirme için son Convolution katmanı kullanılmıştır. GradCAM görselleri için ısı haritaları elde edilmiş ve bu ısı haritaları orijinal resim üzerine bindirilerek süper piksel görseller oluşturulmuştur. Böylelikle doğrudan orijinal resim üzerindeki sonuçlar görülebilmiştir. Sonuçlarda her resim için sırasıyla orijinal

resim, temel ısı haritası ve GradCAM süper piksel ısı haritası bir arada bulunmaktadır. Isı haritasındaki mavi renkli alanlar az öneme sahip, kırmızı renkli alanlar ise modelin karar verme süreci için daha çok öneme sahip alanları ifade etmektedir. Bununla birlikte GradCAM çalışma yapısı nedeniyle, sonuçların son convolution katmanı için elde edilmiş olması modelin her adımındaki karar verme sürecini temsil etmemektedir.

Shapley Yöntemi

Shapley yöntemi, makine öğrenmesi ile oyun teorisinde kullanılan bir yaklaşımdır. Bu yöntem, makine öğrenmesi modellerinin açıklanabilirliğinin artırılması amacıyla geliştirilen araçlardan bir tanesidir. Shapley yöntemi, bir modelin tahmin edilmesinde her özelliğin katkısını adil bir biçimde dağıtmayı amaçlamaktadır. Shapley yöntemi, oyun teorisindeki işbirlikçi oyunlar kapsamında geliştirilmiştir (Verdinelli ve Wasserman, 2024). Bu yöntem, makine öğrenmesi modelinde herhangi bir özelliğin tahmine olan katkısının belirlenmesi amacıyla kullanılmaktadır. Shapley yönteminde her bir özellik "oyuncu" olarak ele alınmakta ve bir takımın toplam kazancını, oyuncular arasında eşit bir biçimde dağıtmayı hedeflemektedir. Özetle oyuncular, modelin özellikleri (bireylerin yaşları, eğitim düzeyleri, gelirleri vs.) iken kazanç ise modelin tahmininin toplam sonucudur. Özelliklerin katkıları, ilgili özelliğin tahmine dahil olma durumu ya da dahil olmama durumu arasındaki farkların ortalaması ile hesaplanmaktadır. Bu yöntemin tüm süreçlerinde bütün özelliklerin kombinasyonları dikkate alınmaktadır. Shapley yönteminin formülü Denklem 3'de görülebilmektedir.

$$\emptyset_i(\vartheta) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [\vartheta(S \cup \{i\}) - \vartheta(S)] \quad (3)$$

Denklem 3'de yer alan "$\emptyset_i(\vartheta)$"; modelin özellik i'sinin Shapley değerini, "N"; modeldeki tüm özelliklerin kümesini, "S"; modeldeki i özelliğini içermeyen alt kümeyi ve "$\vartheta(S)$" ise S'nin tahminini ifade etmektedir (Caelen, 2022). Shapley yöntemi kullanıcılarına çeşitli avantajlar sağlamaktadır. Bunlar; yöntemin farklı model türleriyle çalışabilmesi, katkıların adil bir biçimde hesaplanması için matematiksel garanti sağlaması ve özelliklerin modele etkisinin açık bir şekilde elde edilebilmesidir (Fatima vd., 2008).

Çalışma da Shapley metodu uygulanırken veri setinde sahip olunan dört sınıfın tamamı için sonuçlar elde edilmiştir. En başta orijinal resim, yanında ise dört sınıf için Shapley sonuçlarını içeren resimler bulunmaktadır. Isı haritasındaki mavi renkli alanlar az öneme sahip, kırmızı renkli alanlar ise modelin karar verme süreci için daha çok öneme sahip alanları ifade etmektedir.

LIME Yöntemi

LIME yöntemi, makine öğrenmesi modellerinin tahminlerinin açıklanması amacıyla kullanılan bir tekniktir. Modellerden bağımsız olarak çalışmakta ve bireysel tahminlerin açıklanmasına odaklanmaktadır. Özellikle CNN modellerinin karar alma süreçlerinin anlaşılabilmesinin kolaylaştırılmasında ve açıklanabilirliğinin artırılmasında avantajlar sunmaktadır (Vimbi vd., 2024). LIME, karmaşık modellerin yerel bölgelerde nasıl çalıştığını açıklamak amacıyla basit model oluşturmaktadır. Bu model genel olarak doğrusal regresyon ya da karar ağacı gibi açıklaması kolay modellerdendir. Bu durum sonucunda tahminlerin neden o şekilde yapıldığına dair kullanıcılarına bilgi sağlamaktadır (Juscafresa, 2022). LIME yönteminde tahminlerin açıklanabilmesi için çeşitli adımlar izlenmektedir. İlk olarak hedef tahmininin seçilmesi aşamasında; karmaşık modelden tahmin yapılmaktadır. Ardından girişin bozulması ve veri oluşturma adımı gelmektedir. Bu aşamada orijinal giriş verisi bozulmaktadır yani modelin belirli bir tahmini nasıl yaptığına dair açıklamaların üretilmesi amacıyla orijinal girdinin sistematik bir şekilde değiştirilme işlemi yapılmaktadır. Bu, modelin belirli tahmine ulaşmak için hangi özelliklere dayandığının analiz edilmesine yardımcı olmaktadır. Bu bozulmuş girişlerden veri kümeleri oluşturulmaktadır. Özetle makine öğrenmesi modelleri çok sayıda özelliğe dayalı kararlar vermektedir. LIME yönteminin, modelin hangi özelliklere daha fazla önem verdiğini anlamak için orijinal girdileri değiştirerek (bozarak) alternatif versiyonlar oluşturmaktadır. Bu bozulan girişler, modelin nasıl tepki verdiğini analiz etmek amacıyla kullanılmaktadır. Üçüncü aşamada karmaşık modelin çıktıları kullanılmaktadır. Yani bozulan veri noktaları karmaşık modele verilmekte ve tahminleri alınmaktadır. Dördüncü aşamada yerel doğrusal model eğitilmektedir. Bu aşamada orijinal verilerin çevrelerinde basit ve açıklanabilir model eğitilmektedir. Bu basit model, karmaşık modelin yerel davranışını yakalamaya çalışmaktadır. Son olarak özelliklerin katkılarının belirlenmesi yapılmaktadır. Bu aşamada basit modelin ağırlıkları ya da karar kuralları, karmaşık modelin tahmininde hangi özelliklerin önemli olduğunu göstermektedir (Garreau ve Mardaoui, 2021). LIME yöntemi, herhangi bir makine öğrenmesi modeliyle çalışabilmektedir. Ayrıca tahmin edilen veri noktaları için modelin yerel davranışları açıklanabilir ve karmaşık modellerden bağımsız olarak yalnızca seçilen tahminlerin çevrelerindeki bölgeleri analiz edebilmektedir. LIME yönteminin de çeşitli avantajları vardır. Bunlar;

hangi özelliklerin tahmin üzerinde etkili olduğunun kolay bir şekilde anlaşılabilir olması, çeşitli makine öğrenmesi modelleriyle entegre biçimde çalışılabilmesi ve yapılan tahminlerin neden yapıldığının ve hangi özelliklerin daha fazla katkıda bulunduğunun açıklanabilir olmasıdır (Bilekyiğit, 2022).

Çalışmada LIME görselleri için önce orijinal resim üzerinde sonuç alınmış daha sonra ısı haritaları elde edilmiş ve bu ısı haritaları orijinal resim üzerine bindirilerek süper piksel görseller oluşturulmuştur. Böylelikle doğrudan orijinal resim üzerindeki sonuçlar ısı haritası değerlerine göre görülebilmiştir. Sonuçlarda her resim için sırasıyla orijinal resim, LIME sonucu, temel ısı haritası ve LIME süper piksel ısı haritası bir arada bulunmaktadır. LIME sonucunu içeren ikinci sıradaki resimlerde (LIME Explanation) sarı kenar çizgileri içinde yapılan segmentasyon modelin karar verme sürecini en çok etkileyen üç alanı ifade etmektedir. Segmentasyon için quickshift algoritması kernel_size=9 parametresiyle kullanılmıştır. Bu üç alan yeşil veya kırmızı renkle ifade edilebilmektedir. Bu üç alandan birbirine bitişik olan aynı renkteki bölgeler birleştirilmiş şekildedir ve sonuçta bu sebeple üçten daha az alan görülebilmektedir. Bu bazı resimlerde tamamen yeşil yani olumlu etkileyen iki alanı birleşmiş şekilde, bazı resimlerde ise yeşil ve kırmızı olarak olumlu ve olumsuz etkileyen alanları ifade etmektedir. Bunların yanındaki LIME Isı Haritası (Heatmap) farklı renklerden oluştuğu için bu üç alan renk aralığı birbirine yakın dahi olsa seçilebilmektedir.

Nancy ve Sathyarajasekaran (2024) çalışmasında beyin tümörü segmentasyonu gibi karmaşık görevlerin açıklanabilir yapay zeka modellerinin yorumlanabilirliğinin ve etkinliğinin incelenmesi amaçlanmıştır. Çalışmada MACE, Gradient Shap, LIME, GradCAM ve Guided GradCAM yöntemleri kullanılmıştır. Çalışma sonucunda beyin tümörü görüntülerinin etkili ve yorumlanabilirlik açısından en iyi performans gösteren modelin GradCAM olduğu ortaya konulmuştur. Amin vd. (2024) çalışmasında beyin tümörü teşhisinde GradCAM, LIME ve Shapley yöntemlerinin kullanılabilirlik performanslarının incelenmesini amaçlamıştır. Çalışma sonucunda her üç yönteminde doğrulama doğruluğunun %98 olduğu ve beyin tümörlerinde kullanılabilir olduğu ortaya konulmuştur. Gaur vd. (2024) çalışmasında MR görüntü veri setlerinin kullanılarak Menenjiyom, Gliom ve Hipofiz tümörleri gibi beyin tümörlerinin ayrık alt türlerinin tahmin edilmesinin derin öğrenme modelleriyle uygulanabilirliğinin incelenmesi amaçlanmıştır. Çalışma da Shapley, LIME ve GradCAM yöntemleri kullanılmıştır. Çalışma sonucunda CNN modelinin %94,64 doğruluğunun olduğu ve yöntemlerin kullanılabilir olduğu ortaya konulmuştur.

BULGULAR

Bu bölümde GradCAM, LIME ve Shapley yöntemlerine yönelik No Tumor, Glioma, Meningioma ve Pituitary sınıfına ait modelin elde ettiği görseller verilmiştir. Çalışmada çok sınıflı karmaşıklık matrisi kullanılmış olup model değerlendirme metriği olarak Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru kullanılmıştır. Denklem 4'de Doğruluk formülü görülebilmektedir.

$$\text{Doğruluk} = \frac{Doğru\ tahminlerin\ sayısı}{Toplam\ tahmin\ sayısı} \quad (4)$$

Denklem 4'den de görüldüğü gibi Doğruluk oranı; doğru tahminlerin sayısının toplam tahmin sayısına bölümüyle bulunmaktadır. Denklem 5'de Hassasiyet formülü görülebilmektedir.

$$\text{Kesinlik} = \frac{Gerçek\ pozitifler}{Gerçek\ pozitifler + Yanlış\ pozitifler} \quad (5)$$

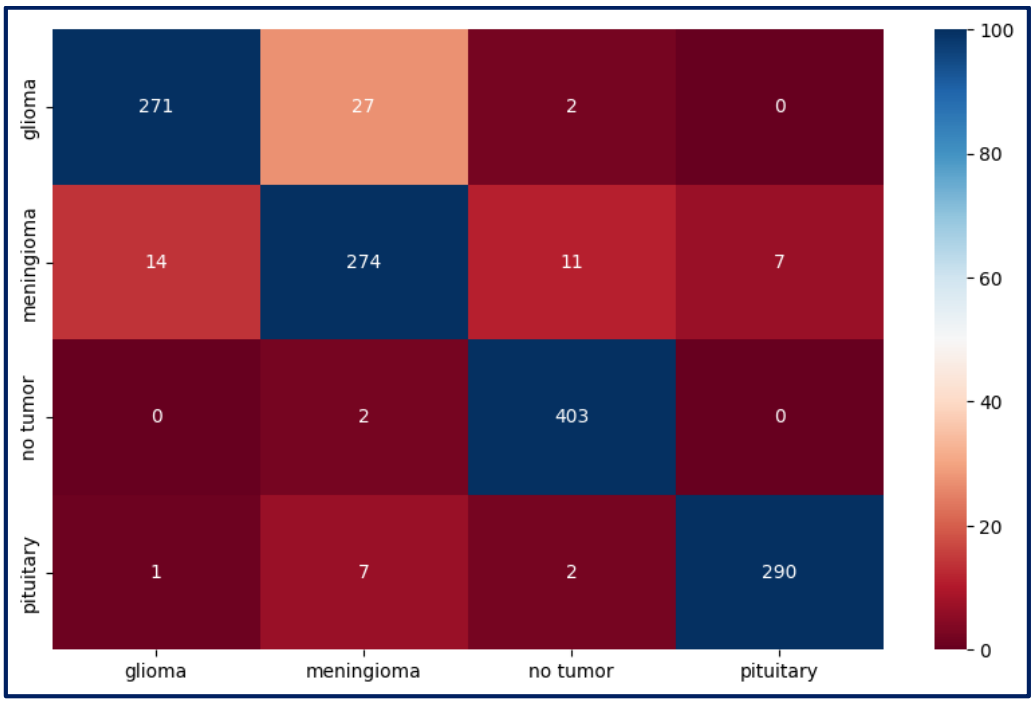
Denklem 5'den de görüldüğü gibi Kesinlik oranı gerçek pozitiflerin gerçek pozitifler ile yanlış pozitiflerin toplamına bölünmesiyle bulunmaktadır. Denklem 6'da Duyarlılık formülü görülebilmektedir.

$$\text{Duyarlılık} = \frac{Gerçek\ pozitifler}{Gerçek\ pozitifler + Yanlış\ negatifler} \quad (6)$$

Denklem 6'dan da görüldüğü gibi Duyarlılık; gerçek pozitiflerin gerçek pozitifler ile yanlış negatiflerin toplamına bölünerek bulunmaktadır. Denklem 7'de F1 skoru formülü görülebilmektedir.

$$\text{F1 skoru} = 2 \times \frac{Hassasiyet x Duyarlılık}{Hassasiyet + Duyarlılık} \quad (7)$$

Denklem 7'den de görüldüğü gibi F1 skoru; hassasiyet ile duyarlılığın çarpımının, hassasiyet ile duyarlılığın toplamına bölünmesi ve iki ile çarpılmasıyla bulunmaktadır. Şekil 3'de CNN modelinin eğitimi sonucunda elde edilen karmaşıklık matrisi verilmiştir. Bu sonuçlar test seti kullanılarak elde edilmiştir. Resimlere ait gerçek sınıfları sol dikey eksen, modelin tahmin ettiği sınıfları ise alt yatay eksen temsil etmektedir.



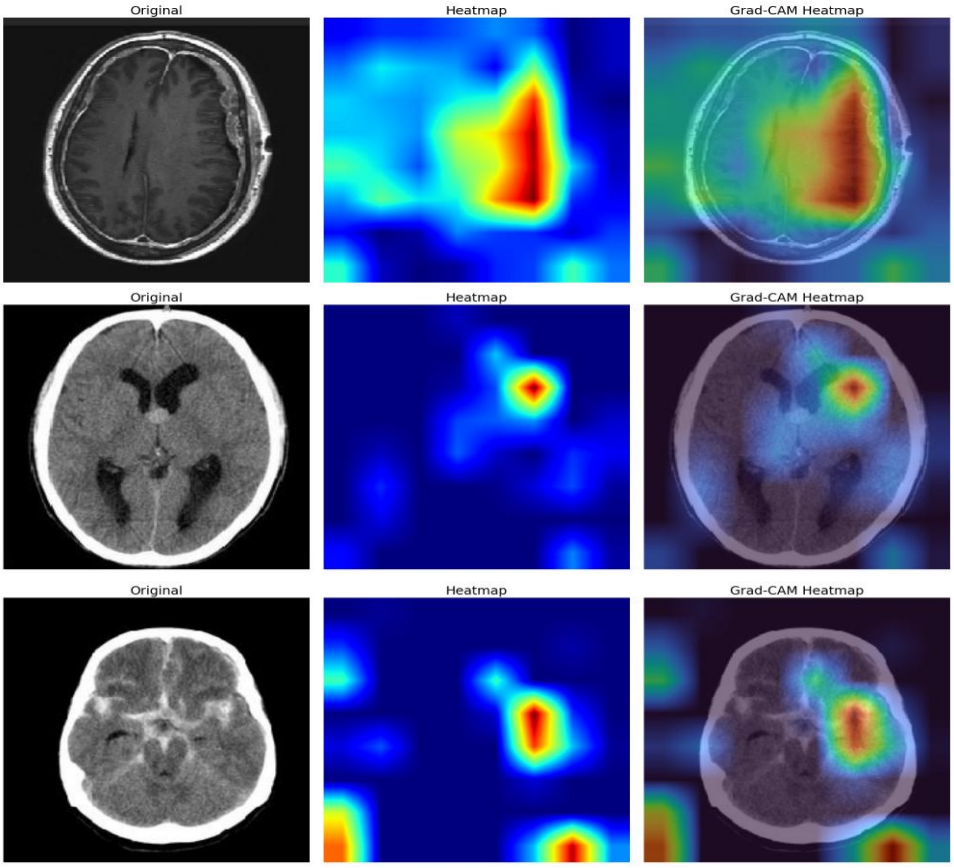
Şekil 3. Karmaşıklık Matrisi

Şekil 3'den de görüldüğü gibi Glioma sınıfına ait resimlerden 271 adet doğru şekilde tespit edilirken, bu sınıftan 27 adet görsel "Meningioma", 2 adet görsel ise "No tumor" sınıfı olarak, Meningioma sınıfına ait resimlerden 274 adet doğru şekilde tespit ederken, bu sınıftan 14 adet görsel "Glioma", 11 adet görsel "No tumor", 7 adet görsel ise "Pituitary" sınıfı olarak, No Tumor sınıfına ait resimlerden 403 adet doğru şekilde tespit edilirken, bu sınıftan 2 görsel "Meningioma" sınıfı olarak, Pituitary sınıfına ait resimlerden 290 adet doğru şekilde tespit edilirken, bu sınıftan 1 görsel "Glioma", 7 görsel "Meningioma", 2 görsel "No tumor" sınıfı olarak tahmin edildiği tespit edilmiştir.

GradCAM Yöntemi ile Görselleştirme

Çalışmada CNN eğitiminden sonra elde edilen sonuçlar bağlamında, modelden elde edilen tahminler öncelikle GradCAM yöntemi kullanılarak görselleştirilmiştir. GradCAM sonuçları Şekil 4'de gösterilmiştir.

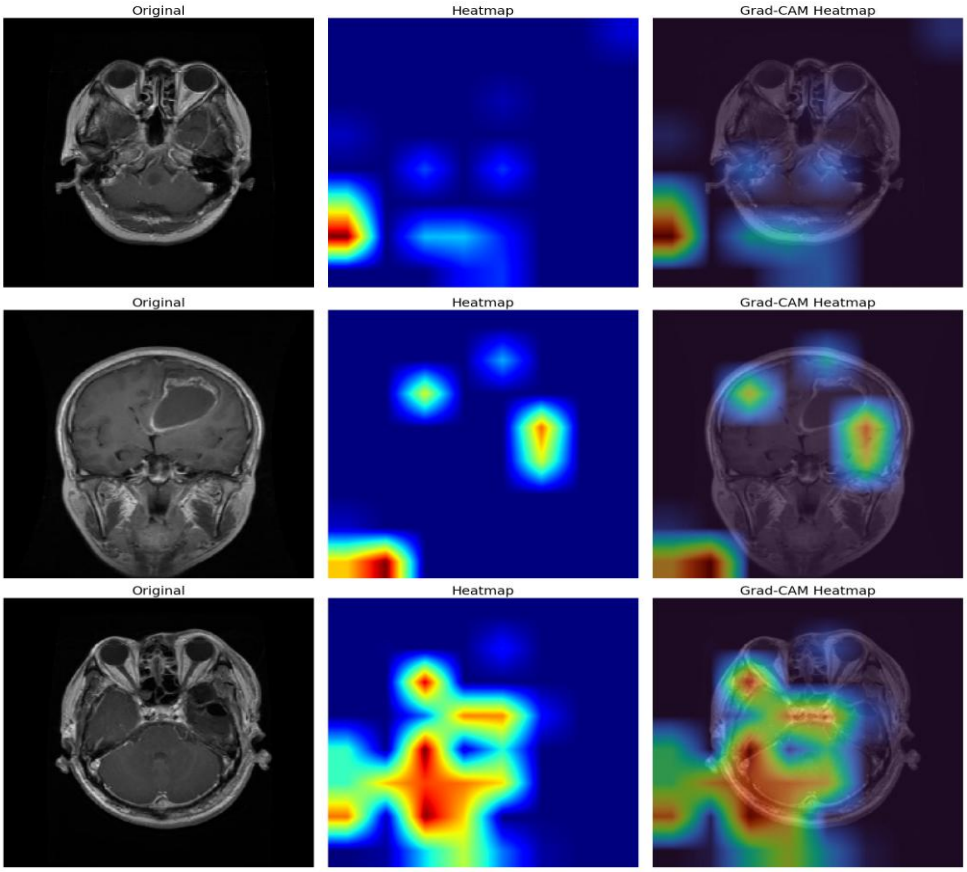
Şekil 4'den de görüldüğü gibi ilk görselde No Tumor sınıfına ait bulgular verilmiştir. Bu modelin geniş bir alanında mavi renkte aktivasyon gösterdiği ve daha kısa fakat yoğun kırmızılı bir alanda yoğun aktivasyon gösterdiği görülmektedir. Geniş mavi alanlar modelin tümör bulunmayan beyin dokusunu geniş bir alanda değerlendirdiğini, dar kırmızı alanlar ise bu bölgede tümör olasılığı ile ilgili yüksek önem taşıyan aktivasyonları ifade etmektedir. Bununla birlikte bu sonuç modelin No Tumor sınıfı için büyük oranda doğru değerlendirme yaptığını göstermektedir. İkinci No Tumor sınıfına ait görselde, yine mavi rengin geniş bir alanda etkin olduğu ve daha küçük bir beyin bölgesinde ise kırmızı rengin ağırlıkta olduğu görülmektedir. Kırmızı aktivasyonun dar bir bölgede yoğunlaşması, modelin potansiyel anormal bölgeleri doğru bir şekilde tespit etme çabasında olduğunu ancak bu bölgenin küçüklüğü göz önüne alındığında, yanlış pozitif tespitlere de yol açabileceğini düşündürmektedir. Üçüncü No Tumor sınıfına ait görselde geniş bir alana yayılmış birkaç mavi aktivasyon bölgesi ve daha küçük birkaç alanda kırmızı aktivasyon bölgesi gözlemlenmektedir. Bu, modelde aktivasyon sırasında farklı tümör tipleri için görselin farklı alanlarının değişken aktivasyonlar ürettiğini göstermektedir. Bu durum, modelin beyin boyunca geniş bir alanı taradığını, No Tumor sınıfını ayırt ettiğini ve potansiyel anomalilerin farkında olduğunu göstermektedir. Şekil 5'de Glioma tümör sınıfından 3 görsel görülebilmektedir.



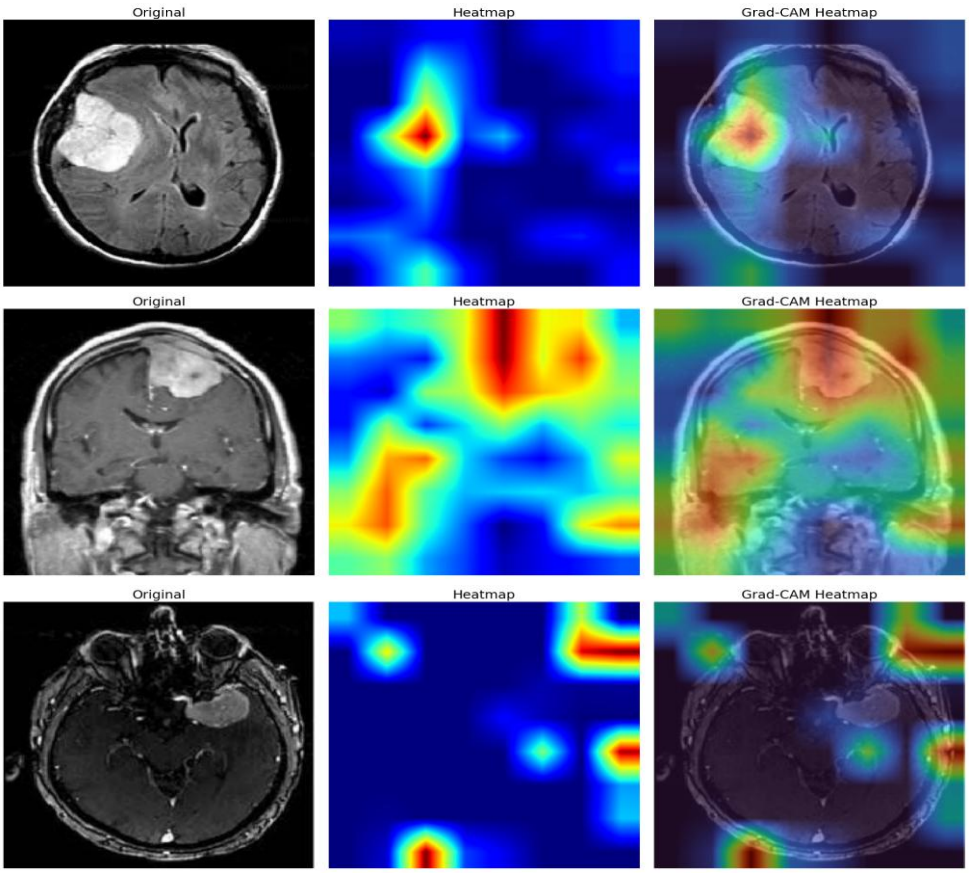
Şekil 4. GradCAM Yöntemi Görselleştirmesi (No Tumor)

Şekil 5'den de görüldüğü gibi ilk Glioma sınıfına ait görselde, modelin aktivasyonları beynin alt ve orta kısımlarında mavi renkte yoğunlaşırken, beynin dışında daha yüksek bir kırmızı aktivasyon gözlemlenmiştir. Bu, modelin beynin dışında tümör olması muhtemel olmayan bir alana odaklanmakta ve dolayısıyla yanlış çalıştığını gösteren yanlış pozitif bir aktivasyon vermektedir. Hatalı aktivasyon, modelin bu resimde yanlış tahminde bulunabileceğini göstermektedir. İkinci Glioma sınıfına ait görselde, modelin beynin üst kısımlarında mavi ve kırmızı renklerde aktive olduğu, ancak beynin dışında bir bölgede de kırmızı aktivasyonun yoğun olduğu bir alan bulunduğu görülmektedir. Bu modelin tümörün olmaması gereken alanlara odaklandığını gösteriyor ve bu görüntü için yanlış sınıflandırmaya olan eğilimini açıklıyor. Hatalı aktivasyon, modelin bu resimde yanlış tahminde bulunabileceğini göstermektedir. Üçüncü Glioma sınıfına ait görselde, modelin beynin sol kısmında ve bunun dışındaki bazı alanlarda büyük miktarlarda kırmızı renkte aktivasyon gösterdiği görülmektedir. Modelin tümörün de bulunabileceği geniş bir beyin alanını yeterince değerlendirdiği gözlemlenebilir; ancak çevredeki alanlara doğru düşük seviyeli aktivasyonlar olduğu dikkat çekmektedir. Bu, modelin geniş alanları glioma sınıfı için değerlendirirken bazı dış bölgelerde gereksiz aktivasyonlar gösterdiğini ifade eder. Şekil 6'da 3 farklı Meningioma sınıfına ait sonuç çıktısı görülebilmektedir.

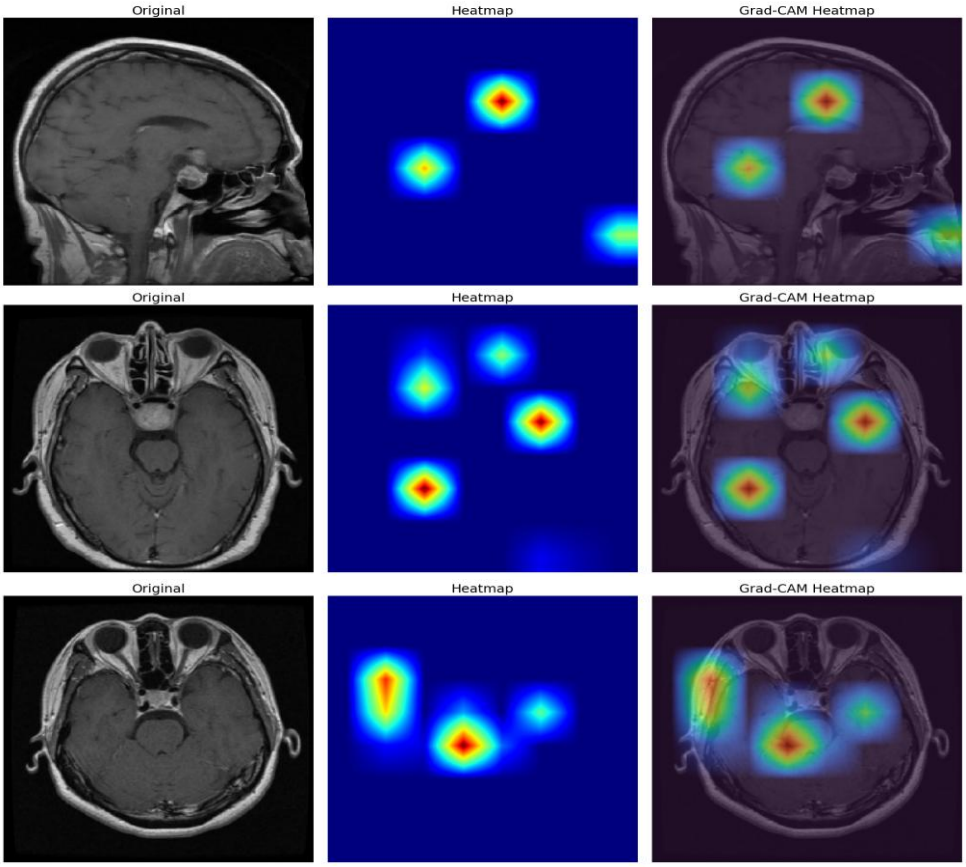
Şekil 6'dan da görüldüğü gibi ilk Meningioma sınıfına ait görselde, beynin zar bölgesine yakın bir alanda modelin aktivasyonlarının yoğunlaştığı görülmektedir Modelin tümör tespiti için doğru bir alan belirleyebildiğine dair bu kırmızı aktivasyonlar, söz konusu görselde etkili bir sınıflandırma tahmini yapabileceğini göstermektedir. İkinci Meningioma sınıfına ait görselde, model beyinin çevresinde, zar bölgesinde kırmızı renkte yoğun aktivasyon göstermiştir. Bu durum modelin tümör tespitinde doğru alanları değerlendirdiğini göstermektedir. Üçüncü Meningioma sınıfına ait görselde, modelin beynin etrafında ve hatta dışında aktive olduğu bazı alanları kırmızı renkte göstermektedir. Burada modelin tümörün bulunabileceği ve tümör olmaması gereken dış alanlara da odaklandığını görülmektedir. Bu hatalı aktivasyonların bulunması, modelin bu resimde yanlış tahminde bulunabileceğini göstermektedir. Şekil 7'de 3 farklı Pituitary tümör sınıfına ait sonuç çıktısı görülebilmektedir.



Şekil 5. GradCAM Yöntemi Görselleştirmesi (Glioma Tümör)



Şekil 6. GradCAM Yöntemi Görselleştirmesi (Meningioma Tümör)



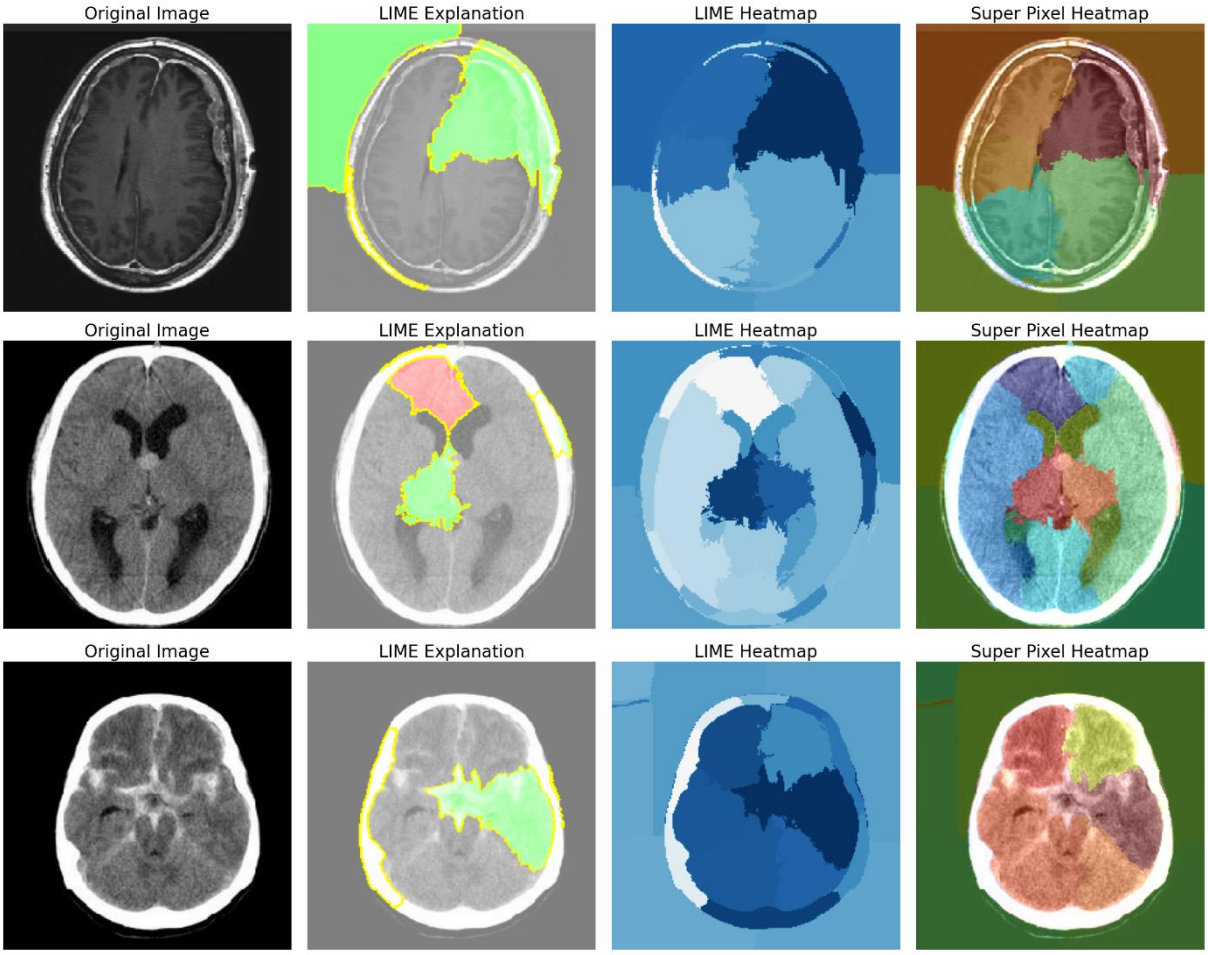
Şekil 7. GradCAM Yöntemi Görselleştirmesi (Pituitary Tümör)

Şekil 7'den de görüldüğü gibi ilk Pituitary sınıfına ait görselde, modelin beynin orta ve alt bölümlerinde kırmızı, ayrıca burun bölgesinde mavi renkte aktivasyon gösterdiği gözlemlenmiştir. Modelin beynin iç bölgesinde doğru aktivasyonlar göstermesi, tümör tespiti için uygun alanlara odaklandığını ortaya koymaktadır. Ayrıca, burun bölgesinde mavi aktivasyon göstermesi, modelin bu görselde gereksiz bölgelere önem vermediğini göstermektedir. İkinci Pituitary sınıfına ait görselinde, modelin sol ve sağ beyin bölgelerinde kırmızı, göz çevresinde ise mavi renkte aktivasyonlar gözlemlenmiştir. Modelin beynin iç bölgelerinde doğru aktivasyonlar göstermesi, tümör tespiti için uygun alanları değerlendirdiğini gösterirken, göz çevresinde düşük öneme sahip mavi aktivasyonların olması, modelin tümör bulunmaması gereken bölgelerde daha az odaklandığını, yani dikkatli bir sınıflandırma yapabileceğine işaret etmektedir. Üçüncü Pituitary sınıfına görselde, modelin beynin orta ve sol bölümlerinde kırmızı, sağ tarafında ise mavi renkte aktivasyon gösterdiği tespit edilmiştir. Bu aktivasyonlar, modelin tümör bulunabilecek beyin bölgelerinde doğru alanlara odaklandığını ve bu görselde tutarlı bir sınıflandırma gerçekleştirebileceğini ortaya koymaktadır.

LIME Yöntemi ile Görselleştirme

Şekil 8'de Lime yöntemi No Tumor sınıfına ait 3 farklı bulgu görülebilmektedir.

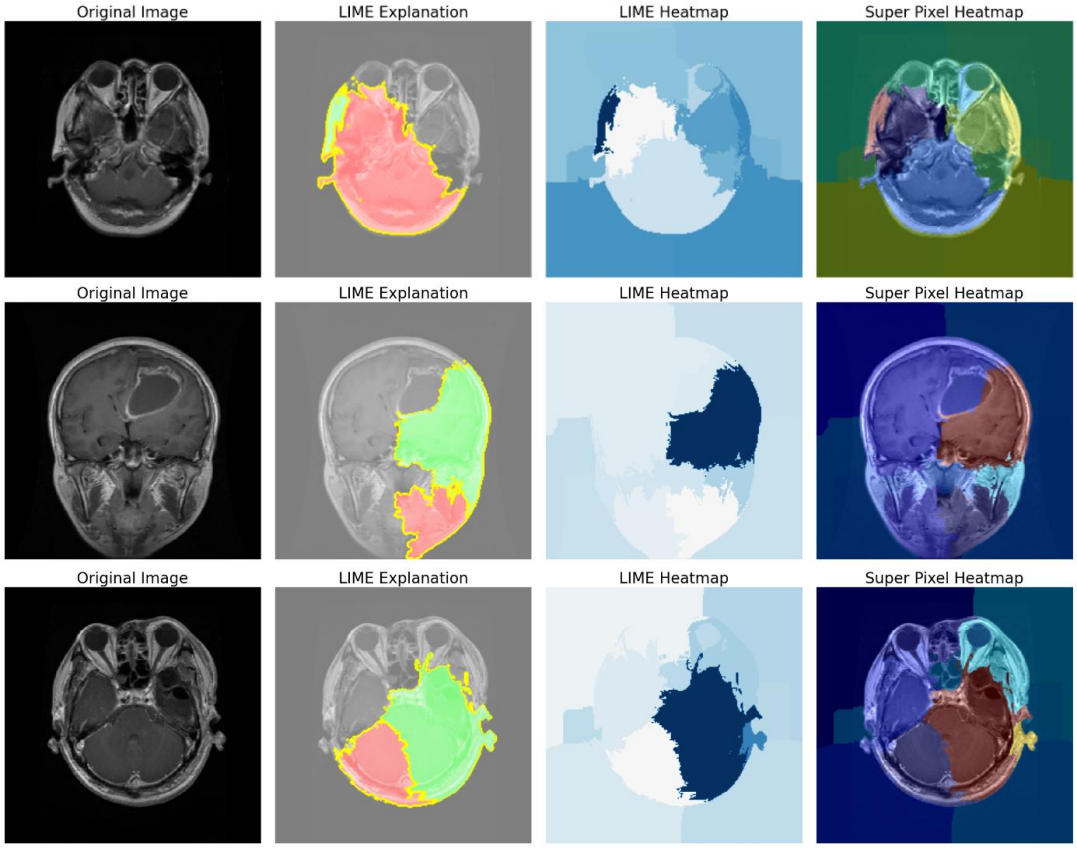
Şekil 8'den de görüldüğü gibi ilk No Tumor sınıfına ait görselde, modelin aktivasyonlarının beyinin sağ bölümünde, sol alt kenarda ve beynin dışında sol bölümde yeşil renkte yoğunlaştığı gözlemlenmiştir. Modelin geniş bir alanda aktivasyon göstermesi, tümör bulunmayan bölgelerde de değerlendirme yaptığını ortaya koymaktadır. Bu durum, modelin gereksiz bölgelere odaklanarak yanlış pozitif tahmin yapabileceğini göstermektedir. İkinci No Tumor sınıfına ait görselde, modelin beyinin orta bölümünde ve sağ kenarda yeşil, beynin üst bölümünde ise kırmızı renkte aktivasyon gösterdiği tespit edilmiştir. Bu aktivasyonların, modelin tümör bulunabilecek alanlarda odaklandığını göstermesi, sınıflandırmanın doğru tahmin edilebileceğini göstermektedir. Üçüncü No Tumor sınıfına ait görselde, modelin beyinin sağ bölümünde, sağ ve sol kenarlarda yeşil renkte aktivasyonlar gösterdiği görülmüştür. Bu aktivasyonların, modelin potansiyel tümör alanlarına odaklandığını göstermektedir. Bu durum farklı tümör tipleri için modelin görselde farklı alanları değerlendirdiğini göstermektedir. Şekil 9'da Glioma sınıfa ait LIME görselleştirmeleri görülebilmektedir.



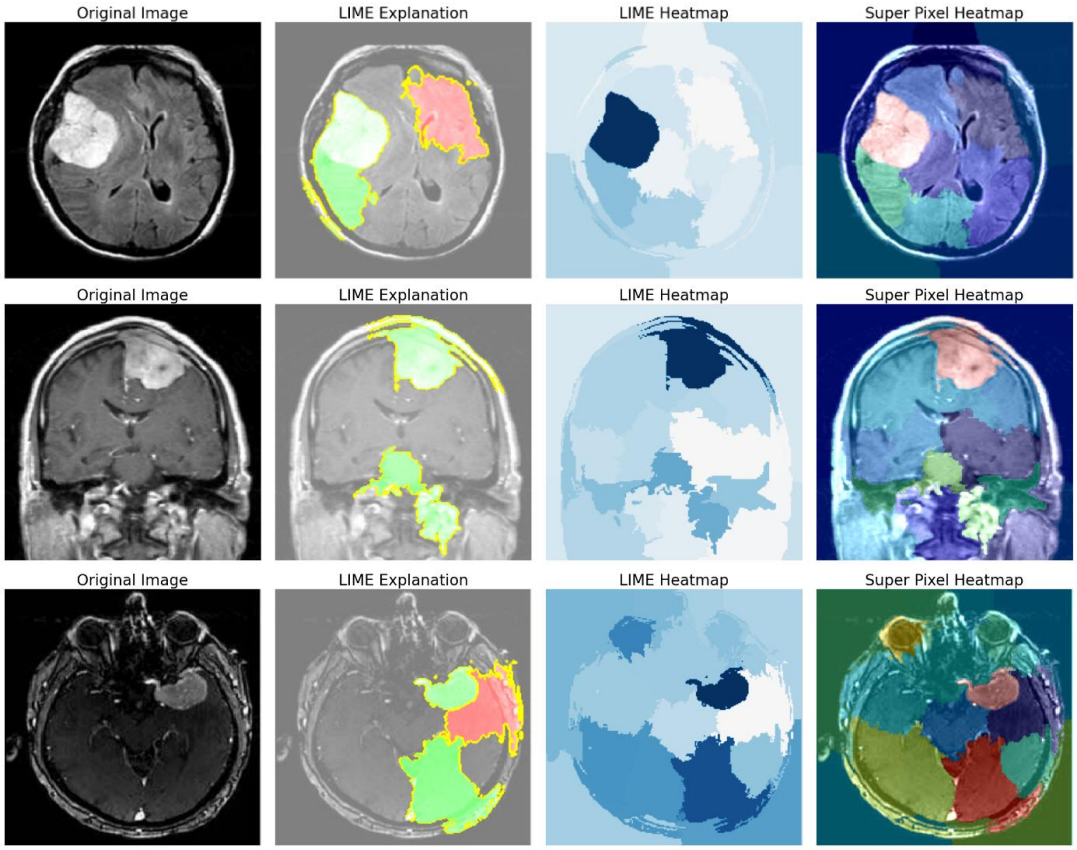
Şekil 8. LIME Yöntemi Görselleştirmesi (No Tumor)

Şekil 9'dan da görüldüğü gibi ilk Glioma sınıfına ait bir görselde, modelin beynin sol kenar bölümünde yeşil, beynin büyük bölümünde ise kırmızı renkte aktivasyonlar gösterdiği tespit edilmiştir. Bu durum, modelin tümör bulunabilecek alanlarda düşük öncelikli aktivasyonlar oluşturduğunu ve bu görselde hatalı bir şekilde odaklandığını göstermektedir. Kırmızı aktivasyonun geniş bir alana yayılması, modelin bazı bölgelerde gereksiz yüksek önem verdiğini ve bu nedenle doğruluk oranının olumsuz etkilenebileceğini ortaya koymaktadır. İkinci Glioma sınıfına ait görselde, modelin beyinin sağ bölümünde ve sağ alt kısmında yeşil, ayrıca sağ alt bölümde kırmızı renkte aktivasyon gösterdiği gözlemlenmiştir. Model, tümör bulunabilecek alanlara odaklanırken, alt bölgedeki düşük önem seviyesine sahip aktivasyonlar ile tümör bulunma ihtimali az olan bölgelerde de odaklanma göstermektedir. Bu durum, modelin genel olarak doğru bir sınıflandırma yapabileceğini göstermektedir. Üçüncü Glioma sınıfına ait görselde, modelin beyinin sağ bölümünde ve sağ kenarda yeşil, sol alt bölümünde ise kırmızı renkte aktivasyonlar gösterdiği gözlemlenmiştir. Modelin, tümör bulunabilecek alanlara odaklanma eğiliminde olduğu anlaşılmaktadır. Bu durum modelin bu görselde tümör tespiti için belli bir alana odaklandığını göstermektedir. Şekil 10'da Meningioma sınıfına yönelik yapılan LIME görselleştirme sonuçları verilmiştir.

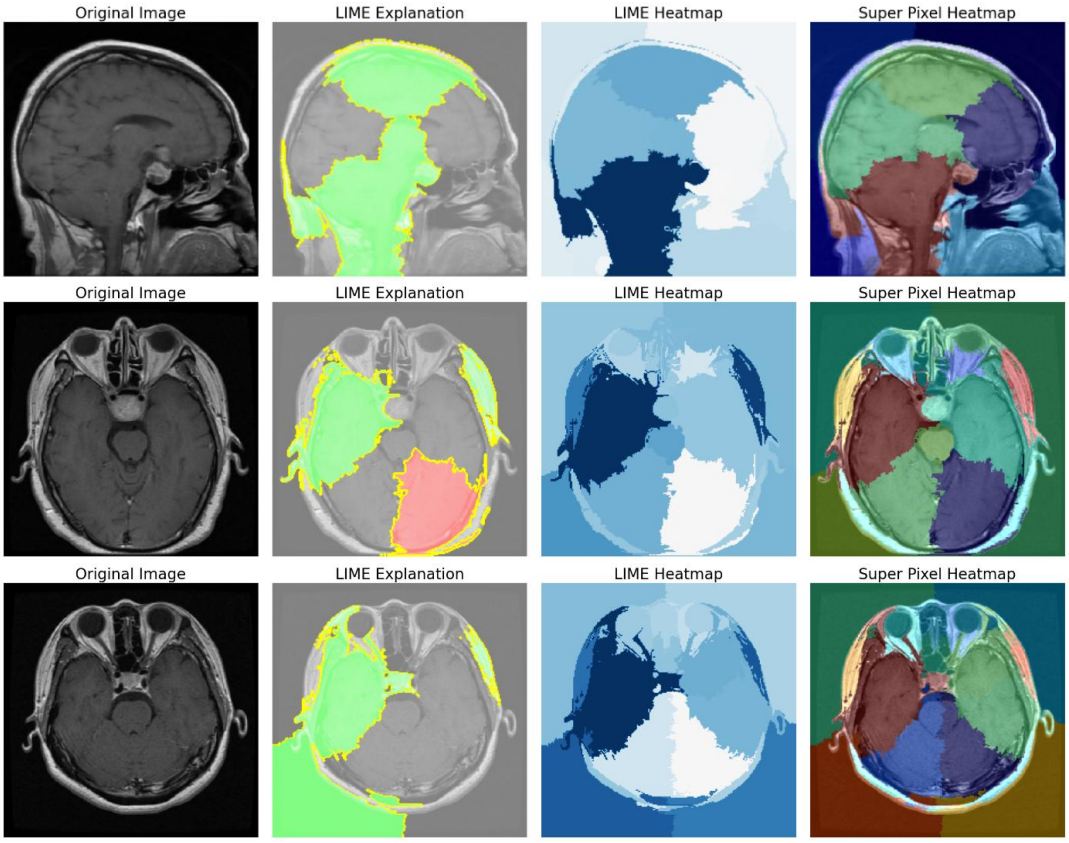
Şekil 10'dan da görüldüğü gibi ilk Meningioma sınıfına ait görselde, modelin beyinin sol alt ve sol üst bölümlerinde yeşil, sağ üst bölümünde ise kırmızı renkte aktivasyon gösterdiği tespit edilmiştir. Modelin, tümör bulunabilecek sol bölgelere yoğunlaşması doğru bir odaklanmayı gösterirken, sağ üst bölümdeki düşük öncelikli kırmızı aktivasyonlar, modelin bu bölgede de dikkat verdiğini ancak daha az öneme sahip olduğunu ortaya koymaktadır. Bu durum modelin bu görselde tümör tespiti için doğru alana odaklandığını göstermektedir. İkinci Meningioma sınıfına ait görselde, modelin beyinin üst sağ ve orta alt bölümlerinde yeşil renkte aktivasyonlar gösterdiği gözlemlenmiştir. Bu durum, modelin tümör bulunabilecek doğru alanlara odaklandığını ve bu görselde tutarlı bir sınıflandırma gerçekleştirdiğini göstermektedir. Üçüncü Meningioma sınıfına ait görselde, modelin beyinin sağ orta ve sağ alt bölümlerinde yeşil, sağ bölümünde ise kırmızı renkte aktivasyon gösterdiği tespit edilmiştir. Modelin sağ bölgedeki düşük öncelikli kırmızı aktivasyonlara rağmen, tümör bulunma olasılığı yüksek alanlara odaklanmış olması, sınıflandırma tahmininin doğru olabileceğini göstermektedir. Şekil 11'de Pituitary sınıfına yönelik yapılan LIME görselleştirme sonuçları verilmiştir.



Şekil 9. LIME Yöntemi Görselleştirmesi (Glioma Tümör)



Şekil 30. LIME Yöntemi Görselleştirmesi (Meningioma Tümör)



Şekil 41. LIME Yöntemi Görselleştirmesi (Pituitary Tümör)

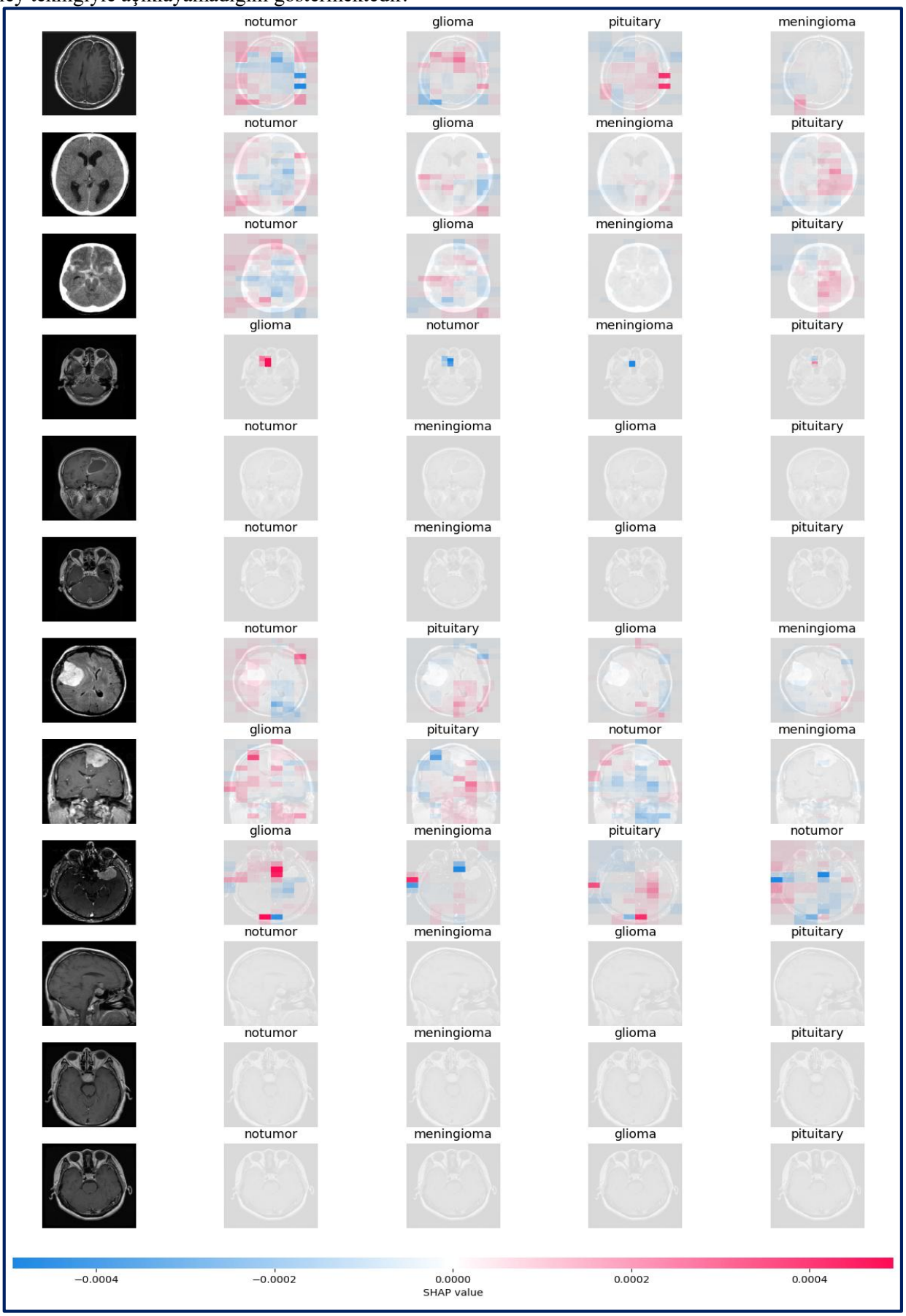
Şekil 11'den de görüldüğü gibi ilk Pituitary sınıfına ait görselde, modelin beynin üst, orta ve sol alt bölümlerinde yeşil renkte aktivasyon gösterdiği tespit edilmiştir. Modelin, tümör bulunma olasılığı yüksek geniş alanlarda aktivasyonlar göstermesi, bu görselde doğru alanları değerlendirdiğini ortaya koymaktadır. İkinci Pituitary sınıfına ait görselde, modelin beynin sol ve sağ kenar bölümlerinde yeşil, sağ alt bölümünde ise kırmızı renkte aktivasyonlar gözlemlenmiştir. Bu aktivasyonların dağılımı, modelin tümör bulunabilecek alanlara odaklandığını ve sağ alt bölgedeki düşük öncelikli aktivasyonların, bu alanlarda daha az önem verildiğini göstermektedir. Bu, modelin doğru bir sınıflandırma için belirli alanlara odaklandığına işaret eder. Üçüncü Pituitary sınıfına ait görselde, modelin beynin sol ve sağ kenarları ile beyin dışındaki sol alt bölgede yeşil renkte aktivasyon gösterdiği tespit edilmiştir. Bu durumda, modelin hem beynin içindeki doğru bölgelerde hem de tümör bulunmaması gereken dış alanlarda aktivasyon göstermiştir. Bu durum %100 doğruluk elde edilmeyen modelin bu görsel için hatalı aktivasyona sahip olabileceğini göstermektedir.

Shapley Yöntemi ile Görselleştirme

Şekil 12'de Shapley yöntemine ait her bir resim için orijinal ve her dört sınıf tahmininde aktivasyon görülen sonuçlar görülebilmektedir.

Şekil 12'den de görüldüğü gibi No Tumor Sınıfına ait görselde modelin beyin genelinde geniş bir aktivasyon alanı oluşturduğu gözlemlenmiştir. Beynin farklı bölgelerinde görülen bu aktivasyonlar, modelin tümör tespiti için beynin geniş alanlarını değerlendirdiğini ve çoğunlukla doğru alanlara odaklandığını göstermektedir. Ancak, bazı durumlarda aktivasyonların beyin dışına kayması, modelin sınıflandırma kararlarını etkileyebilecek potansiyel hatalara yol açabileceğini göstermektedir. Bu durum, modelin No Tumor sınıfı için genellikle güvenilir bir sınıflandırma yaparken, nadiren de olsa yanlış pozitiflere yol açabileceğini göstermektedir. Glioma Sınıfına ait görsellerde yalnızca bir görselde belirgin bir aktivasyon gözlemlenmiştir. Bu görselde model, tümör bulunabilecek bir bölgeye odaklanarak doğru bir alanda değerlendirme yapmıştır. Ancak, diğer iki görselde herhangi bir aktivasyonun bulunmaması, modelin bu sınıf için her zaman tutarlı bir performans sergileyemediğini göstermektedir. Meningioma Sınıfına ait görsellerin bazılarında, modelin beynin kenar kısımlarında ve genelinde aktivasyonlar gösterdiğini ortaya koymuştur. Modelin doğru alanlara odaklanarak etkili bir sınıflandırma yaptığı anlaşılmakla birlikte, bazen beyin dışına da kayması, yanlış pozitif sonuçlara neden olabilmektedir. Pituitary Sınıfına ait

görsellerin bazılarında ise herhangi bir belirgin aktivasyon gözlemlenmemiştir. Bu durum bu sınıftaki resimleri Shapley tekniğiyle açıklayamadığını göstermektedir.



Şekil 12. Shapley Yöntemine Ait Bulgular

Yapılan çalışma ile literatürdeki çalışmaların (Nancy ve Sathyarajasekaran, 2024; Amin vd., 2024; Gaur vd., 2024) doğruluklarının paralel sonuçlar verdiği ancak kullanılacak model önerilerinin farklılık gösterdiği bildirilebilmektedir. Bu farklılığın literatürdeki çalışmalardan farklı bir sonuç vermesinin en temel sebepleri; yapılan çalışmada karmaşık modelin kullanılmasından ve model mimarisinin özellik haritasının anlamlandırılmasının zor olmasından kaynaklı olabileceği düşünülmektedir.

SONUÇ

Çalışma sonucunda her üç teknik de modeli açıklamaya çalışırken model metrikleri ve sınıflandırma tahminleri ne olursa olsun, seçilen resimler üzerinde tam performans gösteremediği tespit edilmiştir. GradCAM yöntemi, modelin son katmanlarındaki aktivasyonları görselleştirerek, tümör bulunma olasılığı yüksek alanların belirlenmesinde önemli bir araç olarak öne çıkmıştır. Bu teknik, modelin doğru alanlara odaklandığını açık bir şekilde göstermiş ancak zaman zaman modelin tümör bulunmayan bölgelere de odaklanabildiğini ortaya koymuştur. LIME yöntemi ise, modelin her bir sınıflandırma kararını daha detaylı bir şekilde açıklayarak, modelin hangi alanlara daha fazla önem verdiğinin anlaşılmasına olanak tanımıştır. LIME analizleri, modelin genellikle doğru alanlara odaklandığını ancak bazı durumlarda yine gereksiz aktivasyonlar üretebildiğini göstermiştir. Shapley yöntemi, her bir sınıf tahmini için modelin aktivasyon dağılımını inceleyerek, modelin genel performansının değerlendirilmesine yardımcı olmuştur. Shapley analizi, modelin bazı sınıflarda (No Tumor ve Meningioma) genellikle doğru alanlara odaklandığını ancak Glioma ve Pituitary sınıflarında performansın tutarlı olmadığını göstermiştir. Sonuç olarak, her üç teknik de modelin karar mekanizmalarını farklı açılardan inceleyerek, modelin güçlü ve zayıf yönlerini belirlemekte önemli bir rol oynamıştır. GradCAM, modelin genel odaklanma alanlarını belirlemekte etkili olmuş; LIME, modelin kararlarının detaylı bir açıklamasını sunmuş; Shapley ise modelin genel performansını ve eksikliklerini ortaya koymuştur.

Açıklanabilir yapay zeka tekniklerinin Meningioma sınıfındaki başarısının diğer sınıflara göre daha yüksek olduğu görülmüştür. GradCAM yöntemiyle yapılan görselleştirmelerde, modelin Meningioma tümörlerini genel olarak doğru alanlara odaklanarak tespit ettiği ve aktivasyonların beynin zar bölgesinde yoğunlaştığı görülebilmektedir. Ancak bazı görsellerde yanlış altivasyonlar olsa da genel olarak doğru tahmin yapabildiği belirtilebilmektedir. LIME yönteminde yapılan analizlerde, modelin Meningioma tümörlerinin bulunduğu alanlara güçlü bir şekilde odaklandığı ve genel olarak doğru sınıflandırmalar yaptığı görülebilmektedir. Shapley yöntemiyle yapılan incelemelerde ise modelin Meningioma sınıfında beynin doğru bölgelerine odaklandığı ancak bazen beyin dışına da aktivasyon kaydırarak yanlış pozitif sonuçlara sebep olduğu söylenebilmektedir. Bu durum Meningioma tümörlerinin MR görüntülerinde daha belirgin olmasıyla ilişkili olabilmektedir. Bunun sebebi Meningioma'nın beynin zar kısmında yer alan bir tümör türü olması ve bu nedenle kontrast farkı diğer bazı tümörlere kıyasla daha belirgin olabilmektedir. Ayrıca tümör sınırları daha net görünebilir ve bu da modelin görselleştirme yöntemleriyle daha doğru aktivasyonlar oluşturmasını sağlayabilir. Glioma ve Pituitary sınıflarında model bazen düşük aktivasyon göstermiş ya da tutarsız tahminler yapmışlar. Bu da Meningioma'nın diğerlerine kıyasla daha kolay tespit edilebildiğini düşündürmektedir. Meningioma tümörlerinin net görünebilir olması, açıklanabilir yapay zeka yöntemlerinin bu sınıfta daha başarılı olmasını sağlamış olabilir. Ancak, modelin gerçekten başarılı olup olmadığını değerlendirmek için sayısal metriklerin (Accuarcy (Doğruluk), Precision (Kesinlik), Recall (Duyarlılık), F1 Skoru) de incelenmesi gerekmektedir.

Açıklanabilir yapay zeka tekniklerinin incelenmesi için seçilen kesitlerde diğer tümörlü sınıflardan tümörün net bir şekilde görülebildiği kesitler seçilseydi, açıklanabilir yapay zeka tekniklerinin başarımının muhtemelen daha yüksek olacağı söylenebilmektedir. Açıklanabilir yapay zeka teknikleri, modelin odaklandığı bölgeleri insan yorumuna sunmaktadır. Ancak, eğer bir görüntüde tümör belirgin değilse ya da çok düşük kontrasta sahipse, modelin doğru tahmin yapması ve bunu anlamlandırması zorlaşabilmektedir. Glioma ve Pituitary sınıflarında tümörün konumu net olmadığı için modelin aktivasyonları da düşük ya da yanlış olabilmektedir. Bu da Shapley, GradCAM ve LIME gibi yöntemlerin tümörlü bölgeyi doğru bir şekilde vurgulayamamasına sebep olabilmektedir. Eğer tümör net olarak seçilebilen kesitler kullanılsaydı, modelin tahmin yaparken daha doğru bölgelere odaklanması sağlanırdı. Bu durumda açıklanabilir yapay zeka tekniklerinin gösterdiği aktivasyon haritaları daha anlamlı olabilir ve modelin başarımı artırılabilirdi. Bu tekniklerin birlikte kullanılması, modelin daha güvenilir ve etkili bir şekilde çalışmasını sağlamak için daha fazla veri sağlanmasına veya gerekli iyileştirmelerin yapılmasına olanak tanımaktadır.

Yapılan çalışma açıklanabilir yapay zeka yöntemlerini (GradCAM, LIME ve Shapley değerleri) kullanarak MR görüntülerinden beyin tümörü tespiti yapmaktadır. Bu yöntemlerin birlikte kullanılması, modelin kararlarını şeffaflaştırarak tıbbi alanda güvenilir yapay zeka modellerinin geliştirilmesine katkı sağlayabilmektedir. Literatürde genel olarak tek bir açıklanabilirlik yöntemi kullanılırken, bu çalışmada birden fazla açıklanabilir yapay zeka

yönteminin kullanılmış ve modellerin karar verme süreçleri daha kapsamlı bir şekilde incelenmiştir. Çalışma, beyin tümörü teşhisi konusunda doktorlara daha iyi karar destek sistemleri sunabilecek bir yapay zeka modeli geliştirmeye yönelik katkı sunmuştur. Açıklanabilirlik teknikleri, modelin hangi özelliklere dayalı olarak karar verdiğini göstererek, tıp uzmanlarının modelin doğruluğunu değerlendirmesine yardımcı olabilmektedir. Bu da klinik uygulamalarda yapay zeka kullanımının benimsenmesini hızlandırabilmektedir. Ayrıca gelecekte yapılacak çalışmalara; daha geniş ve dengeli veri setleriyle çalışmaları, farklı tümör tipleriyle alt türlerinin tespit edilmesine yönelik özelleştirilmiş modeller geliştirmeleri, MR görüntüleme tekniklerinin yanı sıra farklı tıbbi görüntüleme tekniklerini kullanarak çok modaliteli modeller geliştirmeleri önerilmektedir. Modelin eksiklikleri ve gelecekte yapılacak çalışmalara öneriler ise şu şekilde sıralanabilmektedir; veri setinin, beyin tümörlerinin tüm türlerini ve varyasyonlarını tam olarak kapsamadığı için gelecekte yapılacak çalışmaların daha büyük ve çeşitli veri setleri kullanarak modelin genelleme kapasitesini ve performansını artırmaları, doktorlarla iş birliği yapılarak gerçek dünyada klinik ortamlarda modelin gerçek hasta verileriyle test edilmesi, farklı açıklanabilirlik yöntemlerinin etkinliğinin daha kapsamlı analiz edilmesi ve modelin daha fazla hiperparametre optimizasyonuyla geliştirilmesi ve transfer öğrenme gibi çeşitli tekniklerin kullanılarak performansının artırılması önerilmektedir.

KAYNAKLAR

Aamir, M., Rahman, Z., Dayo, Z. A., Abro, W. A., Uddin, M. I., Khan, I., & Hu, Z. (2022). A deep learning approach for brain tumor classification using MRI images. *Computers and Electrical Engineering*, 101, 108-145. https://doi.org/10.1016/j.compeleceng.2022.108105

Abdusalomov, A. B., Mukhiddinov, M., & Whangbo, T. K. (2023). Brain tumor detection based on deep learning approaches and magnetic resonance imaging. *Cancers*, 15(16), 1-29. https://doi.org/10.3390/cancers15164172

Amin, K. H., Saleh, Z. S., & Deo, C. (2024). An explainable aı framework for artificial ıntelligence of medical things, https://arxiv.org/pdf/2403.04130 03.01.2025'de erişildi.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley*, 11(5), 1-13. https://doi.org/10.1002/widm.1424

Aslan, E. (2024). LSTM-ESA Hibrit modeli ile MR görüntülerinden beyin tümörünün sınıflandırılması. *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, 11(22), 63-81.

Baran, F. D. (2024). Belirli nöropsikolojik rahatsızlıkların yapay zeka temelli sınıflandırılması. Yüksek Lisans Tezi. Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Denizli 134s.

Bilekyiğit, S. (2022). Kalp yetmezliği riskinin makine öğrenmesi yöntemleri ile analiz edilmesi. Yüksek Lisans Tezi. Karamanoğlu Mehmetbey Üniversitesi Fen Bilimleri Enstitüsü Mühendislik Bilimleri Anabilim Dalı, Karaman 152s.

Caelen, O. (2022). What is the Shapley value?.. https://medium.com/the-modern-scientist/what-is-the-shapley-value-8ca624274d5a 03.01.2025'de erişildi.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *WACV*, 839-847. https://doi.org/10.1109/WACV.2018.00097

Ellah, M. K., Awad, A. I., Khalaf, A. A., & Hamed, H. F. (2019). A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned. *Magnetic Resonance Imaging*, 61, 300-318. https://doi.org/10.1016/j.mri.2019.05.028

Fatima, S. S., Wooldridge, M., & Jennings, N. R. (2008). A linear approximation method for the Shapley value. *Artificial Intelligence*, 172(14), 1673-1699. https://doi.org/10.1016/j.artint.2008.05.003

Garreau, D. ve Mardaoui, D. (2021). What does LIME really see in images? https://proceedings.mlr.press/v139/garreau21a/garreau21a.pdf 03.01.2025'de erişildi.

Gaur, L., Bhandari, M., Razdan, T., Mallik, S., & Zhao, Z. (2022). Examining the prediction of discrete subtypes of brain tumors with deep learning models 2022. file:///C:/Users/admin/AppData/Local/Microsoft/Windows/INetCache/IE/3DRHLJ80/fgene-13-822666[1].pdf 03.01.2025'de erişildi.

Gülle, K., Özdemir, D., & Temurtaş, H. (2024). Derin öğrenme yöntemleri kullanılarak böbrek hastalıklarının tespiti ve çoklu sınıflandırma. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, 5(1), 19-28.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., & Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1), 45-74.

Juscafresa, A. (2022). An introduction to explainable artificial intelligence with LIME and SHAP.. https://diposit.ub.edu/dspace/bitstream/2445/192075/1/tfg_nieto_juscafresa_aleix.pdf 03.01.2025'de erişildi.

Kaggle. Datasets. (2025). https://www.kaggle.com/datasets 06.03.2025 'de erişildi.

Karakaya, A. (2024). Meme kanseri tahmininde makine öğrenmesi algoritmaları ve AutoML. Yüksek Lisans tezi. Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Denizli 98s.

Khan, H. A., Jue, W., Mushtaq, M., & Mushtaq, M. U. (2020). Brain tumor classification in mrı image using convolutional neural network. *Math. Biosci. Eng*., 17, 6203-6216.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework For Modeling Biological Vision And Brain İnformation Processing. *Annual Review of Vision Science*, 1(1), 417-446. https://doi.org/10.1146/annurev-vision-082114-035447

Kumar, S., Abdelhamid, A. A., & Tarek, Z. (2023). Visualizing the unseen: Exploring GRAD-CAM for interpreting convolutional image classifiers. *Full Length Article*, 4(1), 34-42. https://doi.org/10.54216/JAIM.040104

Manne, R. & Kantheti, S. C. (2021). Application of artificial intelligence in healthcare: chances and challenges. *Current Journal of Applied Science and Technology*, 40(6), 78-89.

Marmolejo, J. A. & Kose, U. (2024). Numerical Grad-Cam based explainable convolutional neural network for brain tumor diagnosis. *Mobile Networks and Applications*, 29(1), 109-118.

Nancy, A. M. & Sathyarajasekaran, K. (2024). Multi-modal explainability evaluation for brain tumor segmentation: Metrics MSFI. *International Journal of Intelligent Systems and Applications in Engineering*, 12, 341–347.

Orman, A. (2021). Brain Tumor Detection Via Explainable Convolutional Neural Networks. *El-Cezeri Journal of Science and Engineering*, 8(3), 1323-1337. http://doi.org/10.31202/ecjse.924446

Pannu, A. (2015). Artificial Intelligence And Its Application In Different Areas. *Artificial Intelligence*, 4(10), 79-84.

Pillai, V. (2024). Enhancing Transparency And Understanding In AI Decision-Making Processes. *Iconic Research and Engineering Journals*, 8(1), 168-172.

Rahman, A. (2019). Statistics-Based Data Preprocessing Methods And Machine Learning Algorithms For Big Data Analysis. *International Journal of Artificial Intelligence*, 17(2), 44-65.

Reddy, S. (2018). Use of artificial intelligence in healthcare delivery. London: IntechOpen.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128, 336-359.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6), 1-19. https://doi.org/10.3390/jimaging6060052

Turay, T. & Vladimirova, T. (2022). Toward performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey. *IEEE Access*, 10, 14076-14119.

Verdinelli, I. & Wasserman, L. (2024). Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4), 623-636. https://doi.org/10.1214/24-STS937

Vimbi, V., Shaffi, N., & Mahmud, M. (2024). Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer's disease detection. *Brain Informatics*, 11(1), 1-29.