T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

AÇIKLANABİLİR VE YORUMLANABİLİR YÜZ DUYGU TANIMA

YÜKSEK LİSANS TEZİ

Elif NASIR TOKMAK

1002010011

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Bilgisayar Mühendisliği

Tez Danışmanı: Doç. Dr. Fatma Patlar Akbulut

HAZİRAN 2023

T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

AÇIKLANABİLİR VE YORUMLANABİLİR YÜZ DUYGU TANIMA

YÜKSEK LİSANS TEZİ

Elif NASIR TOKMAK

1002010011

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Bilgisayar Mühendisliği

Tez Danışmanı: Doç. Dr. Fatma Patlar Akbulut

Jüri Üyesi: Prof. Dr. Özgür Koray Şahingöz

Jüri Üyesi: Doç. Dr. Akhan Akbulut

HAZİRAN 2023

ÖNSÖZ

Tez çalışmamdaki katkıları, desteği ve bana yol göstermesinden dolayı değerli danışman hocam Doç. Dr. Fatma PATLAR AKBULUT'a çok teşekkür ederim. Her zaman daha çok ilerlemem için beni yüreklendiren aileme, hep yanımda olup beni destekleyen sevgili eşim Ahmet Vedat TOKMAK'a ve oyun vaktinden feragat eden kızım Ela'ya teşekkürü borç bilirim.

Haziran 2023

Elif NASIR TOKMAK

İÇİNDEKİLER

KISALTMALAR

XAI	: Açıklanabilir Yapay Zekâ (Explainable Artificial Intelligence)
CNN	: Konvolüsyonel Sinir Ağı (Convolutional Neural Network)
LIME	: Yerel Yorumlanabilir Model-Bağımsız Açıklamalar (Local Interpretable Model-Agnostic Explanations)
SHAP	: Shapley Toplam Açıklamaları (SHapley Additive exPlanations)
GradCAM	: Gradyan-ağırlıklı Sınıf Aktivasyon Haritalama (Gradient weighted Class Activation Mapping)
VGG	: Görsel Geometri Grubu (Visual Geometry Group)
ResNet	: Kalıntı Ağı (Residual Network)

TABLO LİSTESİ

ŞEKİL LİSTESİ

Üniversite : **T.C. İstanbul Kültür Üniversitesi**
Enstitü : **Lisansüstü Eğitim Enstitüsü**
Anabilim Dalı : **Bilgisayar Mühendisliği**
Program : **Bilgisayar Mühendisliği**
Tez Danışmanı : **Doç. Dr. Fatma PATLAR AKBULUT**
Tez Türü ve Tarihi : **Yüksek Lisans – Haziran 2023**

ÖZET

AÇIKLANABİLİR VE YORUMLANABİLİR YÜZ DUYGU TANIMA

Günümüzde birçok sektörde otonom sistemler artmaktadır. Bu sistemler ile insanlar arasındaki etkileşimin artmasıyla bu sistemlerin doğru kararlar vermesi oldukça zorlaşmıştır. Otonom sistemlerin karar alırken insanların duygularını dikkate almaması, yanlış kararlar almasına sebep olmaktadır. Oluşturulan sistemlerin kişinin duygu durumuna göre karar vermesi gereken durumlarda, bu sistemlerin kişilerin duygularını yüz ifadelerinden hızlı ve doğru bir şekilde tahmin etmelerini gerekli kılmıştır. Bu çalışma yüz ifadesi tanımak için yapılan çalışmalarda oluşturulan yapay zekâ modellerinin aldığı kararların açıklanması ve eksikliklerinin belirlenmesi konularında yol gösterici olacaktır. İlk olarak Literatürde bu alanda yer alan çalışmalar incelenerek bu çalışmada kullanılacak model, veriseti ve teknikler belirlenmiştir. Bu çalışmada daha önceden eğitilmiş modeller olan VGG, ResNet ve Inception modelleri üzerinde iyileştirmeler yapılarak yeni üç model oluşturulmuştur. Oluşturulan modeller Fer2013 veriseti ile eğitilmiş ve modeller ilk olarak eğitim ve test verileri üzerinden karşılaştırılmıştır. Daha sonra ise modeller yedi farklı duygu durumu için seçilen yedi görüntüyü tahmin etmeleri sağlanarak ikinci defa karşılaştırılmıştır. Her iki karşılaştırmada da en iyi model VGG olmuştur. Modellerin yaptıkları tahminlerin doğruysa neden doğru ve yanlışsa neden yanlış olduğunu açıklanması için beş farklı Açıklanabilir Yapay Zekâ (XAI) tekniği kullanılmıştır. Kullanılan XAI teknikleri GradCAM, GradCAM++, Saliency haritası, SHAP ve LIME teknikleridir. Her modelin yedi farklı duygu durumunu tahmin etmesi sağlandıktan sonra bu tahminin beş farklı teknikle açıklaması yapılmıştır. En açıklayıcı bilgiler GradCAM++ ve Saliency haritası teknikleriyle elde edilmiştir.

Anahtar Kelimeler: GradCAM, Saliency Haritası, SHAP, LIME, Açıklanabilir Yapay Zekâ

University	**: T.C. İstanbul Kültür University**
Institute	**: Institute of Graduate Studies**
Department	**: Computer Engineering**
Program	**: Computer Engineering**
Thesis Advisor	**: Assoc. Prof. Dr. Fatma PATLAR AKBULUT**
Degree Awarded And Date	**: MA – June 2023**

ABSTRACT

Explainable and Interpretable Facial Emotion Recognition

In today's world, autonomous systems are increasingly prevalent in various industries. As the interaction between these systems and humans increases, it becomes more challenging for these systems to make accurate decisions. One of the reasons for the systems making incorrect decisions is that they do not take into account human emotions when making decisions. In situations where these systems need to make decisions based on a person's emotional state, it has become necessary for these systems to quickly and accurately predict people's emotions from their facial expressions. This study aims to provide guidance on explaining the decisions made by artificial intelligence models developed for facial expression recognition and identifying their shortcomings. Firstly, existing studies in this field were examined to determine the model, dataset, and techniques to be used in this study. In this study, three new models were created by improving the pre-trained models VGG, ResNet, and Inception. These models were trained using the Fer2013 dataset, and initially compared using training and test data. Subsequently, the models were compared again by having them predict seven images selected for seven different emotional states. In both comparisons, the VGG model performed the best. To explain why the models made correct or incorrect predictions, five different eXplainable Artificial Intelligence (XAI) techniques were used. The XAI techniques employed were GradCAM, GradCAM++, Saliency Map, SHAP, and LIME. After each model predicted the seven different emotional states, the predictions were explained using the five techniques. The most informative results were obtained using GradCAM++ and Saliency Map techniques.

Keywords: GradCAM, Saliency Map, SHAP, LIME, Explainable Artificial Intelligence

1. GİRİŞ

İnsanlar duygularını hareketlerle, sözlerle veya yüz ifadeleriyle olmak üzere farklı biçimlerde gösterirler. Yüz ifadeleri, hareketler ve sözlere göre karşımızdaki kişinin duygularını anlamada daha etkilidirler. Yüz ifadeleri insanların duygularını daha net gösterdikleri için yüz ifadesine göre karşımızdaki kişinin duygularını daha net anlayabiliriz. Psikolog Albert Mehrabian'ın "7-38-55" kuralına göre mesajların %7'si sözlü olarak, %38'i konuşmayla ve %55'i ise yüz ifadeleri ile iletilir [1]. Bu da yüz ifadelerinin kişinin duygularını anlamada ne kadar önemli olduğunu göstermektedir. Farklı durumlara karşı verdiğimiz tepkilerin birçoğu doğuştan gelir yani sonradan öğrenilmezler. Korktuğumuz ve şaşırdığımız zaman verdiğimiz tepkiler ve yüzümüzde oluşan ifade doğaldır. Ekman ve Friesen kültürler arası bir çalışmayla öfke, iğrenme, korku, mutluluk, üzüntü ve şaşkınlık olmak üzere altı temel duygu tanımlamışlardır [2]. Yaptıkları bu çalışma farklı kültürlerde insanların olaylara karşı benzer yüz ifadeleri ile duygularını gösterdiklerini ortaya koymuştur.

Günümüzde güvenlik, eğlence, sağlık, eğitim ve pazarlama gibi birçok sektörde kullanılmak üzere yüz ifadelerinden duygu tanıma üzerine birçok çalışma yapılmıştır ve hala yapılmaya devam edilmektedir [3][4]. Ayrıca yüz ifadeleri kişinin duygularını yansıttığı için yüz tanıma terimi ile duygu tanıma birbirinin yerine kullanılmaktadır [5]. İnsan bilgisayar etkileşiminin her geçen gün artmasıyla oluşturulan yeni sistemlerin başarılı olması için duyguların mümkün olduğunca yüksek oranda otomatik olarak tanınması gerekmektedir. Duygu tanıma genelde yüzü algılama, yüzün özelliklerini çıkarma ve ifadelerin sınıflandırılması olmak üzere üç adımda gerçekleşmektedir [6]. Duygu tanımlamanın ilk adımında kişinin ilk olarak yüzünün algılanması gerekmektedir. Daha sonra tanımlanan yüze ait özellik çıkarımı mümkün olduğu kadar iyi bir şekilde yapılmalıdır. Son olarak iyi bir sınıflandırma yapabilmek için her ne kadar farklı kültürlere ait yüz ifadeleri benzer özellik taşısa da oluşturulan modelin birçok kültürü kapsayan geniş bir veriseti ile eğitilmesi gerekmektedir.

1.1.Problemin Tanımı

İnsan bilgisayar etkileşimin artmasıyla duygu tanıma üzerine yapılan çalışmalarda birçok farklı yöntem ve algoritma kullanılmıştır. Duygu tanımada Konvolüsyonel Sinir Ağları (CNN) yaygın olarak kullanılmıştır. ResNet, VGG ve AlexNet gibi birçok CNN modeli elde ettiği başarılı sonuçlardan dolayı duygu tanıma için önerilmiştir [7]. Her ne kadar duygu tanımada kullanılan CNN modelleri başarılı olsa da yorumlanabilirlik ve açıklanabilirlik konusunda eksikliklerinden dolayı siyah kutu modelleri olarak adlandırılırlar. Siyah kutu modelleri karmaşık makine öğrenmeleridir ve yapılan tahmin konusunda bilgi vermediği için yaptıkları tahminlerin güvenirlilik sorunu ortaya çıkmaktadır. Bu modeller verileri alır, işler ve tahminde bulunurlar. Fakat işleyiş ve süreç hakkında herhangi bir bilgi vermezler. Kullanıcılar ise yapılan tahminlerin dayanaklarını bilmek isterler. Örneğin, Duygu tanımada modelimizi eğittikten sonra test ettiğimizde bize sadece verdiğimiz resimde yer alan kişinin yüz ifadesinden yola çıkarak tahminini iletir. Fakat yaptığı bu tahmini, doğru veya yanlış olsun, yüzün hangi noktalarını dikkate alarak yaptığı konusunda bilgi vermez. Bu da yapılan tahminin doğruluğunun sorgulanmasına sebep olmaktadır. Yapılan tahminin yorumlanabilmesi veya açıklanabilmesi modelin güvenirliliğini artıracaktır ve modeldeki eksikliklerin tespit edilmesinin ve modelin geliştirilmesini kolaylaştıracaktır.

Bu çalışmada motivasyonumuz öncelikle Konvolüsyonel Sinir Ağları (CNN) kullanarak duygu tanıma yapmak için iyi bir model geliştirmektir. Geliştirdiğimiz modeller ile elde ettiğimiz sonuçların güvenirliliğini arttırmak için LIME, SHAP, GradCAM, GradCAM++, Saliency haritası gibi Açıklanabilir Yapay Zekâ (XAI) yöntemleri kullanarak elde ettiğimiz tahminlerin dayanaklarını göstererek yorumlamak ve açıklamaktır. Bu sayede bir yandan yapılan tahminlerin güvenirliliği artacak, diğer taraftan modeldeki eksiklikler ortaya çıkacaktır. Ayrıca açıklanabilirlik ve yorumlanabilirlik sayesinde modelin doğru çalışıp çalışmadığını ve geliştirmeye ihtiyacı olup olmadığı görmemizi sağlayacaktır.

1.2.Tezin Amacı

Bu çalışmanın amacı, farklı duygu belirtileri gösteren yüz ifadeleri için oluşturulan derin ögrenme modellerinin sonuçlarını açıklamak için farklı tekniklerin etkinliğini araştırmaktır. GradCAM, GradCAM++, Saliency haritası, SHAP ve LIME

gibi tekniklerin kullanıldığı çalışmada, yüz ifadesi tanıma modelinin yaptığı tahminlerin açık ve yorumlanabilir açıklamaları sağlanarak modelin karar verme sürecinde güven oluşturmaya yardımcı olunması hedeflenmiştir. Bu bağlamda VGG, ResNet ve Inception gibi farklı CNN mimarileri Fer2013 veriseti ile eğitilerek 3 farklı model oluşturulmuştur. Çalışma ayrıca bu farklı tekniklerin etkinliğini karşılaştırmayı ve yüz ifadesi tanıma modellerini açıklamak için hangilerinin en yararlı olduğunu belirlemeyi amaçlamıştır. Bu bilgi, gelecekte daha doğru ve yorumlanabilir modellerin geliştirilmesine rehberlik etmek için kullanılabilir. Genel olarak, bu çalışmanın amacı muhtemelen yüz ifadesi tanıma modellerinin nasıl çalıştığına dair anlayışımızı geliştirmek ve tahminlerini açıklamak için daha iyi teknikler geliştirmektir.

Bu çalışmada, bu tekniklerin yüz ifadesi tanıma modellerinin yorumlanabilirliğini arttırmadaki etkinliğini araştırmayı amaçlıyoruz. Ek olarak, doğru duygu sınıflandırmasına katkıda bulunan temel yüz ipuçlarını ortaya çıkarmak için ön işleme ve veri arttırma teknikleri, önceden eğitilmiş modellerin ince ayarının etkisi ve açıklanabilirlik tekniklerinin kombinasyonu ile ilgili hipotezleri araştırıyoruz. Bu hipotezleri inceleyerek, modelin davranışına ilişkin anlayışımızı derinleştirmeyi, performansını iyileştirmeyi ve yüz ifadesi tanımayı sağlayan önemli özellikler hakkında fikir edinmeyi amaçlıyoruz.

H1: GradCAM, GradCAM++, Saliency haritaları ve GradCAM+Saliency haritası gibi açıklanabilirlik tekniklerinin kullanımı, yüz ifadesi tanıma modellerinin yorumlanabilirliğini önemli ölçüde artıracak ve modelin karar verme sürecini açıklayan daha net içgörüler verecek.

H2: FER2013 veri kümesine uygulanan ön işleme ve veriyi artırma ömrünü, yüz ifadesini tanıma modelinin performansını iyileştirerek daha yüksek doğruluk ve görünmeyen gösterge üzerinde daha iyi genelleme sağlar.

H3: FER2013 veri kümesinde önceden öğrenmiş bir modelde ince ayar yapmak, geleneksel olarak bilinen modelin öğrenilen temsillerinden yararlanarak bir modeli sıfırdan eğitmeyle karşılaştırma ve gelişmiş yüz ifadesi tanıma performansını sağlar.

H4: Yüz ifadesi tanımada açıklanabilirlik tekniklerinin sonuçlanması, doğru duygunun algılaması için önemli olan yüzün bölgelerini veya özelliklerini vurgulayarak, geçmişten kalan yüz ifadelerine ilişkin anlayışımızı geliştirecektir.

H5: GradCAM ve Saliency harita tekniklerinin (GradCAM+Saliency haritası) bir derecelendirmenin kullanımı, her iki tekniğin tek başına kullanımına kıyasla daha geniş kapsamlı ve bilgilendirici açıklamalar sağlayacaktır.

1.3. Tezin Organizasyonu

Bu tez altı bölümden oluşmaktadır. Birinci bölümde, problemin tanımı yapılmış ve araştırmanın amacı ve önemi anlatılmıştır. İkinci bölümde, Kullanılan modeller, temel kavramlar ve duygu tanıma üzerine daha önce yapılan çalışmalardan bahsedilmiştir. Üçüncü bölümde, açıklanabilir yapay zekâ teknikleri ve veriseti anlatılmıştır. Dördüncü bölümde elde edilen deneysel sonuçlar gösterilmiştir. Beşinci bölümde yapılan modellerin karşılaştırılması ve tartışma yer almış; altıncı bölümde ise elde edilen bulgular üzerinden sonuçlar ele alınmıştır.

2. TEMEL KAVRAMLAR

Bu bölümde ilk olarak Açıklanabilir Yüz İfadesi Tanımaya giriş yapılmış sonrasında bu konunun önemi ve kısa tarihçesi anlatılmıştır. Daha sonra Açıklanabilir Yüz İfadesi Tanıma konusunda karşılaşılan zorluklara ve bu konudaki yaklaşımlara yer verilmiştir. Kullanılan veriler ve verisetleri ile değerlendirme metrikleri açıklanmıştır. Son olarak daha önce bu alanda yapılan çalışmalardan ve Açıklanabilir Yüz İfadesi Tanınmasında gelecekteki yönelimler bahsedilmiştir.

2.1. Açıklanabilir Yüz İfadesi Tanımaya Giriş

Yapay Zekâ günümüzde başta eğitim, sağlık, güvenlik, finans ve üretim alanlarında rağbet görmektedir. Fakat Yapay Zekayı oluşturan modellerin ve algoritmaların arka planda ne yaptığının bilinmemesi kullanıcılarda güven sorununun oluşmasına sebep olmaktadır. Kullanıcılar, her ne kadar doğru kararlar alsa da yapay zekanın bu kararı almasındaki nedenleri bilmek ister. Alınan kararların şeffaflığı bu kararların güvenirliliği için gereklidir. Bu gereklilikler Açıklanabilir Yapay Zekâ kavramının ortaya çıkmasına sebep olmuştur. XAI, yapay zekâ modelinin karar verme sürecini yorumlayarak kullanıcıların kafasında oluşan soruları cevaplamaya çalışır yani kısaca sonuçları anlamasını ve analiz etmesini sağlar [8]. Sonuç olarak arka planında karmaşık süreçler işleyen yapay zekâ modellerinin yaptıkları tahmin ve aldıkları kararların şeffaf olması için bunların açıklanabilir olması gerekir. Bu doğrultuda Mayıs 2018'de Avrupa Parlamentosu, şirketlerin makine öğrenimi ve derin öğrenme kullanılarak alınan herhangi bir kararı "açıklamasının" zorunlu olduğu bir yasa oluşturmuştur [9].

Açıklanabilir yüz ifadesi tanımanın amacı ise kullanılan yapay zekâ modelinin yüz ifadelerini tanımlarken yüzün hangi özelliklerine odaklandığını, yüzün hangi bölgelerini dikkate aldığını ve modelin verdiği kararı anlamaktır. Açıklanabilirlik yapay zekâ modellerinin başarısını artıracak ve bu modellerin karar verme sürecinin daha iyi anlaşılabilmesini sağlayacaktır. Açıklanabilir Yüz İfadesi Tanıma

kullanıcılara modellerin verdiği kararlara güvenebilmeleri için daha fazla bilgi sağlayacaktır.

2.2. Açıklanabilir Yüz İfadesi Tanımasının Önemi

Yüz ifadesi tanıma genelde bir kişinin duygusal durumunu ve bu durum doğrultusunda verebileceği kararları tahmin etmek için kullanılır. Oluşturulan yapay zekâ modelinin verdiği kararların doğruluğu bu modellerin geçerliliği ve güvenilirliği açısından oldukça önemli olmakla beraber kullanıcıların kafasındaki soruları tam olarak açıklamaya yetmemektedir. Kullanıcılar alınan kararların nasıl ve hangi özelliklere göre alındığını bilmek ister. Bu noktada açıklanabilirlik kullanıcının bu sorularına cevaplar sağlar ve kullanıcın kararlara olan güvenini artırır. Ayrıca açıklanabilirlik modeldeki sorunların ve eksikliklerin tespitin ve modelin geliştirilmesini de sağlar. Bu sebeplerden dolayı Açıklanabilir Yüz İfadesi Tanıma, kullanılan yapay zekâ modellerinin verdiği kararların güvenilirliği ve modellerin geliştirilmesi için çok önemlidir [9].

2.3.Açıklanabilir Yüz İfadesi Tanımasının Kısa Tarihi

Açıklanabilir Yüz İfadesi Tanıma, insanların yüzlerini algıladıktan sonra duygularına göre sınıflandırarak insan duygularını anlamaya çalışan bir araştırma alanıdır. Bu alandaki ilk çalışmalar ifadeleri belirlemek için belirli özellik ve kuralların tanımlandığı kural tabanlı sistemler ile yapılmıştır. Kural tabanlı sistemlerde ifadenin belirlenmesi için üzgün insanın dudakları aşağı bükülür, sinirli insanın kaşları çatılır gibi kurallar belirleniyordu. Bu yöntem karmaşık duygu durumlarında sınırlı kalıyordu. Bir sonraki aşamada ise yüz ifadelerinin istatistiksel analizinin yapıldığı istatistiksel yaklaşımlar kullanılmıştır. İstatistiksel yaklaşımlardan sonraki dönemlerde, derin öğrenme ve sinir ağları gibi daha karmaşık ve veriye dayalı yöntemler bu alanda büyük bir etki yaratmıştır. Çok katmanlı yapısı sayesinde derin sinir ağları, ifadelerin sınıflandırılmasında büyük bir başarı sağlamıştır. Bu yaklaşımda büyük veri kümeleri kullanılarak eğitim aşamasında yüz ifadelerinin sınıflandırılması öğrenilir ve sonrasında yeni resimlerde ifade tanıma yapılır [4].

Bu alanda ortaya çıkan son yöntemler ise Degrade Tabanlı Yöntemlerdir. Yüz ifadesi tanımanın güvenilirliğini ve açıklanabilirliğini artırmak için ortaya çıkan bu yaklaşım sinir ağları tabanlı modellerin kararlarını nasıl verdiğini anlamak için

geliştirilmiştir. Bu yöntemlerde, sinir ağı tabanlı modellerin kararlarını açıklamak için Görsel Açıklamalar, Önemli Bölge Vurgusu, Sınıf Söyleyici Ağlar ve GradCAM gibi farklı stratejiler kullanılabilir. Bu stratejilerden biri olan GradCAM bize sinir ağının odaklandığı bölgeleri ve özelliklerin ifade tanıma sürecindeki ağırlıklarını görsel olarak gösterir [9].

2.4.Açıklanabilir Yüz İfadesi Tanımadaki Zorluklar

Açıklanabilir yüz ifadesi tanımanın geliştirilmesi ve güvenilirliğinin artırılması için çözülmesi gereken temel sorunlar şöyledir: Veri kalitesi, Bağımlılık ve İlişkiler, Kişisel farklılıklar ve Dinamik ifadelerdir. Kullanılacak verisetinin doğru etiketlenmiş olması, ifadelerin çeşitliliği, yeterli miktarda verinin olması ve bu verilerin dağılımı modelin doğruluğunu ve performansını etkileyen başlıca etkenlerden biridir. Ayrıca, bir ifadeyi yorumlamak için yüz ifadesinin yanı sıra ses tonu, vücut dilindeki hareketler ve diğer bağlamsal faktörlerde dikkate alınmalıdır. Bununla beraber insanların kişisel farklılıkları, duygularını ifade ederken takındıkları yüz ifadelerini de değiştirebilir. Son olarak yüz ifadeleri zaman içinde değişir yani bir ifadenin başlangıcı bitişi arasındaki süreçte farklılıklar olabilir [4].

Açıklanabilir Yüz İfadesi Tanımada güvenilirliği artırmak için doğruluk ve yorumlanabilirlik arasındaki dengenin de sağlanması gerekir. Bu dengeyi sağlarken karşılaşılan zorluklar ise şöyledir: Doğruluk, oluşturulan modelin yüz ifadesini doğru bir şekilde tanıması ve sınıflandırmasıdır. Karmaşık ve belirgin olamayan ifadeler doğruluğu düşürebilir. Yorumlanabilirlik, modelin karar verme süreci ile odaklandığı noktalar açıklanabilirlik olarak ifade edilirken bu açıklamaların kullanıcı tarafından anlaşılabilmesi yorumlanabilirliktir. Karmaşık, teknik ve anlaşılması zor açıklamalar kullanıcının kararı anlamasını zorlaştırabilir. Veri dağılımı, kullanılan verisetinin farklı ırk, cinsiyet, yaş ve kültürlere ait insanların yüz ifadeleri içermesi demektir. Verisetinde bu alandaki eksiklikler modelin farklı gruplara ait ifadeler karşısındaki performansını etkileyecektir. İfade Çeşitliliği, bireysel farklılıklar insanların duygularını ifade etme şekillerine de yansımaktadır. Örneğin üzgün bir ifade ile hayal kırıklığı ifadesi arasındaki fark çok azdır ve karıştırılabilir. Daha iyi bir Açıklanabilir Yüz İfadesi Tanıma geliştirmek için bu zorlukların hepsini üstesinden gelinmesi gerekmektedir [6].

2.5.Açıklanabilir Yüz İfadesi Tanıma Yaklaşımları

Duygular yaşamımızda içinde bulunduğumuz durumları yansıtma şeklimizdir. Duygularımızı sessimizle, beden hareketlerimizle ve en çokta yüzümüzü kullanarak ifade ederiz. Mevcut duygu durumumuz o anki kararlarımız üzerinde çok etkilidir bundan dolayı yapacağımız birçok işi yaparken duygu durumumuza çok dikkat etmeliyiz. Otonom sistemler de insanlar gibi hareket ettiği için onların da karşısındaki insanın durumuna göre hareket etmeleri gerekmektedir. Örneğin sınıfta bir öğretmen ders anlatırken öğrencilerin o anki duygu durumlarını dikkate alarak ders anlatır çünkü buna dikkat edilmemesi dersin verimini ciddi ölçüde düşürecektir. Korkmuş, sinirli veya üzgün bir öğrenciye ne anlatırsanız anlatın öğrenci o konuyu tam olarak anlayamaz. Online öğrenme için tasarlanacak bir sanal sınıf sisteminde de yüksek verim için sistemin karşısındaki öğrencilerin yüz ifadelerine göre hareket etmesi gerekmektedir. Bu gereklilik beraberinde insanların duygularını anlamak için insan sesleri, beden hareketleri ve yüz ifadeleri üzerine birçok çalışma yapılmasını sağlamıştır. Üzerinde en fazla çalışılan alan ise insanların yüz ifadeleri üzerinden duygu tahminidir [6].

Yüz ifadeleri üzerinden duygularını tahmin etmek için yapılan çalışmalarda birçok farklı yapay zekâ modeli geliştirilmiş ve bu modellerle yüksek başarılar elde edilmiştir. Fakat bu modeller çalışırken arka planda çok karmaşık bir süreç işlemektedir ve model kullanıcıya süreç hakkında bilgi vermeden sadece sonucu söylemektedir. İnsanlar bir kişinin yüzündeki ifadeye baktıklarında kendi deneyimlerinden yola çıkarak bu kişini duygu durumunu ifade ederler. Örneğin sinirli bir kişinin genelde kaşlarına bakarız, kaşları çatılmış ise bu kişi sinirli deriz. Yapay zekâ modellerinin eğitiminde bizim deneyimlerimize benzer bir verisetleri vardır. Modeller de insanlar gibi deneyimlerine göre karar vermek üzere geliştirilirler. Bu nedenle modellerin karar verme sürecinin bilinmesi hem modelin daha ileriye taşınması hem de kullanıcının karar verirken dikkat edilen noktaları bilmesiyle kullanıcının verilen kararı daha fazla benimsemesini sağlayacaktır. Açıklanabilir yapay zekâ kavramı tam olarak bu nedenle ortaya çıkmış olup Açıklanabilir yüz ifadesi tanıma yaklaşımları ise kullanıcıların sınıflandırmanın karar verme sürecini ve sonucunu anlayabilmelerini amaçlar. Yaygın olarak kullanılan Açıklanabilirlik yöntemlerinden bazıları şunlardır [10]:

- **Sınıf Aktivasyon Haritaları (Class Activation Maps- CAM)**: Bu yöntem, oluşturulan modelin sınıflandırma kararını destekleyen görüntüdeki önemli bölgeleri belirlemek için kullanılır.
- **Grad-CAM (Gradient-weighted Class Activation Mapping):** CAM yöntemine dayanan bu yöntem görüntüdeki önemli noktaları belirlemek için gradyanları kullanır.
- **Grad-CAM++ (Grad-CAM Plus Plus)**: Grad-CAM yönteminin geliştirilmiş bir şeklidir.
- **SHAP (Shapley Additive Explanations)**: SHAP, özelliklerin karara olan katkılarını tahmin etmek için Shapley değerlerini kullanan oyun teorisi temelli bir yöntemdir.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME, modelin girdiye olan tepkisi ve anlamlı özelliklerin çıkarılmasını birlikte kullanan bir yöntemdir.
- **Saliency Haritası**: Modelin karar sürecinde odaklandığı önemli noktaları gösteren bir yöntemdir.

Çok farklı açıklanabilirlik yöntemleri olmakla beraber her yöntemin artıları ve eksileri vardır. Bu tez çalışması kapsamında kullanılan açıklanabilirlik yöntemleri üçüncü bölümde detaylı olarak anlatılmıştır.

2.6.Açıklanabilir Yüz İfadesi Tanıma için Veriler ve Verisetleri

Açıklanabilir Yüz İfadesi Tanıma için en temel bileşenlerden biri kaliteli bir verisetidir. Veriseti ne kadar iyi olursa oluşturulan modelde o kadar başarılı olacaktır. İyi bir veriseti için verisetinde bulunan verilerin çeşitliliği ve doğru bir şekilde etiketlenmesi çok önemlidir. Duygu durumunu ifade eden resimlerin "üzgün", "mutlu" gibi doğru ifadeyle etiketlenmesi gerekir. Ayrıca verisetinde bulunan verilerin ırk, yaş, kültür ve cinsiyet gibi farklılıklara dikkat ederek olabildiğince çeşitlilik oluşturacak şekilde ve yeterli sayıda olmasına dikkat edilmelidir. Bununla beraber verilerin sınıflara dengeli dağılımı sağlanmalıdır. Ayrıca yanlış etiketlemelere karşı mutlaka veri düzeltme yapılmalıdır. Gerekli olması durumunda veri artırma yöntemi de uygulanabilir. Model eğitilirken veri bölümleme yapılmalı ve eldeki veriler eğitim ve test için belirli oranlarda ikiye ayrılmalıdır. Son olarak yapılan etiketlemelerin doğruluğu bir insan uzman tarafından gözden geçilebilir [3].

Açıklanabilir Yüz İfadesi Tanıma için verisetleri yorumlanabilir etiketleme ve özellik çıkarma gibi yöntemlerle hazırlanabilir. Yorumlanabilir etiketleme, verilerin "mutlu", "üzgün", "kızgın" gibi etiketlenmesi yerine daha ayrıntılı etiketler kullanarak veriyi daha iyi açıklar. Örneğin bir veri "sinirli" olarak etiketlemek yerine az sinirli çok sinirli gibi derecelendirerek etiketlenebilir. Bu derecelendirme 0-1 arasında bir sayı veya bir aralık olarak ifade edilebilir. Özellik çıkarma ise gözlerin durumu, ağız şekli gibi verilerdeki belirli özelliklerin belirlenmesini sağlar. Bu özellikler ise kararların anlanmasını ve açıklanmasını sağlar. Bu işlem bir uzman tarafından elle yapılabileceği gibi önceden eğitilmiş bir model ile otomatik olarak da yapılabilir. Özellik çıkarma yöntemi modelin karar verme sürecinde odaklandığı noktaları vurgular. Tez çalışması kapsamında deneylerde kullanılacak olan Fer2013 verisetinin kullanılmış olduğu son 7 yıldaki bazı çalışmalar Tablo 2-1'de listelenmiştir.

Tablo 2-1 Son 7 yılda veriseti olarak Fer2013 kullanılan bazı çalışmalar

Yazar ve Yayın	Yöntem	Doğruluk
Facial Expression Recognition usingConvolutional Neural Networks: State of the Art Pramerdorfer, C., & Kampel, M. (2016)	VGG Inception ResNet	72.7 71.6 72.4
Facial Expression Recognition with Convolutional Neural Networks Singh, S., & Nasoz, F. (2020, January)	CNN	61.7
Attention mechanism-based CNN for facial expression recognition Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z. (2020)	LBP (Yerel ikili Modeller) LBP'siz	72.56 67.73
Kids' Emotion Recognition Using Various Deep-Learning Models with Explainable AI Rathod M., Dalvi, C., Kaur K., Patil, S., Gite, S., Kamat P.,& Gabralla L. A. (2022)	VGG19 DenseNet201 ResNet152 V2 InceptionV3	89.90 64.75 76 74.04
Interpretable Explainability in Facial Emotion Recognition and Gamification for Data Collection Shingjergji, K., Iren, D., Böttger, F., Urlings, C., & Klemke, R. (2022, October)	CNN	32.80
Deep Learning-Based Facial Emotion Recognition for Driver Healthcare Sahoo, G. K., Das, S. K., & Singh, P. (2022, May)	SqueezeNet CNN	61.09
A Smart Virtual Tutor with Facial Emotion Recognition for Online Learning Suddul, G., Lillmond, C., & Armoogum, S. (2022, May)	CNN	62.04
Real-Time Facial Emotion Recognition for Visualization Systems Özkara, C., & Ekim, P. O. (2022, September)	GoogleNet Alexnet VGG-19 CNN-LSTM	60.70 57.67 62.66 60.34
EmotiTEA: A visual monitoring module based on the recognition of facial emotions with CNN Oliveira, I., Silva, J. L., Quispe, F. P., & Alvarez, A. B. (2021, October)	CNN-A CNN-B	65 67
Emotion Recognition and Visualization using Grad-CAM Araf, T. A., Siddika, A., Karimi, S., & Alam, M. G. R. (2022, April)	OpenCV'nin harcasscade sınıflandırıcısı	85.68
Deep Learning Based Facial Emotion Recognition using Multiple Layers Model Sandra, L., Heryadi, Y., Suparta, W., & Wibowo, A. (2021, December)	ResNet50	60
Sentiment analysis on images using different transfer learning models, Gaurav Meena and Krishna Kumar Mohbey (2023, January)	XceptionNet VGG19 Inception-V3	77,92 65,41 73,09
Deep facial emotion recognition model using optimal feature extraction and dual-attention residual U-Net classifier Belhassen Akrout, (2023, April)	GoogleNet VGG	65,2 66,3
Role of Zoning in Facial Expression Using Deep Learning T. Shahzad, K. Iqbal, M. A. Khan, Imran and N. Iqbal, (2023, February)	VGG	65

2.7.Açıklanabilir Yüz İfadesi Tanıma için Değerlendirme Metrikleri

Açıklanabilir Yüz İfadesi Tanıma için oluşturulan modellerin performanslarının değerlendirilmesi ve karşılaştırma yapılabilmesi için modele ait belli metriklere ihtiyaç vardır. Performans değerlendirmesi için yaygın olarak kullanılan metrikler şunlardır [3]:

- **Doğruluk (Accuracy):** Modelin doğru sınıflandırma yapma yeteneğini ölçen temel metriktir ve doğru sınıflandırılan örneklerin toplam örneklere oranı olarak hesaplanır. (Doğruluk = (Doğru Sınıflandırılan Örnekler) / (Toplam Örnekler))
- **Hassasiyet (Precision):** Pozitif olarak tahmin edilen örneklerin gerçekten pozitif olan örneklerin oranını ölçer. (Precision = (True Positives) / (True Positives + False Positives))
- **Duyarlılık (Recall):** Gerçekten pozitif olan örneklerin doğru bir şekilde pozitif olarak tahmin edilme oranını ölçer. (Recall = (True Positives) / (True Positives + False Negatives))
- **F1-Skoru (F1-Score):** Hassasiyet ve duyarlılığın harmonik ortalamasını temsil eden bir metriktir. F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
- **Hassaslık (Specificity):** Gerçekten negatif olan örneklerin doğru bir şekilde negatif olarak tahmin edilme oranını ölçer. Specificity = (True Negatives) / (True Negatives + False Positives)

2.8.Açıklanabilir Yüz İfadesi Tanımasında Gelecekteki Yönelimler

Açıklanabilir Yüz İfadesi Tanımanın gün geçtikçe popülaritesi artmaya devam etmektedir. Gelecekte bu alanda yapılan çalışmalar için farklı yönelimler olacaktır. Bunların başında hiç kuşkusuz insan merkezli yaklaşımlar ve Açıklanabilir Yapay Zekâ temelli yaklaşımlar gelecektir. İnsan merkezli yaklaşımlarda kullanıcıların bu sistemleri daha etkin kullanmaları için kullanıcı dostu arayüzler ve Açıklanabilirlik özellikleri olacaktır. Bu özellikler kullanıcının sisteme olan güvenini artıracak ve sistemin kullanımını kolaylaştıracaktır. Gelecekte Açıklanabilir Yapay Zekâ yöntemleri dahada gelişecek ve yüz ifadesi tanıma sistemlerinin işleyişi daha şeffaf hale gelecektir. Bu şeffaflık sisteme olan güveni artıracaktır [2][7].

Bu yönelimlerin dışında bu alanda daha geniş verisetleri ve özellik çıkarmayla birlikte etiketleme ve öğrenme tekniklerinin geliştirilmesi üzerine yapılacak çalışmalar sistemlerin doğruluğunu daha da artıracaktır. Bu gelişmeler beraberinde yüz ifadesi tanımanın farklı alanlarda kullanımını da artıracaktır. Çevrimiçi öğrenme, robotlar ve sanal asistanlar gibi teknolojilerde yüz ifadesi tanıma için gerçek zamanlılık ve ölçeklenebilirlik alanlarındaki çalışmalara da yönelim olacaktır. Yüz ifadesi tanımanın bu alanlardaki kullanımların artmasıyla daha fazla ölçeklenebilirliğe ve daha fazla hıza da ihtiyaç artacaktır. Bu ihtiyaçlar bu alanlara yönelimi de artıracaktır [9].

2.9. Açıklanabilir Yüz İfadesi Tanıma Üzerine Daha Önce Yapılan Çalışmalar

İnsan bilgisayar etkileşiminin birçok alanda artması beraberinde bilgisayar karşısındaki insanların yüz ifadelerinden o anki duygularının otomatik olarak tespit edilmesini sağlayan sistemlerin varlığını da gerekli kılmıştır. Literatürde birçok araştırmacı yüz ifadelerinden duygu tanıma üzerine farklı yöntemler kullanarak hızlı ve doğruluğu yüksek modeller geliştirmek için birçok çalışma yapmışlardır. Yapılan çalışmalar kullanılan verisetleri, modeller ve açıklanabilirlik açısından incelenmiştir.

Literatürde yer alan çalışmalar kullanılan verisetleri olarak incelendiğinde en çok kullanılan verisetleri şunlardır: Fer2013, 48*48 piksel boyutlarında, yedi farklı ifadeden oluşan 28.709 eğitim, 3.589 doğrulama ve 3.589 test görüntüsünden oluşmaktadır [11]. JAFFE, 60 kadın denek tarafından kontrollü bir ortamda pozlanan, yedi farklı duyguyu ifade eden 256*256 boyutlarında 213 görüntüden oluşmaktadır [12]. MMI, 43 farklı kişiye ait hem statik hem de video görüntülerinden oluşan 1280 görüntü dizisi ve 250 görüntü içermektedir [1]. Cohn–Kanade (CK), 97 kişiye ait 486 adet 640×480 veya 640×490 piksel çözünürlüğe sahip çerçeveli video bölümünden oluşur. CK+, 66 kişinin 122 spontane gülümseme ve 593 pozlu ifade içerir. CK ve CK+, kapalı yüzleri içermez [13]. Binghamton Üniversitesi 3B yüz ifadesi (BU-3DFE), 56 Kadın 44 Erkek olmak üzere 100 kişiden alınan 606 yüz ifadesini içermektedir [14]. PIE, 15 bakış açısı ve 19 aydınlatma koşulu altında 337 deneğe ait altı ifadeden biriyle etiketlenmiş 755.370 görüntüden oluşur [5]. Oulu-CASIA, 80 denekten toplanan altı ifadeden biriyle etiketlenmiş 2.880 görüntü dizisinden oluşmaktadır [15]. Vahşi Doğada Harekete Geçirilmiş Yüz İfadesi (AFEW), spontane ifadeler, çeşitli baş pozları, oklüzyonlar ve aydınlatmalarla farklı filmlerden derlenmiş video kliplerden oluşurken SFEW ise AFEW'den statik görüntü çerçeveleri çıkarılarak

türetilmiştir, yedi farklı ifadeyle etiketlenmiş 700 resimden oluşmaktadır [16]. Gerçek Dünya Efektif Yüz Veritabanı (RAF-DB), internetten indirilen 29.672 çok çeşitli yüz görüntüsünü içeren gerçek dünyaya ait bir verisetidir [17]. LIRIS, nötr duygu dışında beş evrensel ifade içeren web kamerasıyla spontane olarak kaydedilmiş, 30 fps hızında 24.000 duygusal çerçeveye sahip 12 öznenin 206 video klibini içermektedir [4].

Literatürde yer alan çalışmalar kullanılan yöntemler bakımından incelendiğinde öne çıkan çalışmalar şunlardır: Sikkandar ve Thiyagarajan (2021) yaptıkları çalışmada insan yüz ifadesi tanımlamak için yeni bir yaklaşım olan Geliştirilmiş Cat Swarm Optimizasyonu (ICSO) algoritmasını önermişlerdir. Bu yaklaşımda yüz ifadesinden özellik çıkarımında Derin Evrişim Sinir Ağı (DCNN) ve yüz ifadelerinin sınıflandırılmasında Neural Network (NN) ve Support Vector Machine (SVM) algoritmaları kullanılmaktadır [5]. Yao, L. Ve ark. (2022), yaptığı çalışmada otomatik yüz ifadesi tanımayı gerçekleştirmek için alan ve kanala dayalı hibrit bir dikkat mekanizması olan Yükseklik Performans Modülü Uygulaması (HPMI) dikkat mekanizması önermektedir. VGG-16 ağına gömülü olan HPMI modülü modelin başarısını yaklaşık yüzde 4 artırmıştır [13]. S. Shaees ve ark. (2020) önceden eğitilmiş en iyi CNN'lerden biri olan AlexNet ile Destek Vektör Makinesi (SVM) arasında karşılaştırma yapmaktadır [18]. Ahuja ve Vishwakarma (2022), Bu çalışmada, CNN ve çekirdek Extreme Learning Machine (KELM) in beraber kullanıldığı (CNN-KELM) şeklinde yüz tanıma için yeni bir öğrenme çerçevesi tanıtmaktadır [19]. Wu ve ark. (2019), yüz ifadesi tanımanın bir sınıflandırıcısı olarak Grafik üzerinde evrişimli sinir ağı (GCN) ağını yönsüz oluşturmak için sabit noktaları rastgele noktalarla birleştirmenin yeni bir yöntemini önerirler [20]. Akhand, M. A. H ve ark. (2021), Son derece hassas bir duygu tanıma sistemi geliştirmek için bu çalışmada, yoğun üst katmanlarını duygu tanıma ile uyumlu değiştirerek önceden eğitilmiş bir DCNN modelinin benimsendiği Transfer Öğrenimi (TL) tekniği aracılığıyla çok Derin bir CNN (DCNN) modellemesi önermektedir [21]. Sahoo G. K. ve ark. (2022), Bu çalışma araç içi uygulamada için duygu tanıma sürecini, yüz görüntüsü elde etme, görüntü ön işleme, önceden eğitilmiş SqueezeNet CNN modelini kullanarak özellik çıkarma ve sınıflandırmayı kapsamaktadır [22]. Tegani, S., ve Abdelmoutia, T. (2021), Covid-19 sürecinde hayatımıza giren maskelerden sonra maskelenmiş bir yüzde duygu tanıma yapabilmek için Derin Konvolüsyonel Sinir Ağları (DCNN) algoritmasının uygulanmasını önermektedir [23].

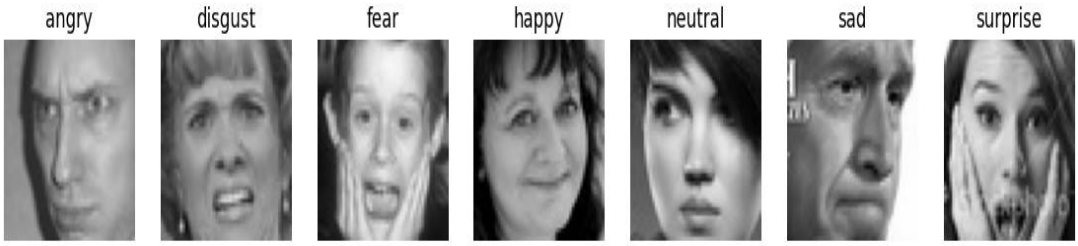
Literatürde yer alan çalışmalar yorumlanabilirlik ve Açıklanabilirlik bakımından incelendiğinde öne çıkan çalışmalar şunlardır: Araf T. A. Ve ark. (2022), Duygu tanıma için Cascade Classifier ve model tahmininin görselleştirilmesi için Grad-CAM kullanmışlardır [24]. Rathod M. Ve ark. (2022), oluşturdukları CNN modellerinin Grad-CAM, Grad-CAM++ ve SoftGrad kullanarak veri kümesindeki duyguları nasıl tanıdığına odaklanmışlardır [4]. Mery D. (2022), Bu çalışmada Duygu tanıma modelinin içinde yüz analizi yaklaşımını açıklamak için MinPlus adlı bir belirginlik haritası yöntemi sunulmakta ve MinPlus ile AVG, LIME ve RISE gibi diğer yöntemler karşılaştırılmaktadır [25].

3. YÖNTEM

Bu bölüm veri kümesinin kalitesini ve niceliğini artırmak için kullanılan ön işleme adımlarını ve veri artırma tekniklerini, evrişimli sinir ağı (CNN) modelimizin tasarımını ve uygulamasını, eğitim ve doğrulama sürecini, model performansını değerlendirmek için kullanılan değerlendirme ölçütlerini ve modelin tahminlerini yorumlamak için kullanılan Açıklanabilirlik teknikleri (LIME, SHAP, GradCAM, GradCAM++ ve Saliency Haritası) barındırmaktadır. Ayrıca çalışmamızda kullanılan istatistiksel analiz yöntemleri de detaylandırılmıştır. Bu yöntemlerle, yüz ifadesi tanıma modelimizin geliştirilmesi ve değerlendirilmesine ilişkin kapsamlı bir anlayış sağlamayı amaçlıyoruz.

3.1. Ön İşleme ve Veri Büyütme

Deneysel çalışmalarımızda kullanmış olduğumuz Fer2013 veriseti toplam 35887 adet 48x48 boyutunda gri-seviye yüz ifadesinden oluşmaktadır. Verisetinde bulunan yüz ifadeleri kızgın, tiksinti, korku, mutlu, üzgün, sürpriz, nötr olarak etiketlenmiş olup her ifade için yeterli sayıda görüntü mevcuttur. Veriseti dördüncü bölümde detaylı olarak açıklanmıştır. Verisetimize ait örnek resimler Şekil 3-1'de gösterilmiştir.



Şekil 3-1 Fer2013 örnek resimler

Oluşturulan modellerin daha iyi genelleme yapabilmesi verisetinin çeşitliliğinin olabildiğince yüksek olması gerekir. Veri ön işleme ve veri artırma daha fazla çeşitlilikle beraber modelin farklı açılardan ve ölçeklerden veri öğrenmesini sağlar. Deneysel çalışmalarımızda kullandığımız VGG ve ResNet modellerinde verisetimizdeki resimlerin orijinal boyutu olan 48x48 kullanılırken Inception modelinde bu model için geçerli minimum boyut olan 75x75 için resimlerin boyutu 75x75 olarak değiştirilmiştir. Verilerin uyumunu artırmak ve eğitim sürecinin performansını yükseltmek için görüntülerin pixel değerleri 0-1 arasında ölçeklendirilmiştir.

Modellerimizin daha iyi genelleme yapması ve aşırı uyumu azaltmak için verisetimizde bulunan verileri çeşitlendirerek veriseti genişletilmiştir. Kullanılan veri artırma teknikleri rastgele döndürme, yatay/dikey kaydırma, yakınlaştırma ve yansıtmadır. Verisetimizde veri ön işleme ve veri artırma işlemleri için oluşturmuş olduğumuz ImageDataGenerator nesnesine ait parametreler ve bu parametre değerleri şunlardır:

- **rescale:** Görüntü piksel değerlerini ölçeklendirmek için 1./255 olarak ayarlanmıştır.
- **rotation_range:** Görüntülerin rastgele döndürmek için 10 derece olarak ayarlanmıştır.
- **width_shift_range ve height_shift_range:** Görüntülerin %10 oranında yatay ve dikey olarak kaydırılmak için 0.1 olarak ayarlanmıştır.
- **zoom_range:** Görüntülerin rastgele olarak %10 oranında yakınlaştırmak ve uzaklaştırmak için 0.1 olarak ayarlanmıştır.
- **horizontal_flip:** Görüntülerin yatay olarak simetrisi alınmasını için True olarak ayarlanmıştır.

3.2. Model Mimarisi: Yüz İfadesi Tanıma İçin Evrişimli Bir Sinir Ağı Tasarımı

Evrişimli Sinir Ağları: David Hubel ve Torston Wiesel tarafından 1950'lerde yapılan biyolojik deneyler sırasında ortaya çıkmıştır [26]. CNN'ler günümüzde nesne tanıma, resim sınıflandırma ve duygu tanıma problemlerinde yaygın olarak kullanılmaktadır. CNN modelinin çalışma süreci şu şekilde işler: İlk adımda girdi olarak verilen resmin evrişim katmanında (Convolutional Layer) özellikleri çıkarılır. İkinci adım olan havuzlama (Pooling) katmanında bir önceki katmanda çıkarılan önemli özelliklere daha çok odaklanabilmek için gereksiz özellikler çıkarılır. Üçüncü adımda matris şeklindeki görsel düzleştirme katmanında (Flattening Layer) düz bir vektör haline getirilir. Son adım olan Tam Bağlantılı katmanda (Fully-Connected layer) yapay sinir ağlarıyla öğrenme gerçekleşir. CNN modelinin en önemli adımı son adımdır çünkü sınıflandırma bu adımda gerçekleşir [27]. Günümüzde yaygın olarak kullanılan bazı popüler CNN mimarileri şunlardır: Alexnet, ResNet, VGGnet, Inception, GoogleNet ve Densnet'dir [28].

Tez çalışması kapsamında deneylerde kullanılmak üzere 13 evrişim ve 3 tam bağlı olmak üzere 16 ana katmadan oluşan VGG [29], VGG ağlarından daha derin fakat

daha az karmaşıklığa sahip ve 152 katmandan oluşan ResNet [30] ve daha derine gitme fikriyle hem daha derin hem de daha geniş bir model olarak tasarlanan Inception [31] mimarileri seçilmiştir.

3.3.Önceden Eğitilmiş Modeller

Deneysel çalışmalarda daha önce başarılı sonuçlar elde edilmiş VGG16, ResNet ve InceptionV3 modelleri kullanılarak oluşturulan 3 farklı model kullanılmıştır. Modeller oluşturulurken daha önce eğitilmiş modellere ek olarak, Dense, MaxPooling2D, BatchNormalization ve Conv2D katmanları eklenmiştir. İlk olarak önceden eğitilmiş modelin çıktı katmanı yerine, önce 128 nöron ve ReLU aktivasyonu kullanan bir Dense katmanı eklenmiştir. Daha sonra ise bir MaxPooling2D katmanı ile bir BatchNormalization katmanı eklenmiştir. Bu katmanlardan sonra 256,512 ve 1024'lük filtrelerden oluşan 3x3 boyutunda 3 adet Conv2D katmanı eklenmiştir. Ayrıca her Conv2D katmanından sonra boyutları azaltmak için bir adet MaxPooling2D katmanı ile aşırı uyumu önlemek için bir adet Dropout katmanı eklenmiştir. Daha sonra veriyi tek boyutlu tensörlere dönüştürmek için bir Flatten katmanı eklenmiştir. Model 256 nörona sahip ve ReLU aktivasyonu kullanan Dense katmanı ile Sınıf sayısına eşit ve aktivasyonu Softmax olan ikinci Dense katmanı ile tamamlanmıştır. Daha önceden eğitilmiş modellere toplam 9 yeni katman eklenmiştir.

3.4.Eğitim ve Doğrulama

Tez çalışmamız kapsamında modellerimizin daha iyi performans göstermeleri için parametre ve hiperparametreleri için ayarlamalar yapılmıştır. Modellerin "input" parametresi temel alınan modele göre yani VGG ve ResNet için 48x48x3, Inception [32] için 75x75x3, "output" parametresi ise modellerin çıktı katmanı olarak ayarlanmıştır. Optimizasyon algoritması, "learning_rate" parametre değeri 0.0001 olan Adam algoritması kullanılmıştır. Kayıp fonksiyonu olarak ise çok sınıflı sınıflandırma problemlerinde kullanılan "categorical_crossentropy" fonksiyonu tercih edilmiştir.

Modellerin eğitim aşamasında eğitim veriseti olarak 28709 görüntüden oluşan eğitim seti ve test veriseti olarak 7178 görüntüden oluşan test seti kullanılmıştır. Modelin eğitileceği toplam epoch sayısı 50 olarak belirlenmiştir. Modelleri eğitiminde iyileşme olmaması durumunda eğitimi erken durdurmak için Erken Durdurma

(EarlyStopping) parametresi, takip edilecek değer test doğruluğu ve sabredilecek tur sayısı 7 olacak şekilde ayarlanmıştır. Sınıflar arasındaki dengesizliği telafi etmek için ise Sınıf Ağırlıkları (class_weights) parametresi ayarlanmıştır. Bu parametre için önce eğitim veri setindeki sınıf etiketlerinin sayısı hesaplanmış daha sonra en fazla örneği olan sınıfın sayısı, tüm sınıfların sayılarının içinde en büyük olan değer olarak belirlenmiştir son olarak ise her sınıfın ağırlığı, en büyük sayıya bölünerek hesaplanmıştır. Hesaplanan değerler sınıf id ve ağırlık olacak şekilde bir sözlükte depolanmıştır. Oluşturulan sözlük modelin eğitimi aşamasında class_weights parametresine değer olarak girilmiştir.

3.5.Değerlendirme Metodolojisi

Açıklanabilir Yüz İfadesi Tanıma için oluşturulan modellerin ve kullanılan hiperparametre değerlerinin karşılaştırılması için standart metrikler olan Doğruluk, Duyarlılık, Hassasiyet, F1_score ve AUC kullanılmıştır. Tüm metrikler dikkate alınsa da Test doğruluğu ile modelimizin duyarlılık ve özgüllük performansını gösteren AUC daha ön planda tutulmuştur. İlk değerlendirme karşılaştırmalar bu metriklere göre yapıldıktan sonra modellerin tekrar değerlendirilmesi için yedi farklı duygu durumunu temsilen yedi görüntü seçilmiş ve modellerin bu görüntüler için yaptıkları tahminler üzerinden modeller ikinci defa karşılaştırılmıştır.

3.6.Açıklanabilirlik Teknikleri

Oluşturulan modellerin standart metrikler ile karşılaştırılması ve değerlendirilmesinden sonra modellerin aldıkları kararlarda odaklandıkları noktalar belirlemek ve modelleri bu yönden de karşılaştırabilmek için GradCAM, GradCAM++, Saliency haritası, GradCAM++ ve Saliency haritası, LIME, SHAP gibi teknikler kullanarak her duygu durumunda modelin odaklandıkları noktalar incelenmiştir. Karşılaştırmalar hem duygu durumuna hem de modele göre yapılmıştır. Modelle birlikte tekniklerde karşılaştırılmıştır. Bu tekniklerin detaylı açıklaması ise şu şekildedir:

3.6.1 LIME (Local Interpretable Model-Agnostic Explanations)

Yerel Yorumlanabilir Model Agnostik Açıklamalar (LIME), yapay zekâ modellerinin yorumlanmasında kullanılan en popüler ve yaygın tekniklerden biridir. LIME, modelin tahminine katkıda bulunan girdinin ana kısımlarının betimler. Bunun

için önce modelin girdilerini bozar daha sonra modelin yeni tahminlerinin nasıl davrandığını gözlemler. Son olarak, karışıklıkların ağırlıklandırılması yoluyla doğrusal bir model kullanarak modelin nasıl çalıştığını öğrenir [8]. Ortaya çıkan açıklamalar küresel olarak geçerli olmasa da yerel olarak doğrudur. Ayrıca LIME model-agnostiktir yani herhangi bir modeli açıklayabilir, modelden bağımsızdır. LIME, görüntü sınıflandırması için kullanıldığında görüntünün bağlı bir alanını kaplayan süper pikselleri veya piksel koleksiyonunu vurgular. LIME, görüntünün bölümlerini sistematik olarak kapatarak en göze çarpan girdi özelliklerini haritalar [33]. Bir görüntünün yanlış şekilde bölümlere ayrılması görüntünün önemli özelliklerinin bölünmesine neden olabileceği için görüntünün bölümlere nasıl ayrıldığı çok önemlidir. LIME tekniğinin amacı görüntü bölümleme tekniği kullanarak görüntüdeki anlamlı bölümleri ortaya çıkarmaktır.

3.6.2 SHAP (SHapley Additive exPlanations)

SHAP, herhangi bir makine öğrenimi modelinin çıktısını açıklamaya yönelik oyun kuramsal bir yaklaşımdır [34]. Toplu bir ödülü elde etmek için birlikte mücadele eden bir grup oyuncunun ödülü paylaşırken adil bir şekilde paylaşabilmesi için her bir oyuncunun ödül için bireysel katkılarının bilinmesi gerekir. 1951'de Lloyd Shapley, bu tip durumlar için "Shapley değeri" terimini ortaya koymuştur. Shapley değeri her bir oyuncunun ödül için bireysel katkısını ifade etmektedir. SHAP, yapay zekâ modellerinin tahminlerini Shapely değerlerini kullanarak yorumlayabilmek için Lundberg ve Lee tarafından 2017 yılında ortaya atılmıştır. SHAP yönteminde ana fikir her bir özelliğin model tahminindeki etkisini bulmak için her özelliğin Shapley değerini hesaplamaktır [35]. Görselleştirmede Shapley değeri yüksek özellikler kırmızı ile ifade edilirken düşük değerler mavi renk ile gösterilir.

3.6.3 GradCAM (Gradient-weighted Class Activation Mapping)

Gradyan ağırlıklı sınıf aktivasyon haritalama (Grad-CAM), CNN tabanlı yapay zekâ modelleri ile yapılan tahminleri açıklamak yani modelleri daha şeffaf hale getirmek için kullanılan tekniklerden biridir [36]. Grad-CAM, görüntü sınıflandırma, duygu tanıma vb. birçok farklı görevde yapay zekâ modellerinin tahmini için görüntülerdeki önemli bölgeleri aslına uygun olarak ısı haritası ile vurgulayabilir [37]. Bu yöntemin uygulanmasında üç ana adım vardır: İlk adımda gradyan hesabı yapılır

daha sonra özellik haritaları üzerinde gradyanın tüm elemanlarının ortalamaları ile havuzlama yapılır. Son adımda ise Grad-CAM ısı haritası oluşturulur. Isı haritasında Sarı-Kırmızı olan noktalar, modelin tahminini yaparken dikkat ettiği noktaları gösterirken Mavi olan noktalar ise açıklıkların en az olduğu yerleri göstermektedir [38].

Modele ait tahminin bir GradCAM gösterimini elde etmek için takip edilmesi gereken adımlar ve bu adımların matematiksel olarak açıklaması şunlardır [39].

1.Adım Gradyan Hesabı, y^{logit} softmax sonucu sınıf skoru ve F^n n adet özellik haritası olmak üzere $\frac{\partial y^{logit}}{\partial F^n}$ şeklinde yapılır.

2.Adım Aktivasyon haritalarını derecelendirmek

Bu aşamada nöron önem ağırlıklarını (α_k^{logit}) elde etmek için Y yükseklik, G genişlik olmak üzere Global Average Poolling ($\frac{1}{Y.G}\sum_i\sum_j$) ve Backprop ile gelen gradyanlar ($\frac{\partial y^{logit}}{\partial F_{i,j}^n}$) kullanılır. Bu adım için kullanılan matematiksel formül, α_k^{logit}= $\frac{1}{Y.G}\sum_i\sum_j\frac{\partial y^{logit}}{\partial F_{i,j}^n}$ şeklindedir.

3.Adım GradCAM ısı haritası

GradCAM ısı haritası ($L_{Grad-CAM}^{logit}$) elde etmek için bir önceki adımdaki formülden yararlanarak elde edilen ağırlıklar ile aktivasyon haritalarının ağırlıklı kombinasyonları alınır. Negatif değerleri elemek için ise RELU kullanılır. Bu adımın matematiksel gösterimi ise $L_{Grad-CAM}^{logit} = ReLU\left(\sum_n \alpha_n^{logit}.F^n\right)$ şeklindedir.

GradCAM ısı haritası, tahmin edilecek sınıflara ait özellik haritasını kullanır. Bu sayede hangi özelliğin hangi sınıf için belirleyici olduğu oluşan ısı haritası ile belirlenebilir. Örneğin mutlu olarak tahmin edilen resimlerde yüzün hangi bölümüne odaklanıldığı bu ısı haritası ile tespit edilebilir. Odaklanılan özellikler dikkate alınarak model ve yapılan tahmin için olumlu veya olumsuz yorumlar yapılabilir ve modelin çalışma şekli açıklanabilir.

3.6.4 GradCAM++ (Gradient-weighted Class Activation Mapping Plus Plus)

GradCAM++, GradCAM' in gelişmiş bir versiyonu olmakla beraber aynı işlevi yerine getirir, yani girdideki özelliklerin tahmindeki katkılarını ısı haritası şeklinde gösterir. GradCAM 'den farklı olarak GradCAM++, tahmin için doğru gradyanların yüksekliklerini ve yönlerini ağırlıklandırarak daha geniş bir alanı kapsar. GradCAM++, bir sınıfa ait özelliğin görüntüde olabileceği tüm konumlarında ısı haritaları üretir bu sayede oluşan ısı haritası, görüntüde bir sınıfın tahmini için birden çok özellik bulunduğunda modelin davranışını daha iyi açıklar [40].

GradCAM++, GradCAM 'in daha genelleştirilmiş bir versiyonudur ve GradCAM den farklı olan alanların matematiksel olarak açıklaması şu şekildedir [41]:

1. Belirli bir sınıfı (c) ve aktivasyon haritası (k) için gradyan ağırlıklarını α_{ij}^{kc} elde etmek için GradCAM' den farklı olarak yeni bir formül türetilmiştir ve formülasyonu şu şekildedir:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} + \sum_a \sum_b A_{ab}^k \left\{\frac{\partial^3 Y^c}{\left(\partial A_{ij}^k\right)^3}\right\}}$$

2.Adım GradCAM++ ısı haritası

GradCAM den farklı olarak GradCAM++ belirli bir etkinleştirme haritasının önemini yakalar A^k ve çıkış nöronunun aktivasyonunu bastırmak yerine artıran görsel özellikleri belirtmek için pozitif gradyanları tercih eder. GradCAM++ ısı haritasının (ω_k^c) formülasyonu ise şu şekildedir:

$$\omega_k^c = \sum_i \sum_j \left[\frac{\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} + \sum_a \sum_b A_{ab}^k \left\{\frac{\partial^3 Y^c}{\left(\partial A_{ij}^k\right)^3}\right\}}\right] . relu\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right)$$

GradCAM++ ile GradCAM birbirine benzerdir fakat GradCAM++, sınıf özellik haritasının yanı sıra son özellik haritasının gradyanlarının da hesaplanmasını içerdiğinden dolayı daha ayrıntılı ve kesindir.

3.6.5 Saliency Haritası

Saliency Haritası yani Belirginlik Haritaları bir görüntüde bulunan önemli özelliklerin vurgulanmasını sağlayan tekniklerden biridir. İlk olarak Oxford Üniversitesi araştırmacıları tarafından "Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps" adlı makale ile sunulmuştur. Belirginlik haritaları, görüntüdeki özellikleri ağırlandırarak nereye odaklanıldığını tahmin etmek ve tahmini anlamak için kullanılır. Belirginlik haritaları, görüntüdeki özelliklerin önem derecesini ölçmek için renk, yoğunluk, kenarlar, kontrast, nesne boyutu, şekil gibi birçok özellikle ilişkilendirilebilir. Belirginlik haritaları genel olarak üç adımda oluşturulur. İlk olarak görüntüdeki renk, yoğunluk ve yön gibi temel özellikler görüntüden çıkarılır. Daha sonra bu görüntü özellik haritaları ve Gauss Piramitleri oluşturmak için kullanılır. Son adımda oluşturulan tüm özellik haritalarının ortalaması alınarak Belirginlik Haritaları oluşturulur [42].

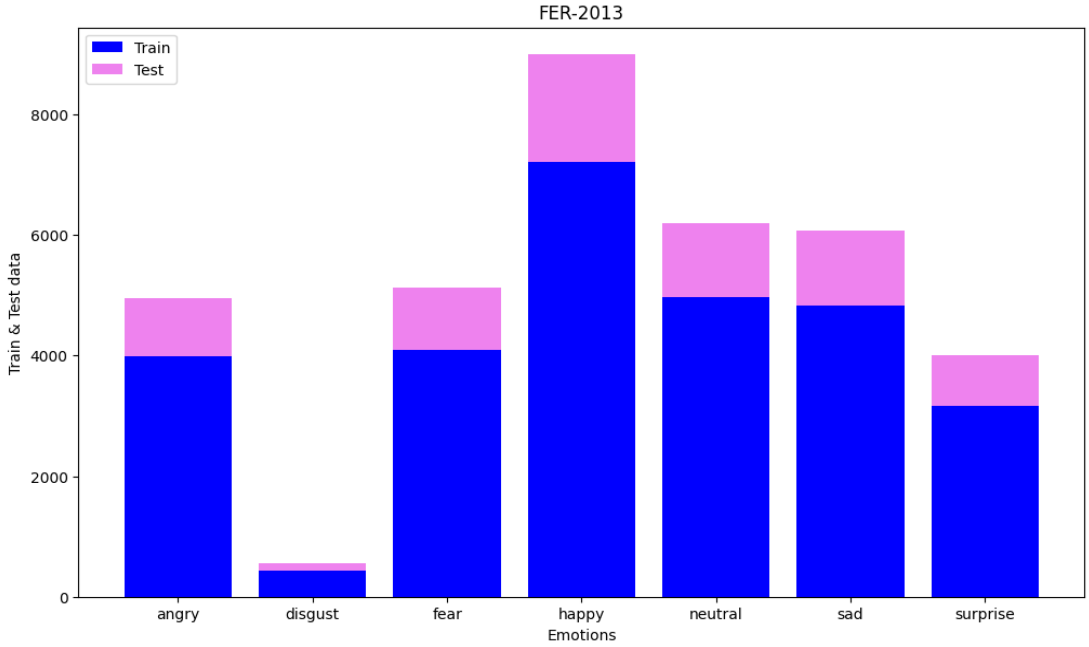
4. DENEYSEL SONUÇLAR

Bu bölümde ilk olarak tez çalışması kapsamında kullanılan veriseti, veri önişleme ve veri artırma teknikleri uygulanarak genişletilmiştir. Daha sonra temel alınan modelleri geliştirmek için ek katmanlar ile iyileştirmeler yapılmıştır. Modeller belirlenen parametreler ve Fer2013 veriseti ile eğitilmiştir. Ayrıca eğitilen her model için elde edilen metrikler gösterilmiştir. Son olarak bu modellerin her duygu durumu için seçilen bir resim üzerinde tahmin yapmaları sağlanmıştır. Yapılan tahminlerde odaklanılan noktalar GradCAM, GradCAM++, Saliency haritası, Shap ve LIME ile görselleştirilmiştir.

4.1 Test Ortamı ve Veriseti

Bu tez çalışması kapsamında yapılan tüm deneyler için kullanılan bilgisayarın özellikleri şunlardır: İşlemci 11.Nesil Intel Core i7-11800H 2.30GHz, Ram 16GB DDR4 3200MHz, Ekran kartı 4GB NVIDIA GeForce RTX 3050 ve kullanılan işletim sistemi Windows 10 Profesyonel (64 Bit) işletim sistemidir. Programlama editörü olarak da Jupyter Notebook [34] tercih edilmiştir.

Oluşturulan modellerin başarısı için modellerin eğitimi için kullanılan veriseti çok önemlidir. Bu tez çalışması kapsamında yapılan deneylerde halka açık Fer2013(Yüz İfadesi Tanıma 2013) [43] veriseti kullanılmıştır. Fer2013 veriseti toplam 35887 adet 48x48 boyutunda gri-seviye yüz ifadesinden oluşmaktadır. Verisetinde bulunan ifadelerin 28709'u eğitim, 3589'u doğrulama ve 3589'u test verisidir. Şekil 4-1'de Fer2013 verisetinde bulunan sınıflar ile bu sınıflar ile etiketlenmiş eğitim ve test verilerinin dağılımı gösterilmiştir. Verisetinde bulunan yüz ifadeleri kızgın, tiksinti, korku, mutlu, üzgün, sürpriz, nötr olarak etiketlenmiş olup tiksinti ifadesinde yaklaşık 600 olmasına rağmen diğer tüm etiketlerde yaklaşık 5000 örnek vardır. Fer2013 verisetine ait örnek resimler Şekil 4-2'de gösterilmektedir.



Şekil 4-1 Fer2013 verisetinde eğitim ve test verisi dağılımı

Şekil 4-2 Fer2013 veriseti örnek resimler

4.2 Modellere Ait Performans Değerleri

Tüm modeller veri ön işleme ve artırma işlemlerinden sonra aynı iyileştirmeler yapıldıktan sonra 50 adımlık eğitime tabi tutulmuştur. Modellere ait performans karşılaştırması için doğruluk, kayıp, duyarlılık, hassasiyet, Auc ve F1_score olmak üzere beş metrik belirlenmiş ve karşılaştırmalar bu metriklere göre yapılmıştır.

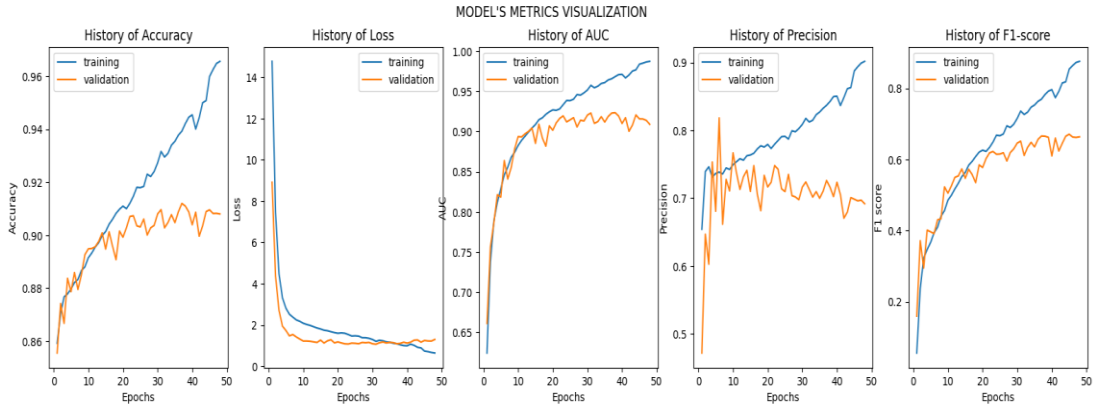
VGG16 modeli kullanarak oluşturulan ilk model 50 adımlık eğitimini daha fazla gelişme göstermediği için 48.adımda sonlandırmıştır. Bu modelin başarısını değerlendirmek için kullanılan metrikler ve bu metriklere için elde edilen değerler: Doğruluk (Accuracy) eğitim 0.945 test 0.912, Kayıp (Loss) eğitim 0.641 test 1.084, Hassasiyet (Precision) eğitim 0.857 test 0.726, Duyarlılık (Recall) eğitim 0.737 test 0.616, AUC eğitim 0.973 test 0.923 ve F1_score eğitim 0.792 test 0.665 şekildedir. Elde edilen değerlere göre modelimiz eğitim verilerine fazlaca uyum sağlamış fakat test verilerinde aynı başarıyı gösterememiştir.

ResNet50 modeli kullanarak oluşturduğumuz ikinci model 50 adımlık eğitimini tamamlamıştır. Bu modelin başarısını değerlendirmek için kullanılan metrikler ve bu metriklere için elde edilen değerler: Doğruluk (Accuracy) eğitim 0.986 test 0.904, Kayıp (Loss) eğitim 0.247 test 1.507, Hassasiyet (Precision) eğitim 0.957 test 0.670, Duyarlılık (Recall) eğitim 0.943 test 0.641, AUC eğitim 0.997 test 0.898 ve F1_score eğitim 0.950 test 0.653 şekildedir. Elde edilen değerlere göre modelimiz eğitim verilerine fazlaca uyum sağlamış fakat test verilerinde aynı başarıyı gösterememiştir.

InceptionV3 modeli kullanarak oluşturduğumuz üçüncü model 50 adımlık eğitimini daha fazla gelişme göstermediği için 38.adımda sonlandırmıştır. Bu modelin başarısını değerlendirmek için kullanılan metrikler ve bu metriklere için elde edilen değerler: Doğruluk (Accuracy) eğitim 0.934 test 0.908, Kayıp (Loss) eğitim 0.826 test 1.163, Hassasiyet (Precision) eğitim 0.851 test 0.734, Duyarlılık (Recall) eğitim 0.654 test 0.560, AUC eğitim 0.961 test 0.913 ve F1_score eğitim 0.739 test 0.635 şekildedir. Elde edilen değerlere göre modelimiz eğitim verilerine fazlaca uyum sağlamış fakat test verilerinde aynı başarıyı gösterememiştir.

Tüm modeller karşılaştırıldığında eğitim verileri en iyi model ResNet50 iken test verilerinde en başarılı model VGG16 olmuştur. VGG16 modeline ait metrikler Şekil

4-3'deki grafiklerde gösterilmiştir. Ayrıca tüm modeller eğitim verilerindeki başarısı test aşamasına aynen yansıtamamıştır. Modellere ait ölçüm sonuçları Tablo 4-1'de verilmiştir.



Şekil 4-3. VGG modeline ait metrikler

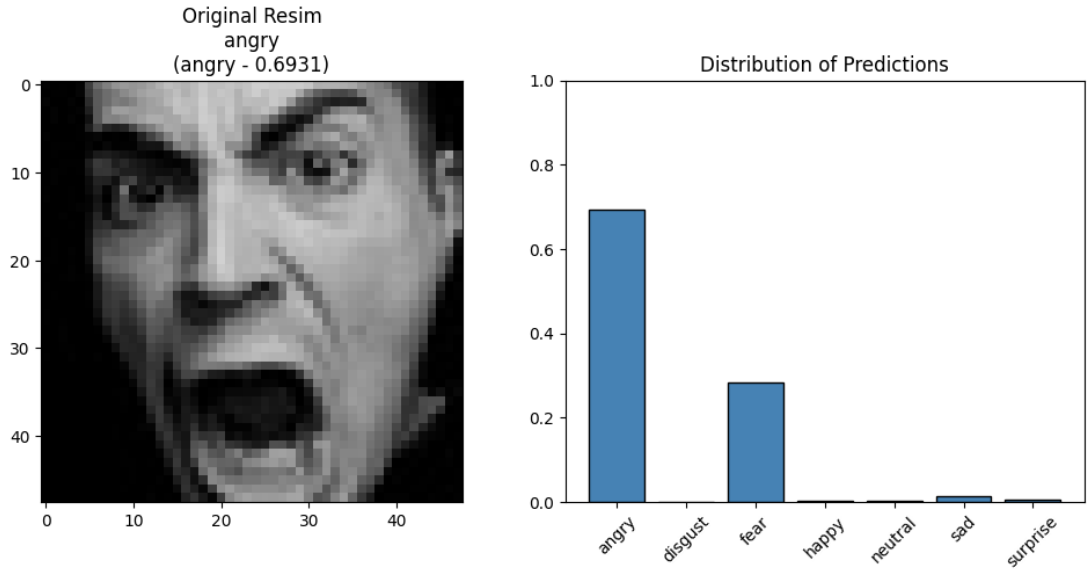
Tablo 4-1 Modellerin değerlendirme metriklerine ait değerler

Metrikler	Modeller	VGG	ResNet	Inception
Doğruluk (Accuracy)	Eğitim	0,945	0,986	0,934
	Test	0,912	0,904	0,908
Kayıp (Loss)	Eğitim	0,641	0,247	0,826
	Test	1,084	1,507	1,163
Hassasiyet (Precision)	Eğitim	0,857	0,957	0,851
	Test	0,726	0,670	0,734
Duyarlılık (Recall)	Eğitim	0,737	0,943	0,654
	Test	0,616	0,641	0,560
AUC	Eğitim	0,973	0,997	0,961
	Test	0,923	0,898	0,913
F1_score	Eğitim	0,792	0,950	0,739
	Test	0,665	0,653	0,635

4.3 Gerçek veriler ile Modellerin Karşılaştırılması

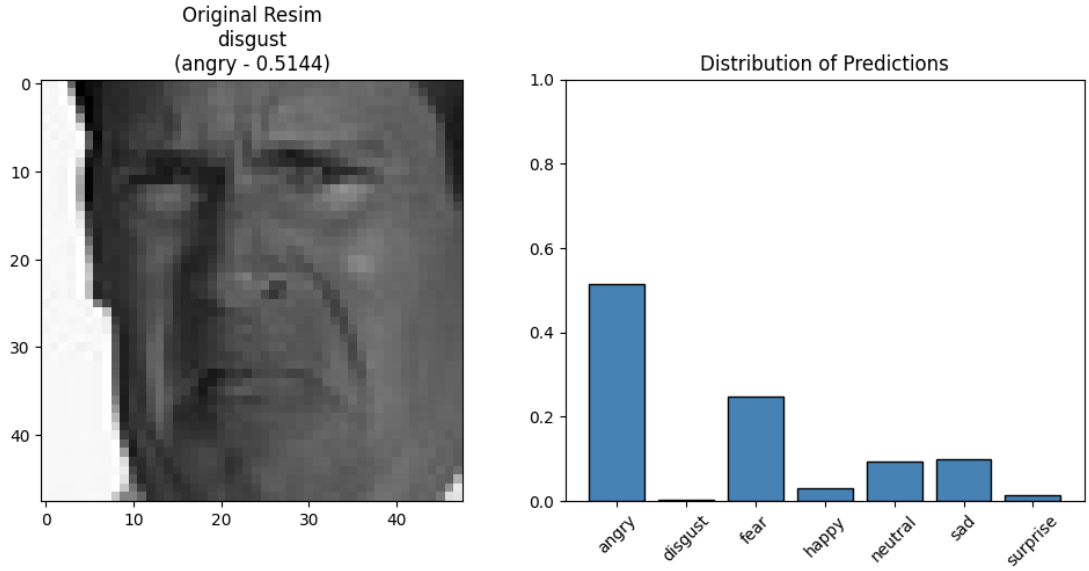
Modellerin ikinci karşılaştırılma aşamasında 7 duygu durumunu temsilen seçilen 7 görüntü üç modele de ayrı ayrı tahmin ettirilmiştir. Modellerin başarısı görüntü için yaptığı en yüksek tahmin kadar diğer tahmin olasılıkları ile de ilgili olduğu için orijinal görüntü ile beraber tahmin dağılımına ait grafiklerde görüntü ile birlikte görüntülenmiştir.

İlk olarak sinirli ifadesi için bir resim seçilmiş ve bu resim üç farklı modelle tahmin edilmiştir. VGG modeli bu resmi 0.6283 oranla sinirli, ResNet modeli 0.6931 oranla sinirli ve Inception modeli 0.5387 oranla sinirli olarak tahmin etmiştir. Üç model de resmi doğru tahmin ederken orijinal resim ve en yüksek oranla doğru tahminde buluna ResNet modeline ait tahmin dağılımı Şekil 4-4'te gösterilmiştir.



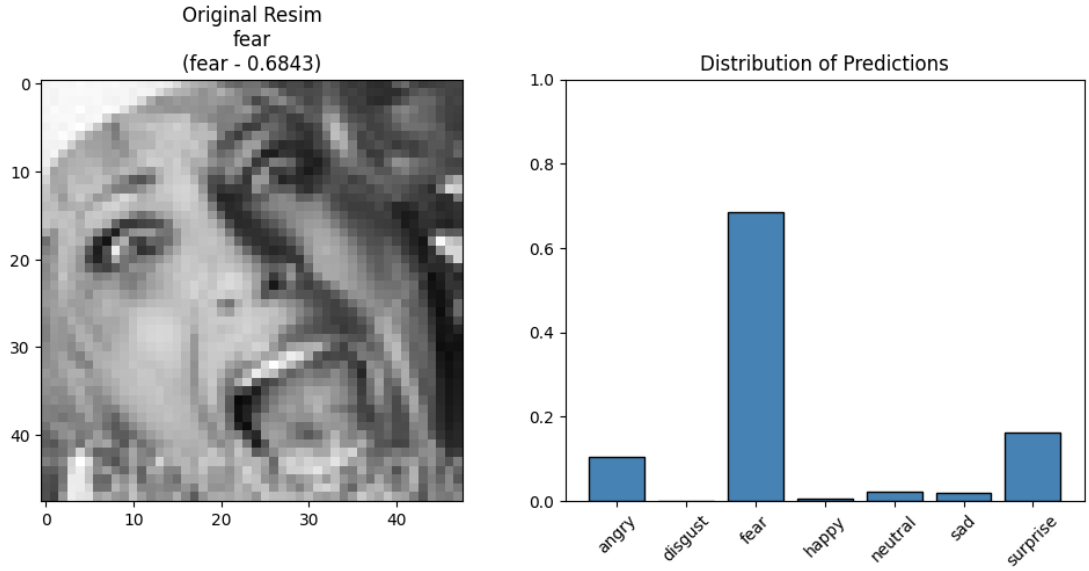
Şekil 4-4. ResNet Modeli ile "Sinirli" duygu durumu tahmini

İkinci olarak tiksinti ifadesi için bir resim seçilmiş ve bu resim üç farklı modelle tahmin edilmiştir. VGG modeli bu resmi 0.5144 oranla sinirli, ResNet modeli 0.4247 oranla mutlu ve Inception modeli 0.4601 oranla üzgün olarak tahmin etmiştir. Üç model de resmi doğru tahmin edemezken üç modelde farklı tahminde bulunmuştur. Üç modelin yaptığı tahmindeki tek benzerlik üç modelin tahmininde de sinirli tahminin birinci veya ikinci sırada olmasıdır. Orijinal resim ve tahmin dağılımı Şekil 4-5'te gösterilmiştir.



Şekil 4-5. VGG ile "Tiksinti" duygu durumu tahmini

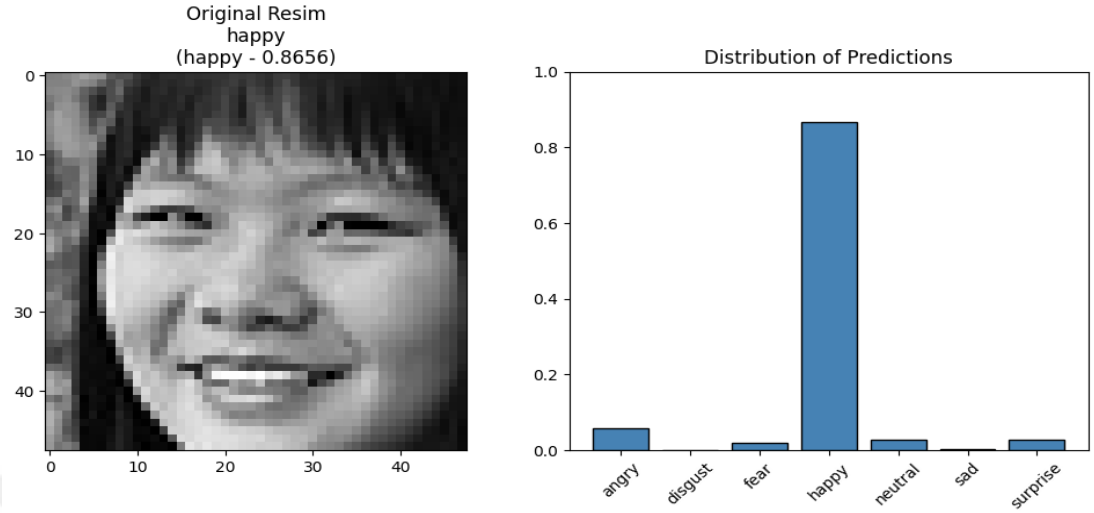
Üçüncü olarak korku ifadesi için bir resim seçilmiş ve bu resim üç farklı modelle tahmin edilmiştir. VGG modeli bu resmi 0.6843 oranla korku, ResNet modeli 0.5543 oranla korku ve Inception modeli 0.6094 oranla sinirli olarak tahmin etmiştir. Orijinal resim ve en yüksek oranla doğru tahminde buluna VGG modeline ait tahmin dağılımı Şekil 4-6'da gösterilmiştir.



Şekil 4-6. VGG modeline ait "Korku" duygu durumu tahmini

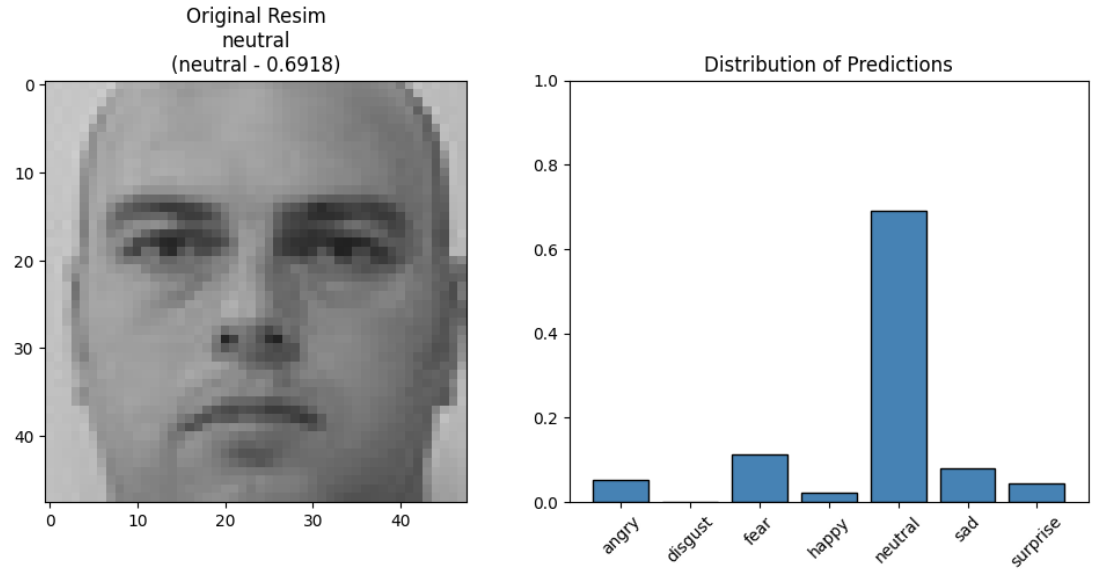
Dördüncü olarak mutlu ifadesi için bir resim seçilmiş ve bu resim üç farklı modelle tahmin edilmiştir. VGG modeli bu resmi 0.8656 oranla mutlu, ResNet modeli 0.5052 oranla sinirli ve Inception modeli 0.4278 oranla mutlu olarak tahmin etmiştir.

Orijinal resim ve en yüksek oranla doğru tahminde buluna VGG modeline ait tahmin dağılımı Şekil 4-7'de gösterilmiştir.



Şekil 4-7 VGG modeli "Mutlu" duygu durumu tahmini

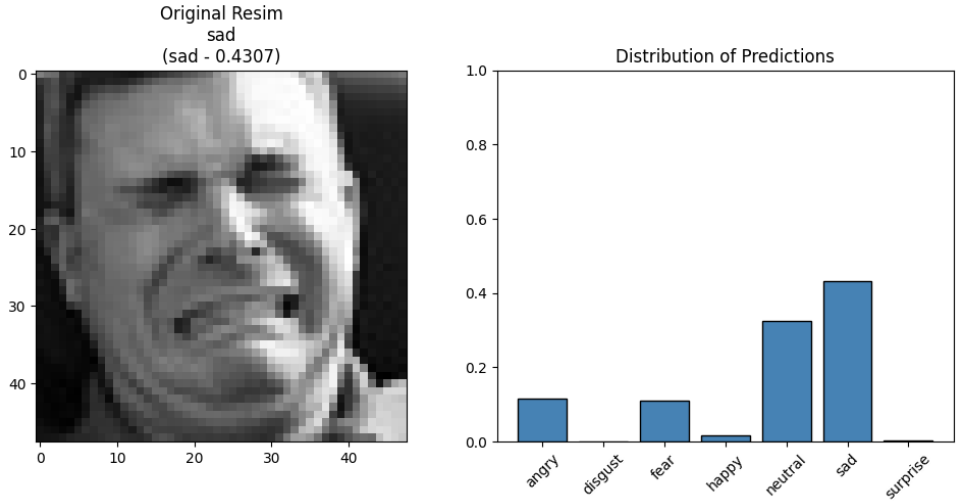
Beşinci olarak nötr ifadesi için bir resim seçilmiş ve bu resim üç farklı modelle tahmin edilmiştir. VGG modeli bu resmi 0.6918 oranla nötr, ResNet modeli 0.6684 oranla korku ve Inception modeli 0.5432 oranla korku olarak tahmin etmiştir. Sadece VGG modeli doğru tahmin ederken orijinal resim ve modele ait tahmin dağılımı Şekil 4-8'de gösterilmiştir.



Şekil 4-8. VGG modeli "Nötr" duygu durumu tahmini

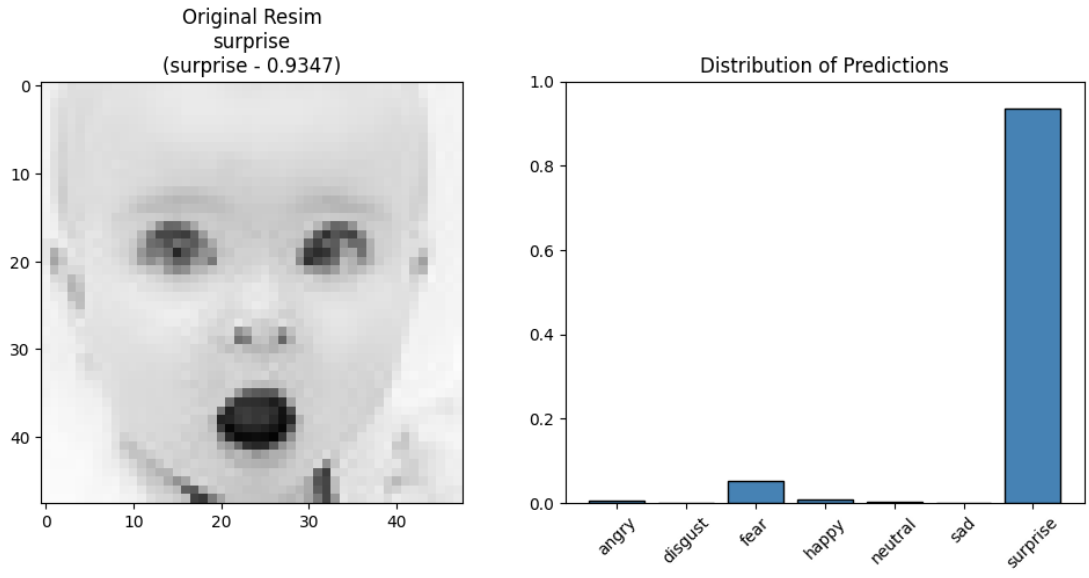
Altıncı olarak üzgün ifadesi için bir resim seçilmiş ve bu resim üç farklı modelle tahmin edilmiştir. VGG modeli bu resmi 0.4307 oranla üzgün, ResNet modeli 0.9846 oranla sinirli ve Inception modeli 0.7364 oranla sinirli olarak tahmin etmiştir.

Sadece VGG modeli doğru tahmin ederken orijinal resim ve modele ait tahmin dağılımı Şekil 4-9'da gösterilmiştir.



Şekil 4-9. VGG modeli ile "Üzgün" duygu durumu tahmini

Yedinci olarak sürpriz ifadesi için bir resim seçilmiş ve bu resim üç farklı modelle tahmin edilmiştir. VGG modeli bu resmi 0.7695 oranla sürpriz, ResNet modeli 0.9347 oranla sürpriz ve Inception modeli 0.4671 oranla sürpriz olarak tahmin etmiştir. Üç model de resmi doğru tahmin ederken orijinal resim ve en yüksek oranla doğru tahminde buluna ResNet modeline ait tahmin dağılımı Şekil 4-10'da gösterilmiştir.



Şekil 4-10. ResNet modeli ile "Sürpriz" duygu durumu tahmini

Yedi duygu durumu için seçilen resimlerin üç farklı modelle tahmin edildiği deneylerde sadece sinirli ve sürpriz ifadesi için seçilen resimler üç model tarafından da doğru tahmin edilmiştir. Ayrıca tiksinti ifadesi için seçilen resim üç farklı model tarafından da yanlış tahmin edilmiştir. En çok doğru tahminde bulunan model altı doğru tahmin ile VGG olurken en başarısız model iki doğru tahminle Inception olmuştur.

4.4 Açıklanabilirlik ve Yorumlanabilirlik Sonuçları

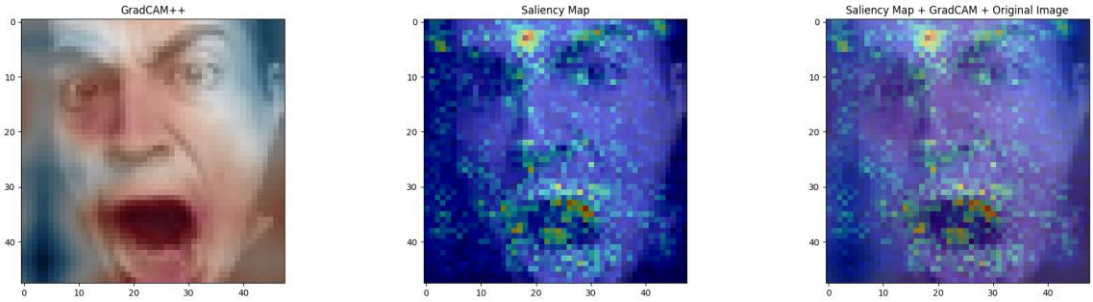
Bu bölümde XAI teknikleri ile her duygu durumu için modeller tarafından yapılan tahminlere ait odak noktaları gösterilmiştir. Her duygu durumu için en başarılı doğru tahmin ile en başarısız tahmine ait GradCAM ile GradCAM++, GradCAM++ ile Saliency haritası, SHAP ve LIME gösterimleri yapılarak modellerin doğru ve yanlış tahminde bulunurken odaklandığı noktalar incelenmiştir.

4.4.1 Sinirli İfadesi Açıklanabilirlik Analizi

Sinirli ifadesi için seçilen resim kullanılan üç farklı model tarafından da doğru tahmin edilmiş olup en yüksek tahminde bulunan model olan ResNet'in tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-11'de, GradCAM++ ile Saliency haritası Şekil 4-12'de, SHAP Şekil 4-13'te ve LIME Şekil 4-14'te gösterilmiştir. GradCAM ve GradCAM++ gösterimlerinde incelendiğinde modelin ağız, burun, gözlerin altı ve sol kaş bölgesine odaklandığı görülmüştür.

Şekil 4-11. ResNet ile "Sinirli" ifadesine ait GradCAM ve GradCAM++ gösterimi

GradCAM++ ve Saliency haritası gösterimleri incelendiğinde ise modelimizin GradCAM++ dan farklı olarak Saliency haritası daha çok ağız ve iki kaşın ortasına odaklandığı görülmektedir.

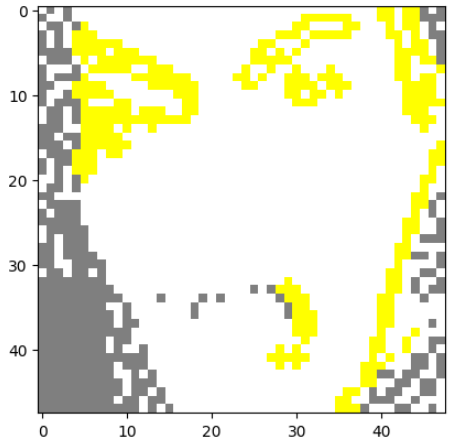


Şekil 4-12. ResNet ile "Sinirli" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

SHAP gösterimi incelendiğinde ise modelin en çok pozitif özellik eşleştirmesini korkuyla ifadesiyle yaptığı ve bu özelliklerin ağız ve yanak bölgelerinde toplandığı korku ifadesiyle eşleşen negatif özelliklerin ise iki kaşın ortasında toplandığı görülmektedir.

Şekil 4-13. ResNet modeli ile "Sinirli" ifadesine ait SHAP gösterimi

LIME gösterimi incelendiğinde ise modelin gözler ve kaşlar ile ağız çevresine odaklandığı görülmüştür. LIME gösterimi Şekil 4-46'dadır.

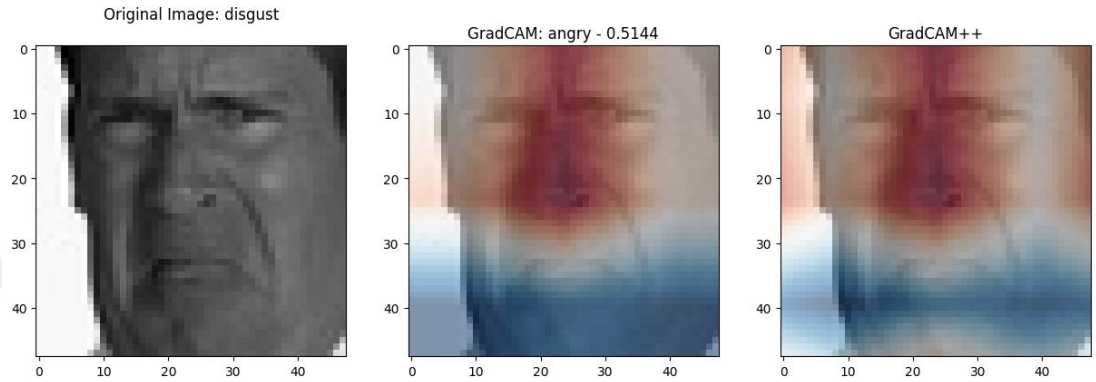


Şekil 4-14. ResNet modeli ile "Sinirli" ifadesine ait LIME gösterimi

4.4.2 Tiksinti İfadesi Açıklanabilirlik Analizi

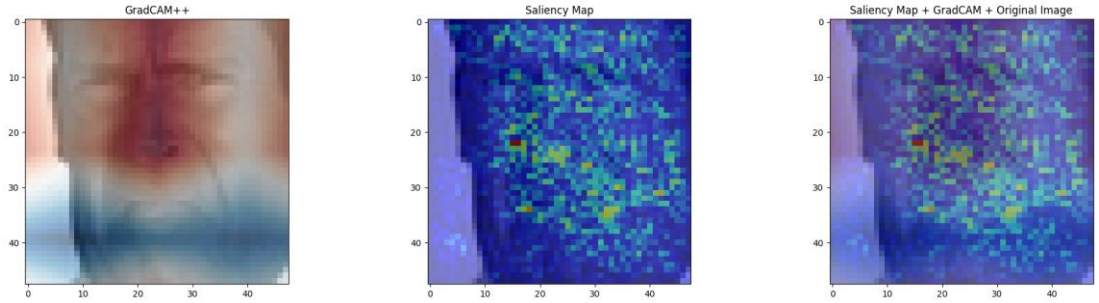
Tiksinti ifadesi için seçilen resim kullanılan üç farklı model tarafından da yanlış tahmin edilmiş olup en yüksek tahminde bulunan model olan ResNet'in

tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-15'te, GradCAM++ ile Saliency haritası Şekil 4-16'da, SHAP Şekil 4-17'de ve LIME Şekil 4-18'de gösterilmiştir. Modelin yaptığı yanlış tahmindeki nedenleri anlamak için modele ait GradCAM ve GradCAM++ gösterimleri incelendiğinde modelin Sinirli ifadesinde olduğu gibi yüzün orta noktası ile iki kaş arasına odaklandığı görülmektedir. Bu bölgeler dikkate alındığında modelin neden Sinirli tahmini yaptığı açıklanabilir.



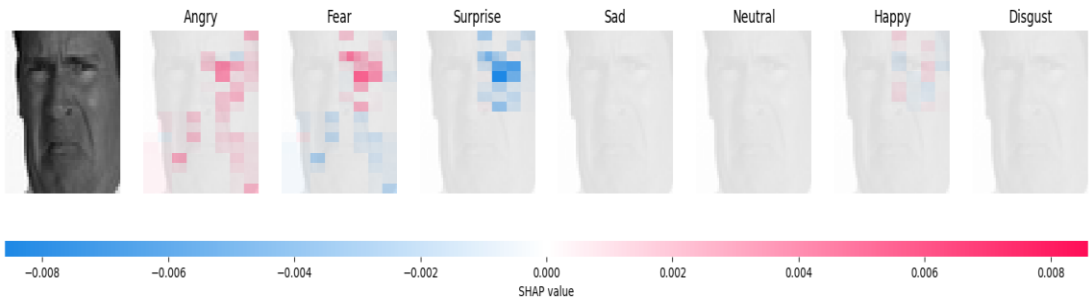
Şekil 4-15. VGG ile "Tiksinti" ifadesine ait GradCAM ve GradCAM++ gösterimi

GradCAM++ ve Saliency haritası gösterimi yan yana incelendiğinde ise modelin yüzün orta noktası ağırlıklı olmak üzere yüzün geneline odaklandığı görülmüştür.



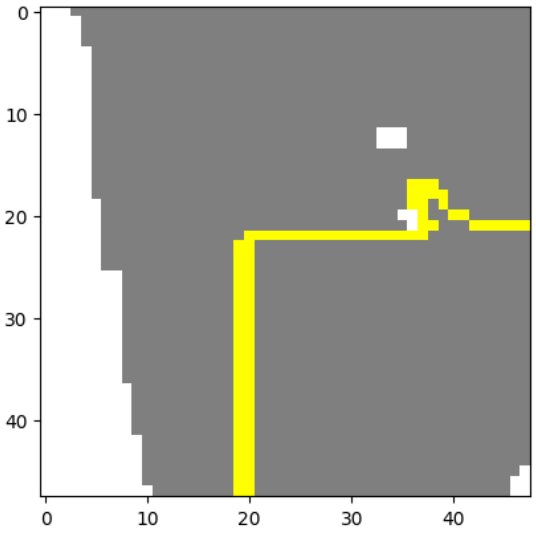
Şekil 4-16. VGG ile "Tiksinti" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

SHAP gösterimi incelendiğinde ise pozitif etkisi olan noktaların yüzün orta bölgesi ile Sol göz bölgesine toplandığı ve Sinirli ile Korku ifadeleriyle ilişkilendirildiği görülmüştür. Negatif etkisi olan noktaları ise sol göz üzerinde toplandığı ve sürpriz ifadesi ile ilişkilendirildiği görülmektedir.



Şekil 4-17. VGG modeli ile "Tiksinti" ifadesine ait SHAP gösterimi

LIME gösterimi incelendiğinde ise net olmamakla beraber sol göz bölgesi ile yüzün ortasına odaklanıldığı görülmektedir.



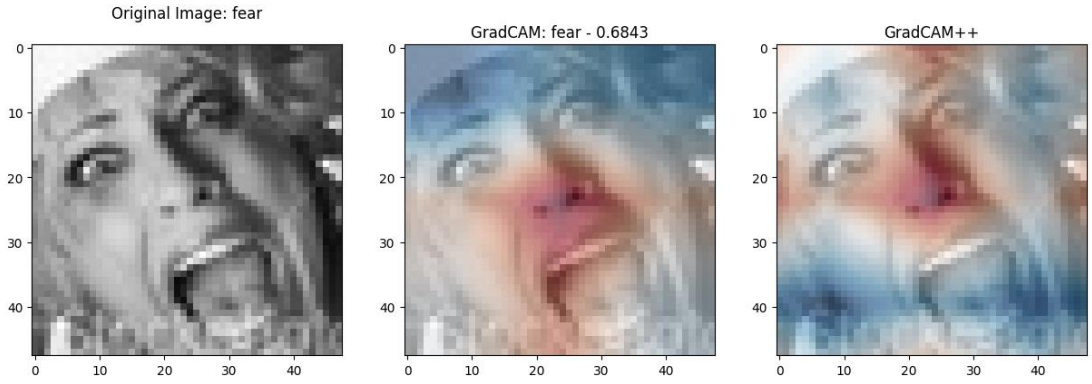
Şekil 4-18. VGG modeli ile "Tiksinti" ifadesine ait LIME gösterimi

4.4.3 Korku İfadesi Açıklanabilirlik Analizi

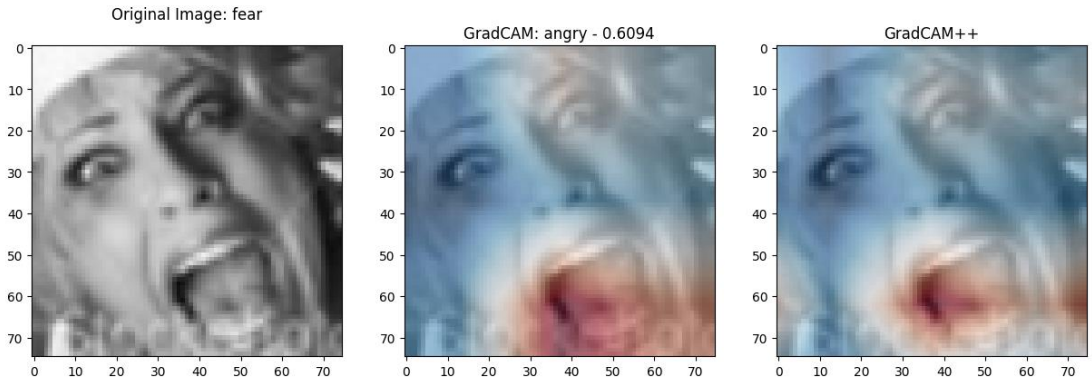
Korku ifadesi için seçilen resim VGG ve ResNet modelleri tarafından da doğru tahmin edilmiş olup en yüksek tahminde bulunan model olan VGG'nin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-19'da, GradCAM++ ile Saliency haritası Şekil 4-21'de, SHAP Şekil 4-23'te ve LIME Şekil 4-25'te gösterilmiştir. Yanlış tahminde bulunan Inception modelinin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-20'de, GradCAM++ ile Saliency haritası Şekil 4-22'de, SHAP Şekil 4-24'te ve LIME Şekil 4-26'da gösterilmiştir.

VGG modeline ait GradCAM gösterimi incelendiğinde modelin yüzün ortası ağırlıklı olmak üzere burun ve üst dudak bölgelerine odaklandığı görülmüştür. GradCAM ++ gösteriminde ise farklı olarak göz altı ve kaş üstü bölgelerine de odaklandığı görülmektedir. Inception modeline ait GradCAM ve GradCAM++

sonuçları incelendiğinde ağız çevresine odaklandığı görülmüştür.

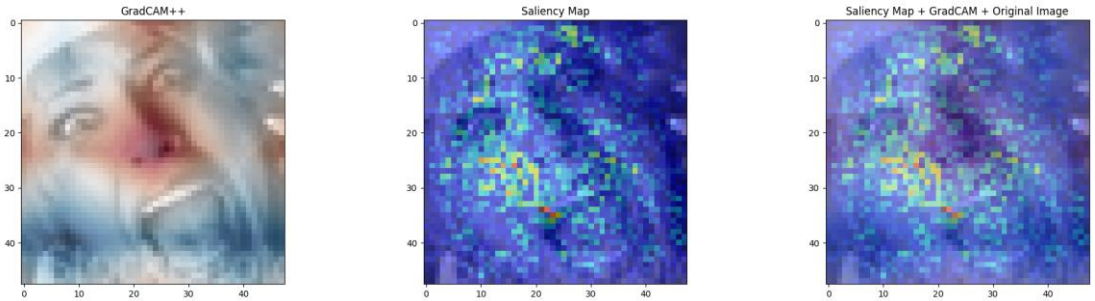


Şekil 4-19. VGG ile "Korku" ifadesine ait GradCAM ve GradCAM++ gösterimi

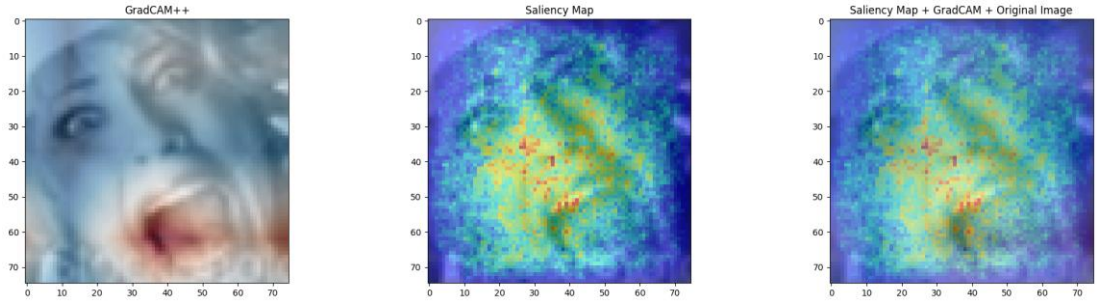


Şekil 4-20. Inception ile "Korku" ifadesine ait GradCAM ve GradCAM++ gösterimi

VGG modeline ait Saliency haritası ile odaklanılan noktalara bakıldığında burun çevresi ağız ve kaş üzeri olduğu görülmektedir. Inception modeline ait Saliency haritası incelendiğinde ise GradCAM++ gösterimindeki ağız çevresine ev olarak yüzün orta kısmına da odaklandığı görülmüştür.

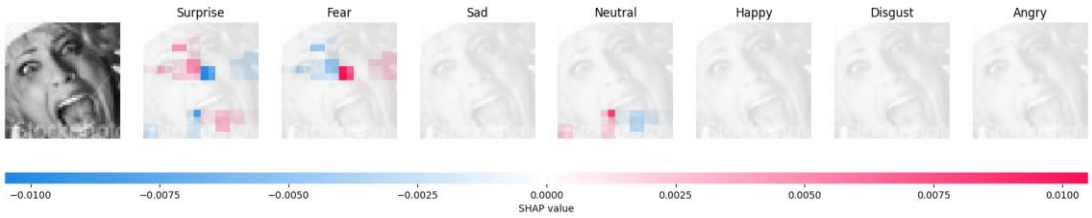


Şekil 4-21. VGG ile "Korku" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

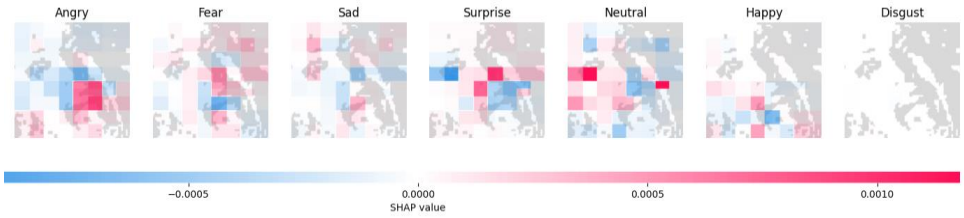


Şekil 4-22. Inception ile "Korku" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

VGG modeline ait SHAP gösterimini incelediğimizde korku duygu durumu için pozitif ilişki bulduğu noktalar burun ve ağız üstünde yoğunlaşmıştır. Sürpriz duygu durumu olarak tahmin etmesine sebep olan noktaların ise burun üstü ve göz çevresinde olduğu görülmektedir. Inception modeline ait SHAP gösterimi incelendiğinde korku ifadesi için pozitif etkisi olan özelliklerin yüzün geneline dağıldığı, negatif etkisi olan özelliklerin ise burun bölgesinde toplandığı görülmektedir. Ayrıca resmin sinirli ifadesi ile ilişkilendirilen özelliklerinin korkuya benzer ve ağız bölgesinde yoğun olduğu görülmüştür.

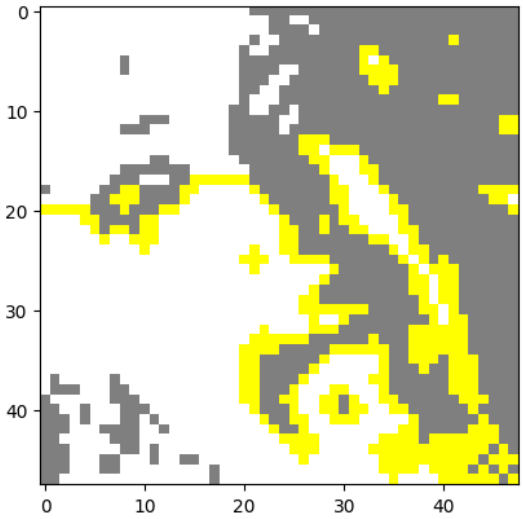


Şekil 4-23. VGG modeli ile "Korku" ifadesine ait SHAP gösterimi

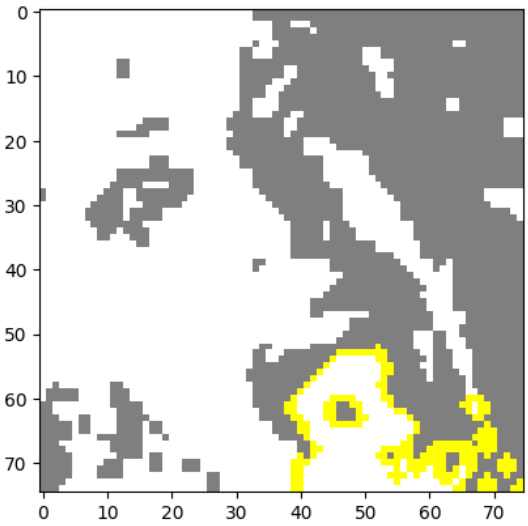


Şekil 4-24. Inception modeli ile "Korku" ifadesine ait SHAP gösterimi

VGG modeline ait LIME gösterimini incelediğimizde göz çevresi, burun ve ağız bölgesine dağıldığı görülmektedir. Inception modeline ait LIME gösterimi incelendiğinde ağız çevresine odaklandığı görülmektedir.



Şekil 4-25 VGG modeli ile "Korku" ifadesine ait LIME gösterimi

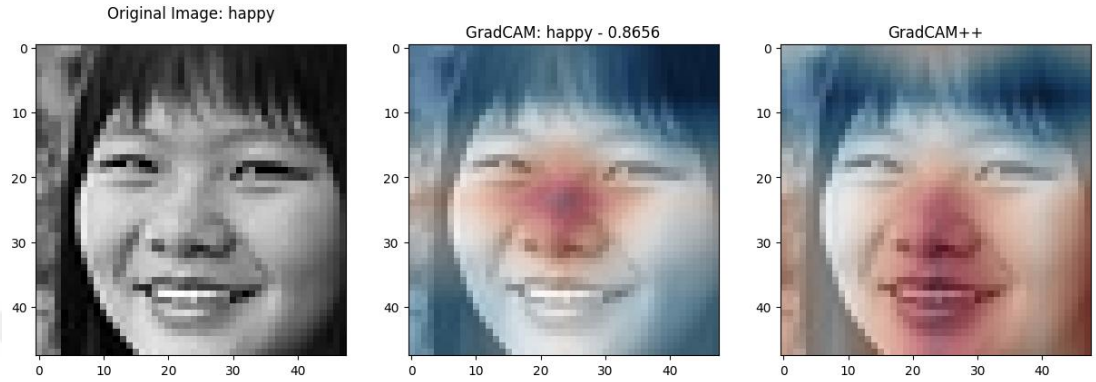


Şekil 4-26 Inception modeli ile "Korku" ifadesine ait LIME gösterimi

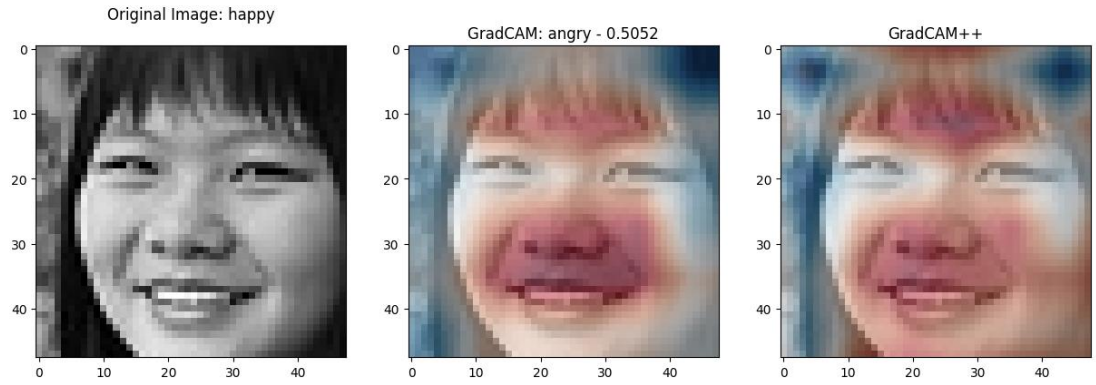
4.4.4 Mutlu İfadesi Açıklanabilirlik Analizi

Mutlu ifadesi için seçilen resim VGG ve Inception modelleri tarafından da doğru tahmin edilmiş olup en yüksek tahminde bulunan model olan VGG'nin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-27'de, GradCAM++ ile Saliency haritası Şekil 4-29'da, SHAP Şekil 4-31'de ve LIME Şekil 4-33'te gösterilmiştir. Yanlış tahminde bulunan ResNet modelinin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-28'de, GradCAM++ ile Saliency haritası Şekil 4-30'da, SHAP Şekil 4-32'de ve LIME Şekil 4-34'te gösterilmiştir.

VGG modeline ait GradCAM gösteriminde modelin odaklandığı noktanın burnu merkezi ile göz altlarının burun tarafı olduğu görülmüştür. GradCAM++ da ise farklı olarak modelin ağız bölgesine de odaklandığı görülmektedir. ResNet modeline ait GradCAM ve GradCAM++ gösterimleri incelendiğinde modelin ağız ve alın bölgelerine odaklandığı net bir şekilde görülmektedir.

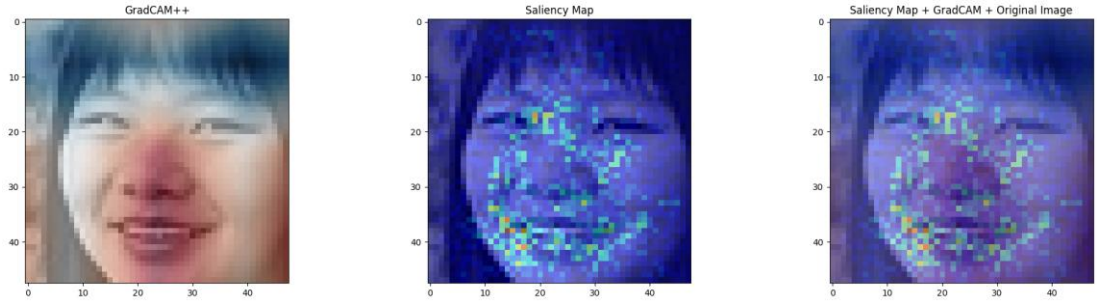


Şekil 4-27. VGG modeli ile "Mutlu" ifadesine ait GradCAM ve GradCAM++ gösterimi

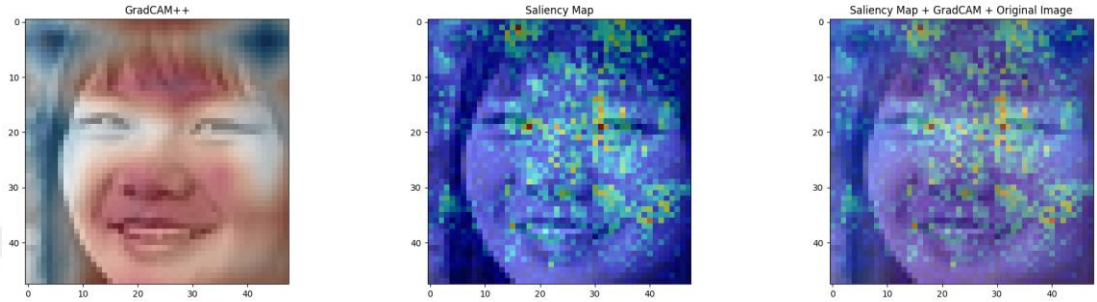


Şekil 4-28. ResNet ile "Mutlu" ifadesine ait GradCAM ve GradCAM++ gösterimi

VGG modeline ait GradCAM++ ve Saliency haritası birlikte görüntülendiğinde ise Saliency haritası gösteriminde de modelin GradCAM++'da olduğu gibi ağız bölgesi ağırlıklı olmak üzere yüzün merkezine odaklandığı görülmüştür. ResNet modeline ait Saliency haritası gösteriminde ise modelin GradCAM++ dan farklı olarak yanak gözler ve burun bölgesine odaklandığı görülmektedir.

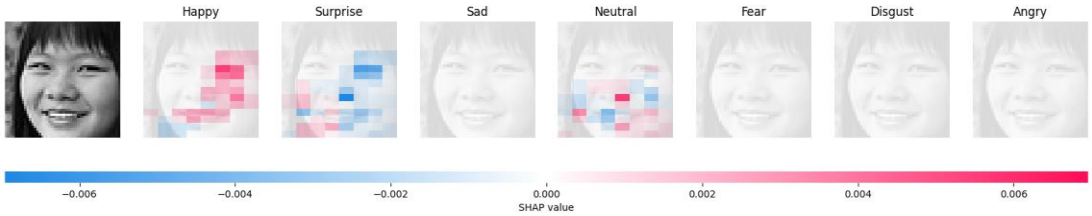


Şekil 4-29. VGG modeli ile "Mutlu" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

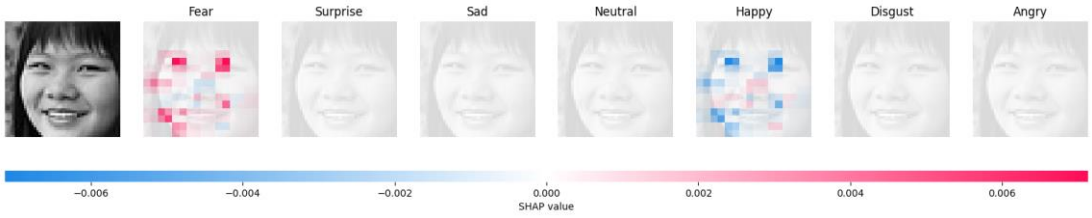


Şekil 4-30. ResNet ile "Mutlu" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

VGG modeline ait SHAP gösterimi incelendiğinde ise pozitif etkisi olan noktaları çoğunlukla ağız ve sol göz bölgesinde toplandığı ve Mutlu ifadesiyle ilişkilendirildiği görülmektedir. ResNet modeline ait SHAP gösteriminde ise modelin pozitif etkisi olan özellikleri çoğunlukla korku ifadesiyle ilişkilendirdiği ve bu özelliklerin ise gözler ve ağız çevresinde toplandığı görülmüştür. Mutlu ifadesinde ise bu bölgeler negatif olarak ilişkilendirilmiştir.

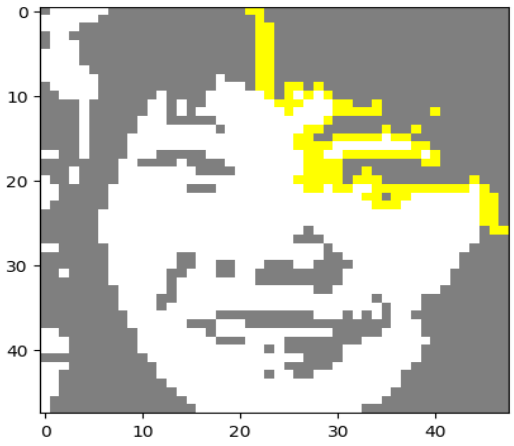


Şekil 4-31. VGG modeli ile "Mutlu" ifadesine ait SHAP gösterimi

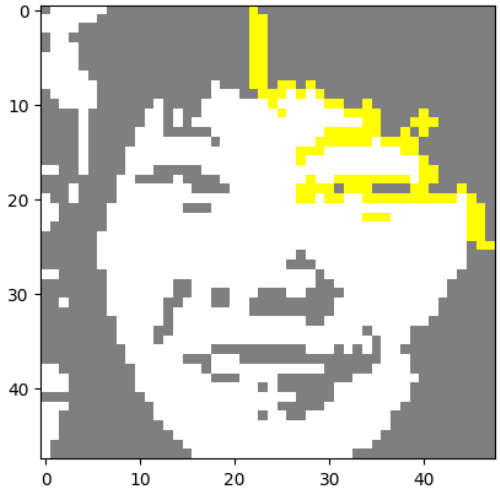


Şekil 4-32. ResNet modeli ile "Mutlu" ifadesine ait SHAP gösterimi

VGG modeline ait LIME gösteriminde ise modelin sol göz bölgesine odaklandığı görülmektedir. ResNet modeline ait LIME gösterimi de bize modelin sol göz bölgesine odaklandığını göstermektedir.



Şekil 4-33. VGG modeli ile "Mutlu" ifadesine ait LIME gösterimi

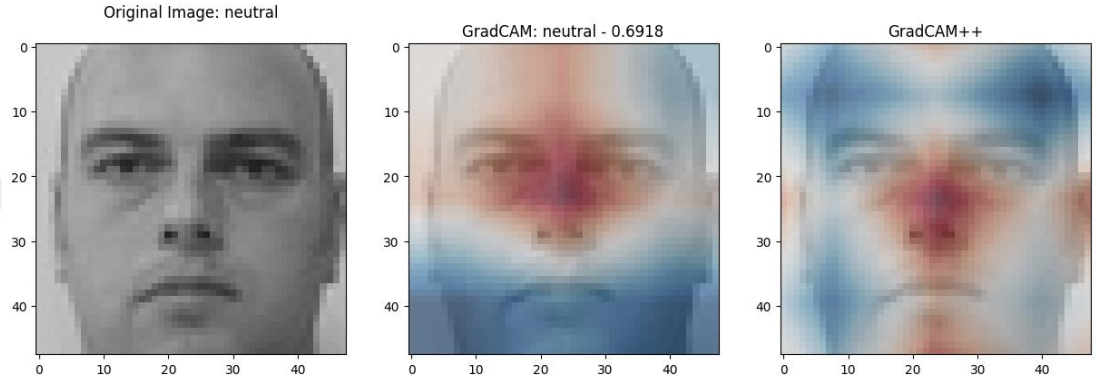


Şekil 4-34. ResNet modeli ile "Mutlu" ifadesine ait LIME gösterimi

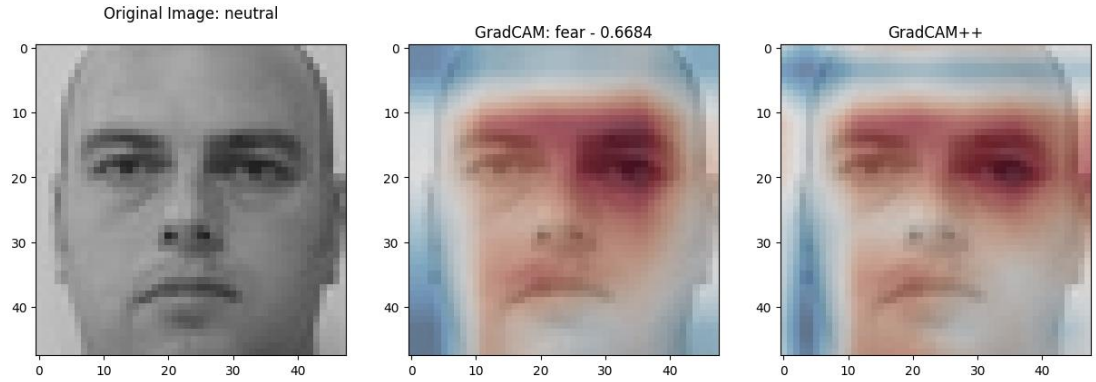
4.4.5 Nötr İfadesi Açıklanabilirlik Analizi

Nötr ifadesi için seçilen resim sadece VGG modeli tarafından da doğru tahmin edilmiş olup modelin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-35'te, GradCAM++ ile Saliency haritası Şekil 4-37'de, SHAP Şekil 4-39'da ve LIME Şekil 4-41'de gösterilmiştir. Yanlış tahminde bulunan ResNet modelinin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-36'da, GradCAM++ ile Saliency haritası Şekil 4-38'de, SHAP Şekil 4-40'ta ve LIME Şekil 4-42'de gösterilmiştir.

VGG modeline ait GradCAM ve GradCAM++ sonuçları incelendiğinde GradCAM tekniğinde burun, göz altı ve alın bölgelerine yoğunlaştığı görülmektedir. GradCAM++ gösteriminde ise burun üzeri ve göz altlarına odaklanmaktadır. ResNet modeline ait GradCAM ve GradCAM++ görüntüleri karşılaştırıldığında GradCAM de burun çevresine daha çok yayılmakla beraber ikisinde de kaş üstü, göz, burun ve ağız çevresine odaklandığı görülmektedir.

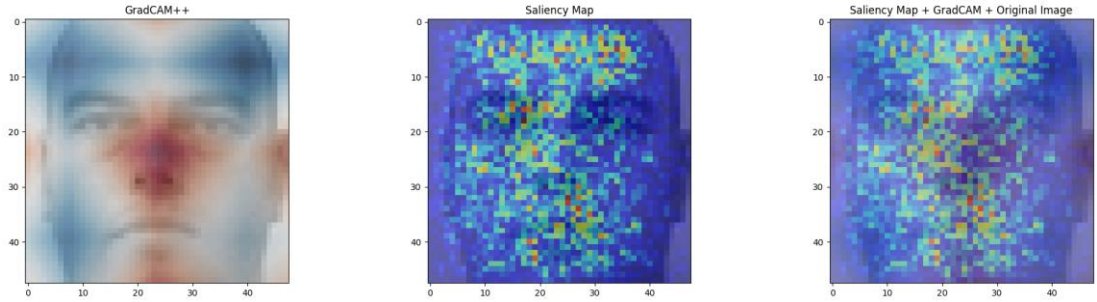


Şekil 4-35. VGG ile "Nötr" ifadesine ait GradCAM ve GradCAM++ gösterimi

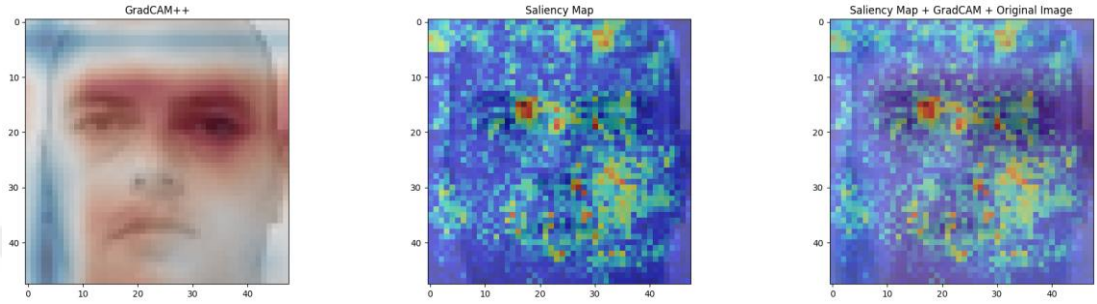


Şekil 4-36. ResNet ile "Nötr" ifadesine ait GradCAM ve GradCAM++ gösterimi

VGG modeline ait Saliency haritası gösterimi incelendiğinde odak noktalarının yüzün geneline yayıldığı görülmektedir. ResNet modeline ait Saliency haritası çalışması incelendiğinde odak noktalarının göz çevresinde yoğunlaşmasıyla beraber yüzün geneline yayıldığı görülmüştür.



Şekil 4-37. VGG ile "Nötr" ifadesine ait GradCAM++ ve Saliency haritası gösterimi



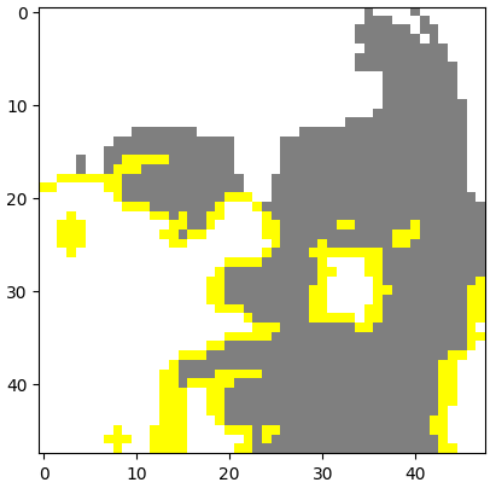
Şekil 4-38. ResNet ile "Nötr" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

VGG modeline ait SHAP tekniğinde ilk sırada nötr tahmininin göz, burun ve ağız çevresine bakılarak yapıldığı görülmektedir. Yine aynı noktaların sürpriz duygu durumu için negatif özellikte olduğu gösterilmiştir. ResNet modeline ait SHAP gösterimi incelendiğinde ilk sırada dudak altındaki noktalardan yola çıkarak mutlu duygu durumu olarak tahminde bulunduğu görülmektedir.

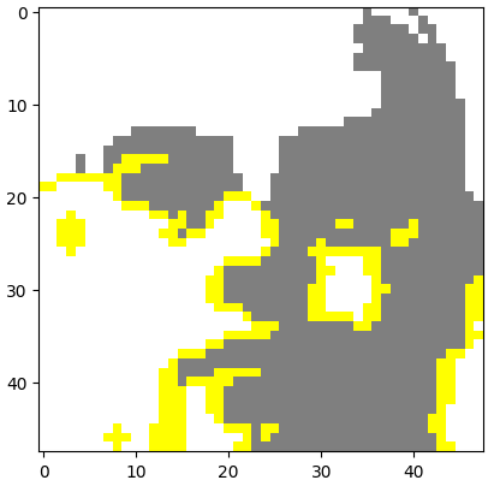
Şekil 4-39. VGG modeli ile "Nötr" ifadesine ait SHAP gösterimi

Şekil 4-40. ResNet modeli ile "Nötr" ifadesine ait SHAP gösterimi

VGG modeline ait LIME gösteriminde modelin göz çevresi, burun ve ağız bölgesine bakarak tahminde bulunduğu görülmektedir. ResNet modeline ait LIME gösterimi incelendiğinde göz altı, burun, yanak ve ağız çevresine dağıldığı görülmektedir.



Şekil 4-41. VGG modeli ile "Nötr" ifadesine ait LIME gösterimi

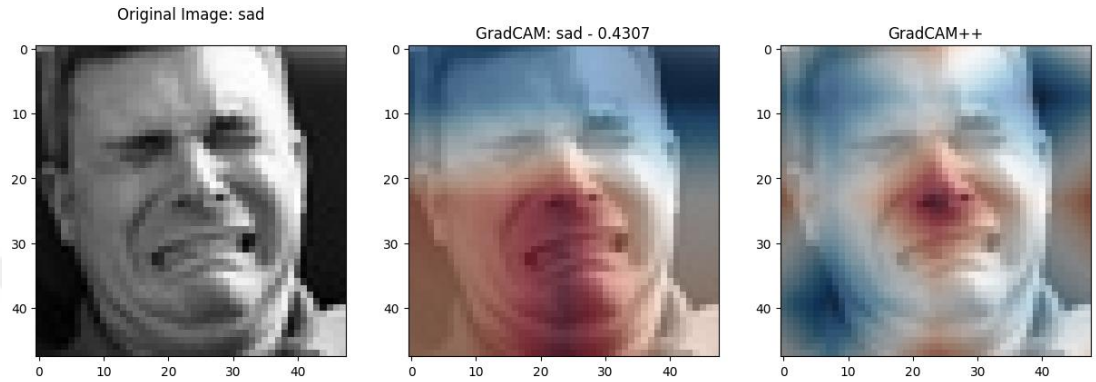


Şekil 4-42. ResNet modeli ile "Nötr" ifadesine ait LIME gösterimi

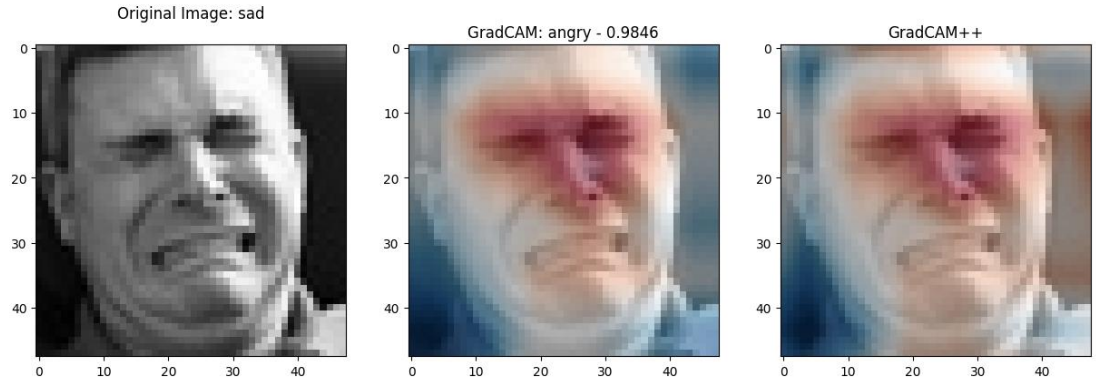
4.4.6 Üzgün İfadesi Açıklanabilirlik Analizi

Üzgün ifadesi için seçilen resim sadece VGG modeli tarafından doğru tahmin edilmiş olup modelin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-43'te, GradCAM++ ile Saliency haritası Şekil 4-45'te, SHAP Şekil 4-47'de ve LIME Şekil 4-49'da gösterilmiştir. Yanlış tahminde bulunan ResNet modelinin tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-44'de, GradCAM++ ile Saliency haritası Şekil 4-46'da, SHAP Şekil 4-48'de ve LIME Şekil 4-50'de gösterilmiştir.

VGG modeline ait GradCAM gösterimi incelendiğinde modelin başta ağız bölgesi olmak üzere yüzün alt bölgesine odaklandığı görülmektedir. GradCAM++ gösteriminde ise modelin çoğunlukla burun bölgesi olmak üzere az da olsa boyun ve alın bölgesine de odaklandığı görülmüştür. ResNet modeline ait GradCAM ve GradCAM++ gösterimlerinin her ikisinde de burun ve göz çevresine yoğunlaştığı görülmektedir.

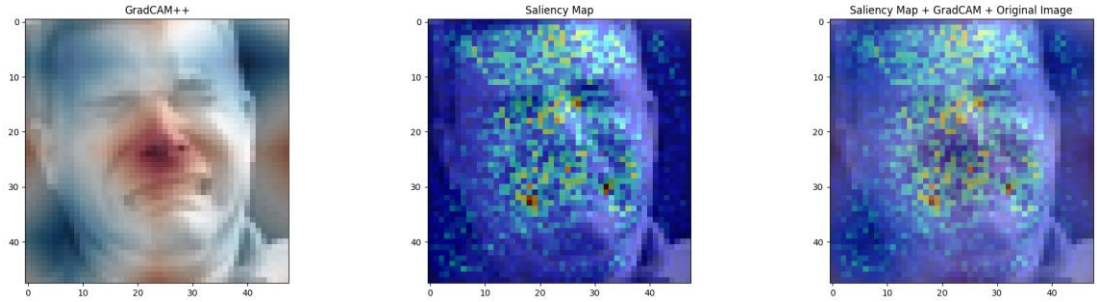


Şekil 4-43. VGG ile "Üzgün" ifadesine ait GradCAM ve GradCAM++ gösterimi

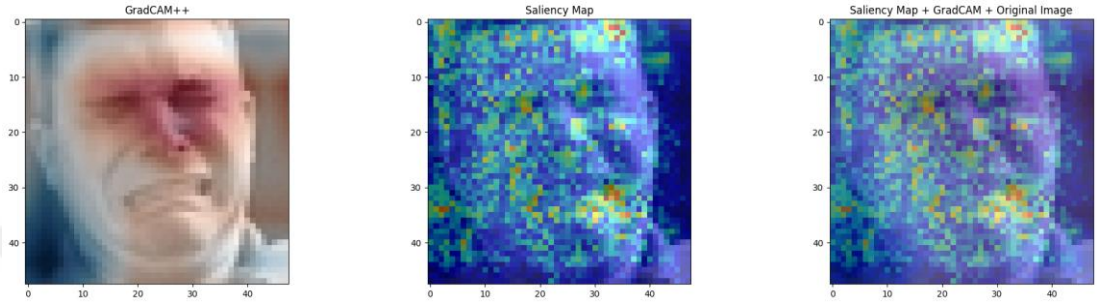


Şekil 4-44. ResNet ile "Üzgün" ifadesine ait GradCAM ve GradCAM++ gösterimi

VGG modeline ait GradCAM++ ile Saliency haritası karşılaştırıldığında ise modelin GradCAM++ dan farklı olarak göz ve alın bölgesine de odaklandığı görülmüştür. ResNet modeline ait Saliency haritası gösteriminde belirgin bir noktaya yoğunlaştığı görülmemekle birlikte ağız çevresine belki daha fazla yoğunlaştığını söylemek doğru olabilir.



Şekil 4-45. VGG ile "Üzgün" ifadesine ait GradCAM++ ve Saliency haritası gösterimi



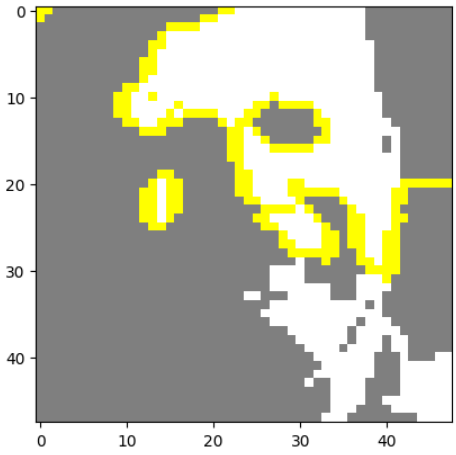
Şekil 4-46. ResNet ile "Üzgün" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

VGG modeline ait SHAP gösterimi incelendiğinde ise en fazla Sinirli ifadesiyle olmak üzere birçok ifade için pozitif etkisi olan özelliklerle ilişkilendirme yapıldığı görülmektedir, Pozitif veya Negatif etkisi olan özelliklerle ilişkilendirilmeyen tek ifadenin tiksinti olduğu görülmüştür. ResNet modeline ait SHAP gösterimi incelendiğinde burun ve göz çevresine pozitif noktaların dağıldığı söylenebilmektedir. Birinci sırada korku olarak tahminleme yaptığı görülmektedir.

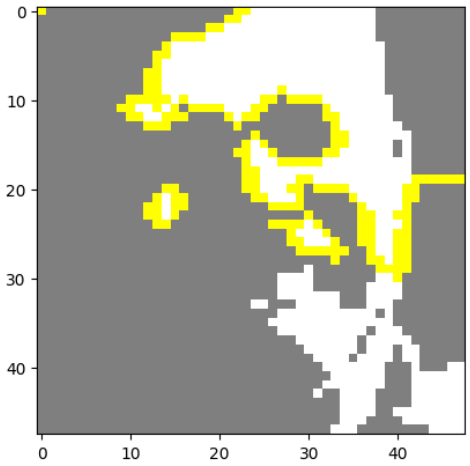
Şekil 4-47. VGG modeli ile "Üzgün" ifadesine ait SHAP gösterimi

Şekil 4-48. ResNet modeli ile "Üzgün" ifadesine ait SHAP gösterimi

VGG modeline ait LIME gösterimine göre ise modelin tahminde göz, alın, burun ve sol yanak bölgesine odaklandığı görülmüştür. ResNet modeline ait LIME gösteriminde modelin alın, göz çevresi, burun ve yanak bölgesine yoğunlaştığı görülmektedir.



Şekil 4-49. VGG modeli ile "Üzgün" ifadesine ait LIME gösterimi

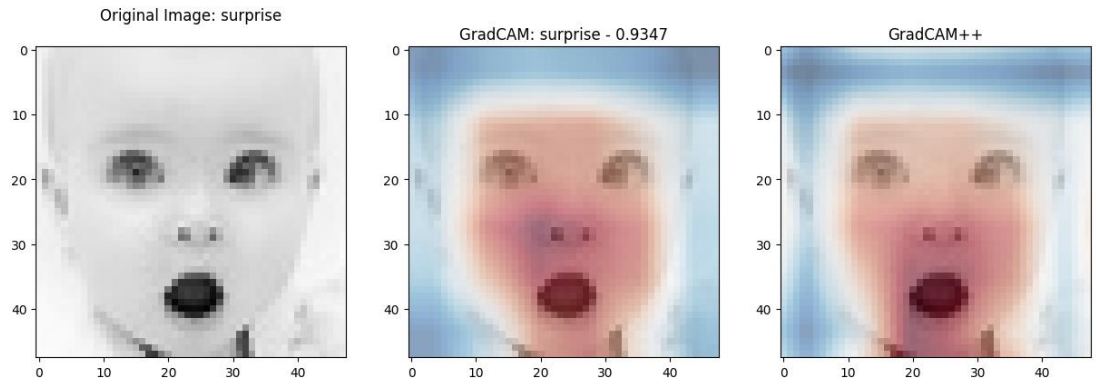


Şekil 4-50. ResNet modeli ile "Üzgün" ifadesine ait LIME gösterimi

4.4.7 Sürpriz İfadesi Açıklanabilirlik Analizi

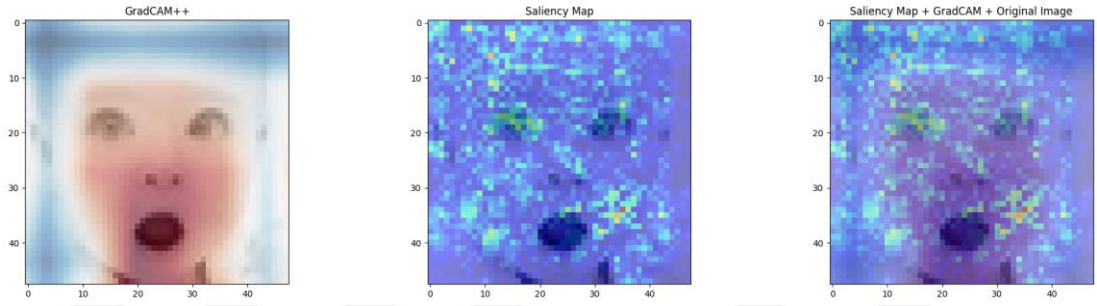
Sürpriz ifadesi için seçilen resim kullanılan üç farklı model tarafından da doğru tahmin edilmiş olup en yüksek tahminde bulunan model olan ResNet'in tahminine ait GradCAM ile GradCAM++ gösterimi Şekil 4-51'de, GradCAM++ ile Saliency haritası Şekil 4-52'de, SHAP Şekil 4-53'te ve LIME Şekil 4-54'te gösterilmiştir.

Sürpriz duygu durumu için modelin yaptığı tahminleri açıklamak için çizdirilen GradCAM ve GradCAM++ gösterimlerinde modelin yüzün geneline odaklandığı görülmüştür.



Şekil 4-51. ResNet i ile "Sürpriz" ifadesine ait GradCAM ve GradCAM++ gösterimi

Modelin sürpriz duygu durumu için Saliency haritası gösterimi incelendiğinde GradCAM++ ile benzer şekilde en az alın bölgesi olmak üzere yüzün geneline odaklandığı görülmüştür.

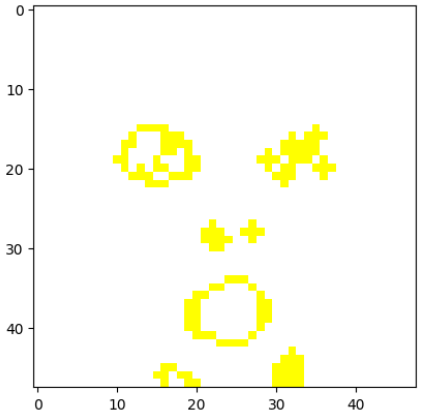


Şekil 4-52. ResNet ile "Sürpriz" ifadesine ait GradCAM++ ve Saliency haritası gösterimi

Sürpriz duygu durumu için modelin SHAP gösterimi incelendiğinde ağız bölgesindeki noktalara odaklanarak ilk olarak mutlu, daha sonra korku olarak tahminlemede bulunmuştur.

Şekil 4-53. ResNet modeli ile "Sürpriz" ifadesine ait SHAP gösterimi

Modelin LIME gösteriminde seçilen örnek sürpriz duygu durumu için yapılan tahminde gözler, burun ve ağız çevresine odaklandığı görülmektedir.



Şekil 4-54. ResNet modeli ile "Sürpriz" ifadesine ait LIME gösterimi

4.5 Modellerin GradCAM++ ile Genel Değerlendirmesi

Modellerimiz farklı duygu durumlarında ve resimlerdeki odak noktalarının tespiti ve karşılaştırılması için 7 duygu durumunu temsilen 49 resim seçilmiş ve modelin tahminde odaklandığı noktalar GradCAM++ ile görselleştirilmiştir. Gösterimin daha net anlaşılabilmesi için gösterimde resimler 9 eş parçaya bölünecek şekilde kılavuz çizgiler ile çizilmiştir.

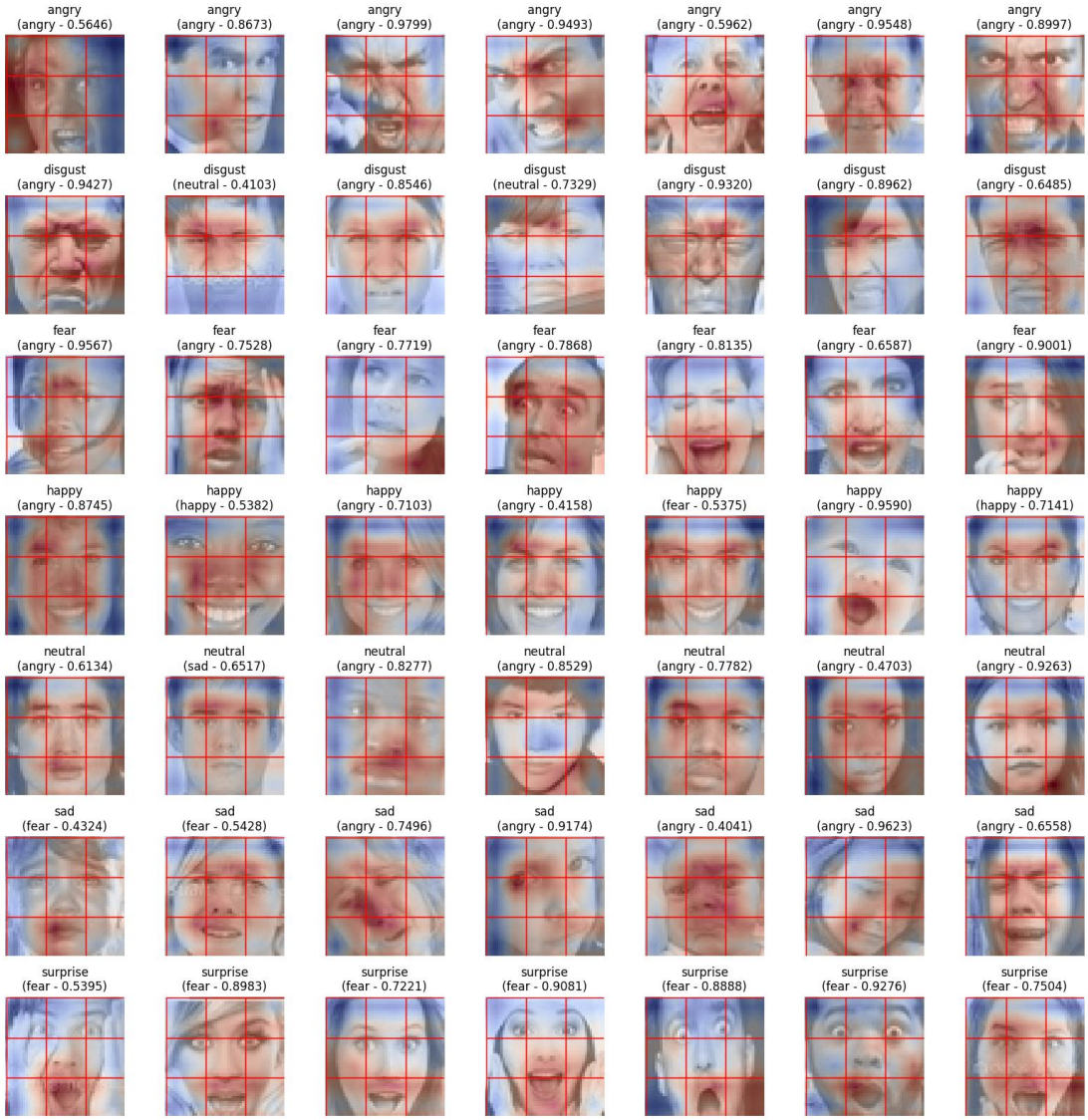
VGG modeli ile yapılan genel değerlendirme de Sinirli ifadesi için seçilen 7 resimden üçü doğru tahmin edilmiş olup bu resimlerde model resimlerin üst bölgesine yani alın bölgesine odaklanmıştır. Korku ifadesi olarak yapılan 3 yanlış tahminde ise alın bölgesine ek olarak resimlerin tam ortasına da odaklanıldığı görülmektedir. Mutlu ifadesi olarak yapılan yanlış tahminde ise resmin sağ alt köşesine odaklanma olmuştur. Tiksinti için seçilen resimlerde ise model resimlerin tamamında burun bölgesine odaklanmış olup hiç doğru tahminde bulunamamıştır. Korku ifadesi için seçilen resimlerde de model doğru tahminde bulunamamış olup alın bölgesine odaklanılan resimlerde sinirli, burun bölgesine odaklanılanlarda nötr, sürpriz ve mutlu olarak yanlış tahminde bulunmuştur. Mutlu ifadesi için seçilen resimlerde ise genelde ağız ve burun bölgesine odaklanmış olup 7 resmin dördünü doğru tahmin etmiştir. Nötr ifadesi için seçilen resimlerde de resmin merkezine odaklanmış olup 7 resmin dördünü doğru ikisini mutlu olarak tahmin etmiştir. Üzgün ve Sürpriz ifadeleri için seçilen resimlerde de modelimizin doğru tahminde bulunamadığı görülmüştür. Duygu durumları için GradCAM++ ile görselleştirilen tahmin matrisi incelendiğinde modelin alın bölgesine odaklandığı resimleri Sinirli, sadece merkeze odaklandığı resimleri nötr, merkez ve ağız bölgesine odaklandığı resimleri mutlu, merkez ve alın bölgesine odaklandığı

resimleri ise korku ifadesi olarak tanımladığı görülmektedir. VGG modeline ait GradCAM++ tahmin matrisi Şekil 4-55'dedir

Şekil 4-55. VGG modeli tahmin matrisi

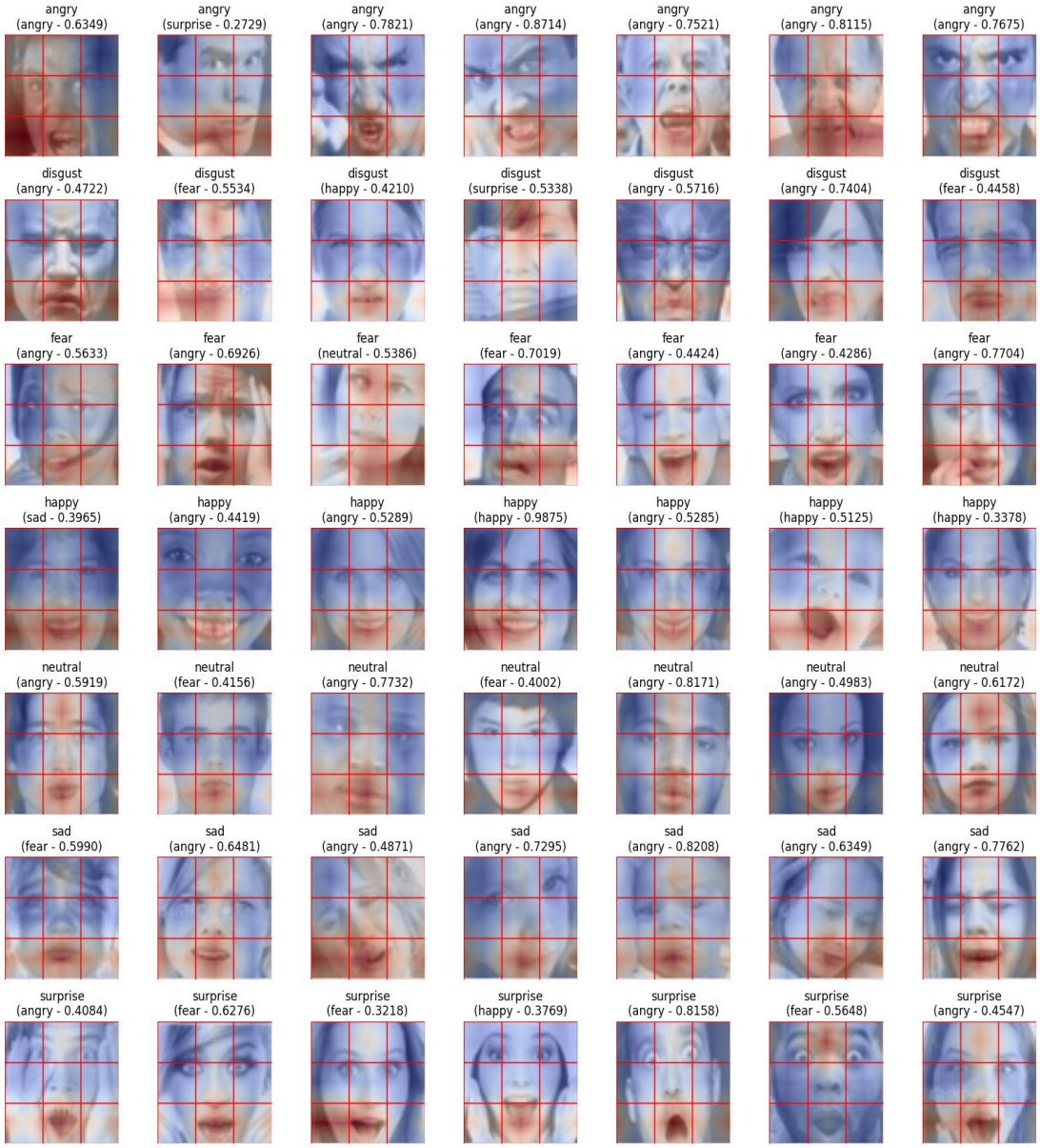
ResNet modeli ile yapılan değerlendirmede Sinirli ifadesi için seçilen 7 görüntünün tahmini doğru olarak yapılmış olup tahminlerde modelin genel olarak ağız burun ve kaşlara odaklandığı görülmektedir. Tiksinti için seçilen görüntülerde ise model resimlerin merkez ve alın bölgesine odaklandığında sinirli, yalnızca alın bölgesine odaklanılan resimlerde ise nötr olarak yanlış tahminde bulunmuştur. Korku ifadesi için seçilen resimlerde de model doğru tahminde bulunamamış olup alın resimlerin genel olarak ağız, kaş ve burun bölgesine odaklanan model sinirli olarak tahminde bulunmuştur. Mutlu ifadesi için seçilen resimlerde ise genelde ağız ve burun

bölgesine odaklanmış olup 7 resmin ikisini doğru tahmin etmiştir. Yanlış tahminlerde ise modelin Kaş ve gözlere de odaklandığı görülmektedir. Nötr ifadesi için seçilen resimlerde de resmin merkezine ve alın bölgesine odaklanmış olup 7 resmin altısını doğru birini üzgün olarak tahmin etmiştir. Üzgün ifadesi için seçilen resimlerde yüzün merkezi ile ağız ve alın bölgesine odaklanmış olup 7 üzgün ifadenin beşi sinirli ve ikisi korku olarak tahmin edilmiştir. Sürpriz için seçilen tüm ifadelerde modelin ağız bölgesine odaklandığı ve tamamını korku olarak yanlış tahmin ettiği görülmektedir. Duygu durumları için GradCAM++ ile görselleştirilen tahmin matrisi incelendiğinde modelin burun ve alın bölgesine odaklandığı resimleri Sinirli, ağız bölgesine odaklandığı resimleri Korku, merkeze odaklanmadığı resimleri ise nötr olarak tahmin ettiği görülmektedir. ResNet modeline ait GradCAM++ tahmin matrisi Şekil 4-56'dadır.



Şekil 4-56. ResNet modeli tahmin matrisi

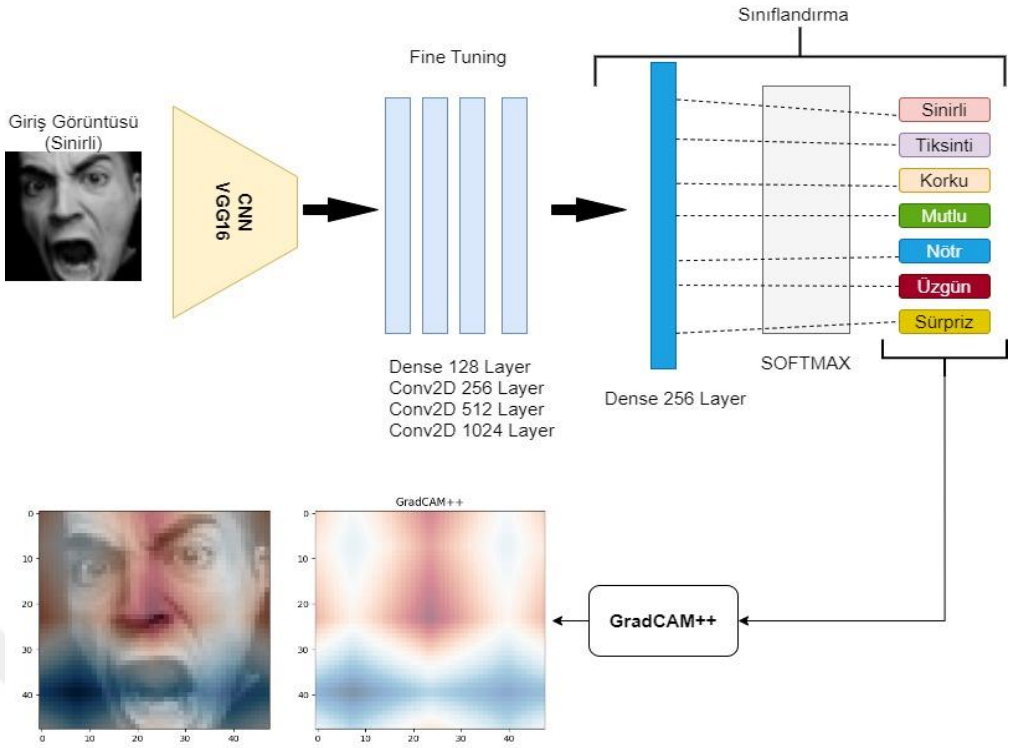
Inception modeli ile yapılan değerlendirmede Sinirli ifadesi için seçilen 7 görüntünün altısının tahmini doğru olarak yapılmış olup tahminlerde modelin genel olarak ağız bölgesine odaklandığı görülmektedir. Tiksinti için seçilen görüntülerde ise model resimlerin ağız bölgesine odaklandığında sinirli ve korku olarak yanlış tahminde bulunmuştur. Korku ifadesi için seçilen resimlerde de model sadece bir doğru tahminde bulunmuş olup ağız ve alın bölgesine odaklandığı resimleri için sinirli olarak tahminde bulunmuştur. Mutlu ifadesi için seçilen resimlerde ise genelde ağız bölgesine odaklanmış olup 7 resmin üçünü doğru ve üçünü de sinirli olarak yanlış tahmin etmiştir. Nötr, Üzgün ve Sürpriz ifadesi için seçilen resimlerde de resmin ağız ve alın bölgesine odaklanan modelimiz doğru tahminde bulunamamış resimleri korku veya sinirli olarak yanlış tahmin etmiştir. Sürpriz için seçilen tüm ifadelerde modelin ağız bölgesine odaklandığı ve tamamını korku olarak yanlış tahmin ettiği görülmektedir. Duygu durumları için GradCAM++ ile görselleştirilen tahmin matrisi incelendiğinde modelin ağız ve alın bölgesine odaklandığı resimleri genel olarak Sinirli ve Korku ifadesi olarak tahmin ettiği görülmektedir. Inception modeline ait GradCAM++ tahmin matrisi Şekil 4-57'dedir.



Şekil 4-57. Inception modeli tahmin matrisi

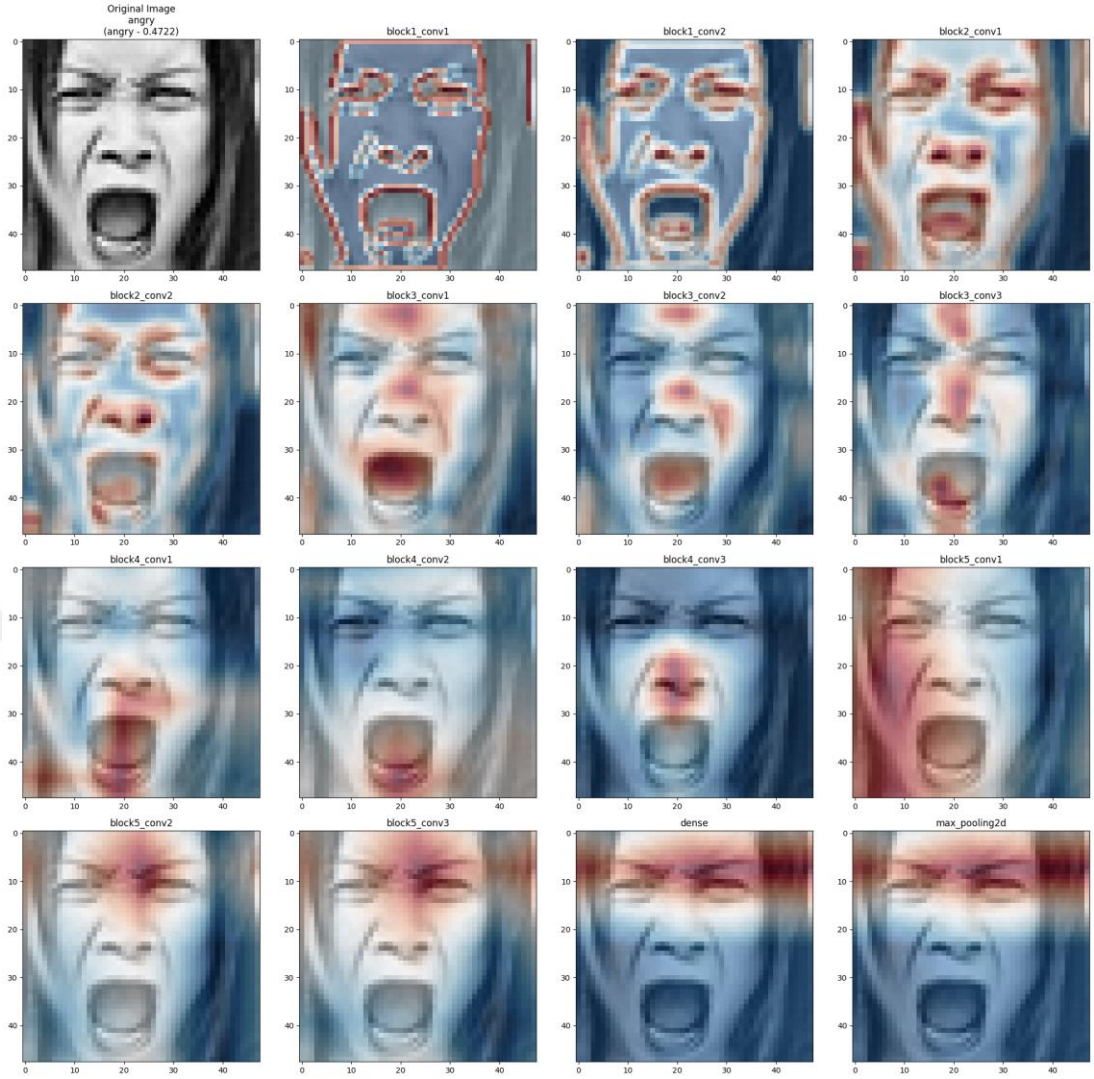
4.6 VGG Model Mimarisi

Deneylerde en başarılı olan VGG modeli ile yapılan tahminleri GradCAM++ ile açıklaması yapılmıştır. Deneylerde giriş görüntüsü geliştirilen VGG modelinden geçtikten sonra bir tahmin oluşmakta ve sonrasında tahmine ait GradCAM++ gösterimi ile modelin odaklandığı noktalar açıklanmaktadır. Model mimarisi Şekil 4-58'de gösterilmiştir.



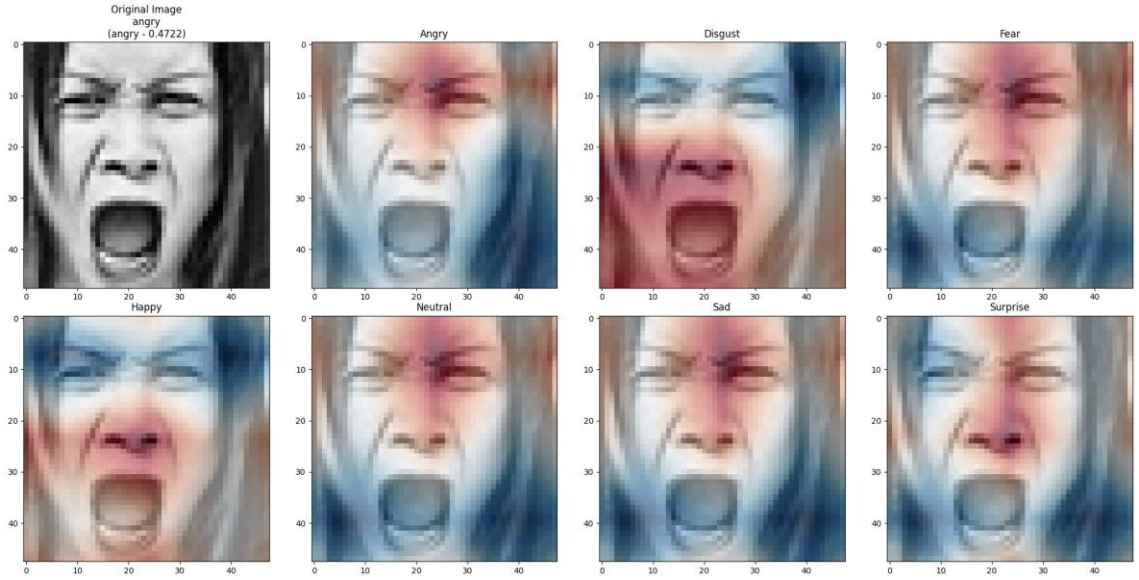
Şekil 4-58. Model Mimarisi

Modellerin tahmin süreci için modelimizin oluşturulan katmanlarda odaklandığı noktaları tespit etmek için GradCAM++ yöntemi kullanılarak modelin tahmin sürecinde odaklandığı noktalar tespit edilmiştir. Görüntünün genelinden başlayan odaklanmanın son katmanlara doğru alın bölgesinde yoğunlaştığı görülmüştür. Şekil 4-59'da VGG modelinin sinirli ifadesi için tahmin süreci GradCAM++ ile gösterilmiştir.



Şekil 4-59. VGG modeli tahmin sürecinin GradCAM++ ile gösterimi

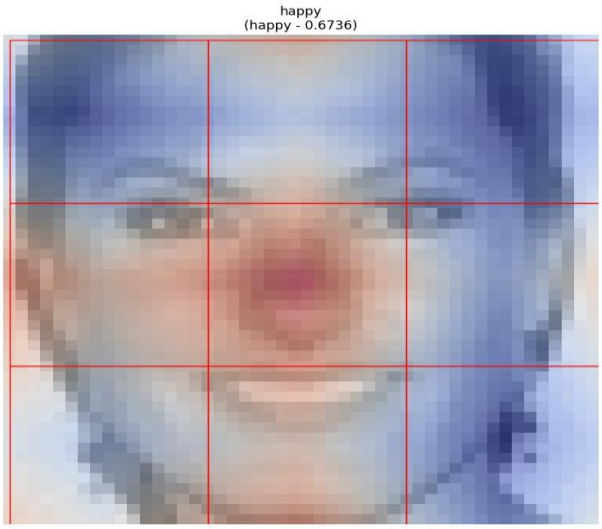
Her görüntü tahmin aşamasında yedi farklı duygu durumunun özelliklerini taşıma oranına göre değerlendirilerek karar verir. Açıklanabilirlik teknikleri modellerin duygu durumları ile görüntü üzerinde ilişkilendirilen noktaları tespit etmemizi sağlar. Şekil 4-60'da görüldüğü gibi modelin test edilen görüntü için farklı duygularda hangi noktalara bakıldığı bilgisi tahminin açıklanmasını kolaylaştırmıştır.



Şekil 4-60. VGG modeline ait tahminin Duygulara göre odak noktaları

4.7 Duygulara Göre Odak Noktalarının İstatistiksel Analizi

Bu bölümde yüz ifadelerinin tespiti yapılırken odaklanılan noktaların istatistiksel analizi yapılmıştır. İlk olarak duygulara göre yapılacak istatistiksel analizin verisetini ve oluşturulan modeli daha iyi ifade etmesi için verisetinden her ifade için 100 adet olmak üzere toplam 700 adet resimden oluşan yeni bir test veriseti seçilmiştir. Her resim Şekil 4-61'de görüldüğü gibi 9 eş parça olacak şekilde bölümlere ayrılmış ve sol üstten başlanarak sağa doğru bölümler numaralandırılmıştır. GradCAM++ ile ısı haritası oluşturulmuştur. Odaklanılan noktalar bölümlere göre gruplandıktan sonra en çok ağırlığa sahip beş bölge istatistiksel analiz için kaydedilmiş ve analizler bu veriler üzerinden yapılmıştır.



Şekil 4-61 Resimlere Ait İstatistiksel Analiz Bölümleri

Duygu bazlı olarak yapılan incelemede 100 resim üzerinde en çok odaklanılan bölgelerin 1. ve 7. bölgeler olduğu tespit edilmiştir. Resimler arasında 2. olarak odaklanılan bölgelerin 2,4 ve 8. bölgeler olduğu görülmüştür. Tablo 4-2'de Odak sıralarına göre en çok odaklanılan bölgeler verilmiştir. Tabloda görüldüğü gibi en çok odaklanılan bölge sayısı arttıkça modelin diğer odaklandığı bölgelerin sayısı da artmakta ve modelin tahmini üzerinde etkisi olan diğer bölgeler de görülmektedir.

Tablo 4-2 İlk 3 sırada yer alan odaklanılan bölgeler

Orijinal Duygu	**Odak1 Değerleri**			**Odak2 Değerleri**			**Odak3 Değerleri**						
	1	**3**	**7**	**2**	**4**	**8**	**1**	**2**	**3**	**4**	**7**	**8**	**9**
Sinirli	61	0	39	38	41	21	12	17	15	33	6	6	11
Tiksinti	75	0	25	48	37	15	8	19	18	34	8	2	11
Korku	77	1	22	54	33	13	6	18	18	42	6	4	6
Mutlu	85	1	14	60	36	4	8	16	34	26	10	3	3
Nötr	86	0	14	63	29	8	1	19	25	41	4	5	5
Üzgün	85	0	15	47	45	8	4	27	22	27	11	3	6
Sürpriz	86	0	14	58	35	7	4	19	22	40	9	3	3

Her resim için en çok odaklanılan 5 bölge alındığında ise her duygu durumu için 500 adet veriye ulaşılmıştır. Bu veriler Tablo 4-3 'te gösterilmiştir. En çok odaklanılan Odak 1 noktasında 1. ve 7. Bölgeler ağırlıktayken odaklanılan diğer noktaların da eklenmesiyle diğer bölgelerin de tahmin üzerindeki etkisi görülmüştür.

Tablo 4-3 En çok odaklanılan ilk 5 bölgeye ait toplam veriler

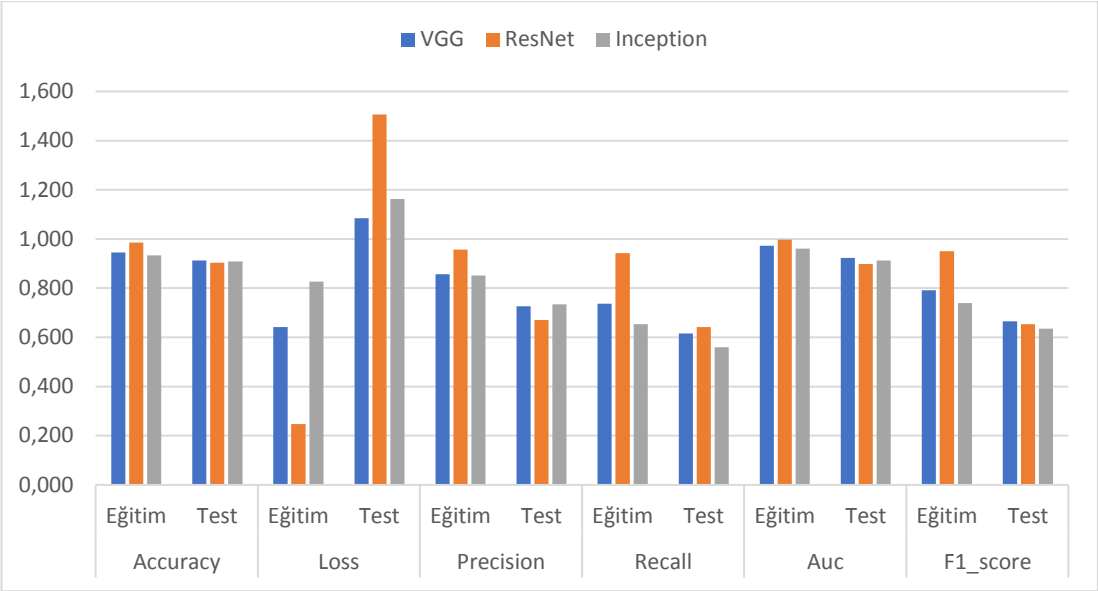
Duygu/Bölge	1	2	3	4	5	6	7	8	9
Sinirli	79	61	38	100	100	0	62	39	21
Tiksinti	85	74	48	100	100	0	52	26	15
Korku	87	78	55	99	99	1	46	22	13
Mutlu	96	86	60	99	100	1	40	14	4
Nötr	92	86	63	100	100	0	37	14	8
Üzgün	92	85	47	99	100	1	53	15	8
Sürpriz	93	86	58	100	100	0	42	14	7

Duygulara göre odak noktalarının analizi için Tablo 4-3'teki veriler incelenerek her ifade için en çok odaklanılan bölgeler hakkında analiz yapılmıştır. Sinirli ifadesi için odaklanılan noktalar incelendiğinde 4. ve 5. bölgelerin %20 ile en çok odaklanılan noktalar olduğu görülmüştür. Bu bölgeleri yaklaşık %16 ile 1. bölge ve yaklaşık %12 ile 2. ve 7. bölgeler takip etmiştir. Tiksinti ifadesi için odaklanılan noktalar incelendiğinde 4. ve 5. bölgelerin %20 ile en çok odaklanılan noktalar olduğu onları %17 ile 1. bölge ve %15 ile 2. bölge takip etmiştir. Korku ifadesi için odaklanılan noktalar incelendiğinde 4. ve 5. bölgelerin %19 ile en çok odaklanılan noktalar olduğu görülmüş, onları %18 ile 1. bölge ve %16 ile 2. bölge izlemiştir. Mutlu ifadesi için odaklanılan noktalar incelendiğinde 5. bölgenin %20 ile en çok odaklanılan nokta olmuş, onu %19 ile 1. ve 4. bölgeler ve %17 ile 2. bölge takip etmiştir. Tiksinti ifadesi için odaklanılan noktalar incelendiğinde 4. ve 5. bölgelerin %20 ile en çok odaklanılan noktalar olduğu, onları %17 ile 1. bölge ve %15 ile 2. bölgenin takip ettiği görülmüştür. Nötr ifadesi için odaklanılan noktalar incelendiğinde 4. ve 5. bölgelerin %20 ile en çok odaklanılan noktalar olduğu, onları %18 ile 1. bölge ve %17 ile 2. bölgenin izlediği görülmüştür. Üzgün ifadesi için odaklanılan noktalar incelendiğinde 4. ve 5. bölgelerin %20 ile en çok odaklanılan noktalar olmuş, onları %18 ile 1. bölge ve %17 ile 2. bölge izlemiştir. Sürpriz ifadesi için odaklanılan noktalar incelendiğinde 4. ve 5. bölgelerin %20 ile en çok odaklanılan noktalar olduğu, onları %19 ile 1. bölge ve %17 ile 2. bölgenin takip ettiği görülmüştür. Bu veriler modelimizin seçilen resimlerde en çok odaklandığı bölgelerin genel olarak resimlerin burun bölgesine denk gelen 5. bölgesi ile sağ göze denk gelen 4. bölgesi olduğunu göstermiştir.

5. TARTIŞMA

Açıklanabilir Yüz İfadesi Tanıma üzerine yapılan bu çalışmada üç farklı model üzerinde açıklanabilirlik teknikleri ile modellerin tahmin süreci ve odaklanılan noktalar incelenmiştir. Çalışmanın ilk aşamasında önceden eğitilmiş ve başarılar elde etmiş VGG, ResNet ve Inception modelleri üzerinde daha rahat karşılaştırma yapabilmek için iyileştirmeler yapılmıştır. Bu modellere aynı katmanlar eklenmiş, bu modellerde parametrelere ve hiperparametrelere aynı değerler verilmiştir. Modeller değerlendirilirken doğruluk, kayıp, hassasiyet, duyarlılık, AUC ve F1_score ölçümlerine göre karar verilmiştir.

Oluşturulan üç model arasında eğitim için en iyi doğruluk oranını 0.986 ile ResNet modeline ait iken test için en iyi doğruluk oranı VGG modeline aittir. Bu sonuçlar ResNet modelinin eğitim aşamasında daha başarılı olmasına rağmen bu sonuçları test aşamasına tam olarak yansıtamadığını göstermektedir. Kayıp değerlerine baktığımızda da eğitimde en iyi sonuç yine ResNet modeline ait iken test aşamasında ise en kötü sonuç ResNet modeline aittir. Kayıp değerlerinde test aşamasında en iyi sonuç VGG modeline aittir. Hassasiyette ise eğitim aşamasında en iyi sonuç ResNet modelinde iken test aşamasında Inception modeli en iyi hassasiyeti elde etmiştir. Duyarlılıkta ise eğitim ve test aşamasında en iyi sonuç ResNet modeline aittir. AUC değerleri incelendiğinde eğitim aşamasında ResNet, test aşamasında VGG modeli en iyi sonuçları elde etmiştir. F1_score değerlerine bakıldığında eğitim aşamasında ResNet, test aşamasında ise VGG en iyi sonuçları elde etmiştir. Genel olarak bakıldığında ResNet modelinin eğitim aşamasında elde ettiği başarıyı test aşamasına taşıyamadığı görülmektedir. VGG ve Inception modelleri ise eğitim ve test verileri arasındaki farklar bakıldığında daha tutarlı sonuçlar elde etmişlerdir. Test verilerine bakıldığında en iyi modelin VGG modeli olduğu görülmüştür. Modellerin karşılaştırması Şekil 5-1'de gösterilmiştir.



Şekil 5-1 Modellerin karşılaştırılma grafiği

Oluşturulan modellerin yedi duygu durumunu temsilen seçilen yedi görüntüyü tahmin etmeleri sağlanarak modellerin bu tahminler üzerinden de karşılaştırılması yapılmıştır. Sinirli ifadesi için seçilen görüntüyü üç modelimiz de doğru tahmin etmiş olup en yüksek doğruluk ResNet modeline aittir. Tiksinti ifadesi için seçilen görüntüyü ise VGG sinirli, ResNet Mutlu ve Inception Üzgün olarak yanlış tahmin etmiştir. Korku ifadesini ise VGG ve ResNet doğru olarak tahmin etmiş olup en yüksek değer VGG modeline aittir. Mutlu ifadesini ise VGG ve Inception doğru tahmin etmiş olup en iyi sonuç VGG modeline aittir. Nötr ve Üzgün ifadelerini ise sadece VGG modeli doğru tahmin etmiştir. Sürpriz ifadesini de üç model doğru olarak tahmin etmiş olup en iyi sonuç ResNet modeline aittir. Modellere ait tahmin sonuçları Tablo 5-2'de verilmiştir. 7 ifadenin altısını doğru tahmin eden VGG modeli burada da en başarılı model olmuştur.

Tablo 5-1 Test edilen ifadelere ait tahmin tablosu

Model İfade	Sinirli	Tiksinti	Korku	Mutlu	Nötr	Üzgün	Sürpriz
VGG	0.6283 Doğru	0.5144 Sinirli	0.6843 Doğru	0.8656 Doğru	0.6915 Doğru	0.4307 Doğru	0.7695 Doğru
ResNet	0.6931 Doğru	0.4247 Mutlu	0.5543 Doğru	0.5052 Sinirli	0.6680 Korku	0.9846 Sinirli	0.9347 Doğru
Inception	0.5387 Doğru	0.4681 Üzgün	0.6094 Sinirli	0.4278 Doğru	0.5432 Korku	0.7364 Sinirli	0.4671 Doğru

Açıklanabilir yüz ifadesi tanımanın temelinde modelin doğru veya yanlış tahminde bulunmasından ziyade model bu kararı nasıl ve neye dayanarak verdi düşüncesi vardır. Sonuç olarak yapay zekâ modelleri öğrenme yaparken onlara verdiğimiz verilere muhtaçlar. Verilerin çeşitliliği ve doğru etiketlenmiş olması modelin performansında ana belirleyici unsurlardır. Modellerin dayanak noktaları bilinirse hatanın nedenini bulmak ve düzeltmek, bunlarla beraber modeli daha ileriye taşımak da kolaylaşacaktır. Modellerimizin karar verme süreçlerini anlamak için XAI teknikleri olan GradCAM, GradCAM++, Saliency Haritası, SHAP ve LIME teknikleri kullanılarak modellerin her duygu durumu için yaptıkları tahminler açıklanmaya çalışılmıştır.

Açıklanabilirlik teknikleri karşılaştırılınca GradCAM++ gösteriminin GradCAM gösterimine göre daha açıklayıcı ve kapsayıcı olduğu, modelin odak noktalarını daha iyi gösterdiği belirlenmiştir. Yapılan diğer bir karşılaştırmada ise GradCAM++ ile yine benzer bir teknik olan Saliency haritası karşılaştırılmıştır. Genelde benzer noktaları gösteren bu tekniklerden Saliency haritası tekniğinin bazı modellerde ve duygu durumlarında daha küçük etki noktalarını da gösterdiği için daha iyi olduğu söylenebilir. SHAP tekniği ise bize her duygu durumu için pozitif ve negatif etkisi olan noktaları vermesi bakımından bir adım öne çıkmakla beraber bazı duygu durumlarında bu noktaların tam olarak oluşmaması Açıklanabilirlik açısından kabul edilebilir bir şey değildir. LIME tekniği de bize açıklanamayan, nötr ve odaklanılan noktaları göstermekte olup bu teknikte bazı duygu durumlarında bize tam olarak doğru

bilgileri verememiştir. Farklı her model ve tahmin için kesin doğru sonucu verecek bir açıklama tekniği bulmak imkansızdır. Açıklanabilir yüz ifadesi tanıma çalışmamızda farklı modeller ve farklı ifade durumlarında yaptığımız deneyler bizlere modellerin ve tekniklerin farklı durumlarda farklı sonuçlar verebileceğini göstermiştir. Yapılan çalışmalarda en iyi model konusunda VGG ön plana çıkarken, En açıklayıcı teknik konusunda ise GradCAM++, Saliency haritası ve SHAP teknikleri ön plana çıkmıştır.

6. SONUÇ

Açıklanabilir yüz ifadesi tanıma konusunda yapılan bu çalışmada öncelikle daha önceden eğitilmiş modeller arasından üç model seçilmiştir. Seçilen modellerin eğitilmesi için günümüzde popüler olan ve yedi duygu durumu için yeterince etiketlenmiş görüntü barındıran Fer2013 veriseti seçilmiştir. Verisetimizde bulunan veriler üzerinde veri ön işleme ve veri arttırma teknikleri uygulandıktan sonra modellerimiz üzerinde iyileştirmeler yapılmıştır. Seçilen modellerimiz olan VGG, ResNet ve Inception modelleri aynı parametreler ve hiperparametre değerleri ile eğitilmiştir. Eğitim verilerine göre en başarılı model ResNet olurken test verilerine göre en başarılı model VGG olmuştur.

Modellerin verdiği kararlar kadar bu kararları alırken dayanak noktaları modellerin güvenilirliği için çok önemlidir. Modellerin karar verme süreci ile odaklandığı noktaların bilinmesi hem modelin güvenilirliğini artıracak hem de olası hataların tespit edilmesini sağlayarak modelin geliştirilmesine katkı sağlayacaktır. Bu aşamada yedi duygu durumu için seçilen birer resim üç model tarafından tahmin edilmiştir. Tahminlerde en başarılı model yedi resmin altısını doğru tahmin eden VGG modeli olmuştur. Daha sonra doğru ve yanlış fark etmeksizin yapılan tüm tahminlerin XAI teknikleri olan GradCAM, GradCAM++, Saliency haritası, SHAP ve LIME teknikleri ile görselleştirilmeleri sağlanarak odaklanılan noktalar tespit edilmiştir. Odak noktaları incelenerek modellerin neden doğru veya neden yanlış kararlar verdiği açıklanmıştır.

Kullanılan farklı XAI teknikleri ile modelin karar verme süreci ve odaklanılan noktalar konusunda açıklanabilirlik için gerekli bilgiler alınmıştır. Her teknikte farklı model ve ifadelerde benzer sonuçlar elde edilememiştir. Alınan bilgilerin bu şekilde farklılıklar göstermesi her model ve ifade için mükemmel bir tekniğin olmadığını göstermiştir. Oluşturulan modeller ile yapılan testlerde alınan kararları açıklayabilmek için kullanılan XAI teknikleri arasında benzer şekilde çalışan GradCAM, GradCAM++ ve Saliency haritası birbirleriyle karşılaştırıldığında GradCAM++'ın GradCAM e göre daha açıklayıcı bilgiler verdiği görülmüştür. Saliency haritasının ise tahminde etkisi olan pozitif, negatif ve Nötr noktalar konusunda daha detaylı bilgiler verdiği tespit edilmiştir. SHAP tekniği ile tüm tahminler için pozitif ve negatif etkisi yüksek olan noktalar görüntülenmiştir. LIME tekniği ise diğer teknikler kadar olmasa

da açıklayıcı bilgiler vermiştir. Tahminler konusunda en açıklayıcı bilgiler ise GradCAM++ ve Saliency haritası teknikleriyle elde edilmiştir.

Açıklanabilir yüz ifadesi tanıma çalışması bu alanda oluşturulan yapay zekâ modellerinin tahmin süreci ve odaklandığı noktaları göstererek kararların doğruysa neden doğru ve yanlışsa neden yanlış olduğunun açıklanmasına yardımcı olmuştur. Ayrıca bu çalışma yapay zekâ modellerinin yaptığı tahminlerin güvenilirliğini artırma ve modellerin geliştirilmesi konusunda XAI tekniklerin gerekliliğini göstermiştir. Günümüzde her geçen gün artan insan bilgisayar etkileşimi, otonom sistemlerin insanların duygularını tanımlamada otomatikleşmeyle beraber daha doğru tahminler yapmasını da zorunlu kılmıştır. Bu gereklilikten dolayı ifade tanıma konusunda gelecekte daha çok çalışma yapılacaktır. Farklı kültürler, yaşlar, cinsiyet ve karmaşık duygu durumları ifade tanımada karşılaşılan temel sorunlar olsa da modellerdeki eksikliklerin tespit edilmesinde ve geliştirilmesinde açıklanabilir yüz ifadesi tanıma alanında çalışmalar gelecekte de devam edecektir.

KAYNAKÇA

[1] Dixit, A. N., & Kasbe, T. (2020, February). A survey on facial expression recognition using machine learning techniques. In *2nd international conference on data, engineering and applications (IDEA)* (pp. 1-6). IEEE.

[2] Vo, Q. N., Tran, K., & Zhao, G. (2019, September). 3D facial expression recognition based on multi-view and prior knowledge fusion. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)* (pp. 1-6). IEEE.

[3] Raghuvanshi, A., & Choksi, V. (2016). Facial expression recognition with convolutional neural networks. *CS231n Course Projects*, *362*.

[4] Rathod, M., Dalvi, C., Kaur, K., Patil, S., Gite, S., Kamat, P., ... & Gabralla, L. A. (2022). Kids' Emotion Recognition Using Various Deep-Learning Models with Explainable AI. *Sensors*, *22*(20), 8066.

[5] Sikkandar, H., & Thiyagarajan, R. (2021). Deep learning based facial expression recognition using improved Cat Swarm Optimization. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 3037-3053.

[6] Rajan, S., Chenniappan, P., Devaraj, S., & Madian, N. (2019). Facial expression recognition techniques: a comprehensive survey. *IET Image Processing*, *13*(7), 1031-1040.

[7] Deramgozin, M., Jovanovic, S., Rabah, H., & Ramzan, N. (2021, August). A Hybrid Explainable AI Framework Applied to Global and Local Facial Expression Recognition. In *2021 IEEE International Conference on Imaging Systems and Techniques (IST)* (pp. 1-5). IEEE.

[8] del Castillo Torres, G., Roig-Maimó, M. F., Mascaró-Oliver, M., Amengual-Alcover, E., & Mas-Sansó, R. (2022). Understanding How CNNs Recognize Facial Expressions: A Case Study with LIME and CEM. *Sensors*, *23*(1), 131.

[9] Hailemariam, Y., Yazdinejad, A., Parizi, R. M., Srivastava, G., & Dehghantanha, A. (2020, December). An empirical evaluation of AI deep explainable tools. In *2020 IEEE Globecom Workshops (GC Wkshps* (pp. 1-6). IEEE.

[10] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022, April). Explainable AI methods-a brief overview. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML*

2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers (pp. 13-38). Cham: Springer International Publishing.

[11] Oliveira, I., Silva, J. L., Quispe, F. P., & Alvarez, A. B. (2021, October). EmotiTEA: A visual monitoring module based on the recognition of facial emotions with CNN. In *2021 IEEE Engineering International Research Conference (EIRCON)* (pp. 1-4). IEEE.

[12] Uçar, A. (2017, July). Deep Convolutional Neural Networks for facial expression recognition. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 371-375). IEEE.

[13] Yao, L., He, S., Su, K., & Shao, Q. (2022). Facial expression recognition based on spatial and channel attention mechanisms. *Wireless Personal Communications*, *125*(2), 1483-1500.

[14] Jiao, Y., Niu, Y., Zhang, Y., Li, F., Zou, C., & Shi, G. (2019, December). Facial attention based convolutional neural network for 2D+ 3D facial expression recognition. In *2019 IEEE Visual Communications and Image Processing (VCIP)* (pp. 1-4). IEEE.

[15] Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z. (2020). Attention mechanism-based CNN for facial expression recognition. *Neurocomputing*, *411*, 340-350.

[16] Ramalingam, S., & Garzia, F. (2018, October). Facial expression recognition using transfer learning. In *2018 International Carnahan Conference on Security Technology (ICCST)* (pp. 1-5). IEEE.

[17] Shingjergji, K., Iren, D., Böttger, F., Urlings, C., & Klemke, R. (2022, October). Interpretable Explainability in Facial Emotion Recognition and Gamification for Data Collection. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1-8). IEEE.

[18] Shaees, S., Naeem, H., Arslan, M., Naeem, M. R., Ali, S. H., & Aldabbas, H. (2020, September). Facial emotion recognition using transfer learning. In *2020 International Conference on Computing and Information Technology (ICCIT-1441)* (pp. 1-5). IEEE.

[19] Ahuja, B., & Vishwakarma, V. P. (2022, December). Convolutional Neural Network and Kernel Extreme Learning Machine for Face Recognition. In *2022 8th International Conference on Signal Processing and Communication (ICSC)* (pp. 329-333). IEEE.

[20] Wu, C., Chai, L., Yang, J., & Sheng, Y. (2019, July). Facial expression recognition using convolutional neural network on graphs. In *2019 Chinese control conference (CCC)* (pp. 7572-7576). IEEE.

[21] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. *Electronics*, *10*(9), 1036.

[22] Sahoo, G. K., Das, S. K., & Singh, P. (2022, May). Deep Learning-Based Facial Emotion Recognition for Driver Healthcare. In *2022 National Conference on Communications (NCC)* (pp. 154-159). IEEE.

[23] Tegani, S., & Abdelmoutia, T. (2021, December). Using covid-19 masks dataset to implement deep convolutional neural networks for facial emotion recognition. In *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)* (pp. 1-5). IEEE.

[24] Araf, T. A., Siddika, A., Karimi, S., & Alam, M. G. R. (2022, April). Real-Time Face Emotion Recognition and Visualization using Grad-CAM. In *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-5). IEEE.

[25] Mery, D. (2022). True black-box explanation in facial analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1596-1605).

[26] Hailemariam, Y., Yazdinejad, A., Parizi, R. M., Srivastava, G., & Dehghantanha, A. (2020, December). An empirical evaluation of AI deep explainable tools. In *2020 IEEE Globecom Workshops (GC Wkshps* (pp. 1-6). IEEE.

[27] Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, *5*(23), 495.

[28] Sapijaszko, G., & Mikhael, W. B. (2018, August). An overview of recent convolutional neural network algorithms for image recognition. In *2018 IEEE 61st International midwest symposium on circuits and systems (MWSCAS)* (pp. 743-746). IEEE.

[29] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 818-833). Springer International Publishing.

[30] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[31] Saha, S., Khabir, K. M., Abir, S. S., & Islam, A. (2019, May). A newly proposed object detection method using faster R-CNN inception with ResNet based on Tensorflow. In *Real-Time Image Processing and Deep Learning 2019* (Vol. 10996, pp. 246-256). SPIE.

[32] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

[33] Shah, S. S., & Sheppard, J. W. (2020, July). Evaluating explanations of convolutional neural network image classifications. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[34] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[35] Shapley, L. S. (1953). 17. A value for n-person games. *Contributions to the theory of games (AM-28), Volume II*, 307-318.

[36] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[37] Chen, L., Chen, J., Hajimirsadeghi, H., & Mori, G. (2020). Adapting grad-cam for embedding networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2794-2803).

[38] Alhawiti, K. M. (2015). Advances in artificial intelligence using speech recognition. *Int. J. Comput. Inf. Eng*, *9*, 1432-1435.

[39] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[40] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations

for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 839-847). IEEE.

[41] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 839-847). IEEE.

[42] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[43] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016, March). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.