

T.C.
BOLU ABANT İZZET BAYSAL ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM
DALI



MAKİNE ÖĞRENMESİ KULLANILARAK HİDROLOJİK
MODELLEME: FIRTINA DERESİ ÖRNEĞİ

YÜKSEK LİSANS TEZİ

ENSAR EFENDİ CELEPCİ

TEZ DANIŞMANI
DOÇ.DR. ALPER BAYRAK
İKİNCİ TEZ DANIŞMANI
DR.ÖĞR.ÜYESİ ŞAFAK KAYIKÇI

BOLU, AĞUSTOS - 2021

KABUL VE ONAY SAYFASI

ENSAR EFENDİ CELEPCİ tarafından hazırlanan “**MAKİNE ÖĞRENMESİ YÖNTEMİ KULLANILARAK HİDROLOJİK MODELLEME YAPILMASI**” adlı tez çalışması jürimiz tarafından Elektrik Elektronik Mühendisliği Anabilim Dalı’nda Yüksek Lisans Tezi olarak oy birliğiyle kabul edilmiştir. 9/08/2021

Jüri Üyeleri

İmza

Danışman
Doç. Dr. Alper BAYRAK
Bolu Abant İzzet Baysal Üniversitesi

Üye
Prof. Dr. Nusret KARAKAYA
Bolu Abant İzzet Baysal Üniversitesi

Üye
Doç. Dr. Abdullah Talha KABAKUŞ
Düzce Üniversitesi

Lisansüstü Eğitim Enstitüsü Onayı

Prof. Dr. Osman GÖRÜR
Lisansüstü Eğitim Enstitüsü Müdürü

ETİK BEYAN

Bolu Abant İzzet Baysal Üniversitesi, Lisansüstü Eğitim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu bildirir,

aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Teze ilişkin 01/09/2021 tarihinde Turnitin adlı intihal tespit programından enstitü müdürlüğünce belirlenen filtrelemeler uygulanarak alınmış olan benzerlik raporuna göre, tezin benzerlik oranı % 7 olarak tespit edilmiştir.

.....
ENSAR EFENDİ CELEPCİ

ÖZET

**“MAKİNE ÖĞRENMESİ KULLANILARAK HİDROLOJİK
MODELLEME: FIRTINA DERESİ ÖRNEĞİ”
YÜKSEK LİSANS TEZİ
ENSAR EFENDİ CELEPCİ
BOLU ABANT İZZET BAYSAL ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI**

**(TEZ DANIŞMANI:DOÇ.DR. ALPER BAYRAK)
(İKİNCİ DANIŞMAN:DR.ÖĞR.ÜYESİ ŞAFAK KAYIKÇI)
BOLU, AĞUSTOS - 2021
XII + 53**

Su yeryüzündeki yaşamın temelini oluşturan maddedir. Suyun var olmaması dünya üzerindeki canlıların çok büyük bölümünün yaşamasını imkansız hale getirmektedir. Su canlılar ve insan için bu kadar önemliyken suya kolay ulaşım imkanı da aynı derecede önemli olmaktadır. Ancak suya yakın olmak bazı sorunları da beraberinde getirmiştir. Nehir kenarlarına kurulan şehirler, taşkınlar, su yolunun değişimi, gel-git gibi sorunlarla baş etmek zorunda kalmıştır. Bu sorunların sonucu olarak yerleşim alanları ve tarım arazileri yok olan insanlar yerleşim yerlerini terk etmek zorunda kalmıştır. Ek olarak maddi ve manevi zararları da beraberinde getirmiştir. Bölge analizlerinin yapılması ile elde edilecek doğru debi tahminleri sayesinde bu zararların önüne geçilebilir.

Bu çalışmada makine öğrenimi ve derin öğrenme modelleri vasıtasıyla debi tahminleri gerçekleştirilmiştir. Makine Öğrenimi programlama dilleri vasıtasıyla geliştirilen ve bilgisayar sistemlerinin öğrenme algoritmalarını kullanarak tahminler üretebilen yöntemdir. Bu çalışmada XGBoost makine öğrenimi, LSTM ve GRU derin öğrenme eğitim algoritmaları kullanılmıştır ve bu algoritmaların karşılaştırılması yapılmıştır.

Çalışma sahası olarak Fırtına Deresi havzası seçilmiştir. Çalışma sahasının bağımsız değişken veri setleri için CFSR (iklim tahmin sistemi analizi) sisteminden alınan günlük maksimum sıcaklık, minimum sıcaklık, bağıl nem, solar radyasyon ve rüzgar değişkenleri kullanılmıştır. Akım gözlem verileri ise Elektrik İşleri Enstitüsü (EİE) 2232 nolu akım gözlem istasyonu Topluca Köyü mevkiinden (AGİ) kayıtlarından elde edilmiştir. Veriler 1 Ocak 1979 ve 31 Aralık 2002 yılları arasında toplam 8758 günün verisini kapsamaktadır. Temin edilen veriler 21 farklı bağımsız değişken ve 1 adet bağımlı değişken olarak tahmin modellerine tanıtılıp 875 günün günlük akım tahmini yapılmıştır.

Çalışma sonuçları incelendiğinde en başarılı tahmini LSTM modeli yaptığı gözlemlenmiştir. En kötü başarıya sahip model ise XGBoost olduğu görülmüştür. Sonuç olarak; LSTM ve GRU modellerinin debi tahmini yapmakta oldukça iyi performanslar gösterdiği sonucuna varılmıştır. Ek olarak CFSR veri setlerinin makine öğrenimi hidrolojik modellemelerinde kullanılabileceği tespit edilmiştir.

ANAHTAR KELİMELER: Makine Öğrenmesi, LSTM, GRU, XGBoost, Hidrolojik Modelleme

ABSTRACT

**A HYDROLOGİC MODELLİNG BASED ON MACHINE LEARNING:
FIRTINA STREAM CASE
MSC THESIS
ENSAR EFENDİ CELEPCI
BOLU ABANT İZZET BAYSAL UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
(SUPERVISOR: DOÇ. DR. ALPER BAYRAK)
(CO-SUPERVISOR: ASİST. PROF. DR. ŞAFAK KAYIKÇI)
BOLU, AUGUST 2021
XII + 53**

Water is the basis of life on Earth. The lack of water makes it impossible for many of the creatures on Earth to live. So it is very important to be close to water sources since its vital importance. But being close to the water has also brought some problems. The cities set up near water sources have to cope with some problems such as flood, tidal currents and changes in the water lines. People whose residential areas and farmland have been destroyed as a result of these problems have been forced to leave their settlements. They had also faced financial and spiritual damages. These losses can be avoided thanks to the accurate flow predictions to be obtained by performing regional analyzes.

In this study, flow predictions were carried out through machine learning and deep learning models. Machine learning is a method developed through programming languages that can generate predictions using the learning algorithms. In this study, XGBoost machine learning, LSTM and GRU as deep learning training algorithms were performed and compared.

Fırtına River basin has been chosen as the study site. Daily maximum temperature, minimum temperature, relative humidity, solar radiation and wind variables from the CFSR (climate forecast system reanalysis) system were used for independent variable data sets of the study site. Current observation data were obtained from the Institute of Electrical Affairs (EİS) 2232 current observation station at Topluca Village location. The data covers a total of 8758 days between 1 January 1979 and 31 December 2002. The data were used in prediction models by separating into 21 independent and 1 dependent variable and flow predictions of 875 days were performed.

As a result, it was seen that LSTM model had the best performance while XGBoost had the worst. GRU model has shown quite good performance in estimating flow. In addition, it was observed that the CFSR datasets are suitable to use in machine learning hydrological modeling.

KEYWORDS: Machine Learning, LSTM, GRU, XGBoost, Hydrologic Modeling

İÇİNDEKİLER

KABUL VE ONAY SAYFASI	iii
ETİK BEYAN.....	iv
ÖZET.....	Hata! Yer işareti tanımlanmamış.
ABSTRACT	Hata! Yer işareti tanımlanmamış.
İÇİNDEKİLER	vii
ŞEKİL LİSTESİ.....	ix
TABLO LİSTESİ	x
KISALTMA VE SEMBOLLER LİSTESİ	xi
TEŞEKKÜR	xii
1. GİRİŞ	1
2. GENEL BİLGİLER.....	3
2.1 Önceki Çalışmalar	4
2.1.1 Dünya Geneline Yapılan Çalışmalar	4
2.1.2 Türkiye’de Yapılan Çalışmalar.....	5
2.2 Çalışmanın Amacı ve Kapsamı	6
3. MATERYAL VE YÖNTEM.....	9
3.1 Çalışma Sahasının Özellikleri	9
3.2 YSA İçin Kullanılan Veri Seti.....	11
3.3 Kullanılan Yazılımlar	15
3.4 Tahmin.....	15
3.5 Makine Öğrenmesi	16
3.5.1 Aşırı Öğrenme(Over Fitting)	16
3.5.2 Makine Öğrenimi Modelinin Kurulumu.....	17
3.6 Gradyan Artırma (Gradient Boosting).....	18
3.6.1 XGBoost Yöntemi	19
3.6.2 XGBoost Parametreleri.....	23
3.7 Derin Öğrenme (Deep Learning).....	24
3.7.1 Hata Fonksiyonu ve Optimizasyon.....	28
3.7.2 Uzun Kısa Süreli Bellek (LSTM)	30
3.7.3 Geçitli Tekrarlayan Ünite (GRU)	33
3.7.4 Verilerin Normalizasyonu.....	35
3.7.5 Dropout Tekniği.....	35
3.7.6 Derin Öğrenme Parametreleri	36
3.7.7 Modelin Performansının Değerlendirilmesi	37
3.8 Modeli Eğitecek Verilerin Belirlenmesi.....	38
4. BULGULAR.....	40
4.1 XGBoost Modeli ile Fırtına Deresi Debi Tahmin Bulguları	40
4.2 LSTM Modeli ile Fırtına Deresi Debi Tahmin Bulguları	43
4.3 GRU Modeli ile Fırtına Deresi Debi Tahmin Bulguları.....	45
4.4 Modellerin Karşılaştırılması.....	47
5. SONUÇ VE ÖNERİLER.....	49

6.	KAYNAKLAR.....	51
-----------	-----------------------	-----------



ŞEKİL LİSTESİ

Sayfa

Şekil 3.1. Havzanın Türkiye haritasında konumu.....	9
Şekil 3.2. Fırtına Havzası konum koordinatları.....	10
Şekil 3.3. Fırtına havzasında veri noktalarının konumları.....	12
Şekil 3.4. Fırtına havzasında debi gözlem noktasının harita üzerinde gösterimi.....	13
Şekil 3.5. Düzgün ve aşırı öğrenmiş modelin grafikleri.....	17
Şekil 3.6. Makine öğrenimi model akış diyagramı.....	18
Şekil 3.7. YSA nöron yapısı.....	25
Şekil 3.8. Sigmoid ve Tanh fonksiyonlarının grafiği.....	26
Şekil 3.9. Yapay sinir ağının yapısı.....	27
Şekil 3.10. Geri yayılım algoritmasının görselleştirilmesi.....	28
Şekil 3.11. Örnekleştirilmiş hata değerleri grafiği(Dickson, 2020).....	29
Şekil 3.12. LSTM nöronunun yapısı (Olah,2015).....	31
Şekil 3.13. GRU nöronunun yapısı (Anderson,2019).....	33
Şekil 3.14. Örnek Dropout regülüzasyonunun çalışma şekli (Srivastana vd. ,2014).	36
Şekil 3.15. Korelasyon ısı haritası.....	39
Şekil 4.1. XGBoost dağılım grafiği.....	41
Şekil 4.2. Gerçek ve XGBoost tahmin değerlerinin grafiği.....	42
Şekil 4.3. LSTM dağılım grafiği.....	43
Şekil 4.4. Gerçek ve LSTM tahmin değerlerinin grafiği.....	44
Şekil 4.5. GRU dağılım grafiği.....	46
Şekil 4.6. Gerçek ve GRU tahmin değerlerinin grafiği.....	46
Şekil 4.7. Modellerin hata değerlerinin karşılaştırma grafiği.....	48
Şekil 4.8. Modellerin teknik açıdan karşılaştırılması.....	49

TABLO LİSTESİ

Sayfa

Tablo 1.1. Türkiye’de 1988-1998 yılları arasında gerçekleşen taşkınların özellikleri (Kılıçer ve Özgüler, 2002).	2
Tablo 3.1. Veri noktalarına ait verilerin istatistiksel özellikleri.....	14
Tablo 3.2. NSE değer aralıkları ve performans açıklamaları (Moriasi vd.,2007). 38	
Tablo 4.1. 30 günlük XGBoost tahmin ve gerçek değerleri.....	42
Tablo 4.2. 30 günlük LSTM tahmin ve gerçek değerleri.	44
Tablo 4.3. 30 günlük GRU tahmin ve gerçek değerleri.	47



KISALTMA VE SEMBOLLER LİSTESİ

BP	: Geri Yayılım
CFSR	: İklim Tahmin Sistemi Analizi
CRU	: Hava Araştırma Birimi Tekrarlayan Ünite
D-YSA	: Dalgacık Yapay Sinir Ağı
GB	: Gradyan Artırma
GBDT	: Gradyan Artırılmış Karar Ağaçları
GRU	: Geçitli Tekrarlayan Ünite
LSTM	: Uzun Kısa Süreli Bellek
MAE	: Ortalama Mutlak Hata
MGM	: Meteoroloji Genel Müdürlüğü
MSE	: En Küçük Kareler Regresyonu
MÖ	: Makine Öğrenimi
NSE	: Nash-Sutcliffe Katsayısı
RMSE	: Ortalama Karekök Hatası
RNN	: Tekrarlayan Sinir Ağı
TPU	: Tensor İşleme Birimi
TSA	: Tekil Spektrum Analizi
XGBOOST	: Ekstrem Gradyan Artırma
YBGYSA	: Geçitli Tekrarlayan Ünite
YBGYSA	: İleri Beslemeli Geri Yayınım Sinir Ağı
YSA	: Yapay Sinir Ağı

TEŞEKKÜR

Tez yazım sürecimde yardımlarını esirgemeyen danışmanlarım Doç. Dr. Alper BAYRAK ve Dr. Öğr. Üyesi Şafak KAYIKÇI hocamlarıma, engin tecrübe ve bilgi birikimiyle bana yol gösterici olarak her zaman yanımda olan hocam Prof. Dr. Nusret KARAKAYA'ya en içten duygularıyla teşekkür ediyorum. Her koşulda bana destek olan anneme, babama, kardeşlerime ve çalışma arkadaşlarıma teşekkür ediyorum. Ek olarak bu zamana kadar yetişmemde katkıda bulunmuş tüm öğretmenlerime desteklerinden dolayı çok teşekkür ediyorum.

Bu tez TÜBİTAK tarafından desteklenen 116Y447 nolu proje kapsamında hazırlanmıştır. Bu nedenle TÜBİTAK'a desteklerinden dolayı teşekkür ederim.

1. GİRİŞ

Su yeryüzündeki yaşamın temelini oluşturan maddedir. Suyun var olmaması dünya üzerindeki canlıların çok büyük bölümünün yaşamasını imkânsız hale getirmektedir.

Ortalama yetişkin bir insan vücudunun %60'ı sudan oluşmaktadır. Bu sebeple insanlık için susuz bir yaşamın var olması imkânsızdır. Su insan için bu kadar önemliyken suya yakın ve kolay ulaşımının olmasında aynı derecede önemli olmaktadır. Bu yüzden geçmişteki medeniyetlerin büyük çoğunluğu her zaman suya yakın veya suya ulaşımın kolay olduğu yerlerde kurulmak zorunda kalmıştır. Bunun en büyük örnekleri Fırat ve Dicle nehirleri arasında kurulan Mezopotamya ve Nil Nehri kıyısına kurulan Mısır uygarlıklarıdır. Ancak suya yakın olmak bazı sorunları da beraberinde getirmiştir. Nehir kenarlarına kurulan şehirler; taşkınlar, su yatağının değişimi, gel-git gibi sorunlarla baş etmek zorunda kalmıştır. Bu sorunların sonucu olarak yerleşim alanları ve tarım arazileri yok olan insanlar yerleşim yerlerini terk etmek zorunda kalmıştır. Ek olarak maddi ve manevi zararları da beraberinde getirmiştir. Dünyada ve ülkemizde bu sorunlar nedeniyle meydana gelen can ve mal kayıplarında her geçen gün artış yaşanmaktadır (Kılıçer ve Özgüler , 2002).

Taşkınlar ve seller dünyada aşırı bölgesel yağışlardan ya da toplu kar erimeleri nedeniyle oluşmaktadır. Taşkınlar, bölgenin iklim koşullarına, topoğrafik ve jeoteknik özelliklerine bağlı olarak oluşan bir doğal afettir. Ülkemizde taşkın felaketleri, depremlerin ardından en büyük ekonomik kayıplara neden olan doğal felaketlerdir. Var olan verilere bakıldığında taşkınların sebep olduğu ekonomik zarar her yıl ortalama 100 milyon ABD dolarını ulaşmaktadır. Taşkınlarla mücadele için genelde yapısal önlemler alınmaktadır. Sürdürülen proje faaliyetleri için ayrılan yatırım miktarı ise yılda ortalama 30 milyon ABD dolarına ulaşmaktadır. Türkiye’de 1988-1998 yılları arasında oluşan taşkınların karakteristik özellikleri Tablo 1.1’de gösterilmiştir (Kılıçer ve Özgüler ,2002).

Tablo 1.1. Türkiye’de 1988-1998 yılları arasında gerçekleşen taşkınların özellikleri (Kılıçer ve Özgüler, 2002).

Yıl	Taşkın Sayısı	Ölü Sayısı	Su altında kalan alan (ha = 10000 m ²)	Toplam parasal zarar (US\$)
1989	10	1	9500	1.900.000
1990	26	57	7450	206.000.000
1991	23	23	15770	14.000.000
1992	14	1	690	11.000.000
1993	2	-	60	43.000
1994	9	4	1680	1.200.000
1995	20	164	201100	1.100.000.000
1996	4	1	11000	12.000.000
1997	1	-	1390	60.000
1998	2	57	7000	600.000.000
Toplam	111	306	255640	1.935.400.000

Doğru yapılan debi tahminleri öncelikli olarak insan hayatına mal olabilecek taşkınların önlenmesi, maddi zararların önlenmesi, çevrenin korunması, hidroelektrik üretimi, sulama planlaması, su temini ve su kaynaklarının sürdürülebilir gelişimi gibi su ile ilgili mevcut uygulamaları desteklemektedir.

Yapılan araştırmalarda Antik Mısır medeniyetlerinde taşkınları ve gel git gibi su olaylarını doğru tahmin etmek için matematik, geometri ve yıldızların konumunun sıkça kullanıldığı görülmüştür. Yaptıkları doğru tahminlerin sonuçlarını da oldukça verimli bir şekilde kullanmışlardır. Örneğin taşkın sonucunda nehrin geride bıraktığı verimli alüvyon toprakları tarım için kullanmışlardır. Fakat günümüzde kullanılan debi tahmin metotları teknolojinin ve bilgisayar sistemlerinin gelişmesiyle geçmişe nazaran çok daha üstün hale gelmiştir.

Uzun yıllardır debi tahmini yapmak için birçok yağış akış modeli kullanılmıştır. Geliştirilen çoğu model deneysel, kavramsal veya fiziksel temellere dayanmaktadır. Ancak son zamanlarda veriye dayalı modeller daha popüler hale gelmiştir. Yeni verimli algoritmalara ek olarak veriye dayalı modellerin arkasındaki ana etmenler, veri kullanılabilirliği ve bilgisayar işleme gücüdür. Her ikisi de son yıllarda büyük ölçüde gelişmiş ve veriye dayalı modellerin popüleritesindeki artışa sebep olmuştur.

2. GENEL BİLGİLER

Akarsu havzaları ölçülebilen belirli değişkenlere sahip doğal sistemlerdir. Sistem içinde bulunan değişkenler; nem, yağış, iklim koşulları, bitki örtüsü, jeomorfolojik özellikler gibi birçok değişken sayesinde akıma dönüştüğü bir döngüyü barındırmaktadır. Akarsu havza sistemlerinde bu değişkenler kullanılarak hazırlanan hidrolojik modeller vasıtasıyla giriş değişkenlerinden çıkış akımının tahmin edilebilmesini mümkün kılmaktadır (Singh and Fiorentino, 1996; Özcan, 2007).

Hidrolojik modelleme, alt yapı sistemlerinin (su, kanalizasyon, drenaj ve taşkın) hidrolik olarak analizi yapılarak bu sistemlerin matematiksel bir modelinin oluşturulma işlemidir. Hidrolojik bir modelin oluşturulabilmesi için sahadaki mevcut tesislerin sayısallaştırılması çok önemlidir. Hidrolojik modelleme sayesinde sistemin izlenebilirliği ve erken uyarı sistemleriyle devamlılığı daha hızlı ve doğru bir şekilde yapılabilir. Modellemenin sonucu olarak sistemin ekonomik olarak devamlı olması sağlanır.

Hidrolojik süreci oluşturan parametreler oldukça karmaşıktır. Ayrıca parametre değişimlerinin tam olarak tahmin ve kontrol edilmesi oldukça zordur. Bu sebeplerden ötürü elde edilen veriler eksiksiz ve doğru ölçümlü yapılsada hidrolojik süreçlerin daha basite indirgenerek değerlendirilmesi daha sağlıklı sonuçların alınmasını sağlamaktadır (Chow vd, 1988; Cüceloğlu 2013).

Hidrolojik modeller;

- Ölçüm verisi olmayan havzalarda hidrolojik veri üretilmesi,
- Havzalarda olabilecek değişimlerin havzaya etkisinin ortaya çıkarılması,
- Global iklim durumlarının modellenmesi,
- Depolama tesislerinde işletim, planlama ve gelişim konularında ihtiyaç olabilecek ileriye dönük verilerin tahmin edilmesi,

amaçlarıyla kullanılmaktadırlar (Wheater vd., 2007; Uğurluoğlu, 2017).

2.1 Önceki Çalışmalar

2.1.1 Dünya Geneline Yapılan Çalışmalar

Kratzert vd. 2018 yılında yaptıkları çalışmada yağış-akış modellemesini Uzun Kısa Süreli Bellek (LSTM) ağını kullanarak yapmışlardır. Test etmek için sonuçları iyi bilinen Sacramento Toprak Nemi Hesaplama Modeli (SAC-SMA) ile Snow-17 kar rutini ile karşılaştırmak için serbestçe kullanılabilen CAMELS veri setinin 241 havzasını kullanmışlardır. Sonuç olarak LSTM yapılarının performansını ortaya koymuşlardır.

Le vd. 2019 yılında yaptıkları araştırmada taşkın tahmini için günlük deşarj ve yağışın girdi verisi olarak kullanıldığı bir LSTM derin öğrenme modelini kullandılar. Bir günlük, iki günlük ve üç günlük ileri akış tahmini için LSTM modelinde tahminler oluşturdular. Sonuç olarak tahmin yeteneği oldukça yüksek bir model ettiler.

Ni vd. 2019 yılında yaptıkları çalışmada akarsu akış ve yağış tahmini için LSTM'ye dayalı iki hibrit model geliştirdiler. Modelleri değerlendirmek için iki akış veri seti ve iki yağış veri seti kullandılar. Çalışma sonucunda LSTM'nin zaman serisi tahmini için uygulanabilir olduğunu gösterdiler.

Han vd. 2019 yılında yaptıkları çalışmada referans buharlaşmayı Ekstrem Gradyan Artırma (XGBOOST) algoritmasıyla tahmin modeli geliştirdiler. 1991-2015 yılları arasında üç istasyondan (Datong, Yinchuan ve Taiyuan) gelen meteorolojik verileri kullandılar. Model eğitimi ve testi için, ortalama hava sıcaklığı, maksimum hava sıcaklığı, minimum hava sıcaklığı, rüzgâr hızı, bağıl nem gibi parametreleri kullandılar. Çalışma sonucunda oldukça yüksek performans değerleri elde ettiler.

Ayzel ve Heistermann 2021 yılında yaptıkları araştırmada GR4H hidrolojik modeli ile derin öğrenme modeli olan LSTM ve Geçitli Tekrarlayan Ünite (GRU) arasındaki performans oranlarını incelemişlerdir. Veri seti olarak CAMELS veri setini kullanmışlardır. Çalışma sonucu olarak derin öğrenme mimarilerinin yağış akış modellemesindeki performansını ortaya koydular.

Tyralis vd. 2021 yılında yaptıkları çalışmada günlük akış tahmini için çeşitli makine öğrenimi algoritmaları ile hidrolojik modelleri karşılaştırdılar. Yapılan çalışmada XGBOOST algoritmasının oldukça iyi sonuçlar verdiği gözlemlenmiştir.

2.1.2 Türkiye’de Yapılan Çalışmalar

Türkiye’de hidrolojik modelleme için Yapay Sinir Ağı (YSA) ile yapılan birçok çalışma mevcuttur. Fakat LSTM ile yapılan çalışmalar oldukça azdır.

Yurdusev vd. 2008 yılında Akarçay bölgesindeki aylık akım değerlerinin tahmini için YSA ile bir çalışma yapmışlardır. Çalışma için veriler bölgede mevcut olan yağış gözlem ve akım istasyonlarından alınan veriler kullanılmıştır. Bu çalışmada, YSA ve Makine Öğrenmesi (MÖ) tahmin modelleri karşılaştırılmıştır. Sonuç olarak YSA modelinin veri tahmininde daha yüksek bir performans ortaya koyduğu gözlemlenmiştir.

Terzi ve Köse 2010 yılında yaptıkları çalışmada YSA yöntemi ile Göksu Nehri’nin akım tahminini yapmışlardır. YSA için akım gözlem istasyonlarından alınan günlük akım değerleri kullanılmıştır. Çalışma sonucunda performans değerleri, YSA yönteminin akım tahmininde kullanılabilir olduğunu göstermiştir.

Gümüş vd. 2013 yılında yaptıkları çalışmada Orta Fırat Havzasında bulunan istasyonlardan elde edilen akım ve yağış verileri arasındaki bağlantıyı İleri Beslemeli Geri Yayılım Sinir Ağı (YBGYSA) metodu ile Çoklu Doğrusal Regresyon (ÇDR) yöntemini karşılaştırmışlardır. Çalışma sonunda YBGYSA yönteminin ÇDR yöntemine göre daha iyi sonuç verdiği gözlemlenmiştir.

Kızılaslan vd. 2014 yılında yaptıkları çalışmada Sakarya Nehri üzerinde yer alan akım gözlem istasyonu verilerini kullanarak akım tahmini için model geliştirmişlerdir. Çalışmada yapay sinir ağlarının YBGYSA kullanılmış oluşturulan tahmin modelin tahmin performansı oldukça yüksek sonuçlar verdiği gözlemlenmiştir.

Terzi ve Barak 2015 yılında yaptıkları çalışmada Kızılırmak Nehri üzerinde bulunan Söğütluhan akım istasyonuna ait akım tahmin modeli geliştirmişlerdir. Veri seti için Sivas meteoroloji istasyonuna ait veriler kullanılmıştır. Dalgacık-

yapay sinir ağı (D-YSA) modellerinin YSA modellerinden daha iyi performans gösterdiğini ortaya koymuşlardır.

Latifoğlu ve Nuralan 2020 yılında yaptıkları çalışmada LSTM ağları ile nehir akım tahmini gerçekleştirmişlerdir. Tahmin performansını yükseltmek için Tekil Spektrum Analizi (TSA) kullanılmıştır. Önerilen TSA-LSTM modelini sentezleyerek yüksek performanslı tahmin verisi elde etmişlerdir.

2.2 Çalışmanın Amacı ve Kapsamı

Debi tahmininde kullanılan kavramsal ve fiziksel hidrolojik modeller, fiziksel yasalara, yani kütlenin ve enerjinin korunması yasasına dayanır. Bununla birlikte, debiye etki eden bazı etmenler henüz tam olarak anlaşılamamıştır. Ek olarak bazı etmenlerin de denklemlere çevrilmesi oldukça zordur. Bu bilgi boşluklarını doldurmak için varsayımların yapılması gerekmektedir. Bu varsayımlar ve bilinmeyen etmenler tahminlerin doğrusallığını olumsuz olarak etkilemektedir. Veriye dayalı tahminler tam olarak bu noktada devreye girmektedir. Temelde veriye dayalı tahminler modellemeye daha objektif ve daha doğrusal bir yaklaşım sağlamaktadır.

Veriye dayalı tahminlerin en büyük ihtiyacı kullanılabilecek nitelikli verilerdir. Uydular, radarlar ve gözlem istasyonları gibi uzaktan algılama teknikleriyle büyük miktarlarda nitelikli veri elde edilebilir hale gelmiştir. Coğrafi bilgi sistemleri ile bu büyük miktardaki verilerin işlenmesi kolaylaşmıştır. Ayrıca ölçüm cihazları da her geçen gün gelişim göstermektedir. Her tür hidrolojik parametre sahada daha doğru bir şekilde ölçülebilir hale gelmiştir. Örnek vermek gerekirse, dere akışı, meteorolojik veriler ve yeraltı suyu seviyesi sürekli olarak uzaktan izlenebilir hale gelmiştir. Tüm bu verilerin uygun bir süre içinde işlenmesi için daha fazla hesaplama gücüne ihtiyaç duyulmuştur.

Elde edilen verilerin işlenmesi için Google'ın geliştirdiği yeni bir teknoloji olan Tensor İşleme Birimi (Tensor Process Unit-TPU) sayesinde daha kısa sürede daha fazla veri işlemeyi mümkün kılmıştır. Bu işlem gücü kullanılarak muazzam miktarda veriyi çok kısa sürelerde işlemek mümkün hale gelmiştir.

Veriye dayalı tahminde MÖ yöntemi kullanılmaktadır. MÖ, verilerden öğrenen, kalıplar belirleyen ve tahminlerde bulunan otomatikleştirme sistemleri veya algoritmalarla ilgili bir yapay zekâ alt alanıdır.

MÖ'nin geçmişine bakıldığında kökeni 1950'lere kadar uzanmaktadır. Ancak son zamanlarda uygulama alanları bir hayli gelişmiştir. Günlük alanlarda kullanılan MÖ örnekleri;

- E-posta spam ve kötü amaçlı yazılım filtreleme işlemleri,
- Çevrimiçi müşteri desteği,
- Çevrimiçi dolandırıcılık algılama,
- Ürün önerileri,
- Sanal kişisel asistanlar
- Otonom araçlar

MÖ doğrusal olmayan tahmin ve sınıflandırma problemlerinde başarılı sonuçlar verebilmektedir. Giriş katmanı verilerinin vektörel ve fiziksel değerlerine bakılmaksızın matematiksel modelleme yapılarak veri giriş katmanı ve çıkışı arasında modelleme yapabilmektedir.

Günümüzde regresyon için kullanılabilen birçok makine öğrenimi algoritması vardır. Bu algoritmalara, destek vektör makineleri, rastsal ormanlar, yapay sinir ağları, gradyan artırılmış karar ağaçları (GBDT) ve Navie Bayes örnek olarak verilebilir.

Bu tezde bir GBDT alt algoritması olan XGBOOST, derin öğrenme algoritmaları olan tekrarlayan bir sinir ağı LSTM ve GRU algoritmaları kullanılmıştır.

XGBOOST, 2016 yılında yayımlandığı için nispeten diğer makine öğrenim tekniklerine göre daha yeni bir MÖ algoritmasıdır (Chen vd. 2016).

Genel olarak XGBOOST modelinin birçok avantajı vardır bunlar (Robert , 2014):

- Modeller kolaylıkla yorumlanabilir.
- Veri eğitim süresi oldukça hızlıdır.

- Sürekli ve kategorik verilerle ilgilenebilir.
- Otomatik olarak eksik verilerle ilgilenir.
- Doğrusal olmayan ilişkileri yakalayabilir.

Her sistemde oluğu gibi XGBOOST algoritmasında bazı dezavantajları vardır;

- Her sınıflandırıcının bir önceki sınıflandırmahatasını düzeltmesi ve aykırı değerlere duyarlı olması gerekir. Bu nedenle XGBOOST yöntemi aykırı değerlere çok bağımlıdır.
- Her tahmin doğruluğunu önceki tahmin değerlerine dayandırır. Böylece modelin akıcı hale gelmesi zorlaşır.

XGBOOST'un aksine, LSTM modelleri uzun yıllardır kullanılmaktadır. LSTM modelleri ve YSA'nın birçok avantajı vardır. Bunlar (Gonzalez ve Wen Yu , 2018):

- Gürültülü verilere ve ölçüm hatalarına karşı dayanıklıdır.
- Eğitim verileri değiştikçe zamanla uyum sağlar.
- Doğrusal olmayan davranışı modellemede iyi sonuçlar verir.
- Eğitim süresini azaltmak için birden çok işlemci üzerinde eğitilebilirler.
- Parametre tahminindeki hataları ortadan kaldırır.

LSTM modelinin bu tezde uygulanmasının ana nedeni, doğrusal olmayan davranışı modelleyebilmesi ve zamana bağlılıkla başa çıkabilmesidir. Ancak YSA'ların da kendi sınırlamaları aşağıda verilmiştir (Yildirim , 2018) :

- Eğitilmiş parametreler fiziksel hiçbir şeyi temsil etmez; girdi parametrelerinin önemini otomatik olarak göstermezler.
- Eğitim almak için genellikle çok fazla veriye ihtiyaç duyarlar.
- Sinir ağı mimarisini seçmenin bir yolu yoktur.

Bu çalışmada Fırtına Deresinin günlük ortalama debi (m^3/s) değerlerinin tahmini yapılmıştır. Tahmin verileri için Fırtına Deresi ortalama günlük akım değerleri, 4 farklı istasyondan elde edilen, maksimum sıcaklık, minimum sıcaklık, yağış, rüzgâr hızı, bağıl nem ve solar radyasyon verileri (toplamda 25 farklı veri) XGBOOST, LSTM ve GRU için girdi verisi olarak kullanılmıştır.

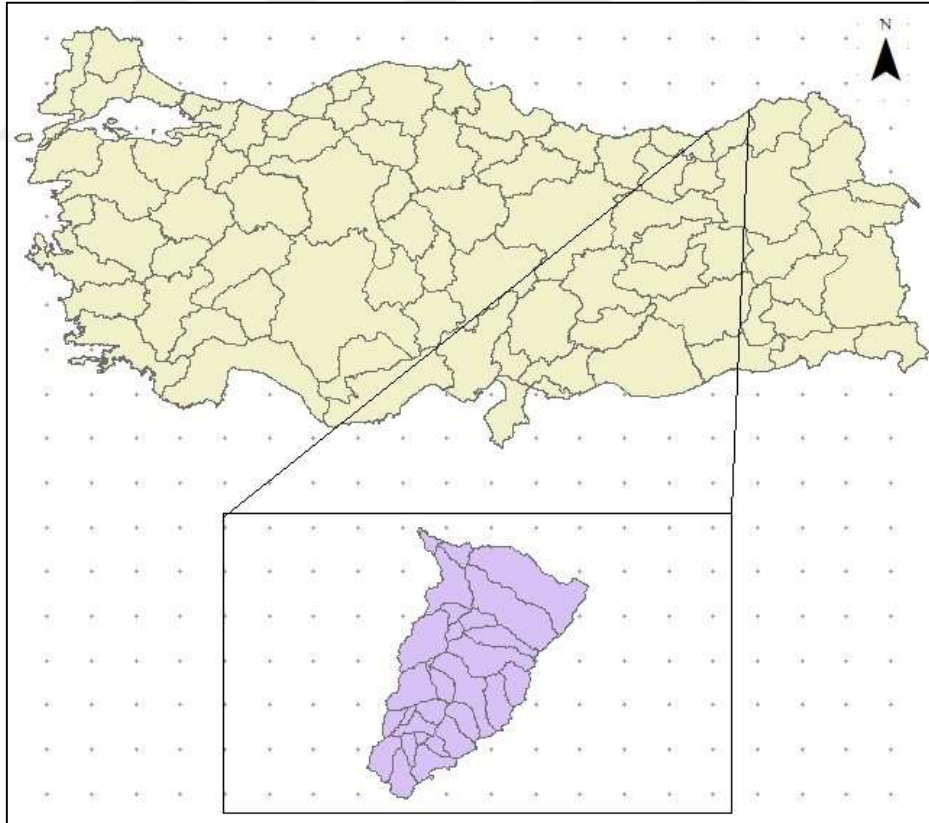
3. MATERYAL VE YÖNTEM

Çalışmanın bu bölümünde kullanılan materyal ve yöntemler sunularak ilk bölümde modellenmiş su havzasının tanımı yapılmıştır.

İkinci bölümde değişkenlerin tanımı ve verilerin nereden elde edildiği açıklanmıştır. Son bölümde kullanılan yazılımların tanımı yapılmıştır.

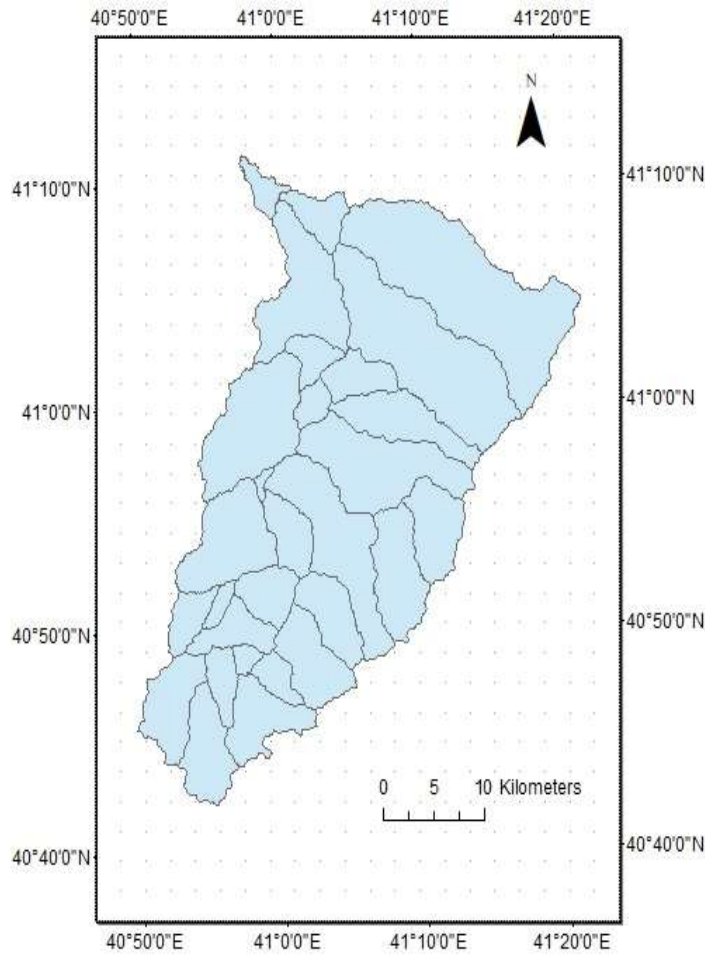
3.1 Çalışma Sahasının Özellikleri

Fırtına Havzası; Türkiye'nin kuzeydoğusunda $40^{\circ} 30'$ ile $41^{\circ} 20'$ kuzey enlemleri ve $40^{\circ} 22'$ ile $41^{\circ} 28'$ doğu boylamları arasında konumlanmaktadır. Havzanın doğusunda Yusufeli (Artvin) Fındıklı (Rize), güneyinde İspir (Erzurum), Batısında Hemşin, Pazar, Güneysu (Rize) ve kuzeyinde Karadeniz'e kıyısı bulunmaktadır ve havzanın dreanaj alanı yaklaşık 1155 km^2 'dir.



Şekil 3.1. Havzanın Türkiye haritasında konumu.

Havzada deniz kıyısından hemen sonra dağlar yükselmeye başlar. Kuş uçuşu yaklaşık 40 km mesafede yükseklik 3900 metrelere yaklaşır. Türkiye'nin 4. büyük zirvesi olan Kaçkar Dağı 3932 m yükseklik ile çalışma sahasının en yüksek noktasıdır. Bölgede bulunan Altıparmak Dağı (3562 m), Verçenik Dağı (3711 m), havzayı güneyden sınırlayan diğer yüksek kütlelerdir. Bölge doğuda Çağlayan Deresi havzası, batıda ise Ortaköy Deresi havzası ile çevrelenmiştir (Bayrakdar, 2006).



Şekil 3.2. Fırtına Havzası konum koordinatları.

Bölgenin içine ilerledikçe daha yüksek kademeli yamaçlar görülür. 2000 metrenin üzerinde kayalıklara rastlanmaktadır. Ayrıca 2500 metre yükseklikten sonra büyük glasiyal vadilere rastlanmaktadır. Daha yüksek noktalarda ise sivri tepeler, buzullarla ve daima karlarla kaplı dağlar yer almaktadır. Tektonik

deformasyonlar akarsu şebekesinin oluşumunu etkilemiştir. Ayrıca deformasyonlar bölgenin bazı noktalarında 300 metreye ulaşan ani su düşüşlerine ve eğim kırıklıklarına neden olmuştur. Alçak yüksekliklerde ve platolarda okyanusal iklim şartlarına benzeyen bir yağış etkinliği görülmektedir. Fırtına deresi havzasında, eğim Doğu Karadeniz dağlarının zirvelerinden Karadeniz’e doğru yönelmektedir (Bayrakdar, 2006).

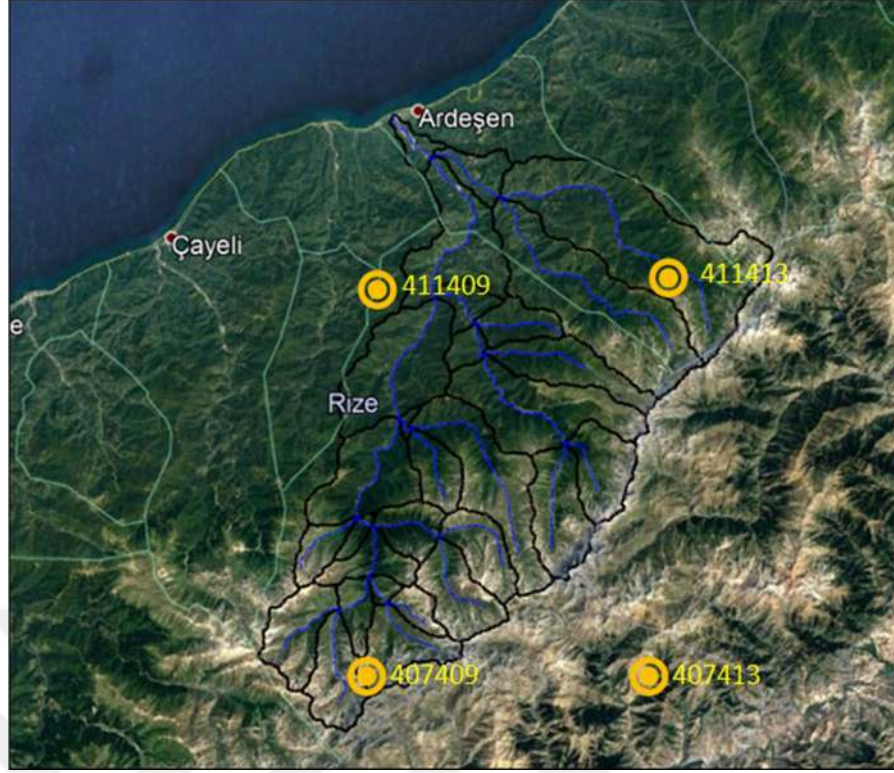
3.2 YSA İçin Kullanılan Veri Seti

Fırtına havzasının zor koşulları, yer istasyonlarının sürekli veri barındırmaması ve daha önce yapılan çalışmalarda yer istasyonlarına göre daha yüksek performans gösterebilmesi nedeniyle CFSR (İklim Tahmin Sistemi Analizi) sisteminden elde edilen veri noktaları kullanılmıştır.

CFSR sistemi dünya çapında modelleme yapabilen ve bu modelleme sonucu hava veri seti üretebilen bir sistemdir. Sistem tüm dünyayı 0.25°’lik grid (kare) bölmektedir. Her bir grid 30kmx30km’lik alanı kapsamaktadır. CFSR sistemi veri seti üretmek için meteoroloji balonlarını, uçakları ve ACARS (Hava Aracı İletişim Adresleme ve Raporlama Sistemi) gözlemlerini, US-NWS(ABD Ulusal Hava Merkezi gözlemleri, sabit yörüngeli (geostationary)) uydu gözlemleri ve yer gözlemlerini içeren gözlem değerlerini kullanmaktadır (Saha vd., 2010, Saha vd. , 2014).

Cüceoğlu vd, (2017) yılında yaptığı modelleme çalışmalarında CFSR, CRU(Climatic Research Unit) ve Meteoroloji Genel Müdürlüğü (MGM) veri setlerinin performans değerleri üzerinde çalışmıştır. Yapılan çalışmanın sonucunda CFSR ve CRU veri setlerinin MGM veri setine göre daha yüksek bir performans gösterdiğini ortaya koymuştur.

CFSR sistemi fırtına havzası için 4 farklı veri noktası belirlemektedir. Şekil 3.3 ’te veri noktalarının bölge içinde konumları haritalandırılmıştır.



Şekil 3.3. Fırtına havzasında veri noktalarının konumları.

CFSR sisteminden alınan 4 farklı veri noktaları 1 Ocak 1979-31 Aralık 2002 yılları arasında toplamda 8758 günün verisini sağlamaktadır. Her veri noktası günlük olarak maksimum sıcaklık, minimum sıcaklık, yağış, rüzgâr, bağıl nem ve solar radyasyon verilerine erişim sağlamıştır. Toplamda 24 farklı değişken elde edilmiştir.

Debi gözlem verileri ise Elektrik İşleri Enstitüsü (EİE) idaresinde bulunan 2232 nolu akım gözlem istasyonundan alınmıştır. Akım gözlem istasyonu Topluca Köyü mevkiinde bulunmaktadır. Veriler 1 Ocak 1979 ve 31 Aralık 2002 yılları arasında toplam 8758 günün verisini kapsamaktadır. Şekil 3.4'te akım gözlem istasyonunun konumu gösterilmiştir.



Şekil 3.4. Fırtına havzasında debi gözlem noktasının harita üzerinde gösterimi.

CFSR veri setleri YSA için bağımsız değişkenler olarak kullanılmıştır. Debi gözlem verileri ise YSA için bağımlı değişken olarak kullanılmıştır.

Her veri noktası 6 adet bağımsız değişken barınmaktadır. Bu değişkenler; yağış, bağıl nem, maksimum sıcaklık, minimum sıcaklık, rüzgâr, solar radyasyon verileridir. Toplamda 24 farklı bağımsız değişken elde edilmiştir. Bu bağımsız değişkenler bağımlı değişken olan debiyle bağdaştırılmıştır.

Tablo 3.1. Veri noktalarına ait verilerin istatistiksel özellikleri.

Veri noktasının adı	İstatistiksel Parametreler	En Yüksek Sıcaklık C°	En Düşük Sıcaklık C°	Yağış (mm)	Rüzgâr (m/sn)	Bağıl Nem (%)	Solar Radyasyon(W/m²)
411413' nolu veri noktası	Veri Sayısı	8758	8758	8758	8758	8758	8758
	Ortalama	11,02	2,68	4,35	2,36	0,85	14,07
	En Küçük Değer	-11,40	-35,01	0,00	0,64	0,23	0,00
	En Yüksek Değer	32,79	20,13	72,77	7,27	0,99	33,57
	Standart Sapma	7,98	6,64	7,31	0,82	0,11	8,98
411409' nolu veri noktası	Veri Sayısı	8758	8758	8758	8758	8758	8758
	Ortalama	12,39	4,00	4,18	2,35	0,83	14,10
	En Küçük Değer	-9,55	-28,50	0,00	0,76	0,22	0,00
	En Yüksek Değer	33,91	22,57	73,68	7,53	1,00	33,17
	Standart Sapma	7,61	6,17	7,70	0,83	0,13	8,87
407409' nolu veri noktası	Veri Sayısı	8758	8758	8758	8758	8758	8758
	Ortalama	7,98	-0,19	6,85	2,03	0,88	13,60
	En Küçük Değer	-16,04	-37,30	0,00	0,48	0,21	0,00
	En Yüksek Değer	33,28	19,93	102,03	6,20	1,00	34,24
	Standart Sapma	8,94	7,07	9,21	0,69	0,13	9,02
407413' nolu veri noktası	Veri Sayısı	8758	8758	8758	8758	8758	8758
	Ortalama	8,09	-1,02	6,25	1,99	0,87	13,86
	En Küçük Değer	-16,36	-37,35	0,00	0,45	0,22	0,00
	En Yüksek Değer	33,57	17,67	83,43	5,91	1,00	34,41
	Standart Sapma	9,31	7,45	8,14	0,67	0,13	8,91

3.3 Kullanılan Yazılımlar

Bu çalışmada açık kaynak kodlu yazılım olan Python 3.8 makine öğrenimi eğitimi için kullanılmıştır. Programda makine öğrenim modelini eğitmek için çeşitli kütüphaneler kullanılmıştır. Kullanılan kütüphanelerden Tensorflow ve Keras kütüphaneleri makine öğrenimi modelini geliştirecek kütüphaneler olarak kullanılmıştır. Numpy kütüphanesi çok boyutlu dizi ve matrisler üzerinde matematiksel işlevler gerçekleştirmesini sağlayan kütüphane olarak kullanılmıştır. Pandas, veri işlemesi ve analizi için yazılmış olan bir yazılım kütüphanesidir. Yapılan programlama işlemleri Google'ın geliştirdiği sanal programlama imkânı sağlayan Colaboratry ortamında yapılmıştır.

3.4 Tahmin

Tahmin (forecast) anlam olarak geçmiş ve şimdiki verilere dayanarak sonraki süreçleri veriye dayalı olarak tahmin etme işlemidir. Örnek olarak belirli bir gelecek zamanda bazı değişkenler kullanılarak başka bir değişkenin tahmini olabilir.

Zaman serisinde birden fazla noktanın tahmini yapılacaksa iki temel yöntem vardır. Bunlar doğrudan tahmin (direct forecast) ve özyinelemeli (recursive forecast) tahmindir.

Doğrudan tahmin kısaca tahmini yapılacak olan her zaman için ayrı bir model oluşturmaktadır. Örnek vermek gerekirse geçmiş veri seti kullanılarak gelecekte 5 ayrı noktanın tahmini yapılacaksa doğrudan tahmin işlemi her nokta için ayrı bir model oluşturarak bu noktaların değerlerini tahmin edecektir. Matematiksel denklem olarak eşitlik (3.1)'de verilmiştir.

Özyinelemeli tahmin ise kısaca tahmini yapılacak her zaman için bir önceki yapılan tahmini kullanmaktadır. Örnek olarak geçmiş verilerden 5 ayrı noktanın tahmini için öncelikle ilk noktanın değer tahmini yapılır. Bu yapılan tahmin sonraki tahmin noktası için değişken olarak kullanılır. Basitçe bir önceki tahmin sonraki tahmin için bir değişken oluşturmaktadır. Matematiksel denklem olarak eşitlik (3.2)'de verilmiştir.

$$\hat{y}_{t+h} = f_h(y_t, y_{t-1}, \dots, y_{t-M+1}) \quad 1 \leq h \leq H \quad (3.1)$$

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, \dots, y_{t-M+1}) \quad (3.2)$$

$$\hat{y}_{t+2} = f_2(y_{t+1}, y_t, \dots, y_{t-M+2})$$

Denklemden $\hat{y}$ tahmin değerini, f_h tahmin modelini, y_t giriş değerlerini, M giriş değerlerinin sayısını ve son olarak H terimi tahmin edilecek nokta sayısını ifade etmektedir.

Bu çalışmada doğrudan tahmin modeli kullanılmıştır. Bunun ilk sebebi yapılan çalışmada giriş değerlerinin tahmin edilmesine ihtiyaç olmamıştır. İkinci sebebi ise özyineleme tahmin modelinde ilk tahmindeki büyük bir hata sonraki tahminlerde de hatalara sebep olacaktır.

3.5 Makine Öğrenmesi

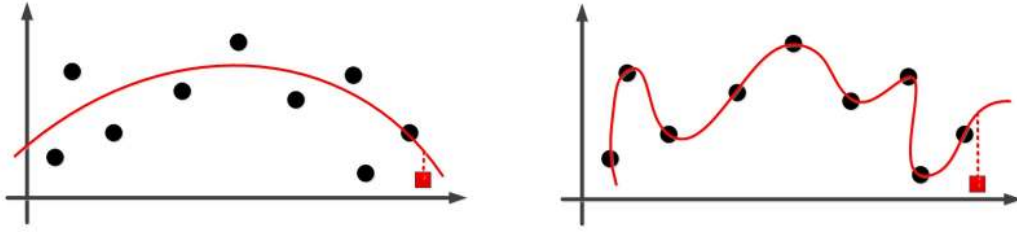
MÖ, çevredeki ortamdan öğrenerek insan zekâsını taklit etmek için tasarlanmış ve gün geçtikçe gelişen bir hesaplama algoritması dalıdır. Makine öğrenimine dayalı teknikler, finans, görüntü işleme, örüntü tanıma, uzay aracı mühendisliği, eğlence ve tıbbi uygulamalara kadar çeşitli alanlarda başarıyla uygulanmaktadır (Naqa ve Murpy ,2015).

MÖ modeli kurulumu ve bu kurulan yöntemlerde dikkat edilmesi gerek en önemli nokta olan aşırı öğrenme (over fitting) konuları iki alt başlıkta incelenmiştir.

3.5.1 Aşırı Öğrenme(Over Fitting)

Aşırı öğrenme kısaca modelin eğitim veri kümesine aşırı oranda uyum göstermesidir diğer bir deyişle giriş veri setini ezberlemesidir. Modelde istenmemesinin nedeni eğitim veri kümesinde var olmayan yeni verilere uyum sağlayamamasıdır (Alpaydın ,2020).

Şekil 3.5’te eğitim verilerine göre üretilmiş iki farklı model oluşturulmuştur. Kırmızı çizgi oluşturulan modeli siyah noktalar ise eğitim veri seti değerlerini ifade etmektedir. Grafiklere bakıldığında eğitim hata fonksiyonu sağ taraftaki modelde diğerine nazaran daha düşük olacaktır. Ancak eğitim veri setinden bağımsız bir test verisi (kırmızı kare) model üzerinde uygulandığında sol taraftaki modelin daha iyi bir performans gösterdiği görülecektir.



Şekil 3.5. Düzgün ve aşırı öğrenmiş modelin grafikleri.

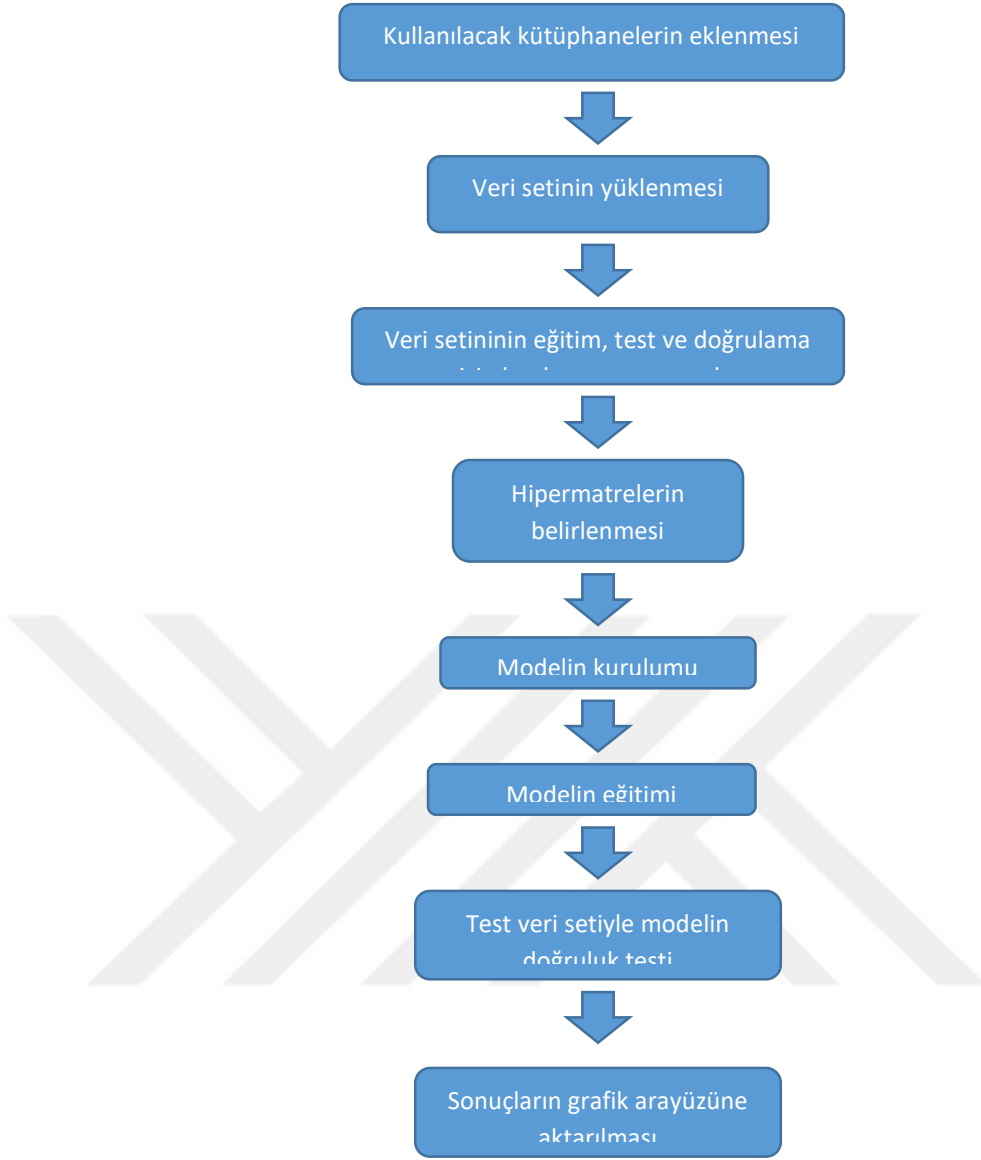
Aşırı öğrenme makine öğrenmesinde ana problemlerden bir tanesidir. Sebebi genelleme ve optimizasyon arasındaki farklardır. Aşırı öğrenme problemi her algoritma için farklılık gösterebilen düzenleme (regülüzasyon) teknikleri uygulanarak çözülebilir.

3.5.2 Makine Öğrenimi Modelinin Kurulumu

MÖ modeli kurulumu oldukça basittir. Öncelikle veri seti 3 ayrı bölüme ayrılmaktadır. Bu çalışma için meteorolojik ve gözlem verileri YSA için eğitim (train), deneme (test) ve doğrulama (validation) işlemleri için bölümlere ayrılmıştır. Bu ayrımın yapılmasının ana sebebi aşırı öğrenmenin önüne geçilmek istenmesidir.

Tüm veri setinden eğitim veri seti için %80 (7005 gün), doğrulama veri seti için %10 (876 gün) ve deneme veri seti için %10 (876 gün) oranında veri ayrımı yapılmıştır.

MÖ modeli algoritmaya verilen eğitim veri seti sayesinde modelin eğitime başlar. Modelin öğrenim şekli verilen algoritmaya göre farklılık göstermektedir. Eğitim sürecinde doğrulama verisi kullanılarak aşırı öğrenimin önüne geçilir. Eğitim işlemi bittikten sonra test verisi sisteme verilir ve sonuç olarak tahmin verisi alınır. MÖ model akışı şekil 3.6' da verilmiştir.



Şekil 3.6. Makine öğrenimi model akış diyagramı.

3.6 Gradyan Artırma (Gradient Boosting)

Bu bölümde MÖ modeli olan Gradyan Artırma (GB) modeli açıklanmıştır. Modelin nasıl çalıştığı, parametrelerinin neler olduğu, bunların nasıl optimize edildiği ve eğitimin nasıl yapıldığı anlatılmıştır.

GB makine öğrenimi hem regresyon hem de sınıflandırma problemleri için kullanılır. Zayıf tahmin yapan Karar Ağacı (Decision Tree) modellerinin bir araya gelmesiyle oluşturulan bir modeldir. Öğrenme algoritmasının amacı, bir kayıp fonksiyonu tanımlamak ve tanımlanan kayıp fonksiyonunu en aza indirmektir. Gradyan artırma metodunu en küçük kareler regresyonunu (MSE) kullanarak

açıklamak en kolaydır. $\hat{y} = F(x)$ cinsinden bir fonksiyon varsayımı yapılırsa. Fonksiyona bakıldığında burada F değeri tahmin modelini, $\hat{y}$ yapılan tahmini, x giriş verilerini ifade etmektedir. Sistem sonucunda elde edilen hata değerlerini minimize etmek için eşitlik (3.3) uygulanır.

$$\frac{1}{n} \sum_{i=0}^n (\hat{y}_i - y_i)^2 \quad (3.3)$$

Denkleme bakıldığında $\hat{y}_i$ yapılan tahmin değerlerini, y_i gerçek değerleri ve n değeri ise yapılan örnekleme değerini temsil etmektedir. Sonrasında sistemi geliştirmek için eşitlik (3.4) uygulanır.

$$F_{m+1} = F_m + h \quad (3.4)$$

Geliştirme işlemi ikinci ağaç modelinden başlar. Her m yinelemesi için geliştirilebilecek kusurlu bir F_m modeli vardır. Kullanılan algoritma öncekinden daha iyi bir model oluşturabilmek için sisteme bir h tahmincisi ekler. Denklemden F_{m+1} değeri geliştirilmiş olan ağacı simgelemektedir. h değerini bulmak için tahmin değerleri ile gözlem değerleri arasındaki fark alınır eşitlik (3.5)'te yapılan işlem gösterilmiştir.

$$F_m + h = y \text{ veya } h = y - F_m \quad (3.5)$$

GB işleminde her yineleme, en son modelin hatasını düzeltmeye çalışır. Model, hata sıfıra yakın olana veya kayıptaki azalma belirli bir eşiğin altına düşene kadar ağaçları büyötmeye devam eder. Sistem belirli bir eşiği aştığında algoritma eğitim verilerini kopyalamayı öğrenir. Buda aşırı öğrenmeye sebep olur.

3.6.1 XGBoost Yöntemi

XGBoost, Friedman tarafından geliştirilen GB yöntemi temellerine dayanan bir makine öğrenimi algoritmasıdır. XGBoost yöntemi günümüzde reklam eşleştirme sistemleri, spam mail tespiti, dolandırıcılık tespit sistemleri gibi alanlarda başarılı bir şekilde kullanılmaktadır (T. Chen ve Guestrin, 2016, Abar

2020). Regresyon ve sınıflandırma ağaçları için boosting algoritmalarının kullanımı etkin sonuçlar elde edilmesini sağlamaktadır. XGBoost algoritması başarılı performans değerlerinin elde edilmesinin yanında diğer MÖ yöntemlerine göre hesaplama kolaylığına da sahiptir. Bu özellikleri nedeniyle büyük veri analizlerinde oldukça etkili şekilde kullanılabilir (Abar 2020).

XGBoost algoritması için ilk olarak bir ağaç topluluğu modeli, $\hat{y}_i$ çıktısını tahmin etmek için K toplama fonksiyonlarını kullanır. Eşitlik (3.7)'de kullanılan fonksiyon gösterilmiştir. Tahmin değeri tüm ağaçların sonuçlarının toplamıdır.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (3.7)$$

Eşitlik (3.8)'de verilen bir eğitim setini öğrenmek için amaç fonksiyonu L minimize edilir. l türevlenebilir kayıp fonksiyonudur. Eşitlik (3.9)'a bakıldığında γ düzenlemeyi kontrol eden parametredir. γ değeri ne kadar yüksek olursa düzenleme oranı o kadar yüksek olur. T değeri ağaçlardaki yaprak miktarıdır. Λ terimi ağırlıklar üzerindeki L2 düzenlemesi (Ridge regresyonu) terimidir. w terimi yaprak ağırlıklarını temsil eder kısaca ağacın sonuçlarıdır.

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_{k=1} \Omega(f_k) \quad (3.8)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda ||w||^2 \quad (3.9)$$

Bir sonraki adım Gradyan Artırma denklemiyle (3.5) XGBoost denkleminin (3.8) birleştirilmesidir. Eşitlik (3.10)'da birleşim ifadesi gösterilmiştir. Yeni tahminin eski tahmine eşit olduğu yerde ağaç fonksiyonunu düzenlemek için f_t terimi eklenir. İşlemin amacı eşitlik (3.10)'da belirtilen amaç fonksiyonunu en aza indirebilen yeni bir ağaç bulmaktır.

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \quad (3.10)$$

Bir sonraki adım, kayıp fonksiyonunun türevini almaktır. Karmaşıklığı azaltmak için XGBoost kayıp değerini için ikinci dereceden Taylor genişlemesini hesaplar eşitlik (3.11). Burada g_i ve h_i , sırasıyla bir önceki yinelemede tahminlere göre bir önceki iterasyondaki kaybın Taylor açılımının birinci ve ikinci mertebesidir.

$$\begin{aligned}\mathcal{L}^t &\simeq \sum_{i=1}^n [l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \\ g_i &= \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i &= \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})\end{aligned}\tag{3.11}$$

Sabit kayıp terimi çıkarıldıktan sonra, t. adımda sabit terim çıkarılarak aşağıdaki daha basit bir amaç fonksiyonu elde edilir eşitlik (3.12).

$$\hat{\mathcal{L}}^{(t)} = \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)\tag{3.12}$$

f_t 'nin T yaprak düğümleri olduğu varsayıldığında, I_j j düğüme ait örnekler kümesidir ve w_j bu düğüm için tahmindir. Buna göre I_j 'ye ait bir i örneği için $f_t(x_i) = w_j$ fonksiyonu yazılabilir. Bununla $f_t(x_i)$, eşitlik (3.12)'deki ilgili tahminlerle ikame edilebilir. Bu, eşitlik (3.13) ile sonuçlanır.

$$\begin{aligned}\hat{\mathcal{L}}^{(t)} &= \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \\ &= \sum_{j=1}^T [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i) w_j^2 + \frac{1}{2} \lambda w_j^2] + \gamma T \\ &= \sum_{j=1}^T [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T\end{aligned}\tag{3.13}$$

Sabit bir ağaç yapısı $q(x)$ için, amaç fonksiyonunun türevini her bir yaprak düğümün ağırlığına göre sıfıra eşitleyerek her yaprak için optimal ağırlık W_j hesaplanabilir. Bu, eşitlik (3.14) ile sonuçlanır.

$$\begin{aligned} \frac{d\hat{\mathcal{L}}^{(t)}}{dw_j^*} &= 0 \\ \sum_{i \in I_j} g_i + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \cdot 2 \cdot w_j^* &= 0 \\ w_j^* &= - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \end{aligned} \quad (3.14)$$

Eşitlik (3.14) ile verilen en uygun ağırlıklar eşitlik (3.13)'te yerine koyulursa, eşitlik (3.15)'te verildiği gibi, K düğümlü sabit bir ağaç yapısı eklemek için en iyi global model kaybına ulaşılır.

$$\hat{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (3.15)$$

Böylece algoritma, Taylor işlevini kullanarak hesaplama yükünü azaltır. Ek olarak algoritma sabit bir ağaç yapısı için en uygun ağırlıkların nasıl bulunacağını öğrenmiş olur. Son olarak algoritmanın gerçekleştirmesi gereken işlem olası tüm ağaç yapılarından doğru yapının nasıl seçileceğidir. Bu işlemi yapmak için basit bir ağaç oluşturmaya başlar. Ardından en yüksek kayıp azalmasını sağlayan bölünme seçilir. Kısaca bir dal iki adet dala ayrılmış olur. Bölme sonucu elde edilen sağ ve sol düğümün I_L ve I_R olduğu varsayılırsa bölme işleminin bilgi kazancı ($\mathcal{L}_{split}$) bölünmeden önce ve sonra amaç fonksiyonları karşılaştırılarak eşitlik (3.16)'daki gibi hesaplanabilir (ABAR, 2020).

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (3.16)$$

3.6.2 XGBoost Parametreleri

Tüm makine öğrenimi modellerinde model parametreleri mevcuttur. Parametreler benzer görülebilir fakat aralarında belirgin farklar vardır. Bu parametreler genellikle eğitilen modelin bir parçası olarak kaydedilir ve modelin tahmin performansına etkide bulunur. Parametreler eğitim sürecinin dışındadır ve modele belirtilmesi gerekir. Parametre ayarlama işlemi ızgara araması (Gridsearch), rastgele arama veya Bayesian optimizasyon yöntemleri kullanılarak yapılabilir.

Gridsearch, manuel olarak tanımlanmış tüm konfigirasyon setlerini deneyerek modeli tekrar tekrar eğitir. Bu yöntem küçük aramalar için uygun olsa da, parametre sayısı arttıkça model eğitim hızı yavaşlamaktadır.

Rastgele arama, belirtilen kümelerden rastgele değerlerin seçilmesidir. Sinir ağları için bu yöntemin parametre optimizasyonu Gridsearch ile karşılaştırıldığında daha verimli olduğu deneysel ve teorik olarak gösterilmiştir (Bergstra ve Bengio,2012). Rastgele örneklemenin dezavantajı ise kötü seçilmiş bir parametre aralığının modelin kötü performans göstermesine neden olabilmesidir. Bu nedenle uygun parametre aralıklarını dikkatlice seçmek bu yöntemin en önemli noktasıdır.

Bayes optimizasyon yöntemleri yeni bir parametre seçerken geçmiş örnekleri hesaba katarak bir sonraki çalıştırma için daha iyi bir tahmin sağlar. Gridsearch ve rastgele arama daha verimli olmalarına rağmen, yöntemi anlamak ve uygulamak çok fazla çaba gerekmektedir.

Rastgele arama ve ızgara aramaları genellikle iyi bir başlangıç noktası sağlar.

XGBoost'un birçok parametresi vardır, bunlar genel parametreler, güçlendirici parametreler ve görev parametreleri olarak gruplandırılabilir. Genel parametreler, modelin çalıştırılacağı paralel iş parçacığı sayısı, kullanılan güçlendirici (örneğin ağaçlar veya doğrusal işlevler), ayrıntı düzeyi ve eğitim sırasında geri bildirim sağlamak için diğer seçeneklerden oluşur. Bu parametreler nihai performansı etkilemez, ancak onu geliştirmek için seçenekler sunar.

Booster parametrelerinden önemli olanları:

- Öğrenme oranı (n); Öğrenme oranı ne kadar düşükse, hataları düzeltmek için gelecekteki ağaçlara o kadar fazla yer verilir bunun içinde daha fazla ağaç yapısına ihtiyaç duyulmaktadır.

- Bir ağacın maksimum derinliği; ağaç ne kadar derinse, model o kadar karmaşıktır ve aşırı öğrenmeye o kadar yatkındır.

- Minimum bölünme kaybı (γ); bir ağaçta ek bir bölme yapmak için gereken minimum kayıp azaltma sayısıdır. Yüksek bir gama değeri modeli daha kapalı hale getirecektir.

- En küçük yaprak aralığı; düğüme ait örnekleme boyutu belirli bir eşiğin altına düştüğünde düğüm bölünme işlemine son verir. Bu işlem her verinin kendine ait bir yaprağı olmasını engeller. Bu parametre ne kadar büyükse, model o kadar kapalı halde olacaktır.

- Alt örnek; belirli bir ağacı eğitmek için kullanılan tüm örneklerin kesirini ifade eder. Bu parametrenin düşük olması aşırı öğrenmeyi azaltır.

- Ağaca göre sütun örneği; Her bir ağacı oluştururken sütunların alt örnek oranıdır. Ağaç başına kullanılan özelliklerin miktarını azaltmak, model varyansını azaltır.

- L2 düzenlemesi (λ); Ağırlıklar üzerinde L2 düzenleme terimi, değer ne kadar yüksekse model o kadar kapalı olacaktır.

Görev parametreleri, amaç fonksiyonunu ve değerlendirme ölçütünü belirtir. Hatanın karesi, kayıp fonksiyonu olarak kullanılır. Ortalama karekök hatası, aşırı veya eksik uyumu kontrol etmek için eğitim ve doğrulama setini karşılaştırmak için değerlendirme ölçütü olarak kullanılır.

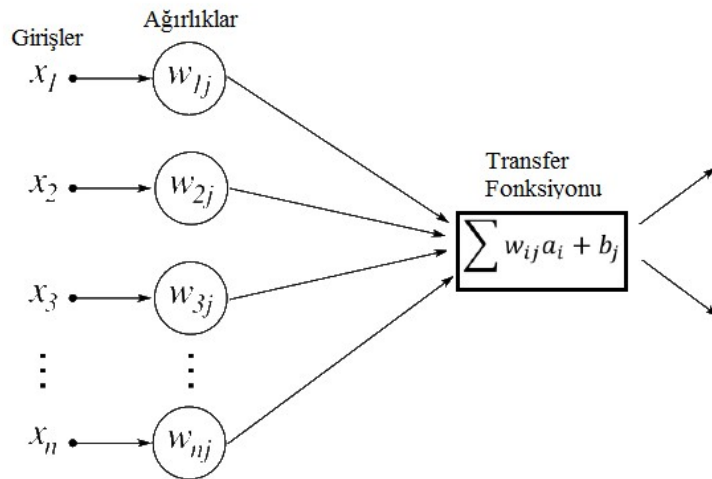
Parametreler rastgele arama yöntemi ile ayarlanmıştır. Her model rastgele parametre değerleriyle 100 kez çalıştırılır. En iyi yinelemenin parametreleri seçilir.

3.7 Derin Öğrenme (Deep Learning)

Derin öğrenme bir makine öğrenmesi alt alanıdır. Diğer makine öğrenimi algoritmalarından farkı birden fazla katmandan oluşan yapay sinir ağlarını kullanan bir yapıda olmasıdır.

Yapay sinir ağları insanın beyin yapısından esinlenilerek geliştirilmiş bir bilgisayar programıdır. Barındırdıkları hücreler ağırlıklı bağlantılar aracılığıyla birbirleriyle iletişim halindedir. Ayrıca her hücre kendi belleğine sahiptir. Kısaca yapay sinir ağları insanın biyolojik sinir ağını taklit edebilen ve bu sinir ağları sayesinde tahmin yürütebilen bilgisayar programlarıdır (Elmas, 2003).

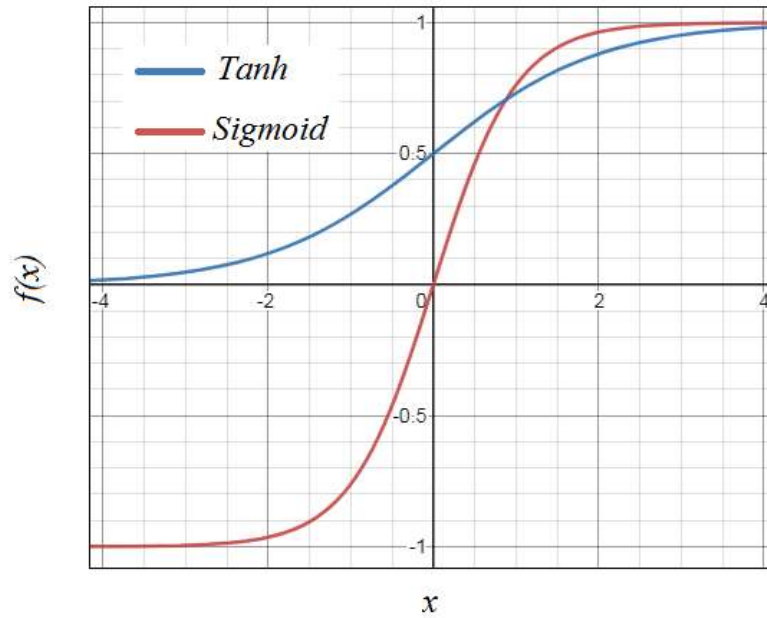
Örnek bir YSA nöron yapısı Şekil 3.7’de gösterilmiştir. Şekile bakıldığında YSA’nın nöron ve sinapslardan oluştuğu görülmektedir. Nöron kısmında hesaplama işlemi gerçekleştirilir. İşlem sonucu sinapslar vasıtasıyla iletilir. İletim diğer nöronlara veya sonuca doğru gerçekleştirilir.



Şekil 3.7. YSA nöron yapısı.

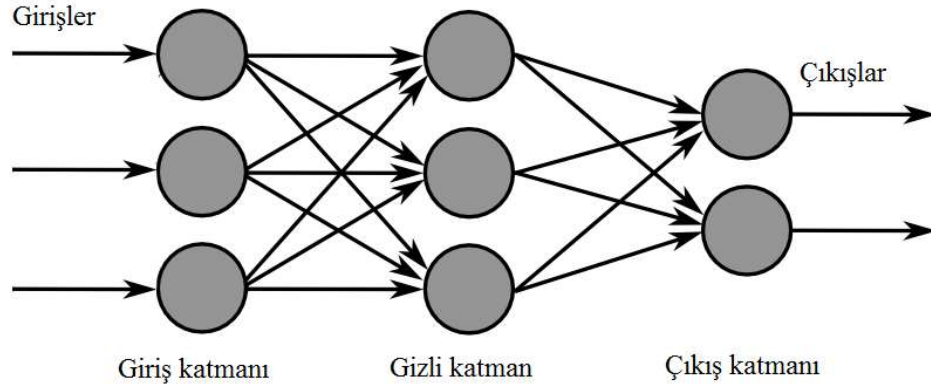
Yapılan işlem giriş x değerlerinin w ağırlıklarıyla çarpılması şeklindedir. Her giriş için belirli bir ağırlık çarpımı gerçekleştirilir ardından b (bias) değeri ile toplanır. Bias değeri, bilginin bir sonraki nörona iletilmesi için ulaşılması gereken bir eşik olarak düşünülebilir. Bütün girişler için bu işlem gerçekleştirilir. Ardından, işlem sonucu bir aktivasyon fonksiyonundan geçirilir ve bir sonraki nörona girdi olarak iletilir.

Aktivasyon fonksiyonu kısaca girdi verilerinin ağırlıklı toplamının bir düğümünden ya da ağ katmanındaki düğümlerden bir çıktıya nasıl dönüştürüleceğini tanımlar. Kısaca yapılan işlem sonucunu belirli bir değerde sınırlandırma ve işlem sonucunu çıktıya dönüştürülme olayıdır. Nöron içinde gerçekleştirilen işlem sonucu aktivasyon fonksiyonuna yerleştirilir. Ardından bir çıktı verir. Bu çıktı başka bir sinir hücrene ya da sonuca tahmin olarak verebilir. Tekrarlayan sinir ağlarında (RNN) tanh veya sigmoid fonksiyonu kullanılır. Şekil 3.8’de Sigmoid ve Tanh fonksiyonunun grafiği verilmiştir. LSTM için daha iyi bir performans gösterdiği için Tanh fonksiyonu sisteme aktivasyon fonksiyonu olarak verilmiştir.



Şekil 3.8. Sigmoid ve Tanh fonksiyonlarının grafiği.

Yapay sinir ağı üç tip katmandan oluşur. Bunlar giriş katmanı, gizli katman ve çıkış katmanıdır. Şekil 3.9’da bir yapay sinir ağının yapısı gösterilmiştir.



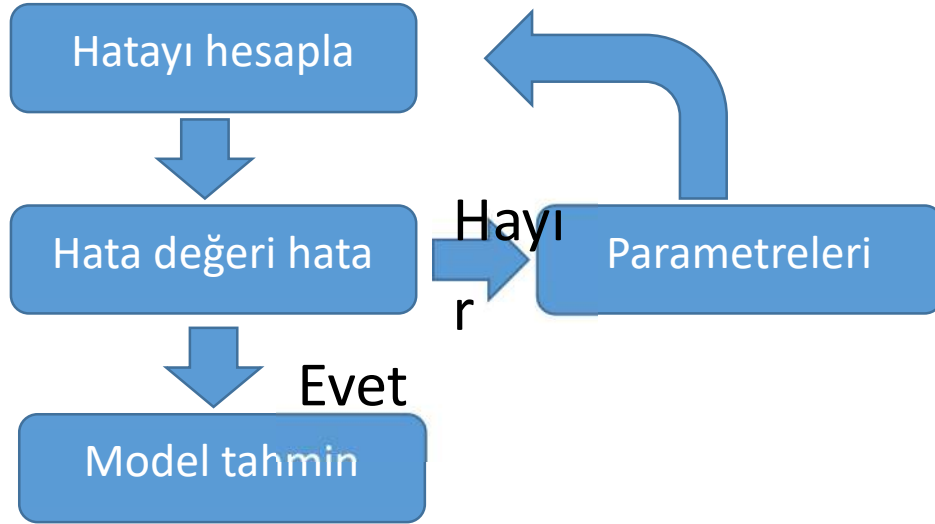
Şekil 3.9. Yapay sinir ağının yapısı.

Giriş katmanı, giriş değişkenleri için kullanılır, burada veri ağıya tanıtılır. Ardından veri gizli katmana ya da çıkış katmanına gönderilir.

Gizli katman basitçe, giriş ve çıkış katmanı arasındaki bir katmandır. Bir ağda birden çok gizli katman olabilir. Bazı modellerde gizli katman bulunmaz. Ağ giriş katmanı ve çıkış katmanı arasında da kurulabilir.

Çıkış katmanı ağ için bir çıktı (tahmin) üreten son nöron katmanıdır. Ağın sonunda bir tahmin değeri çıkacaktır. Bu tahmin değeri gerçek değerle karşılaştırılır ve ikisi arasındaki fark hesaplanır. Bu fark kayıp olarak adlandırılır modelin iyi bir performans gösterip göstermediği kayıptan anlaşılabilir. Bir modeli eğitmenin amacı, kaybı en aza indirmektir ve bu işlem geri yayılım algoritması (backpropagation) kullanılarak yapılır.

Geri yayılım algoritması gradyan iniş (gradyan decrease) ya da delta kuralı tekniğini kullanır. Bu teknikler ağırlıklarda değişimler yaparak hata değerinin minimize edilmesini sağlar. Hata değerindeki her düşüş modelin performansında artış sağlamaktadır. Ağırlık değiştirildiğinde hata çok az azalırsa ağırlık sabit bırakılır. Fakat ağırlık değişimiyle kayıp azalırsa ağırlık yeni değeriyle değiştirilir böylece hata oranında düşüş gerçekleştirilir. Bu işlem hata değeri minimize olana kadar tekrar edilir. Şekil 3.10’da algoritmanın çalışma şeması gösterilmiştir.



Şekil 3.10. Geri yayılım algoritmasının görselleştirilmesi.

Öncelikle model ağırlıklara rastgele değerler atar. Daha sonrasında eğitim seti verilerinden hata değeri elde edilir. Hata değeri elde edildikten sonra geri yayılım algoritması devreye girer. Hatayı önceki değeriyle kıyaslama yapar. Kıyaslanan değer öncekine göre daha yüksek bir hata değeriye sistem parametreleri güncellenir ve tekrardan hata değeri elde edilir. Kıyaslanan değer önceki değere göre daha düşük bir değerse model eğitimi tamamlanır ve model tahmin yapabilecek duruma gelir.

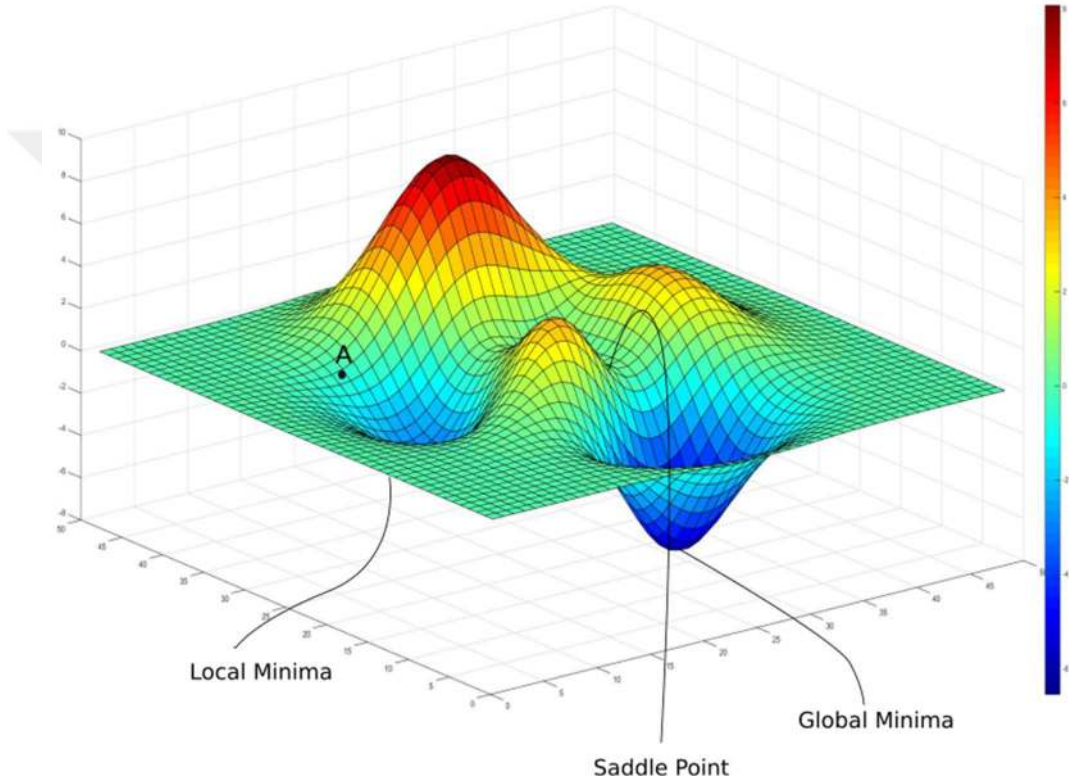
3.7.1 Hata Fonksiyonu ve Optimizasyon

Her Makine öğrenmesi modelinin optimize ettiği bir hata işlevi vardır. Regresyon modelleri genellikle ortalama karekök hatasını (RMSE) kullanır. Sınıflandırma modelleri, sırasıyla ikili ve çok sınıflı sınıflandırma için genellikle lojistik regresyon veya softmax kullanır. Eğitim aşaması sırasında kayıp fonksiyonunu optimize etmek istenir. Yani hatanın mümkün olduğunca küçük olması ve aşırı öğrenme yapmamış olması istenir. Verimli bir şekilde optimize etmek için mevcut birkaç öğrenme yöntemi vardır. Bu yöntemler dış bükey olmayan bir kayıp fonksiyonunda minimum hatayı bulmak için gereklidir.

Bu tezde optimizasyon fonksiyonu stokastik gradyan inişin uzantısı olan Adam fonksiyonu kullanılmıştır. Adam optimizasyon algoritmasının son zamanlarda görüntü işlemede, zaman serilerinde ve doğal dil işlemede derin

öğrenme uygulamalarında geniş bir kullanım alanı vardır. Adam optimizasyon foksionu stokastik gradyan düşüşün bir uzantısıdır.

Stokastik gradyan inişi hatayı minimize etmek için en uygun noktayı bulmayı amaçlar. Metot güçlü ve genel bir optimizasyon metodudur. Örneğin hata değerleri Şekil 3.11'deki grafikleştirildiği varsayılırsa, stokastik gradyan inişin amacı minimum noktayı bulmaktır.



Şekil 3.11. Örnekleştirilmiş hata değerleri grafiği(Dickson, 2020).

Stokastik gradyan inişin formülü eşitlik (3.17)'de gösterilmiştir. Burada η öğrenme oranıdır. Kısaca öğrenme hızı arasındaki adımın uzunluğunu belirten öğrenme oranıdır. Eğer öğrenme oranı çok büyük değer olursa minimum noktayı bulma ihtimali azalır. Öğrenme oranı çok küçük bir olduğunda çok tekrarlama oranı ve hesaplama sürecinde artışa sebep olur.

$$L_{i+1} = L_i - \eta \nabla f(L) \quad (3.17)$$

Şekil 3.11'e bakıldığında 2 farklı minimum noktası görülmektedir. Bunlar yerel (local) minimum ve global minimum noktasıdır. Global minimum noktası model için performansın en yüksek olduğu noktadır. Stokastik gradyan inişin dezavantajı yerel minimum değerinde takılı kalarak modelde daha iyi bir performansın önünü kesebilir. Ek olarak eyer noktalarında (saddle point) gradyan kaybolacaktır. Diğer bir husus algoritmanın her yineleme için eğitim setinin tamamının kullanılmasını gerektirmesidir.

Yerel minimum noktaları ve eyer noktaları sorununun çözümü sisteme momentum tanıtılarak çözülebilir. Bu işlemi geçmiş değerlerin üstel ağırlıklı ortalaması alınarak yapılır. Böylece gürültüyü veya yerel dalgalanmaları azaltır. Eğer algoritma bir eyer noktasına ulaşırsa model momentumu nedeniyle yine de işlemine devam eder. Benzer olay yerel minimum noktasına geldiğinde de geçerlidir. İkinci dezavantajı çözümlerin yolu veri setini küçük veri setlerine ayırmaktır. Momentum ekleme ve veri setini bölme işlemlerine ek olarak Adam'ın statik değil dinamik bir öğrenme hızı özelliği vardır.

Eğim büyük olduğunda öğrenme hızı daha hızlı, eğim küçük olduğunda daha yavaştır. Bu işlem salınımı azaltarak kayıp fonksiyonunun daha hızlı yakınsamasına ve patlayan gradyanların azaltılmasına (exploding gradient decrease) neden olur. Patlayan gradyanların aktivasyon fonksiyonların sonuçlarının geri yayılım algoritmasında kullanılırken stabil sonuçlar üretmemesi olarak tanımlanabilir.

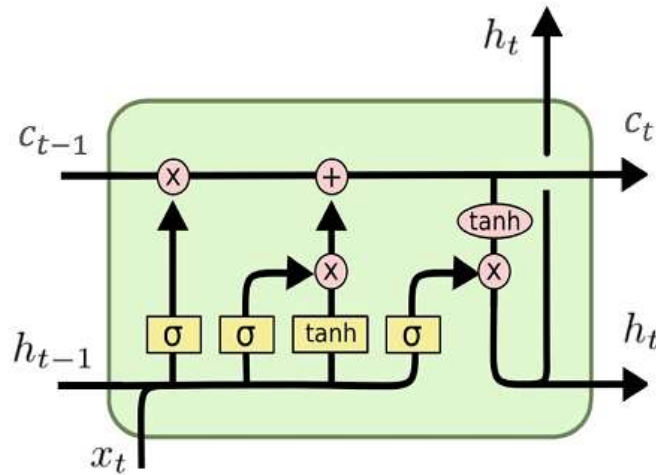
3.7.2 Uzun Kısa Süreli Bellek (LSTM)

RNN zaman bağımlılıklarıyla başa çıkabilen özel bir yapay sinir ağıdır. Hidrolojik değişkenlerin çoğu önceki koşullara bağlı olduğu için zaman bağımlılıklarıyla başa çıkma yeteneği hidrolojik uygulamalar için oldukça önemli bir etmendir. Ancak RNN'lerinde oldukça önemli bir dezavantajı vardır. Kaybolan gradyan probleminden (vanishing gradient problem) etkilenebilir. Kaybolan gradyan probleminin temel sebebi eğitim aşamasında kullanılan geri yayılım algoritmasından kaynaklanır. Geri yayılım algoritmasında birbiriyle bağlı nöronların türevleri alınır sonrasında çarpım işlemi gerçekleştirilir. Kullanılan aktivasyon fonksiyonuna bağlı olarak üretilen sonuçlar -1 ile 1 aralığında

olduğundan türev çarpımları sıfıra doğru yaklaşmaktadır. Bu da eğitimi engellemektedir.

Bu sorunla ilgilenen bir RNN türü LSTM modeli vardır. Normal bir nöron yerine LSTM nöronları hafıza hücresine, giriş kapısına, çıkış kapısına ve bir unutma kapısına sahiptir. Bu kapılar hücrenin içine ve dışına bilgi akışını düzenlemektedir. LSTM nöronunun yapısı, nörona giren ve nörondan çıkan bilgi akışını düzenleyerek kaybolan gradyan problemini çözmektedir. Bu sebepten dolayı bu tezde LSTM modeli kullanılmıştır.

Şekil 3.12’de bir LSTM nöronunun yapısı gösterilmiştir. Şekil kullanılarak bir LSTM nöronu’nun çalışma biçimi açıklanacaktır.



Şekil 3.12. LSTM nöronunun yapısı (Olah,2015).

Bir LSTM nöronunun en önemli noktası C_{t-1} ile C_t arasında bulunan yatay çizgidir. Bu yatay çizgi geçmiş bilgileri hatırlayan bellek durumudur ve C simgesi ile gösterilmiştir. Mevcut diğer kapılar kullanılarak bellek durumu değiştirilebilir. İlk adım hafıza durumuna ait hangi bilgilerin unutulabileceğine karar vermektir. Bu işlemi unutma kapısı karar verir. Bu kapı girdi olarak önceki gizli durum vektörü H_{t-1} ve giriş vektörü olan X_t 'yi kullanmaktadır. Unutma kapısının denklemleri eşitlik (3.18)'de verilmiştir. Sigmoid işlevi (σ) hangi bilgilerin saklandığını veya unutulduğunu filtrelediği için seçilmiştir. Herhangi bir girdi değerini sıfır ile bir arasında bir sayıya dönüştürür. Çıkan değer sıfıra yakınsa önceki bellek bilgisini

tutan C_{t-1} unutulacak ve yeni değeriyle değiştirilecektir. Bire yakınsa bellek önceki değerle sabit tutulacaktır.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3.18)$$

Bir sonraki adım, bellek durumunda hangi yeni bilgilerin saklanması gerektiğine karar verme işlemidir. Bu işlem iki basamakta yapılır. Birinci işlem kullanılan başka bir sigmoid katmanı hangi değerlerin güncelleneceğine karar verir eşitlik (3.19). İkinci işlem hiperbolik tanjant işlevi belleğe eklenecek yeni aday değerleri belirler eşitlik (3.20).

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3.19)$$

$$d_t = \tanh(W_d x_t + U_d h_{t-1} + b_d) \quad (3.20)$$

Sonraki aşama hafıza değerlerini önceki sonuçlarla güncellemektir. Yeni değerler eklenmeden önce eski değerler unutulur. Eklenecek yeni değer girdinin sigmoidi ile bir kez daha ağırlıklandırılır. Hafızanın güncelleme işlemi eşitlik (3.21)'de verilmiştir. Burada $\bullet$ işareti element bazında çarpımını (Hadamard çarpımı) belirtir.

$$C_t = f_t \bullet C_{t-1} + i_t \bullet d_t \quad (3.21)$$

Bellek güncellendikten sonra yeni gizli durum h_t 'ye eşit olan bir tahmin yapılır. Bu, yeni belleğin hiperbolik tanjantının gizli durumun sigmoidi ve girdi değişkenleri ile ağırlıklandırılmasıyla yapılır. Eşitlik (3.22)'de işlem gösterilmiştir.

$$h_t = \tanh(C_t) \bullet \sigma(W_h x_t + U_h h_{t-1} + b_h) \quad (3.22)$$

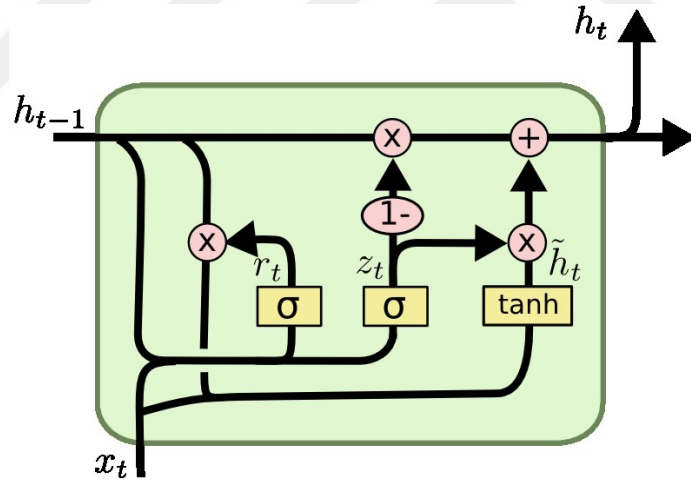
Bellek durumu giriş dizisindeki öğeler arasındaki karşılıklı bağımlılığın kaydını tutar. Giriş kapısı hücreye ne ölçüde yeni bir değer alınacağını düzenler. Unutma kapısı, bir değer hücrede ne kadar süre kalacağını kontrol eder. Çıkış

kapısı, hücrede ne kadar kalacağını belirler. Hafızadaki değer, LSTM nöronu'nun aktivasyon fonksiyonunda kullanılır.

LSTM'nin en önemli avantajı kar etkisi olan su toplama havzalarında depolama etkilerini modellemek için gerekli olan, hidrolojik sisteme sağlanan girdi ve çıktı arasındaki uzun vadeli bağımlılıkları öğrenme yeteneğidir (Kratzert, 2018).

3.7.3 Geçitli Tekrarlayan Ünite (GRU)

GRU kavramı, LSTM yapısına benzer bir yapıdadır. Hesaplamalar verilerdeki zamansal kalıpları hesaba katarak sıralı bir biçimde gerçekleştirilir. Sistemin ana amacı LSTM'i basitleştirmektir. GRU modeli LSTM'in iki durum değişkenini (gizli ve hücre durumu) tek bir gizli durum değişkeni h (RNN'ye benzer) olarak birleştirir. Bu kapıların yerine sıfırlama ve güncelleme olmak üzere iki hesaplamalı kapı kullanılır.



Şekil 3.13. GRU nöronunun yapısı (Anderson,2019).

Şekil 3.13'de GRU nöronunun yapısı gösterilmiştir. h_{t-1} terimi t-1 zaman aralığının gizli durumun temsil eder. x_t ve h_t ifadeleri sırasıyla mevcut t zaman aralığında bulunan giriş verilerini ve çıkış gizli durumunu temsil etmektedir.

Daha önce belirtildiği gibi kapılar, hangi bilgilerin unutulacağını veya tutulacağını tanımlamak için kullanılan farklı sinir ağılarıdır. Şekil 3.13'de gösterildiği gibi, her iki kapı da çıkışını 0 ile 1 arasındaki değerlerde

normalleştirdiği için bu işlemi basitleştirmek için uygulanan bir sigmoidal aktivasyon fonksiyonu çalıştırılır. Böylece 0 ile çarpılan herhangi bir değer unutulacak, 1 ile çarpılan değerler korunacaktır.

h_t değerini hesaplamak için hücre, h_{t-1} ve x_t 'yi birleştirir. Sonrasında elde edilen vektör, unutma ve güncelleme kapılarına gönderilir. Unutma kapısı denklemi eşitlik (3.23) 'de verilmiştir. Güncelleme kapısı denklemi eşitlik (3.24)'de verilmiştir.

$$r_t = \sigma(W_r \bullet [h_{t-1}, x_t] + b_r) \quad (3.23)$$

$$u_t = \sigma(W_u \bullet [h_{t-1}, x_t] + b_u) \quad (3.24)$$

Denklemlerde verilen W_r ve W_{ru} , sinir ağlarının ağırlık matrislerini temsil eder. b_r ve b_u terimleri sinir ağlarının bias (sapma) vektörüdür. Ardından, r_t ve h_{t-1} arasında noktasal bir çarpma yapılır ve sonuç x_t ile birleştirilir. Sonrasında Hiperbolik Tanjant (tanh) aktivasyon fonksiyonu ile üçüncü bir sinir ağına gönderilir. Tanh kullanımı, verileri -1 ile 1 arasındaki normalleştirir, sinir ağının çıktısını düzenler ve yinelemeler arasında verilerin fazla veya küçük olmasını engeller. Bu sinir ağının çıkışı eşitlik (3.25) ile hesaplanır.

$$\hat{h}_t = \tanh(W_o \bullet [r_t * h_{t-1}, x_t] + b_o) \quad (3.25)$$

Eşitlik (3.25)'te verilen W_o ve b_o , sırasıyla sinir ağının ağırlık matrisi ve yanlılık vektörünü temsil eder. Güncelleme geçidinin sonucu u_t iki durum için kullanılır. Birincisi güncelleme sonucunu $\hat{h}_t$ ile çarparak yeni bilginin hangi bölümünün ekleneceğine karar verir. İkincisinde, h_{t-1} 'i $1 - u_t$ ile noktasal olarak çarparak hangi verilerin atılacağını belirler. Son olarak, bu iki çıkış arasında noktasal bir ekleme yapılır ve GRU hücresinin sonucu olan gizli durum h_t üretilir. Eşitlik (3.26) bu durumu açıklar.

$$h_t = (1 - u_y) * h_{t-1} + u_t * \hat{h}_t \quad (3.26)$$

3.7.4 Verilerin Normalizasyonu

YSA için kullanılan verilerin değişken aralıkları birbirinden farklılık göstermektedir. Örneğin bağıl nem sıfır ile bir değer aralığındayken en yüksek sıcaklık değerleri ise -20 C° ile +35C° değer aralığındadır. Bazı makine öğrenmesi algoritmaları için geniş bir aralığa sahip olan değişkenler sonuçları büyük ölçüde etkilemektedir. Örneğin max sıcaklıktaki büyük değişken aralığı bağıl nemdeki küçük değişken aralığına hâkim olacaktır. Bir değişkenin makine öğrenmesi algoritması üzerindeki etkisi değişken aralığına bağlı olmaması arzu edilir. Bu istenmeyen eğilimi çözmek için veriler genellikle kendilerini temsil edecek değerlere dönüştürülür. Bunu yapmanın iki yolu vardır; normalleştirme ve standardizasyon. Standardizasyon işleminde dönüştürülecek veriler normal dağılım yapılmak istenildiğinde kullanılır. Bu çalışmadaki değişkenler için durum böyle değildir. Bu nedenle veriler normalleştirilir. Normalizasyon işleminde eşitlik (3.27) kullanılarak kullanılan tüm değişkenler normalizasyon işleminden geçirilir. İşlem sonucunda her değişken 0 ile 1 arasında değişken aralığına sahip olur. Yapılan bu işleme minimum maksimum normalizasyon işlemi olarak isimlendirilir.

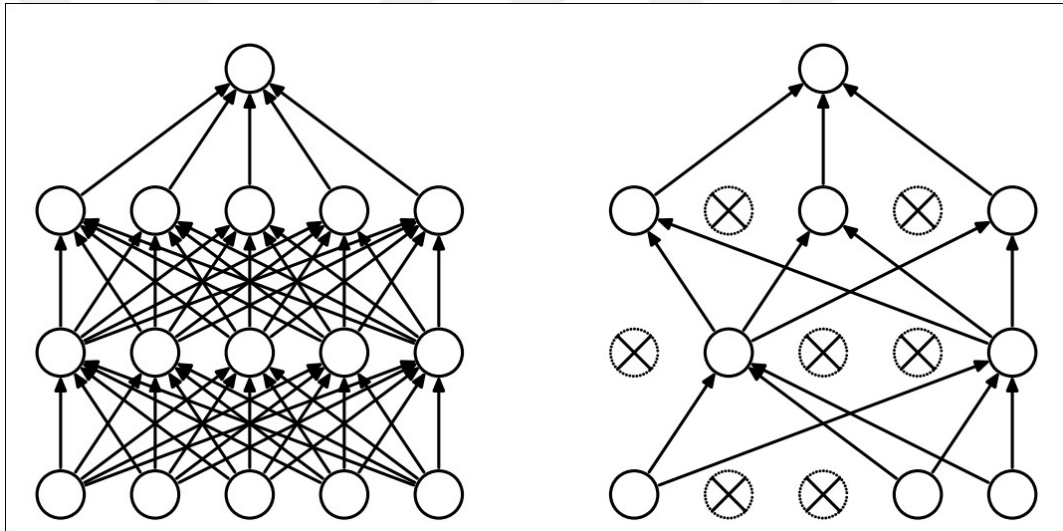
$$X_{scaled} = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (3.27)$$

Verileri normalleştirmek aynı zamanda veriler üzerinde bazı bilgi kayıplarına yol açmaktadır. Yapılacak tahmin eğitim kümesindeki verilerin değer aralığı ile sınırlıdır. Bu nedenle normalleştirilmiş veriler kullanılarak ekstrapolasyon mümkün değildir.

Hedef değişkenlerini (debi verileri) normalleştirmek zorunlu değildir. Hedef değişkenler normalleştirilmezse daha büyük hedeflere öncelik verilecektir.

3.7.5 Dropout Tekniği

Dropout bir regülasyon tekniğidir. Dropout eğitim aşamasında sisteme yapay bir gürültü ekler. Amaç aşırı öğrenmenin önüne geçmektir. İşlem olarak dropout her eğitim tekrarında rastgele bazı nöronları unuttur. Bu işlem giriş, gizli ve çıkış katmanında olabilir. Düğümler arasındaki bazı nöronlar kaybolduğundan ağırlıklar mevcut tüm bağlantılardan daha büyük olacaktır. Bunu geçersiz kılmak için ağırlıklar ilgili her katmanın aynı bırakma olasılığı kullanılarak ölçeklendirilir. Bir katmanın % 20 unutma olasılığı olduğu varsayılırsa o katmanın ağırlıkları 1.20'ye bölünerek ölçeklenir. Bu işlem eğitim ve tahmin arasında gerçekleşir. Eksikler aşırı öğrenmeyi azaltırken, eğitim süresini artırır. Bunun ana nedeni parametre güncellemelerinin gürültülü olmasıdır (Srivastana vd. ,2014). Dropout regülasyonunun nasıl çalıştığına dair görselleştirme Şekil 3.14'te verilmiştir.



Şekil 3.14. Örnek Dropout regülasyonunun çalışma şekli (Srivastana vd. ,2014).

3.7.6 Derin Öğrenme Parametreleri

Bir LSTM ve GRU modelinde manuel olarak ayarlanabilen birçok parametre mevcuttur. Bu parametreler optimizasyon, hesaplama ve performansa yönelik parametrelerdir.

İlk parametre aktivasyon fonksiyonudur. Aktivasyon fonksiyonu giriş verisi işlendikten sonra sonucu çıkış kapısına ileten fonksiyondur. Derin öğrenme için Relu, Tanh ve Sigmoid fonksiyonları gibi aktivasyon fonksiyonları bulunmaktadır. Bu çalışmada aktivasyon fonksiyonu olarak Tanh fonksiyonu kullanılmıştır.

İkinci parametre öğrenme oranının belirlenmesidir. Öğrenme oranı ağırlıkların hangi sırayla güncelleneceğine karar verir. Öğrenme oranının düşük olduğu durumda model performansı artacaktır fakat ağırlıkların tam olarak belirlenmesi içinde daha fazla eğitim tekrarı gerektiğinden hesaplama zamanında bir artış görülecektir.

Diğer bir parametre dropout oranının belirlenmesidir. Kısaca unutulacak nöronların yüzde oranıdır. Parametre bu çalışmada 0,2 oranında belirlenmiştir.

LSTM ve GRU hiperparametreleri toplamda 20 farklı parametreyi içermektedir bunlar:

- Batch size; eğitim ve test sırasında her partide kaç örnek olacağını belirleyen parametredir.
- Units; derin öğrenme iç hücrelerin boyutu.
- Epoch; makine öğrenimi algoritmasının tamamladığı tüm eğitim veri kümesinin geçiş sayısını belirtir.
- Steps per epoch; bir dönem için seçilecek parti sayısını belirtir. Örneğinn 500 adım seçilirse, ağ bir dönemi tamamlamak için 500 grup için eğitim verecektir.
- Loss; modeliniz en aza indirdiği kayıp fonksiyonudur.
- Timesteps; bir dizide kaç değer olduğu anlamına gelmektedir.

3.7.7 Modelin Performansının Değerlendirilmesi

Modellerin performansını değerlendirmek için iki farklı ölçüm tekniği kullanılmıştır. Nash-Sutcliffe katsayısı (NSE) ve R^2 dir.

NSE hidrolojik modellerin performansını değerlendirmek için yaygın olarak kullanılan bir ölçüm tekniğidir (Krause vd. , 2005). Formülü eşitlik (3.28)'de verilmiştir.

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_m^t - Q_o^t)^2}{\sum_{t=1}^T (Q_o^t - \bar{Q}_o)^2} \quad (3.28)$$

$\bar{Q}_0$ terimi gözlenen tüm debi değerlerinin ortalamasıdır. Q_m^t terimi zamana bağlı modellenen debi değerlerini ifade etmektedir. Modelin belirli bir sürede yaptığı tahmin değerlerini temsil etmektedir. Q_0^t terimi ise zamana bağlı gözlemlenen değerleri ifade etmektedir. Gerçek zamanda gözlemlenen değerleri ifade etmektedir. Ayrıca, 1'den çıkarma işlemi gözlemlenen değerlerin ortalama kare hatasının oranı olarak da yorumlanabilir (Nash ve Sutcliffe, 1970).

Nash-Sutcliffe verimliliği $-\infty$ ile 1 değerleri arasında değişir. Tahmin hata değeri sıfıra eşit olduğu zaman model mükemmel tahmin yapabilmektedir. Bu durumda elde edilen Nash-Sutcliffe verimliliği 1'e eşittir. Diğer bir durum NSE değerinin 0 olduğu anda zaman serisinin ortalaması ile aynı tahmin yeteneğine sahip olduğunu gösterir. Sıfırdan küçük bir verimlilik gözlemlenen ortalama değer modelden daha iyi bir tahmin edici olduğunda ortaya çıkar. Tablo 3'te NSE değer aralıklarının açıklamaları verilmiştir (Moriasi vd., 2007):

Tablo 3.2. NSE değer aralıkları ve performans açıklamaları (Moriasi vd., 2007).

Performans	NSE değeri
Çok iyi	$0,75 < NSE \leq 1,00$
İyi	$0,65 < NSE \leq 0,75$
Orta	$0,50 < NSE \leq 0,65$
Kötü	$NSE \leq 0,50$

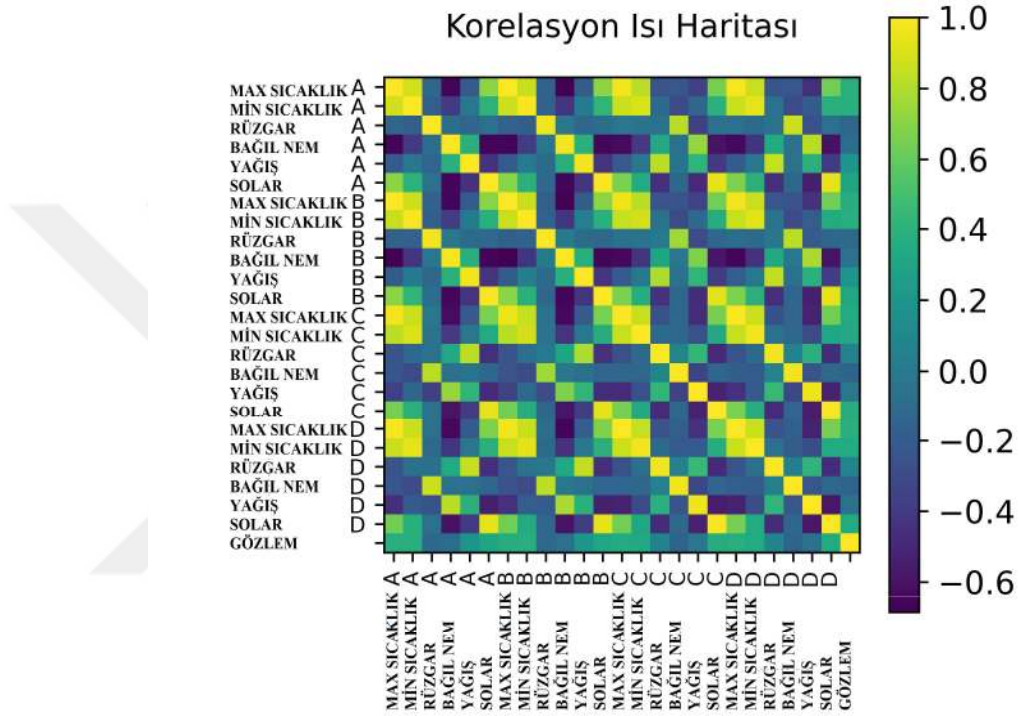
R^2 , bağımlı değişkendeki bağımsız değişkenden tahmin edilebilen varyasyonun oranıdır ve Eşitlik (3.29)'de denklemi verilmiştir. Değer olarak 0 ile 1 aralığında bulunmaktadır. Hesaplama sonucu 0 değeri ise model performansı çok kötüdür, 1 değeri ise mükemmel bir model performansı olduğunu göstermektedir.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - u)^2} \quad (3.29)$$

3.8 Modeli Eğitecek Verilerin Belirlenmesi

Modelleme işlemi başlatılmadan önce hangi girdi değişkenlerinin uygun olduğuna ve en iyi model performansına yol açacağına karar verilmelidir. Bu

çalışmada bu işlem iki ayrı aşamada gerçekleştirilmiştir. İlk aşama korelasyon ısı haritası çıkarımıdır. Bu yöntemde değişkenler arasında herhangi bir istatistiksel ilişki olup olmadığına bakılmaktadır. Şekil 3.15'te değişkenler arasında olan istatistiksel ilişkiler renk dağılımlarına göre haritalandırılmıştır. Görselin sağ tarafında gösterildiği gibi mavi renkte olan değerler birbirinden istatistiksel olarak bağımsız olan değerlerdir. Mavi renkten sarı renge giden değerler orantılı olarak birbiriyle istatistiksel bir ilişkiye sahip olan değerlerdir.



Şekil 3.15. Korelasyon ısı haritası.

Korelasyon ısı haritasına bakıldığında en aykırı değerlerin rüzgar değerleri olduğu ortaya çıkarılmıştır. Fakat değişkenler sadece korelasyon ısı haritasına bakılarak belirlenmemiştir. Bu belirlemede ikinci aşamada bir o kadar etkili olmuştur.

Diğer bir değişken belirleme yöntemi olarak temel değişkenler belirlenir ardından diğer değişkenleri tek tek test ederek bu temel değişkenlere ek en kötü performansı gösteren değişken sistemden atılır.

Yöntemler kullanılarak model performanslarına bakıldığında istasyonlardaki rüzgâr değişkenlerinin performansa olumsuz etkisi olduğu

gözelemlenmiştir. Bu sebepten ötürü model eğitiminde rüzgâr değışkenleri kullanılmamıştır. Geriye kalan değışkenler maksimum sıcaklık, minimum sıcaklık, bağıl nem, yağış ve solar radyasyon değeri giriş değışkenleri olarak kullanılmıştır.

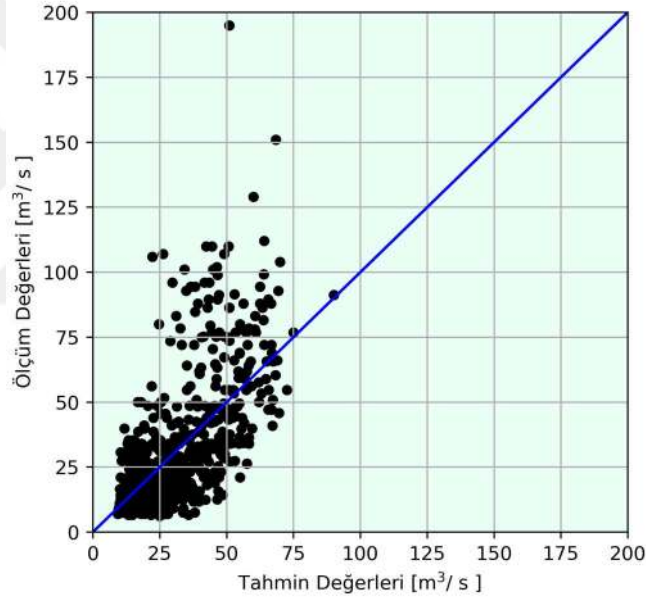


4. BULGULAR

Bu bölümde, CFSR sisteminden elde edilen verilerin MÖ tekniklerinin performansları karşılaştırılmıştır.

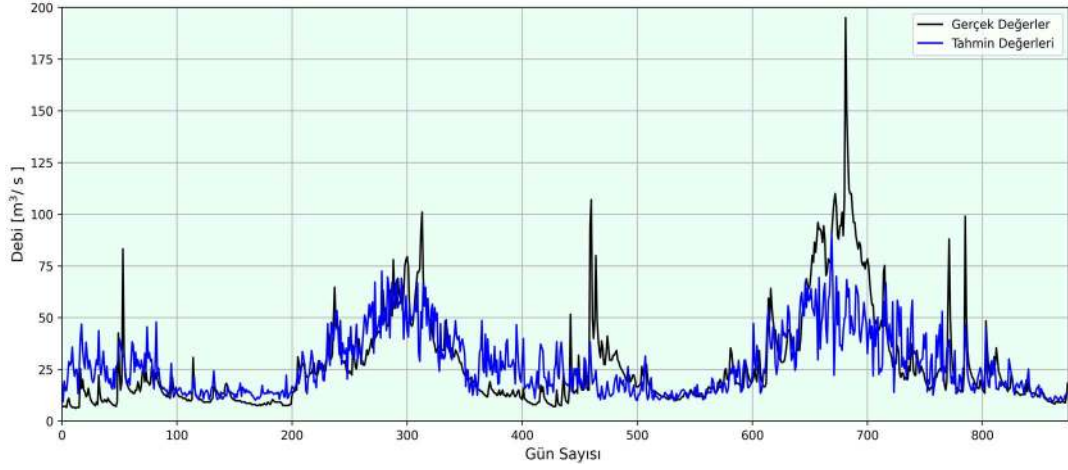
4.1 XGBoost Modeli ile Fırtına Deresi Debi Tahmin Bulguları

Fırtına Deresi günlük akım değerleri bağımlı değişken olarak LSTM eğitim yapısına verildi. CFSR veri noktasından alınan 21 farklı değişken XGBoost modeline bağımsız değişken değeri olarak verildi. Debi ölçüm istasyonundan alınan değişken, bağımlı değişken olarak sisteme tanıtıldı ve sonrasında model eğitimi yapıldı. 875 günlük verinin XGBoost yöntemi vasıtasıyla tahminleri gerçekleştirildi. Model sonucunda çıkan tahminler normalize olmuş verilerdir. Normalize veriler tekrardan normal verilere dönüştürülmüştür. Sonrasında hata değerlerini görmek için NSE ve R^2 değerlerinin hesaplaması yapıldı. Ardından tahmin değerleri ve ölçüm değerlerinin dağılım grafiği oluşturuldu (Şekil 4.1). Gerçek değerler ve tahmin değerleri grafikleştirildi (Şekil 4.2).



Şekil 4.1. XGBoost dağılım grafiği.

Dağılım grafiğine bakıldığı zaman modelin aşırı aykırı değerleri tahmin etmekte güçlük çektiği gözlemlenmiştir. Bunun nedeni ani debi değişimlerinin sistemde aşırı derecede bulunmamasıdır. Model bu değişimleri öğrenememiştir.



Şekil 4.2. Gerçek ve XGBoost tahmin değerlerinin grafiği.

Karşılaştırma grafiğine bakıldığında modelin yükselen ve düşen trendleri gayet başarılı bir şekilde öğrendiği görülmektedir. Fakat ani değişiklikleri yakalayamadığıda bu grafikte net bir şekilde görülmektedir.

Modelin hata değerleri, NSE değeri 0,51, R^2 değeri 0,51 olarak bulunmuştur. Modelin yaptığı tahmin ve gerçek değerlerin son 30 günlük sonuçları Tablo 4.1’ de verilmiştir.

Tablo 4.1. 30 günlük XGBoost tahmin ve gerçek değerleri.

Gün Sayısı	Gerçek Değerler	Tahmin Değerleri
1	13,50	11,59
2	13,50	12,11
3	13,50	13,29
4	12,90	15,79
5	11,80	15,96
6	11,80	17,38
7	13,50	11,57
8	13,50	15,67
9	13,50	14,20
10	13,50	11,77

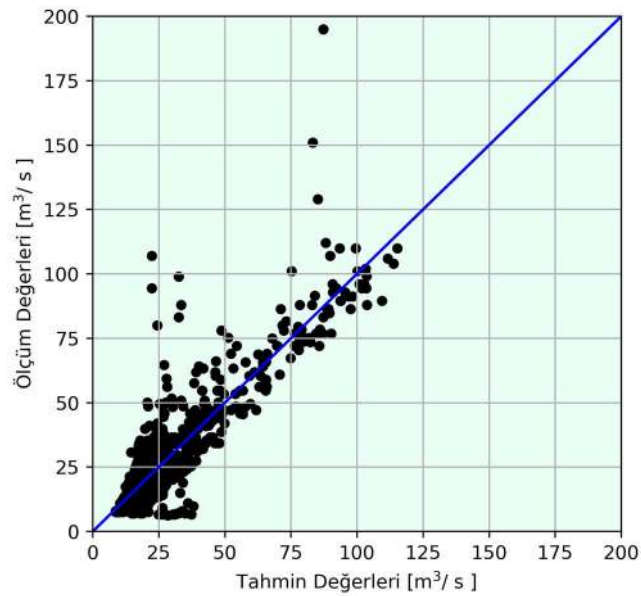
Tablo 4.1 (devam). 30 günlük XGBoost tahmin ve gerçek değerleri.

11	11,80	10,98
12	11,20	10,56
13	10,70	10,15
14	9,51	10,45
15	9,51	10,89
16	9,51	11,26
17	8,94	10,00
18	8,94	9,59

19	8,37	9,89
20	8,37	12,08
21	9,51	11,66
22	9,51	11,05
23	8,94	10,10
24	9,51	10,92
25	8,94	11,88
26	9,51	10,71
27	9,51	9,84
28	8,94	10,90
29	8,94	12,65
30	11,80	12,90

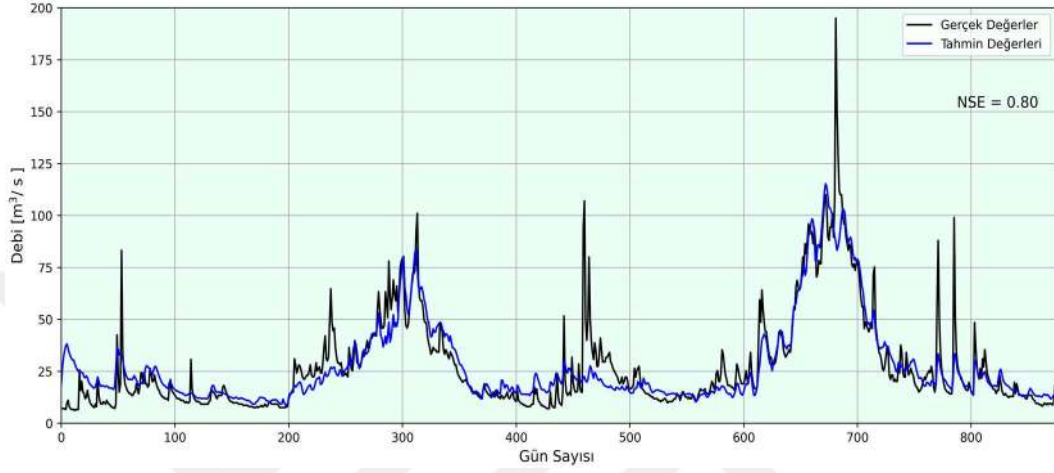
4.2 LSTM Modeli ile Fırtına Deresi Debi Tahmin Bulguları

Fırtına Deresi günlük akım değerleri bağımlı değişken olarak LSTM eğitim yapısına verildi. CFSR veri noktasından alınan 21 farklı değişken LSTM modeline bağımsız değişken değeri olarak verildi. Debi ölçüm istasyonundan alınan değişken, bağımlı değişken olarak sisteme tanıtıldı. Sonrasında model eğitimi yapıldı. 875 günlük verinin LSTM yöntemi vasıtasıyla tahminleri gerçekleştirildi. Model sonucunda çıkan tahminler normalize olmuş verilerdir. Normalize veriler tekrardan normal verileri dönüştürüldü. Sonrasında hata değerlerini görmek için NSE ve R^2 değerlerinin hesaplaması yapıldı. Ardından tahmin değerleri ve ölçüm değerlerinin dağılım grafiği oluşturuldu (Şekil 4.3). Gerçek ve tahmin değerleri grafikleştirildi (Şekil 4.4).



Şekil 4.3. LSTM dağılım grafiği.

Dağılım grafiğine bakıldığı zaman modelin aşırı aykırı değerleri tahmin etmekte güçlük çektiği gözlemlenmiştir. Bunun nedeni ani debi değişimlerinin sistemde aşırı derecede bulunmamasıdır. Model bu değişimleri öğrenememiştir.



Şekil 4.4. Gerçek ve LSTM tahmin değerlerinin grafiği.

Karşılaştırma grafiğine bakıldığında modelin yükselen ve düşen trendleri gayet başarılı bir şekilde öğrendiği görülmektedir. Fakat ani değişimleri yakalayamadığıda bu grafikte net bir şekilde görülmektedir.

Modelin hata değerleri, NSE değeri 0,80, R^2 değeri 0,80 olarak bulunmuştur. Modelin yaptığı tahmin ve gerçek değerlerin son 30 günlük sonuçları Tablo 4.2' de verilmiştir.

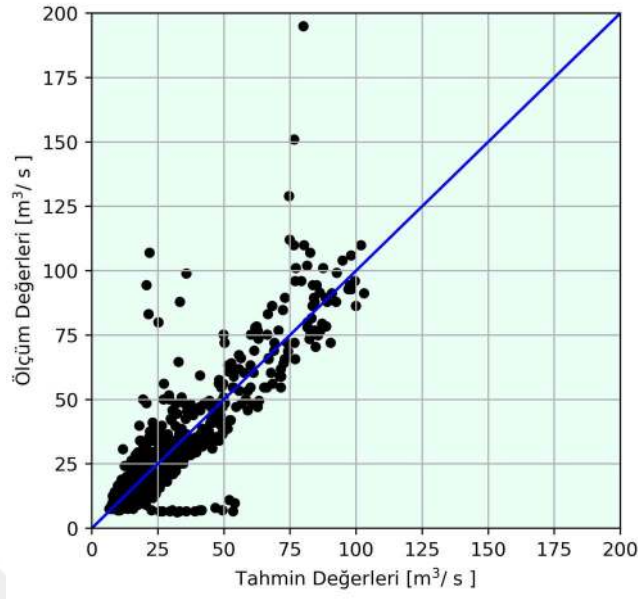
Tablo 4.2. 30 günlük LSTM tahmin ve gerçek değerleri.

Gün Sayısı	Gerçek Değerler	Tahmin Değerleri
1	13,50	13,66
2	13,50	13,39
3	13,50	12,94
4	12,90	13,09
5	11,80	13,36
6	11,80	13,34
7	13,50	13,26
8	13,50	14,05

9	13,50	15,71
10	13,50	16,80
11	11,80	16,00
12	11,20	14,90
13	10,70	14,09
14	9,51	13,48
15	9,51	13,17
16	9,51	12,93
17	8,94	13,10
18	8,94	13,14
19	8,37	12,82
20	8,37	12,63
21	9,51	13,20
22	9,51	13,98
23	8,94	13,81
24	9,51	13,39
25	8,94	13,80
26	9,51	13,31
27	9,51	12,41
28	8,94	11,75
29	8,94	12,08
30	11,80	12,91

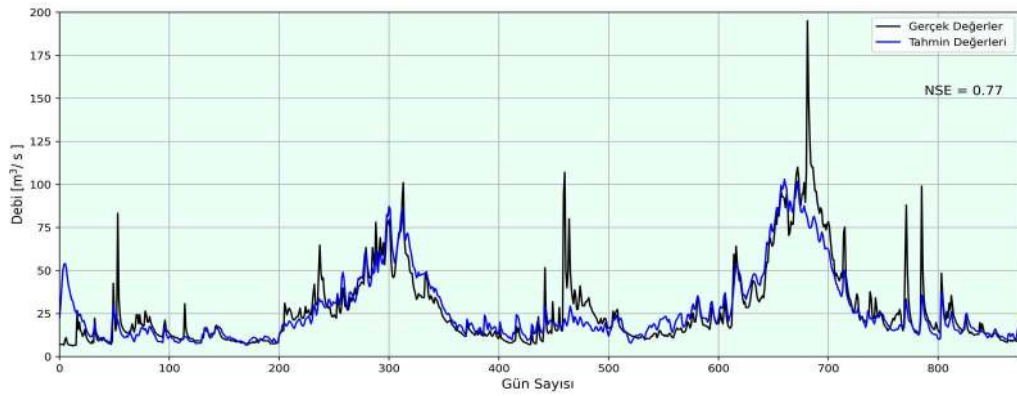
4.3 GRU Modeli ile Fırtına Deresi Debi Tahmin Bulguları

Fırtına Deresi günlük akım değerleri bağımlı değişken olarak GRU eğitim model yapısına verildi. CFSR veri noktasından alınan 21 farklı değişken LSTM modeline bağımsız değişken değeri olarak verildi. Debi ölçüm istasyonundan alınan değişken, bağımlı değişken olarak sisteme tanıtıldı. Sonrasında model eğitimi yapıldı. 875 günlük verinin GRU yöntemi vasıtasıyla tahminleri gerçekleştirildi. Model sonucunda çıkan tahminler normalize olmuş verilerdir. Normalize veriler tekrardan normal verileri dönüştürüldü. Sonrasında hata değerlerini görmek için NSE ve MSE değerlerinin hesaplaması yapıldı. Ardından tahmin değerleri ve ölçüm değerlerinin dağılım grafiği oluşturuldu (Şekil 4.5.). Sonrasında gerçek değerler ve tahmin değerlerini grafikleştirildi (Şekil 4.6).



Şekil 4.5. GRU dağılım grafiği.

Dağılım grafiğine bakıldığı zaman modelin aşırı aykırı değerleri ve debi değerlerinin 25 ile 50 m³/s aralığını tahmin etmekte güçlük çektiği gözlemlenmiştir. Bunun nedeni ani debi değişimlerinin sistemde aşırı derecede bulunmamasıdır. Model bu değişimleri öğrenememiştir.



Şekil 4.6. Gerçek ve GRU tahmin değerlerinin grafiği.

Karşılaştırma grafiğine bakıldığında modelin yükselen ve düşen trendleri gayet başarılı bir şekilde öğrendiği görülmektedir. Fakat ani değişiklikleri yakalayamadığıda bu grafikte net bir şekilde görülmektedir.

Modelin hata deęerleri, NSE deęeri 0,77, R^2 deęeri 0,77 olarak bulunmuştur. Modelin yaptığı tahmin ve geręek deęerlerin son 30 gnlk sonuęları tablo 6’ de verilmiştir.

Tablo 4.3. 30 gnlk GRU tahmin ve geręek deęerleri.

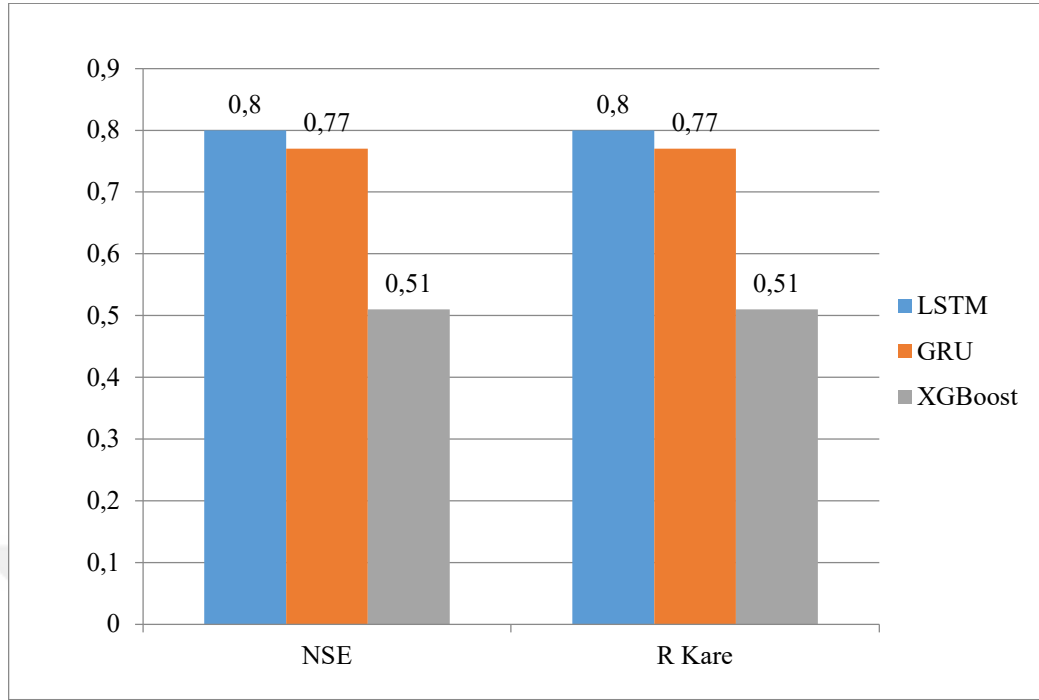
Gn Sayısı	Geręek Deęerler	Tahmin Deęerleri
1	13,50	15,15
2	13,50	14,36
3	13,50	12,98
4	12,90	11,87
5	11,80	10,38
6	11,80	9,12
7	13,50	8,65
8	13,50	8,71
9	13,50	8,67
10	13,50	8,57
11	11,80	8,18
12	11,20	14,37
13	10,70	16,85
14	9,51	14,04
15	9,51	11,48
16	9,51	10,48
17	8,94	11,53
18	8,94	11,56

Tablo 4.3 (devam). 30 gnlk GRU tahmin ve geręek deęerleri.

19	8,37	10,85
20	8,37	10,49
21	9,51	10,00
22	9,51	8,64
23	8,94	8,88
24	9,51	8,41
25	8,94	8,04
26	9,51	8,08
27	9,51	8,83
28	8,94	9,77
29	8,94	10,59
30	11,80	11,04

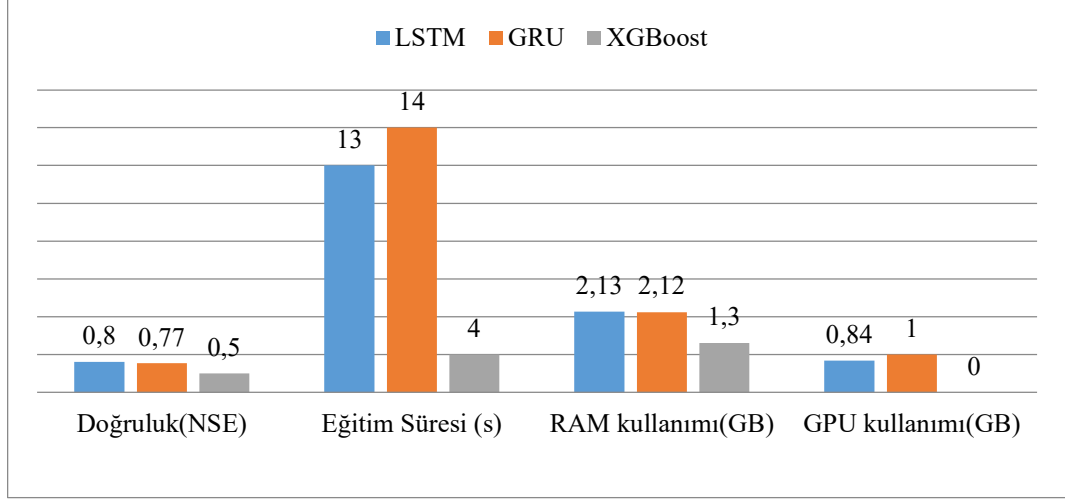
4.4 Modellerin Karşılaştırılması

Model performanslarına bakıldığında LSTM modeli, GRU ve XGBoost modeline gre daha iyi bir performans gsterdiği grlmştr. Őekil 4.7’de modellerin NSE ve R^2 hata deęerlerine gre karşılaştırılması gsterilmiştir.



Şekil 4.7. Modellerin hata değerlerinin karşılaştırma grafiği.

Modellerin doğruluk değerleri, eğitim süreleri, RAM (Random access memory) kullanımları ve GPU (Graphics Processing Unit) kullanımları Şekil 4.8.'de grafik olarak karşılaştırmaları yapılmıştır.



Şekil 4.8. Modellerin teknik açıdan karşılaştırılması.

Grafiğe bakıldığında LSTM ve GRU modellerinin eğitim süreleri, RAM kullanımı, GPU kullanımı XGBoost modeline göre oldukça yüksektir. Bunun nedeni derin öğrenme modellerinin daha çok işlem gücü gerektirmesidir. Veri seti denemeleri ve parametre denemeleri yapıldığı için toplam eğitim süreleri yüksektir.

5. SONUÇ VE ÖNERİLER

Çalışmada Makine Öğrenimi ve Derin öğrenme teknikleriyle Fırtına Deresi günlük ortalama debi değerlerinin tahmini yapılmıştır. Çalışmada diğer çalışmalardan farklı olarak CFSR veri setinden elde edilen veriler kullanılmıştır. Yine diğer çalışmalardan farklı olarak veri setinde bulunan maksimum sıcaklık,

minimum sıcaklık, rüzgâr, bağıl nem ve solar güneşlenme değişkenleri yağış verisinin yanında sisteme bağımsız değişken olarak verilmiştir. Bu değerlerinde model eğitiminde kullanılabileceği ve performansa olan etkileri gösterilmiştir. Ek olarak CFSR veri noktalarının tahmin modellemelerinde kullanılabileceği de gösterilmiştir.

Çalışma sonucunda modellemede elde edilen en iyi tahmin performansının, önceki çalışma sonuçlarıyla karşılaştırıldığında oldukça başarılı olduğu görülmektedir. Veri sayısının artırılması veya kullanılan yöntemlerde yapılacak değişiklikler ile model performans sonuçlarını artırmak mümkündür. Bununla birlikte, modelin tahmin performansı modele daha fazla aykırı değer tanıtılarak modelin daha geniş aralıkta tahmin yapması sağlanabilir. Ayrıca giriş verileri içerisinde kar erimesi değerleri mevcut değildir. Sisteme kar erimesi değerleri eklenirse performans artışı sağlanabilir. Aşırı debi yükseliş zamanlarına bakıldığında bahar aylarında olduğu görülmektedir bunun sebebidir kar erimesinden kaynaklanmaktadır.

Yapılan tahminler 875 günlük tahminleri içermektedir. Tahmin zamanı kısaltılarak daha yüksek performanslar elde edilebilir.

Debi tahmin modellemesine odaklanan birçok çalışma olmasına rağmen LSTM ve GRU model çalışmaları hidrolojik modellerde yeni yeni kullanılmaya başlanmıştır. Yapılan çalışma ilerleyen süreçte yapılabilecek araştırmalara katkıda bulunabilir.

Taşkınların neden olduğu ölümlerin, maddi ve manevi zararların önlenmesi için debi tahminlerinin önemi oldukça yüksektir. Tahmin değerleri kullanılarak erken uyarı sistemi yapılarak bu zararların önüne geçilebilir.

Karşılaştırma grafiğine bakıldığında modelin yükselen ve düşen trendleri gayet başarılı bir şekilde öğrendiği görülmektedir. Fakat ani değişiklikleri yakalayamadığı da bu grafikte net bir şekilde görülmektedir. Bunun nedeninin ani yağışların etkisi olduğun düşünülmektedir. Bu sorun saatlik verilerle çalışarak çözülebilir.

Modelleme sisteminin performansı, daha fazla eğitim tekrarı yapılarak ve model parametreleri değiştirilerek artırılabilir. Daha sonraki çalışmalarda yapılması önerilir.



6. KAYNAKLAR

(Bu tez çalışmasında APA atıf sistemi kullanılmıştır.)

Abar, H. (2020) . XGBOOST ve MARS yöntemleriyle altın fiyatlarının kestirimi. *Ekev Akademi Dergisi*, 83, 427-446

Alpaydin, E. (2020). Introduction to machine learning. *MIT press* (4th Ed.).

- Ayzel, G., & Heistermann, M. (2021). The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset. *Computers & Geosciences*, 149, 104708.
- Bayrakdar, C. (2006). Uygulamalı jeomorfoloji çalışmalarında CBS'nin Kullanımı: Fırtına Deresi havzası (Rize) örneği. Yayınlanmamış yüksek lisans tezi. İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
- B. Dickson (2020), Adversarial AI: Blocking the hidden backdoor in neural networks. 8 Temmuz 2021 tarihinde <https://bdtechtalks.com/2020/04/27/deep-learning-mode-connectivity-adversarial-attacks/> , adresinden alınmıştır.
- C. Robert, Machine learning, a probabilistic perspective (Taylor & Francis, 2014) *MIT press*, 62-63.
- Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*, 61, 304-323.
- Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). Applied hydrology. McGraw-Hill, California, 522.
- C. Olah (2020), Understanding LSTM Networks. 8 Temmuz 2021 tarihinde <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>, adresinden alınmıştır.
- Cüceloğlu, G. (2013). Darlık Havzasının Model Destekli Hidrolojik Analizi (Doctoral dissertation, Fen Bilimleri Enstitüsü).
- De Blij, H. J., Muller, P. O., & Nijman, J. (1997). *Geography: realms, regions, and concepts* (p. 688). New York, NY: Wiley.
- El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *machine learning in radiation oncology* (pp. 3-11). Springer, Cham.
- Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2), 337-407.
- Gümüş, V., Kavsut, M. E., & Yenigün, K. (2011). Yağış Akış İlişkisinin Modellenmesinde Ysa Kullanımının Değerlendirilmesi: Orta Fırat Havzası Uygulaması. *Engineering Sciences*, 6(1), 389-397
- Han, Y., Wu, J., Zhai, B., Pan, Y., Huang, G., Wu, L., & Zeng, W. (2019). Coupling a bat algorithm with XGBoost to estimate reference evapotranspiration in the arid and semiarid regions of china. *Advances in Meteorology*, 2019.
- Kılıçer, Ü., & Özgüler, H. (2002). Türkiye'de Taşkın Durumu. *Türkiye Mühendislik Haberleri*, 420-421.
- Krause, P., Boyle, D. P., & Bâse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in geosciences*, 5, 89-97.
- Latifoğlu, L., & Nuralan, K. B. (2020). Tekil Spektrum Analizi ve Uzun-Kısa Süreli Bellek Ağları ile Nehir Akım Tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, 376-381.
- Le, X. H., Ho, H. V., Lee, G., & Jung, S. (2019). Application of long short-term memory (LSTM) neural network for flood forecasting. *Water*, 11(7), 1387.
- Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885-900.

- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, 10(3), 282-290.
- Özcan, O. (2007). Sakarya Nehri Alt Havzası'nın Taşkın Riski Analizinin Uzaktan Algılama Ve Cbs İle Belirlenmesi (Doctoral dissertation, Bilişim Enstitüsü).
- R. Anderson (2019). RNN, Talking about Gated Recurrent Unit. 8 Temmuz 2021 tarihinde, <https://technopremium.com/blog/rnn-talking-about-gated-recurrent-unit/> adresinden alınmıştır.
- Singh, V. P., & Fiorentino, M. (1996). Hydrologic modeling with GIS. In *Geographical information systems in hydrology* (pp. 1-13). Springer, Dordrecht.
- T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* (ACM, New York, NY, USA, 2016) pp. 785–794.
- Terzi, Ö., & Mehmet, K. Ö. S. E. (2012). Yapay sinir ağırları yöntemi ile Göksu Nehri'nin akım tahmini. *Uluslararası Teknolojik Bilimler Dergisi*, 4(3), 1-7.
- Tyralis, H., Papacharalampous, G., & Langousis, A. (2021). Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. *Neural Computing and Applications*, 33(8), 3053-3068.
- Uğurluoğlu, A., (2017). Su Havzalarında Hidrolojik Modelleme: Manyas Gölü Yukarı Havzası Uygulaması. Uzmanlık Tezi, T.C. Orman ve Su İşleri Bakanlığı Genel Müdürlüğü, Ankara.
- Vo, Q. H., Nguyen, H. T., Le, B., & Nguyen, M. L. (2017, October). Multi-channel LSTM-CNN model for Vietnamese sentiment analysis. In *2017 9th international conference on knowledge and systems engineering (KSE)* (pp. 24-29). IEEE.
- Wheater, H., Sorooshian, S., & Sharma, K. D. (Eds.). (2007). *Hydrological modelling in arid and semi-arid areas*. Cambridge University Press.
- Yildirim, Ö. (2018). A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Computers in biology and medicine*, 96, 189-202.
- Yurdusev, M. A., Müserref, A. C. I., Turan, M. E., & İçağa, Y. (2008). Akarçay nehri aylık akımlarının yapay sinir ağırları ile tahmini. *Celal Bayar University Journal of Science*, 4(1), 73-88.