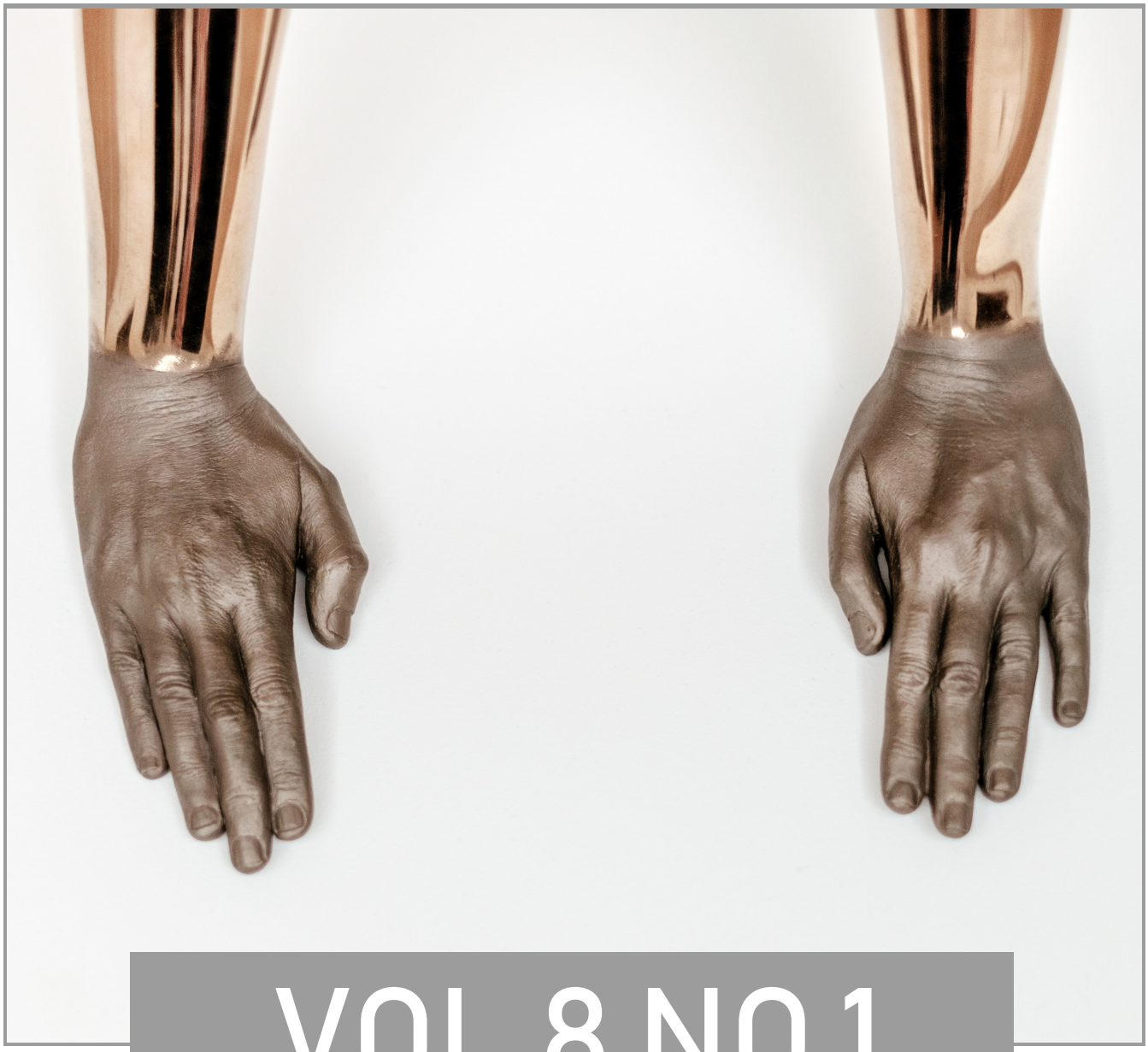
JISTA

Journal of Intelligent Systems: Theory and Applications

MARCH 2025

ISSN: 2651-3927



VOL 8 NO 1

ARTIFICIAL INTELLIGENT > MACHINE LEARNING > DEEP LEARNING
https://dergipark.org.tr/en/pub/jista

Journal of Intelligent Systems:
Theory and Applications
2025 Volume: 8 Issue: 1

Editorial Boards

Journal of Intelligent Systems:
Theory and Applications
2025 Volume: 8 Issue: 1

Language Editor

Editorial Advisory Board

Editorial Assistants

https://dergipark.org.tr/en/pub/jista

Journal of Intelligent Systems: Theory and Applications

2025 Volume: 8 Issue: 1

Contents

Research Articles

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 8(1) (2025) 1-24
DOI: 10.38016/jista.1444481

Makine Öğrenimi Yöntemleri ile Bireylerin Kronik Hastalık Durumlarının Sınıflandırılması: Türkiye İstatistik Kurumu'nun 2023 Gelir ve Yaşam Koşulları Araştırması Üzerine Bir Uygulama

Yunus Emre Gür[1*], Kâmil Abdullah Eşidir[2]

[1] Yönetim Bilişim Sistemleri Bölümü, İktisadi ve İdari Bilimler Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye

[2] Fırat Kalkınma Ajansı, Elazığ, Türkiye

yegur@firat.edu.tr, abdullahesidir@yahoo.com

Öz

Kronik hastalıkların artan prevalansı (görülme sıklığı) ve bunların bireylerin yaşam kalitesi üzerindeki olumsuz etkileri, kamu sağlığı alanında öncelikli meseleler arasında yer almaktadır. Bu hastalıkların erken teşhis ve yönetimi, sağlık hizmetlerine erişimdeki eşitsizlikler ve sosyoekonomik faktörlerle karmaşıklaşan bir süreçtir. Bu bağlamda, makine öğrenimi yöntemleri, büyük ve karmaşık veri kümelerinden bilgi çıkararak tahminlerde bulunma konusunda önemli bir potansiyel sunmaktadır. Özellikle TabNet yöntemi, güçlü tahmin yetenekleri ve karmaşık ilişkileri modelleme kapasitesi ile dikkat çekmektedir. Bu çalışma, Türkiye İstatistik Kurumu'nun 2023 Gelir ve Yaşam Koşulları Araştırması verilerini kullanarak, Yapay Sinir Ağları (YSA), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Destek Vektör Makinesi (DVM), Rastgele Orman, Gradient Boosting ve TabNet gibi yöntemler ile bireylerin kronik hastalık durumlarının sınıflandırılmasını amaçlamaktadır. Bulgular, sağlık hizmetlerine genel erişimin iyi olduğunu, ancak bazı kesimlerin hala erişimde zorluklar yaşadığını; kronik hastalıkların genel sağlık durumu ve istihdam gibi faktörlerle güçlü bir ilişkisi olduğunu ve TabNet yönteminin yüksek doğruluk, kesinlik ve duyarlılık gibi performans metrikleri ile etkili bir sınıflandırma yapabildiğini ortaya koymuştur. Sonuç olarak model, %97 genel doğruluk oranı ile kronik hastalık durumunu başarıyla sınıflandırmıştır. Bu çalışma, sağlık politikalarının geliştirilmesi ve sektörel analizler için stratejik kararlar alınmasında kullanılabilecek değerli bilgiler sunmakta ve makine öğrenimi yöntemlerinin, özellikle TabNet tekniğinin, sağlık verileri analizinde etkin bir şekilde kullanılmasının önemini vurgulamaktadır.

Anahtar kelimeler: Makine Öğrenimi, Veri Sınıflandırma, Kronik Hastalık Yönetimi, TabNet, Sağlık Politikaları Geliştirme.

Classification of Chronic Disease Status of Individuals Using Machine Learning Methods: An Application on The 2023 Income and Living Conditions Survey of the Turkish Statistical Institute

Abstract

The increasing prevalence of chronic diseases and their negative impact on the quality of life of individuals is one of the priority issues in the field of public health. Early diagnosis and management of these diseases is a process complicated by inequalities in access to healthcare services and socio-economic factors. In this context, machine learning methods offer significant potential for making predictions by extracting information from large and complex data sets. In particular, the TabNet method stands out for its strong predictive capabilities and ability to model complex relationships. This study aims to classify the chronic disease status of individuals using methods such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Support Vector Machines (SVM), Random Forest, Gradient Boosting and TabNet using data from the 2023 Income and Living Conditions Survey of the Turkish Statistical Institute. The results showed that overall access to health services is good, but some segments still have difficulty accessing it; chronic diseases have a strong relationship with factors such as general health status and

* Sorumlu yazar.
E-posta adresi: yegur@firat.edu.tr

Alındı : 28 Şubat 2024
Revizyon : 28 Eylül 2024
Kabul : 8 Kasım 2024

employment; and the TabNet method can perform effective classification with performance metrics such as high accuracy, precision and sensitivity. As a result, the model successfully classified chronic disease status with an overall accuracy rate of 97%. This study provides valuable information that can be used to make strategic decisions for health policy development and sectoral analysis, and highlights the importance of using machine learning methods, particularly the TabNet technique, effectively in health data analysis.
Keywords: Machine Learning, Data Classification, Chronic Disease Management, TabNet, Health Policy Development.

1. Giriş (Introduction)

Günümüzde, kronik hastalıkların yaygınlığı ve bunların bireylerin yaşam kalitesi üzerindeki etkileri, kamu sağlığı alanında öncelikli konular arasında yer almaktadır. Kronik hastalıklar, uzun süreli sağlık sorunları olarak tanımlanır ve tedavi edilse bile genellikle tam iyileşme sağlanamamaktadır (Kumsar ve Yılmaz, 2014; Altuntaş vd., 2015). Bu hastalıkların erken teşhisi ve yönetimi, bireylerin yaşam kalitesini önemli ölçüde artırabilirken, sağlık hizmetlerine erişimdeki eşitsizlikler ve sosyoekonomik faktörler bu süreci karmaşıklaştırmaktadır (Küçükberber vd., 2011). Bu bağlamda, makine öğrenimi yöntemleri, büyük ve karmaşık veri kümelerinden bilgi çıkarma ve tahmin yapma konusunda önemli bir potansiyele sahiptir (Sönmez ve Zengin, 2023). Yapay Sinir Ağları (YSA), Uzun Kısa Süreli Bellek ağları (LSTM), Evrişimli Sinir Ağları (CNN), Rastgele Orman, Gradyan Artırma (Gradient Boosting), TabNet ve Destek Vektör Makineleri (DVM) gibi makine öğrenimi modellerinin sağlık hizmetleri sınıflandırma problemlerinde artan kullanımı, tıbbi veri analizi alanında önemli bir eğilimi yansıtmaktadır. Rastgele Orman ve Gradient Boosting, sağlık hizmetleri sınıflandırma görevlerinde kayda değer başarı gösteren topluluk öğrenme yöntemleridir. Rastgele Ormanlar, tahmin doğruluğunu ve aşırı uyuma karşı sağlamlığı artırmak için birden fazla karar ağacını birleştirmektedir. Bu yöntem, yorumlanabilirliği ve çok sayıda özelliğe sahip büyük veri kümelerini işleme yeteneği nedeniyle risk değerlendirmesi ve hasta sınıflandırması dahil olmak üzere çeşitli sağlık senaryolarında yaygın olarak kullanılmaktadır (Luo vd., 2018). Gradient Boosting, önceki modeller tarafından yapılan hataları düzeltmeye odaklanarak modelleri sırayla oluşturur ve bu da karmaşık sağlık hizmeti veri kümelerinde tahmin gücünü artırır. Bu yöntem, güçlü tahmin yetenekleri ve karmaşık ilişkileri modelleme kapasitesi ile sağlık alanındaki sınıflandırma sorunları için giderek daha fazla kullanılan bir yöntem haline gelmiştir (Yaygın, 2019; Özdemir, 2023). YSA'lar, büyük veri kümelerinden öğrenme ve verilerdeki karmaşık örüntüleri belirleme yeteneğine sahip oldukları için sağlık uygulamalarında özellikle etkili olmaktadır. Bu yetenek, YSA'ların hastalık teşhisi ve prognozu gibi görevler için kullanıldığı ve tıbbi verilerin doğasında bulunan doğrusal olmayan ilişkileri işlemedeki etkinliklerini gösteren çeşitli çalışmalarda gösterilmiştir (Ahsan vd., 2023). Öte yandan, uzun vadeli bağımlılıkları hatırlama yetenekleri, LSTM'lerin çeşitli sağlık uygulamalarında geleneksel modellerden daha iyi performans göstermesini sağlamaktadır (Ahsan vd., 2023). Çalışmalar, CNN'lerin de çeşitli sağlık alanı sınıflandırma görevlerinde yüksek doğruluk oranlarına ulaşabildiğini ve böylece erken teşhis ve müdahaleyi kolaylaştırdığını göstermektedir (Almutairi vd., 2022). DVM'ler, sağlık hizmetlerinde makine öğreniminin bir diğer temel taşıdır. DVM'ler özellikle yüksek boyutlu verileri işleme konusunda ustadır ve kanser teşhisi ve genetik veri analizi de dahil olmak üzere çeşitli alanlarda hastalık sınıflandırması için yaygın olarak kullanılmaktadır (Guido vd., 2024). Nispeten daha yeni bir model olan TabNet, özellikleri dinamik olarak seçmek için sıralı bir dikkat mekanizması kullanır ve bu da onu sağlık hizmetleri ortamlarında yaygın olarak bulunan tablo verileri için özellikle etkili kılmaktadır (Arık ve Pfister, 2019). Yüksek performansı korurken yorumlanabilir sonuçlar sağlama yeteneği, TabNet'i modelin karar verme sürecine ilişkin içgörülere ihtiyaç duyan sağlık uygulayıcıları için değerli bir araç haline getirmektedir (Arık ve Pfister, 2019).

Bununla birlikte, gelir dağılımı ve yoksulluk, ekonomik kalkınmanın yanı sıra sosyal adalet ve eşitlik açısından da büyük önem taşıyan konulardır. Gelir dağılımındaki adaletsizlikler ve yoksulluk, bir ülkenin ekonomik başarısının yanı sıra toplumsal dengenin ve istikrarın da önemli göstergeleridir. Bu nedenle, ekonomik açıdan sağlıklı bir değerlendirme yapabilmek için gelir dağılımı ve yoksullukla ilgili verilere dayalı olarak politika ve sosyal programlar geliştirilmelidir. Bu programlar, gelir eşitsizliğini azaltmaya, yoksulluğu önlemeye ve sosyal adaleti sağlamaya yönelik olmalıdır (Yar, 2015). Ayrıca, gelir dağılımı ve yoksullukla ilgili verilerin düzenli olarak izlenmesi ve analiz edilmesi, politika yapıcıların karar alma süreçlerinde daha bilinçli kararlar almasına yardımcı olacaktır (Ersöz, 2003).

Türkiye İstatistik Kurumu'nun (TÜİK) 2023 yılında gerçekleştirdiği Gelir ve Yaşam Koşulları Araştırması (GYKA), Türkiye'deki hanelerin sosyoekonomik durumları ve bireylerin yaşam koşulları hakkında kapsamlı veriler sunmaktadır. Bu araştırma, bireylerin genel sağlık durumları, sağlık hizmetlerine erişimde karşılaştıkları zorluklar, istihdam durumları, eğitim durumları ve sosyoekonomik koşulları gibi önemli göstergeleri içermektedir. Bu çalışma, GYKA veri setini kullanarak, bireylerin kronik bir hastalığının olup olmadığını tahmin etmek için YSA, LSTM, CNN, Rastgele Orman, Gradient Boosting, TabNet ve DVM gibi makine öğrenimi modellerini uygulamaktadır. Araştırma, genel sağlık durumu, sağlık probleminden ötürü faaliyetlerde sınırlılık, doktora başvuramama durumu, istihdam durumu, eğitim seviyesi, sosyal yaşam durumu ve ücretli sosyal faaliyetlere katılım durumu gibi bağımsız değişkenlerin, bireylerin kronik

hastalık durumlarının sınıflandırılmasındaki etkilerini incelemektedir.

Bu çalışmanın amacı, sağlık alanında karar verme süreçlerini desteklemek ve makine öğrenimi tabanlı modellerin potansiyelini göstermek adına bu yöntemleri kullanarak, Türkiye'deki bireylerin kronik hastalık durumlarını sınıflandırmak ve sağlık durumları üzerine derinlemesine bir analiz sunmaktır. Makalenin bundan sonraki bölümlerinde, makine öğrenimi yöntemlerinin sağlık verileri analizindeki rolü hakkında bir literatür incelemesi sunulmuştur. Ardından, çalışmanın metodolojisi anlatılmış ve kullanılan yöntemlerin teorik temelleri ve uygulama süreci detaylandırılmıştır. Daha sonra, çalışmanın bulguları açıklanmış ve makine öğrenimi sınıflandırma algoritmalarının performansı değerlendirilmiştir. Son olarak, makalenin sonuç ve öneri kısmında, elde edilen bulguların sağlık politikaları ve sektör analizleri üzerindeki etkileri ele alınmış ve bu sonuçların gelecekteki araştırmalara yön vermesi için öneriler geliştirilmiştir. Bu değerlendirmeler, sağlık alanında daha bilinçli kararlar alınmasına ve sektörün ihtiyaçlarına yönelik stratejilerin oluşturulmasına katkıda bulunmayı amaçlamaktadır.

2. Literatür İncelemesi (Literature Review)

Makalenin bu bölümünde, makine öğrenimi yöntemlerinin sağlık verileri analizindeki rolü hakkında bir literatür incelemesi gerçekleştirilmiştir. Elde edilen çalışmaların bir kısmı bu bölümde sunulmuştur.

Özkan (2019)'ın çalışmasında, sağlık verileri üzerinden hastalık tanısı koyma sürecinde karşılaşılan problemler ele alınarak, KEEL veri tabanından alınan çeşitli hastalık veri setleri üzerinde torbalama ve artırma algoritmaları karşılaştırmıştır. Veri ön işleme sonrasında, artırma algoritmaları genel olarak torbalama yöntemlerine üstün performans sergilemiştir. Özellikle, Gradient Boosting algoritması, hepatit hastalığının teşhisinde %98.36 doğruluk, %98.68 kesinlik, %98.95 duyarlılık, ve %98.91 F-Skor değerleri ile dikkate değer sonuçlar elde etmiştir. Bu bulgular, veri ön işlemenin ve seçilen algoritmaların doğru hastalık tanısında önemli rol oynadığını göstermektedir.

Hematolojik hastalıklar alanında, Ahmed vd. (2019), mikroskobik görüntülerden lösemi alt türlerini belirlemek için CNN'leri kullanmıştır. Geliştirdikleri model %88,25'lik bir doğruluk oranına ulaşarak derin öğrenme tekniklerinin karmaşık hastalıklar için teşhis doğruluğunu artırma potansiyeline işaret etmiştir. Bu durum, lösemi teşhisi için CNN'lerle derin transfer öğrenmesi kullanan ve modelin kan hücresi görüntülerinden ilgili özellikleri etkili bir şekilde çıkarma yeteneğini vurgulayan Loey vd. (2020) tarafından da desteklenmektedir.

Gaddam ve Pattnaik (2020) tarafından yapılan bir çalışma, YSA kullanarak aritmi tespiti için EKG sinyal sınıflandırmasına odaklanmıştır. Model, kardiyak anormallikleri belirlemedeki etkinliğini göstererek %91'lik bir doğruluk elde etmiştir. Bu yetenek, yaygın kronik durumlar olan kardiyovasküler hastalıkların zamanında teşhis ve tedavisi için çok önemlidir.

Pacci vd. (2021), yaptıkları bir çalışmada, tüp bebek tedavisinde pozitif gebelik sonucunu tahmin etmek amacıyla yapay zekâ tabanlı bir klinik karar destek sistemi geliştirilmişlerdir. Çalışmada, Yeditepe Üniversitesi Hastanesi'nden alınan 1154 tedavi siklusuna ait veriler kullanılarak, beş farklı sınıflandırma yöntemi (DVM, Çok Katmanlı Algılayıcı, Rastgele Orman, XGBoost ve LightGBM) karşılaştırmalı olarak test edilmiştir. En yüksek sınıflandırma performansı, Destek Vektör Makineleri yöntemi ile elde edilmiştir, AUC değeri 0.70 olarak bulunmuş ve karar eşik değerinin optimizasyonu ile gebelik sonucunun %71.7 Doğru Pozitif ve %59.4 Doğru Negatif oranıyla tahmin edilmesi sağlanmıştır.

Tang ve Liu (2021), yapmış oldukları bir çalışmada, Alzheimer hastalığının (AD) ilerleyişini, beyin manyetik rezonans görüntüleme (MRI) verileri kullanılarak çeşitli makine öğrenimi algoritmaları ile sınıflandırmış ve tahmin etmişlerdir. ADNI veri tabanından alınan 560 katılımcı, kognitif normal (CN), erken hafif bilişsel bozukluk (EMCI), geç hafif bilişsel bozukluk (LMCI) ve Alzheimer (AD) olmak üzere dört gruba ayrılmıştır. Rastgele Orman, DVM ve Karar Ağacı algoritmaları kullanılarak bu grupların hastalık ilerleyişi sınıflandırılmıştır. Rastgele Orman algoritması en yüksek doğruluğa (CN-AD için %96.14) ve en yüksek AUC değerine (0.92) ulaşarak diğer modellerden daha başarılı olmuştur. Sınıflandırmada kullanılan MRI özellikleri ile Rastgele Orman modeli, hastalığın erken teşhisinde yardımcı bir araç olarak önerilmiştir.

Gündoğdu (2021)'nun çalışmasında, Kaggle veri tabanından alınan kalp hastalığı veri seti kullanılarak Python aracılığıyla 7 sınıflandırma algoritması (Destek Vektör Makineleri, Gaussian Naive Bayes, Gradient Boosting ve Rastgele Orman) performanslarının karşılaştırılması yapılmıştır. En iyi performans gösteren algoritma, %89.7 F1 skoru ve %90.2 doğruluk ile Rastgele Orman olmuştur. Açlık kan şekeri özelliğinin önem sıralamasında en altta yer alması, sınıflandırma performansı üzerinde minimal etkisi olduğunu göstermiştir. Bu bulgular, kalp hastalığı tahmininde daha etkili ve doğru sistemlerin geliştirilmesine katkı sağlayabilir.

Kim vd. (2021) tarafından gerçekleştirilen "Makine Öğrenimi Tabanlı Kardiyovasküler Hastalık Tahmini Modeli: Kore Ulusal Sağlık Sigortası Hizmeti Verileri Üzerine Bir Kohort Çalışması" isimli araştırmada, Kore Ulusal Sağlık Sigortası Hizmeti'nin sağlık taraması veri setinden en uygun tahmin modelini belirlemek amacıyla bir dizi makine öğrenimi yöntemi uygulanmıştır. Bu yöntemler; Lojistik Regresyon, K-En Yakın Komşu, Karar Ağaçları, Rastgele Orman, Ekstra Ağaçlar, XGBoosting, Gradyan Arttırma, AdaBoost, DVM ve Çok Katmanlı Algılayıcılar içermektedir. Araştırma sonuçlarına göre, XGBoosting, Gradient Boosting ve Rastgele Orman yöntemleri, performans ölçütleri

bazında diğerlerine üstün gelerek en iyi tahmin modellerini oluşturmuştur.

Toğaçar vd. (2021), yapmış oldukları bir çalışmada, deri kanseri tespiti için CNN tabanlı yeni bir model geliştirmişlerdir. Model, Autoencoder, MobileNetV2 ve Spiking Neural Networks (SNN) bileşenlerini kullanarak, iyi huylu ve kötü huylu tümörleri sınıflandırmışlardır. ISIC veri seti kullanılarak yapılan deneylerde, MobileNetV2 modeli ve spiking ağları ile %95.27'lik bir doğruluk oranı elde edilmiştir. Bu sonuçlar, Autoencoder ve SNN'nin MobileNetV2 modelinin performansını artırmada etkili olduğunu göstermiştir. Çalışma, deri kanseri tespiti için yüksek hassasiyetli ve tamamen otomatik bir karar destek aracı sunmaktadır.

Bununla birlikte, Akcan ve Sertbaş (2021)'ın çalışmasında, göğüs kanseri teşhisi için topluluk öğrenme yöntemleri kullanılarak bir dizi makine öğrenimi algoritmasının performansı karşılaştırılmıştır. DVM, K-En Yakın Komşu (KNN), Naive Bayes, Karar Ağaçları ve Rastgele Orman gibi algoritmalar yanında, bagging, boosting ve voting gibi topluluk öğrenme yöntemleri uygulanmıştır. Veri ön işleme ve özellik ölçeklendirme adımları yapıldıktan sonra, bu yöntemlerin doğruluk, kesinlik, duyarlılık, F-Skor ve AUC skorları karşılaştırılmıştır. En yüksek doğruluk oranları Soft Voting, Bagging (SVC) ve XGBoost yöntemleriyle elde edilmiştir, bu da topluluk öğrenme yöntemlerinin bireysel sınıflandırma yöntemlerine göre daha üstün performans sergilediğini göstermiştir.

Purwaningsih (2022), kronik böbrek hastalığının (KBH) tahmini için DVM modeli kullanmış ve ileri özellik seçimi (Forward Selection) kullanarak modeli geliştirmiştir. Çalışmada, DVM modeli farklı çekirdekler (dot, polynomial ve RBF) ile test edilmiştir ve en yüksek doğruluk oranı %98,50 (AUC = 1,000) ile dot çekirdekli DVM'de elde edilmiştir. Ancak, ileri özellik seçimi uygulanarak SVM+FS modeli ile RBF çekirdeği kullanıldığında doğruluk %99,75'e (AUC = 1,000) yükselmiştir. Bu sonuçlar, ileri özellik seçiminin DVM performansını önemli ölçüde artırdığını göstermektedir. Çalışmada ayrıca, hastalık durumunu tahmin etmek için YSA modeli de kullanılmış ve %90,5'lik bir doğruluk elde edilmiştir. YSA'nın klinik ve laboratuvar verilerine dayalı olarak risk altındaki hastaların belirlenmesinde sağlık hizmeti sağlayıcılarına yardımcı olabileceği KBH'de erken tanı ve müdahalenin önemini vurgulamaktadır.

Sevli (2023)'nin çalışmasında, diyabet hastalığının erken teşhisi için Pima Indian Diabetes veri seti üzerinde altı farklı makine öğrenimi yöntemi (Destek Vektör Makinesi, Lojistik Regresyon, K-En Yakın Komşu, Rastgele Orman, AdaBoost, Gradient Boosting) kullanılarak sınıflandırma çalışmaları yapılmıştır. Yeniden örnekleme teknikleri uygulanarak, sınıflandırıcıların başarısı artırılmaya çalışılmıştır. En yüksek performans, Rastgele Orman sınıflandırıcısı ile InstanceHardnessThreshold az örnekleme tekniği kullanılarak elde edilmiş; %96.29 doğruluk, %98.07 kesinlik, %100 geri çağırma, %96.22 F1 Skoru ve %96.29 AUC değerleri raporlanmıştır. Gradient Boosting ve AdaBoost yöntemleri de benzer yeniden örnekleme tekniği ile yüksek performans göstermiştir.

Coşkun ve Yüksek (2023), ölümcül hepatit hastalığının tanısı için öznitelik seçimi yöntemini kullanarak bulanık mantık ve çeşitli makine öğrenmesi yöntemlerinin başarısını karşılaştırmıştır. UCI makine öğrenimi deposundan alınan hepatit veri seti üzerinde öncelikle veri ön işleme ve öznitelik seçimi işlemleri yapılmış, ardından bulanık model ve makine öğrenmesi modelleri test edilmiştir. Bulanık Mantık yöntemiyle %94 doğruluk elde edilirken, Gradient Boosting algoritmasıyla %98.36 doğruluk, %98.68 kesinlik, %98.95 duyarlılık ve %98.91 f-skor değerleri elde edilmiştir. Sonuçlar, Gradient Boosting yönteminin diğer makine öğrenme yöntemleri ve bulanık mantık yaklaşımına göre hepatit hastalığının teşhisinde daha başarılı olduğunu göstermektedir.

Kim vd. (2023), konvolüsyonel sinir ağlarını (CNN), LSTM ile birleştiren çok görevli bir öğrenme çerçevesi kullanarak kronik hastalık tahminini araştırmıştır. Yaklaşımları, birden fazla kronik hastalığı aynı anda tahmin etmede %94,3'lük bir doğruluk elde etmiştir. Bu çalışma, LSTM'nin kronik hastalık tahmini için gerekli olan zaman serisi verilerindeki zamansal ilişkileri modelleme yeteneğini vurgulamaktadır.

Zhang vd. (2023) tarafından yapılan bir başka çalışma, solunum ses dosyalarını kullanarak akciğer hastalığı tespitine odaklanmıştır. LSTM modeli %98,82'lik etkileyici bir doğruluk ve 0,97'lik bir F1 skoru elde ederek ses verilerindeki sıralı örüntüleri yakalamadaki üstün performansını göstermiştir. Bu yüksek doğruluk seviyesi, LSTM'nin öksürük seslerine dayalı kronik solunum yolu hastalıklarını teşhis etmedeki etkinliğinin altını çizmektedir.

Özdemir (2023)'in yapmış olduğu bir çalışmada, aritmilerin sınıflandırılması için makine öğrenmesi algoritmaları kullanılmıştır. Yöntem olarak Bagging Decision Tree, Rastgele Orman, Extra Tree, Gradient Boosting ve DVM algoritmaları kullanılmıştır. Çalışma, MIT PhysioNet veri seti üzerinde gerçekleştirilmiş ve belirli hasta numaraları (203, 208, 210 ve 213) üzerinde odaklanılmıştır. Bulgular, bu algoritmaların aritmilerin sınıflandırılmasında etkili olduğunu ve çeşitli algoritmaların performanslarının karşılaştırılmasıyla elde edildiğini göstermektedir. Özellikle, Random Forest ve Gradient Boosting gibi yöntemlerin yüksek doğruluk oranları sunabileceği belirtilmektedir. Makale, makine öğrenmesi algoritmalarının aritmi sınıflandırmasındaki potansiyelini ve sağlık alanındaki uygulamalarını vurgulamaktadır.

Duyar vd. (2023) bağırsak mikrobiyota verilerini kullanarak kardiyovasküler hastalıkların tespitini araştırmış ve TabNet'i diğer makine öğrenimi modelleriyle karşılaştırmıştır. Çalışma, TabNet'in boosting yöntemlerine kıyasla daha zayıf sonuçlar verdiğini gösterse de, modelin kronik hastalık sınıflandırmasında uygulanabilirliğine ilişkin değerli

bilgiler sağlamıştır. Konuyla ilgili bir başka çalışmada, McLaughlin vd. (2023), TabNet kullanarak pan-kanser varyant arama üzerine odaklanılmıştır. Çalışma, mutasyonları somatik veya germline olarak sınıflandırmayı amaçlamış ve TabNet'in bu sınıflandırma görevinde yüksek performans elde ettiğini bildirmiştir. Çalışma, TabNet'in yorumlanabilirliğini ve kanser genomiğinde kritik öneme sahip olan tablo verilerini işlemedeki etkinliğini vurgulamıştır. Hegde ve Mundada (2020) tarafından yapılan bir çalışmada, diğer makine öğrenimi tekniklerinin yanı sıra TabNet de kullanılmıştır. Çalışma, TabNet'in kronik hastalıkları tahmin etmede %90'lık bir doğruluk elde ettiğini ve sağlık analitiğinde güçlü bir araç olarak rolünü güçlendirdiğini bildirmiştir. Ek olarak, Elkholy vd. (2023) TabNet kullanarak kronik böbrek hastalığı için geliştirilmiş bir optimize sınıflandırma modeli geliştirmiştir. Model %92,5'lik bir doğruluk oranına ulaşarak kronik böbrek hastalığını erken bir aşamada tespit etmedeki etkinliğini ortaya koymuştur. Bu yüksek doğruluk oranı, kronik böbrek hastalığının küresel olarak artan prevalansı göz önüne alındığında özellikle önemlidir.

Choubey vd. (2024), büyük veri madenciliğini kullanılarak kronik hastalıkların tahmini için yeni bir yaklaşım sunulmuştur. Bu yaklaşımda, Özellik Seçimi için Principal Component Analysis (PCA) ve sınıflandırma için eXtreme Gradient Boosting (XGB) algoritmaları kullanılmıştır. PCA, veri boyutunu azaltarak en önemli varyasyonları yakalarken, XGB bu özellikleri kullanarak yüksek doğrulukta hastalık tahmini yapmaktadır. Çalışmanın sonuçları, %98.8 doğruluk, %98 recall ve %98.7 F1-skoru ile yüksek bir performans göstermiştir. Modelin işlem süresi 8 saniye olup, bu hızlı işlem kapasitesi, modeli pratik uygulamalar için uygun hale getirmektedir. Guhan vd. (2024), kronik böbrek hastalığının doğru teşhisinin önemini vurgulamaktadır. Araştırma, proaktif sağlık stratejilerinin iyileştirilmesine yardımcı olmakta ve kronik hastalıkların önlenmesi ve yönetimi konusunda değerli bilgiler sağlamaktadır. Çalışmada, kronik bir hastalık olan böbrek yetmezliği olasılığını tahmin eden Büyük Veri analizine dayalı bir çerçeve geliştirilmiştir. Çerçevede, kronik hastalıklarla ilişkili risk faktörlerini ve korelasyonları tespit etmek için MEG (Mean Decrease Gini), MSE (Mean Square Error), Grid Search, K-fold cross validation gibi gelişmiş Makine Öğrenimi tekniklerini kullanılmıştır. El-Shafeiy vd. (2024), tarafından yapılan bir başka çalışmada, doğurganlık kalitesini tahmin etmek için YSA uygulanmış ve %85'lik bir doğruluk oranı rapor edilmiştir. Bu araştırma, doğru tahminlerin klinik kararlara rehberlik edebileceği ve hasta sonuçlarını iyileştirebileceği üreme sağlığı alanında YSA'nın çok yönlülüğünü göstermektedir. Modelin performansı, karmaşık veri kümelerine dayalı olarak sağlıkla ilgili sonuçların tahmin edilmesinde daha geniş uygulamalar için potansiyeline işaret etmektedir.

Sonuç olarak, YSA, LSTM, CNN, Rastgele Orman, Gradient Boosting, TabNet ve DVM gibi makine öğrenimi modellerinin sağlık hizmetleri sınıflandırma problemlerine entegrasyonu, bu alanda dönüştürücü bir değişimi temsil etmektedir. Literatürdeki çalışmalardan da hareketle, bu modeller yalnızca tahmin doğruluğunu artırmakla kalmamakta, aynı zamanda klinik karar verme sürecini bilgilendirebilecek değerli içgörüler de sağlamaktadır. Araştırmalar ilerlemeye devam ettikçe, bu teknolojilerin hasta sonuçlarını iyileştirme ve sağlık hizmeti süreçlerini kolaylaştırma potansiyeli artacak ve modern tıpta vazgeçilmez araçlar haline gelecektir. Ayrıca, bu çalışmanın özgün değeri, Türkiye İstatistik Kurumu'nun (TÜİK) 2023 yılı Gelir ve Yaşam Koşulları Araştırması (GYKA) veri seti kullanılarak, bireylerin kronik hastalık durumlarının makine öğrenimi yöntemleriyle sınıflandırılmasıdır. Mevcut literatürde bu spesifik veri seti ve yöntemle kronik hastalık durumlarının sınıflandırılmasına dair herhangi bir çalışma bulunmamaktadır. Çalışma, bu kapsamlı ve güncel veri setini kullanarak, sağlık hizmetlerine erişimdeki eşitsizlikler ve sosyoekonomik faktörlerin sağlık durumları üzerindeki etkilerini incelemiş, yüksek doğruluk oranlarıyla sınıflandırma yaparak literatüre önemli bir katkı sağlamıştır. Bu özgün yaklaşım, sağlık politikalarının geliştirilmesinde ve sektörel analizlerde stratejik kararlar alınmasında kullanılabilecek değerli bilgiler sunmaktadır.

3. Çalışmanın Metodolojisi (Methodology of the Study)

Bu çalışmada, Türkiye İstatistik Kurumu'nun (TÜİK) 2023 yılı Gelir ve Yaşam Koşulları Araştırması (GYKA) verileri temel alınarak, Türkiye'deki bireylerin kronik hastalık durumlarını sınıflandırmak amacıyla çeşitli makine öğrenimi yöntemleri kullanılmıştır. GYKA verileri, Türkiye'deki hanelerin sosyoekonomik durumları ve bireylerin yaşam koşulları hakkında kapsamlı veriler sunmaktadır. Bununla birlikte, bireylerin genel sağlık durumları, sağlık hizmetlerine erişimde karşılaştıkları zorluklar ve istihdam durumları gibi önemli göstergeleri de içermektedir. Çalışmada, GYKA araştırmasında kullanılan anket formundan elde edilen yedi bağımsız değişken, kronik hastalık durumu tahminlerinin yapılmasında kullanılmıştır. Yapılan ön analizler sonucunda eksik veriler, veri tabanından çıkarılmıştır. Veri setinde toplamda 64.607 adet hücre verisi kullanılmıştır.

Bu çalışmada, kullanılan bağımsız değişkenler, genel sağlık durumu, faaliyetlerde sınırlama durumu dağılımı, doktora başvuramama durumu, istihdam durumu, eğitim seviyesi, sosyal yaşam durumu ve ücretli sosyal faaliyetlere katılma durumu şeklindedir. Bu değişkenler, araştırmada bireylerin kronik hastalık durumunu (bağımlı değişken) etkileyen faktörler olarak incelenmiştir. Bu bağımsız değişkenler, kronik hastalık durumunu anlamak ve değerlendirmek için kullanılmıştır. Orijinal mikro veri setinde "Genel Sağlık

Durumu" 5 farklı biçimde ifade edilmiştir (1-Çok İyi, 2-İyi, 3-Orta, 4-Kötü, 5-Çok Kötü). Analiz adımlarını sadeleştirmek adına Çok İyi ve İyi sınıfları "0- Sağlığı İyi" şeklinde sadeleştirilmiştir. 3, 4 ve 5 sınıfları ise "1-Sağlığı Kötü" olarak sınıflandırılmıştır.

Tablo 1. Değişkenlere ilişkin tanımlamalar ve açıklamalar (Definitions and explanations of variables)

Değişken Adı	Değişken Tanımı	Açıklama
FS010	Ferdin genel sağlık durumu	1- Çok iyi 2-İyi 3 Orta 4-Kötü 5-Çok kötü
FS030	Sağlık probleminden ötürü faaliyetlerde sınırlama olup olmadığı	1-Evet, çok sınırlandı 2-Evet, sınırlandı 3-Hayır, sınırlanmadı
FS050	Ferdin son 12 ay içerisinde ihtiyaç duyulduğu halde doktora başvuramama durumu	1-Evet, en az 1 kere 2-Hayır, hiç olmadı 3-Hayır, ihtiyaç olmadı
FI010	Ferdin istihdam durumu	1-Tam zamanlı ücretli çalışan 2-Yarı zamanlı ücretli çalışan 3-Tam zamanlı işveren 4-Yarı zamanlı işveren 5-İş arıyor 6-Eğitime devam ediyor 7-Emekli 8-Engelli 9-Ev işleri ile meşgul 10-Diğer
FE030	Ferdin Eğitim Seviyesi	0-Okur-yazar olmayan 1-Bir okul bitirmedi 2-İlkokul 3-İlköğretim 4-Ortaokul ve dengi 5-Genel lise 6-Mesleki veya Teknik lise 7-Yüksekokul 8-Fakülte 9-Yüksek Lisans 10-Doktora
FY050	Ayda en az bir kere arkadaş, aile/akraba ile yemek yemek veya bir şeyler içmek için dışarıda (lokanta, pastane, kafe vb. yerlerde) bir araya gelme durumu	1-Evet 2-Hayır-maddi yetersizlik 3-Hayır-diğer nedenler
FY060	Spor, sinema, konser gibi boş zaman faaliyetlerine (ücret ödeyerek) düzenli olarak katılma durumu	1-Evet 2-Hayır-maddi yetersizlik 3-Hayır-diğer nedenler
FS020 (Bağımlı Değişken)	Ferdin kronik bir hastalığının olup olmadığı	0-Evet 1-Hayır

Tablo 1'de analizlerde kullanılan bağımsız değişkenler ve değişkenlerin açıklama ve tanımlamaları gösterilmiştir. İlgili tabloda yer alan veri seti, içerdiği bilgilerin türünü ve bu bilgilerin nasıl kodlandığını açıklamaktadır. Tablo 1'de analizlerde kullanılan değişkenler ve bunlara ait tanımlamalar ve açıklamalar gösterilmektedir. İlgili tabloda, veri setinin ne tür bilgiler içerdiği ve bu bilgilerin nasıl kodlandığı açıklanmaktadır.

Kronik hastalıkların tahmini için çeşitli makine öğrenmesi sınıflandırma algoritmaları kullanılmıştır ve analiz Python yazılımı aracılığıyla yürütülmüştür. Bu süreçte, veri işleme ve manipülasyonu için Pandas ve NumPy kütüphaneleri; veri görselleştirme amacıyla Matplotlib ve Seaborn; veri bölümlendirme için Scikit-learn'in train_test_split fonksiyonu kullanılmıştır. Veri seti %80'i eğitim ve %20'si test seti olmak üzere ikiye ayrılmıştır. Gradient Boosting, YSA, LSTM, CNN, TabNet, DVM ve Rastgele Orman modellerini kullanarak sınıflandırma işlemi yapmak için çeşitli kütüphaneler gereklidir. Klasik makine öğrenmesi modelleri olan DVM, Rastgele Orman ve Gradient Boosting için scikit-learn kullanılmıştır. Derin öğrenme modelleri (ANN, LSTM, CNN) için ise tensorflow ve keras kütüphaneleri, alternatif olarak PyTorch tabanlı modeller için torch tercih edilmiştir. TabNet modeli için pytorch-tabnet kullanılmış, veri görselleştirme ve model değerlendirme amacıyla matplotlib, seaborn ve scikit-

plot gibi kütüphaneler tercih edilmiştir. Bu araçlar, kapsamlı sınıflandırma işlemlerinin uygulanabilmesini ve performans değerlendirmelerinin yapılabilmesini sağlamaktadır. Bu kütüphaneler, Python'un veri bilimi ve makine öğrenimi projelerinde standart araçları arasında yer almaktadır ve geniş bir kullanım alanına sahiptir (Wade ve Glynn, 2020).

Bununla birlikte, bu çalışmada, makine öğrenmesi sınıflandırma algoritmalarının başarısını değerlendirmek için Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1 Puanı (F1-Score) gibi metrikler kullanılmıştır. Doğruluk, toplam tahminlerin ne kadarının doğru yapıldığını göstermektedir.

$$Doğruluk = \frac{DP+DN}{DP+DN+YP+YN\ (Toplam\ Örnek\ Sayısı)} \quad (1)$$

Kesinlik, pozitif olarak tahmin edilen durumların gerçekte ne kadarının pozitif olduğunu göstermektedir.

$$Kesinlik = \frac{DP}{DP+YP} \quad (2)$$

Duyarlılık veya hassasiyet, gerçek pozitif durumların ne kadarının doğru tahmin edildiğini göstermektedir.

$$Duyarlılık = \frac{DP}{DP+YN} \quad (3)$$

F1 Puanı, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır ve dengeli bir ölçümdür.

$$F1 = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik+Duyarlılık} \quad (4)$$

Bu ifadeler, sınıflandırma modelinin tahmin sonuçlarını temsil etmektedir. DP (Doğru Pozitif) modelin pozitif olarak doğru tahmin ettiği durum sayısını, DN (Doğru Negatif) modelin negatif olarak doğru tahmin ettiği durum sayısını, YP (Yanlış Pozitif) modelin pozitif olarak yanlış tahmin ettiği durum sayısını ve YN (Yanlış Negatif) modelin negatif olarak yanlış tahmin ettiği durum sayısını belirtmektedir. Bununla birlikte, yukarıda açıklanan metriklere ek olarak kullanılan "support" metriği, sınıflandırma raporlarında her sınıf için veri noktalarının toplam sayısını ifade etmektedir. Bu metrik, modele girdi olarak verilen her sınıfın örnek sayısını gösterir ve modelin performans değerlendirmesini yaparken, veri setindeki sınıf dağılımını anlamak için önemlidir. Özellikle dengesiz veri setlerinde, bazı sınıfların diğerlerine göre daha az veya daha fazla örneğe sahip olması durumunda, "support" değeri bu sınıfların analizdeki ağırlığını ve önemini göstermektedir.

3.1. Gradient boosting yöntemi (Gradient boosting method)

Makine öğreniminde karşılaşılan yaygın bir zorluk, veri setlerinden parametrik olmayan regresyon veya sınıflandırma modelleri geliştirmektir. Gerçek dünya senaryolarında, teorik modeller çoğunlukla eksik olur ve araştırmacılar, girdi değişkenleri arasındaki ilişkiler hakkında önceden bilgi sahibi olmadan modeller oluşturmak zorunda kalabilirler. Bu eksiklik, sinir ağları ve destek vektör makineleri gibi parametrik olmayan teknikler kullanılarak, doğrudan verilerden modeller oluşturularak giderilebilir. Modeller genellikle denetimli olarak oluşturulur, bu da hedef değişkenlerin önceden tanımlanması gerektiği anlamına gelir. Pratikte, genellikle güçlü tek bir model yerine, daha güçlü tahminler elde etmek için birçok zayıf modelin birleştirildiği topluluk yaklaşımları tercih edilir. Rastgele ormanlar (Breiman, 2001) ve sinir ağı toplulukları (Hansen ve Salamon, 1990), bu yaklaşımın başarılı örnekleridir (Liu vd., 2004; Shu ve Burn, 2004; Fanelli vd., 2012; Qi, 2012).

Rastgele ormanlar gibi yaygın topluluk teknikleri, topluluktaki modellerin basit ortalamasına dayanmaktadır. Güçlendirme yöntemleri ailesi, farklı, yapıcı bir topluluk oluşturma stratejisini temel almaktadır. Boosting'in ana fikri, topluluğa sırayla yeni modeller eklemektir. Her bir iterasyonda, yeni bir zayıf, temel öğrenici model, o ana kadar öğrenilen tüm topluluğun hatasına göre eğitilmektedir. İlk öne çıkan boosting teknikleri tamamen algoritma güdümlü olmuş, bu da özelliklerinin ve performanslarının ayrıntılı analizini oldukça zorlaştırmıştır (Schapire, 2002). Bu durum, bu algoritmaların neden diğer tüm yöntemlerden daha iyi performans gösterdiğine ya da tam tersine ciddi aşırı uyum nedeniyle uygulanamaz olduğuna dair bir dizi spekülasyona yol açmıştır (Sewell, 2011).

Bu doğrultuda, istatistiksel çerçeve ile bağlantı kurmak için, artırma yöntemlerinin gradyan iniş tabanlı bir formülasyonu türetilmiştir (Freund ve Schapire, 1997; Friedman vd., 2000; Friedman, 2001). Boosting yöntemlerinin bu formülasyonu ve ilgili modeller, gradient boosting makineleri olarak adlandırılmıştır. Bu çerçeve aynı zamanda model hiperparametrelerinin temel gerekçelerini sağlamış ve daha fazla gradient boosting (gradyan artırma) modeli geliştirme için metodolojik temel oluşturmuştur. Gradyan artırma makineleri (GBM) ile öğrenme süreci, ardışık yeni modeller ekleyerek yanıt değişkeninin daha kesin tahminlerini elde etmeyi hedeflemektedir. Bu yöntemin temeli, her yeni temel öğrenicinin, topluluğun genel kayıp fonksiyonunun negatif gradyanına en uygun şekilde eşleşecek biçimde oluşturulmasıdır. Kayıp fonksiyonları çeşitli olabilir; örneğin, klasik karesel hata kaybı kullanıldığında, süreç ardışık hata düzeltmeye yönelik olur. GBM'lerin yüksek derecede özelleştirilebilir yapısı, model tasarımında önemli bir özgürlük sağlamakla birlikte çeşitli uygulama ve araştırma alanlarında etkili çözümler sunmaktadır

(Bissacco vd., 2007; Hutchinson vd., 2011; Pittman ve Brown, 2011; Johnson ve Zhang, 2012).

Gradient Boosting sınıflandırma algoritmasının temel adımları, genel olarak aşağıdaki matematiksel formüllerle açıklanabilir:

Başlangıç Tahmini: İlk adımda, tüm gözlemler için sabit bir başlangıç tahmini yapılmaktadır. Bu, genellikle hedef değişkenin ortalaması olabilir:

$$F_0(x) = \arg min_\gamma \sum_{i=1}^{N} L(y_i, \gamma) \quad (5)$$

Negatif Gradyan Hesaplama: Her iterasyon m için, gerçek değerler ile mevcut tahminler arasındaki kaybın negatif gradyanı hesaplanmaktadır. Bu, modelin hatalarını belirlemektedir:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad (6)$$

Zayıf Öğrenici Fit Etme: Her iterasyonda, hesaplanan negatif gradyanlara (hatalara) en iyi uyan bir zayıf öğrenici (genellikle bir karar ağacı) fit edilir:

$$h_m(x) = \text{Fit model to } r_{im} \quad (7)$$

Adım Boyutu (Öğrenme Oranı) Belirleme: Zayıf öğrenicinin katkısını ayarlamak için bir adım boyutu (α) kullanılır. Bu, modelin her adımda ne kadar "öğreneceğini" kontrol etmektedir.

Model Güncelleme: Model, her iterasyonda zayıf öğrenicinin katkısıyla güncellenmektedir. Böylece, modelin hatalarından öğrenmesi sağlanmaktadır:

$$F_m(x) = F_{m-1}(x) + a \cdot h_m(x) \quad (8)$$

Durma Kriteri: Belirlenen iterasyon sayısına ulaşılana veya başka bir durma kriteri karşılanana kadar adımlar tekrarlanmaktadır. Bu süreç, hedef değişkenin tahmininde kullanılan kümülatif bir model oluşturur. Gradient Boosting, karmaşık tahmin problemlerinde yüksek performans gösteren güçlü ve esnek bir algoritmadır.

3.2. Uzun kısa süreli bellek yöntemi (Long Short term memory method)

LSTM, zaman serisi verileri veya sıralı verilerle (dizi verileri) çalışmak üzere geliştirilmiş bir tür yapay sinir ağıdır. LSTM, özellikle uzun vadeli bağımlılıkları ve karmaşık kalıpları öğrenme yeteneğiyle tanınmaktadır. RNN'lerin (Recurrent Neural Networks - Tekrarlayan Sinir Ağları) geliştirilmiş bir versiyonudur ve zaman içindeki bilgi akışını korumak ve unutmak için "kapılar" kullanarak bilgi kaybını ve gradyan sorunlarını önlemektedir (Vidya ve Hari, 2023).

LSTM, geleneksel RNN'lerden farklı olarak, uzun vadeli bağımlılıkları daha iyi öğrenebilmek için hücre durumu (cell state) ve çeşitli kapılar (gates) kullanmaktadır. Bu kapılar, hücre durumunun ne kadarının güncellenip ne kadarının tutulacağını belirlemektedir. Giriş kapısı, yeni gelen bilginin hücre durumuna ne kadar ekleneceğini kontrol etmektedir. Unutma kapısı, hücre durumundaki bilginin ne kadarının unutulacağını belirlemektedir. Çıkış kapısı ise hücre durumunun hangi kısmının çıkışa aktarılacağını kontrol etmektedir (Dai vd., 2020). Bu kapılar, LSTM'nin sıralı verilere dayanarak kararlar almasını sağlar ve sınıflandırma görevlerinde verinin zamansal özelliklerini dikkate alarak tahminler yaparlar. LSTM'nin sınıflandırma işlemlerinde çalışma şekline bakıldığında, öncelikle LSTM katmanına, sıralı veri girilir. Her veri noktası, belirli bir zamanda (time step) olan veriyi temsil etmektedir. Ardından, LSTM katmanları, zaman içindeki bağımlılıkları ve kalıpları öğrenir. LSTM'den elde edilen özellikler, sınıflandırma katmanına (genellikle dense veya fully connected layer) aktarılır ve son olarak softmax veya sigmoid aktivasyon fonksiyonu ile sınıf olasılıkları hesaplanır (Jongjaraunsuk vd., 2024). LSTM'deki her bir kapının işleyişi Denklem 9-14 'te gösterildiği gibi formüle edilmektedir.

Unutma Kapısı:

$$f_t = \sigma(W_f \cdot [h_{t-1}, \chi_t] + b_f) \quad (9)$$

Burada, f_t, unutma kapısı çıkışıdır. σ, sigmoid aktivasyon fonksiyonudur. W_f ve b_f, ağırlık matrisi ve bias terimidir. h_{t-1}, önceki gizli durumdur. χ_t ise şu anki giriştir.

Giriş Kapısı ve Giriş Adayı:

$$i_t = \sigma(W_i \cdot [h_{t-1}, \chi_t] + b_i) \quad (10)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, \chi_t] + b_C) \quad (11)$$

Burada, i_t, giriş kapısının çıkışıdır. $\tilde{C}_t$, yeni bilgi adayıdır. W_i, W_C ve b_i, b_C, ilgili ağırlıklar ve bias terimleridir.

Hücre Durumu Güncellemesi:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (12)$$

Burada, C_t, güncellenmiş hücre durumudur. C_{t-1}, önceki hücre durumudur.

Çıkış Kapısı ve Gizli Durum:

$$o_t = \sigma(W_o \cdot [h_{t-1}, \chi_t] + b_o) \quad (13)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (14)$$

Burada, o_t, çıkış kapısının çıkışıdır. h_t, güncellenmiş gizli durumdur ve bu, sonraki katmanlara aktarılır. Son olarak, h_t, sınıflandırma katmanına girer ve sınıf tahminleri yapılır.

3.3. Konvolüsyonel sinir ağları yöntemi (Convolutional neural networks method)

CNN, özellikle görüntü işleme ve sınıflandırma gibi alanlarda yaygın olarak kullanılan bir derin öğrenme modelidir. CNN'ler, verilerin mekânsal ve zamansal ilişkilerini anlamak ve yakalamak için katmanlar aracılığıyla özellik çıkarımı yapmaktadırlar. Görüntü sınıflandırma, nesne tanıma ve ses analizi gibi birçok sıralı ve yapısal veriyle çalışabilirler (Maggiori vd., 2017). CNN, genellikle üç ana katman tipinden oluşmaktadır. Bunlardan ilki, evrişim katmanıdır (convolutional layer). Bu katman, girdi verisine küçük filtreler (kernels) uygular. Bu filtreler, veri üzerinde kaydırılarak (stride) özellikler çıkarılır. Çıkan özellik haritaları, girişin mekânsal özelliklerini koruyarak daha derin bir seviyede temsil edilmesini sağlar. İkincisi, havuzlama katmanıdır (pooling layer). Bu katman, evrişim katmanından gelen özellik haritalarının boyutunu azaltmak ve işlem maliyetini düşürmek için kullanılır. Genellikle "Max Pooling" veya "Average Pooling" gibi yöntemler kullanılır. Max Pooling, belirli bir alan içindeki en yüksek değeri alarak boyutu küçültür. Sonuncu ana katman ise tam bağlantılı katmandır (fully connected layer). Bu katman, özellik haritalarını düzleştirerek (flatten) klasik yapay sinir ağına benzer şekilde tüm nöronları birbirine bağlar. Son olarak, çıkış katmanında softmax veya sigmoid aktivasyon fonksiyonu kullanılarak sınıflandırma yapılır (Chan ve Fan, 2022).

Bir CNN, sınıflandırma işlemlerinde belirli adımları takip ederek çalışır. İlk olarak, model bir girdi alır; bu genellikle bir görüntü veya sıralı veri şeklindedir. Ardından, evrişim ve havuzlama katmanları devreye girerek girdi verisinden özellikler çıkarır. Bu katmanlar, verinin mekânsal ve yapısal ilişkilerini kullanarak özellikleri giderek daha soyut bir şekilde temsil eder. Daha sonra, bu özellikler tam bağlantılı katmanlara iletilir ve burada işlenerek sınıf olasılıkları hesaplanır. Son adımda ise, çıktı katmanında her bir sınıf için olasılıklar verilir ve en yüksek olasılığa sahip sınıf, modelin tahmini olarak belirlenir. Bu süreç, CNN'nin veriyi işleyip doğru sınıflandırmayı yapmasını sağlar (Dubey vd., 2023). Bu sürecin matematiksel formülasyonu Denklem 15-19'da gösterilmektedir.

Evrişim Katmanı:

$$Z_{i,j}^{(k)} = (W^{(k)} * X)_{i,j} + b^{(k)} \tag{15}$$

Burada, $Z_{i,j}^{(k)}$, k filtresi için evrişim sonucundaki çıktı değeridir. $W^{(k)}$, k filtresinin ağırlık matrisi (kernel) ve $b^{(k)}$ ise bias terimidir. X, giriş verisidir ve $*$, evrişim işlemini ifade etmektedir.

Aktivasyon Fonksiyonu:

$$A_{i,j}^{(k)} = RELU\left(Z_{i,j}^{(k)}\right) = \max(0, Z_{i,j}^{(k)}) \tag{16}$$

Burada, RELU (Rectified Linear Unit) aktivasyon fonksiyonu kullanılmaktadır.

Havuzlama Katmanı (Max Pooling):

$$P_{i,j}^{(k)} = \max\left(A_{i,j}^{(k)}\right) \tag{17}$$

Burada, havuzlama işlemi belirli bir alan içinde en yüksek değeri alır ve böylece çıktı boyutunu küçültür.

Tam Bağlantılı Katman ve Çıkış:

$$Z = W \cdot A + b \tag{18}$$

Burada, Z, tam bağlantılı katmandaki çıktı vektörüdür. W, ağırlık matrisi ve b bias terimidir. A, düzleştirilmiş özelliklerdir.

Sınıflandırma Çıkışı:

$$P(y = j|x) = \frac{e^{Z_j}}{\sum_{k=1}^{K} e^{Z_k}} \tag{19}$$

3.4. Tabular öğrenme ağı (Tabular learning network-TabNet)

TabNet, tabular veri üzerinde çalışan ve özellikle sınıflandırma ve regresyon görevlerinde başarılı olan bir derin öğrenme modelidir. TabNet, dikkat (attention) mekanizmasını kullanarak önemli özellikleri otomatik olarak seçer ve öğrenme sürecini yönlendirir. Model, her adımda, dikkat mekanizmasını kullanarak hangi özelliklerin daha önemli olduğunu belirlemekte ve bu özelliklere odaklanarak öğrenme sürecini optimize etmektedir. Diğer derin öğrenme modellerinden farklı olarak, TabNet tabular verilerle çalışırken açıklanabilirlik (explainability) ve verimli öğrenme özelliklerini bir araya getirmektedir (Albin Ahmed vd., 2023).

Modelde, veriler, giriş olarak alınır ve doğrusal katmanlar (fully connected layers) ile işlenir. Her giriş özelliği, doğrusal bir dönüşüme tabi tutulur. Model, hangi özelliklerin o adımda önemli olduğunu belirlemek için bir dikkat ağı (attention network) kullanır. Bu katmanlar, her adımda farklı özelliklere odaklanarak bilgiyi verimli bir şekilde öğrenir ve diğer adımlara aktarır. Dikkat mekanizması, verinin bir kısmını maskeler ve geri kalan kısmı seçerek bilgi kaybını minimize eder. Bu sayede, her adımda yeni ve önemli özellikler modele dahil edilir. Dikkat katmanlarından gelen bilgiler, karar katmanlarında işlenir ve bu katmanlar, sınıflandırma veya regresyon görevleri için nihai tahminleri oluşturur (Gao vd., 2022). Bu sürecin matematiksel formülasyonu Denklem 20-23'te gösterilmektedir.

Girdi Dönüşümü:

$$x_t' = W_t \cdot x + b_t \tag{20}$$

Burada, x_t', girişin doğrusal dönüşümünden elde edilen çıktıdır. W_t ve b_t, ağırlık ve bias terimleridir.

Dikkat Mekanizması:

$$M_t = Softmax(W_m \cdot x'_t + b_m) \quad (21)$$

Burada, M_t, o adımda kullanılan maskedir. W_m ve b_m, ağırlık ve bias terimleridir.

Özellik Seçimi ve Maskeleme:

$$x_{t+1} = M_t \odot x'_t \quad (22)$$

Burada, x_{t+1}, maskelenmiş ve seçilmiş özelliklerdir. $\odot$, eleman bazlı çarpma işlemini ifade etmektedir.

Karar Adımları:

$$y_t = f(x_{t+1}) \quad (23)$$

Burada, y_t, her adımda yapılan tahmindir. f, karar adımındaki doğrusal dönüşümler ve aktivasyon fonksiyonları ile ifade edilen bir fonksiyondur. Son aşamada, her karar adımında elde edilen bilgiler birleştirilerek nihai sınıf olasılıkları hesaplanır.

3.5. Yapay sinir ağı-YSA (Artificial neural network-ANN)

YSA, insan beynindeki nöronların çalışma prensiplerinden ilham alınarak geliştirilmiş bir makine öğrenmesi modelidir. YSA'lar, katmanlar halinde düzenlenmiş nöronlardan (node) oluşur ve bu nöronlar, veriler arasındaki karmaşık ilişkileri öğrenerek çeşitli görevleri (sınıflandırma, regresyon, vb.) gerçekleştirebilirler. Görüntü tanıma, metin analizi ve zaman serisi tahmini gibi birçok farklı alanda kullanılabilirler (Huang vd., 2020).

YSA, genellikle üç ana katman tipinden oluşmaktadır: Girdi, gizli ve çıkış katmanı. Girdi katmanı, modelin aldığı ham veriyi temsil etmektedir. Her bir girdi, bu katmandaki bir nöron tarafından işlenir. Girdi katmanından gelen veriler, gizli katmanlarda işlenir. Gizli katmanlar, veriler arasındaki karmaşık ilişkileri öğrenir. Bu katmanlar, doğrusal olmayan (non-linear) aktivasyon fonksiyonları ile donatılmıştır ve bu sayede veriler arasındaki karmaşık kalıpları yakalayabilir (Al-Shamisi vd., 2013). Gizli katmanlardan gelen işlenmiş bilgiler, çıkış katmanında toplanır ve sınıflandırma (veya başka bir görev) için son tahminler yapılır. Çıkış katmanındaki nöron sayısı, sınıflandırılacak sınıf sayısına eşittir. YSA ile sınıflandırma işlemi sürecinde, girdi verileri, giriş katmanına aktarılır ve bu katmandaki nöronlar, veriyi modelin içine alır. Gizli katmanlarda, her nöron bir önceki katmandan gelen veriyi ağırlıklarla çarpar ve bir bias değeri ekleyerek bir toplam elde eder. Bu toplam, aktivasyon fonksiyonu ile işlenir ve doğrusal olmayan bir çıkış üretilir. Bu işlem, katmanlar arasında devam eder. Çıkış katmanına ulaşıldığında, model her sınıf için olasılıklar üretir ve en yüksek olasılığa sahip sınıf modelin tahmini olarak belirlenir (Arkin vd., 2020). Bu sürecin matematiksel formülasyonları Denklem 24-26'da gösterilmiştir.

Ağırlıklı Toplama:

$$z = \sum_{i=1}^{n} w_i x_i + b \quad (24)$$

Burada, z, nöronun toplam ağırlıklı girdisidir. w_i, i girdisinin ağırlığıdır. x_i, i girdisinin kendisidir. b, nöronun bias (sapma) terimidir.

Aktivasyon Fonksiyonu:

$$\alpha = Activation(z) \quad (25)$$

z, değeri bir aktivasyon fonksiyonuna (sigmoid, ReLU vd.) uygulanarak doğrusal olmayan bir dönüşüm yapılır. Burada, α, nöronun aktivasyon çıkışıdır.

Çıkış Katmanı:

$$P(y = j|x) = \frac{e^{Z_j}}{\sum_{k=1}^{K} e^{Z_k}} \quad (26)$$

Burada, $P(y = j|x)$, girişin j sınıfına ait olma olasılığıdır. Z_j, j sınıfı için çıkış katmanından gelen toplamdır. Bu süreç, YSA'nın her nöronunun çıktısını hesaplamak için tekrar edilir ve en sonunda, modelin verdiği en yüksek olasılıklı sınıf, tahmin edilen sınıf olarak belirlenir.

3.6. Rastgele orman yöntemi (Random forest method)

Rastgele Orman, sınıflandırma ve regresyon görevlerinde kullanılan, birden fazla karar ağacının (decision tree) bir araya getirilmesiyle oluşturulan bir makine öğrenmesi modelidir. Temel mantığı, birçok karar ağacını eğiterek bu ağaçların tahminlerini birleştirip (çoğunlukla oylama veya ortalama alma yöntemi ile) daha doğru ve genelleştirilebilir sonuçlar elde etmektir. Bu yöntem, karar ağaçlarının zayıf yönlerini azaltarak daha sağlam ve etkili bir model oluşturur (Liu vd., 2023). Sınıflandırma sürecinde, model, eğitim veri setinden rastgele örnekler seçerek (bootstrap yöntemi) her bir ağaç için farklı alt kümeler oluşturur. Bu yöntem, ormanın her ağacının farklı veri örnekleriyle eğitilmesini sağlar ve çeşitliliği artırır. Her bir ağaç, her düğümde tüm özellikler yerine rastgele bir alt küme özellik kullanır. Bu sayede, her ağaç farklı bir özellik kombinasyonu kullanarak eğitilir ve bu, ağaçlar arasında çeşitliliği artırarak aşırı uyumu (overfitting) önler. Rastgele seçilen veri örnekleri ve özellikler kullanılarak her ağaç eğitilir. Her ağaç, kendi veri kümesi üzerinde en iyi bölünmeyi bulmak için çalışır ve bir karar ağacı oluşturur. Model, yeni bir veri örneği geldiğinde, ormandaki her ağaç bu veri örneği için bir sınıf tahmini yapar. Sınıflandırma için, en çok oyu alan sınıf nihai tahmin olarak kabul edilir (Anjum vd., 2023). Bu sürecin, matematiksel formülasyonu Denklem 27-30'da gösterilmektedir.

Bootstrap Yöntemi:

$$D_i \subset D \quad (27)$$

Burada, D_i, i. karar ağacı için seçilen veri alt kümesidir.

Özellik Alt Kümesi Seçimi:

$$F \subseteq \{X_1, X_2, \ldots, X_M\} \quad (28)$$

Burada, her düğümde, toplam M özellik arasından rastgele m sayıda özellik seçilir. F, seçilen özellik alt kümesidir ve $m \ll M$.

Karar Ağaçlarının Eğitimi:

$$t^* = arg\,max\,(Gain(t)) \quad (29)$$

Burada, her bir karar ağacı, kendi veri kümesi ve seçilen özellikler kullanılarak eğitilir. $Gain(t)$, düğümdeki bilgi kazancını ifade etmektedir (bilgi kazancı veya Gini indeksi).

Tahmin ve Oylama:

$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B) \quad (30)$$

Burada, $\hat{y}$, rastgele ormanın nihai tahminidir. $\hat{y}_i$, her bir ağacın yaptığı tahmindir ve B, ormandaki ağaç sayısıdır. Sınıflandırmada, tüm ağaçların tahminleri alınır ve en sık tekrar eden sınıf seçilir (mode).

3.7. Destek vektör makinesi – DVM (Support vector machine- SVM)

DVM, sınıflandırma ve regresyon gibi makine öğrenmesi görevlerinde kullanılan, veri noktalarını sınıflandırmak için hiper düzlemler oluşturan bir algoritmadır. DVM, doğrusal ve doğrusal olmayan veri kümeleri için çalışabilen, özellikle yüksek boyutlu veri setlerinde etkili sonuçlar veren bir modeldir. Temel amacı, veri noktalarını en iyi şekilde ayıracak ve sınıflar arasındaki mesafeyi maksimize edecek bir hiper düzlem bulmaktır (An ve Liang, 2012).

DVM, veri noktalarını iki farklı sınıf arasında ayıran bir hiper düzlem oluşturur. Bu hiper düzlem, sınıflar arasındaki maksimum marjini (margin) sağlayacak şekilde yerleştirilir. Model, ayırıcı hiper düzleme en yakın olan veri noktalarına (destek vektörleri) dayanarak bu hiper düzlemi optimize etmektedir (Lee vd., 2016). DVM, veri noktalarını özellik uzayında temsil eder. Her bir veri noktası, bu uzayda bir vektör olarak ifade edilir. Algoritma, veri noktalarını ayıracak bir hiper düzlem bulmaya çalışır. Bu hiper düzlem, veri noktalarını doğru şekilde sınıflandıracak ve marjini maksimum yapacak şekilde optimize edilir. Hiper düzleme en yakın olan ve bu düzlemin oluşturulmasında en etkili olan veri noktaları, destek vektörleri olarak adlandırılır. Bu noktalar, hiper düzlemin konumunu ve yönünü belirler. Model, hiper düzleme dayalı olarak her yeni veri noktası için bir karar fonksiyonu kullanarak tahmin yapar. Veri noktası bu fonksiyona göre sınıflandırılır (Xi vd., 2017). Tüm bu süreç, matematiksel formülasyonları ile birlikte Denklem 31-35'te gösterilmektedir.

Hiper Düzlem Denklemi:

$$w \cdot x + b = 0 \quad (31)$$

Burada, w, hiper düzlemin normal vektörüdür. x, veri noktası ve b, bias terimidir.

Marjin Maksimizasyonu:

$$minimize\ \frac{1}{2}\|w\|^2 \quad (32)$$

İki sınıf arasındaki mesafeyi maksimize etmek için hiper düzlem Denklem 32'de gösterildiği gibi optimize edilmektedir. Bu optimizasyon, w vektörünün normunu minimize etmeyi amaçlar ve böylece maksimum marjin elde edilir.

Kısıt Koşulları:

$$y_i(w \cdot x_i + b) \geq 1\ \forall_i \quad (33)$$

Hiper düzlem, her sınıf için veri noktalarının doğru tarafta olmasını sağlayacak şekilde Denklem 33'teki gibi düzenlenir. Burada, y_i, veri noktasının sınıf etiketidir (+1 veya -1). x_i, veri noktasıdır. DVM, marjini maksimize ederken, genellikle dual problem kullanılarak çözülür. Bu aşama Denklem 34'te gösterilmektedir.

Lagrange Çarpanları ve Dual Problem:

$$L(w, b, \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (34)$$

Burada, α_i, Lagrange çarpanlarıdır ve her veri noktası için optimize edilir. Eğitim tamamlandıktan sonra, yeni bir veri noktası x için karar fonksiyonu Denklem 35'teki gibidir.

Karar Fonksiyonu:

$$f(x) = \text{sign}(w \cdot x + b) \quad (35)$$

Burada, sign fonksiyonu, x noktasının hangi sınıfa ait olduğunu belirlemektedir.

4. Bulgular (Findings)

Bu çalışmada, sınıflandırma işlemi için kullanılan tüm makine öğrenimi yöntemlerinin optimal hiperparametrelerinin belirlenmesinde Optuna kütüphanesi kullanılmıştır. Optuna, hiperparametre optimizasyonu için bir otomatikleştirme kütüphanesidir ve makine öğrenimi modellerinin performansını artırmak için en iyi hiperparametreleri bulmayı amaçlamaktadır. Sınıflandırma işlemi için YSA, LSTM, CNN, TabNet, Gradient Boosting, Rastgele Orman (RF) ve DVM modellerinin optimal hiperparametrelerinin belirlenmesinde Python programlama dili kullanılarak Optuna kütüphanesi kapsamlı bir şekilde uygulanmıştır. Optuna, her modelin özel gereksinimlerine uygun olacak şekilde arama alanları (search spaces) tanımlayarak hiperparametre optimizasyonu sağlamıştır. YSA, LSTM ve CNN modellerinde TensorFlow ve PyTorch kütüphaneleri kullanılarak ağ derinliği (katman sayısı), nöron sayısı, öğrenme oranı (learning rate), aktivasyon fonksiyonları ve batch size

gibi parametreler optimize edilmiştir. TabNet için ise, Optuna kullanılarak TabNet'in dikkat (attention) katmanlarının sayısı, öğrenme oranı ve maskeleme oranı gibi spesifik hiperparametreler ayarlanmıştır. Gradient Boosting (GB) modelinde, learning rate, max_depth ve n_estimators gibi ağaç tabanlı yapıların performansını artıran parametreler optimize edilmiştir. Rastgele Orman modelinde, ağaç sayısı (n_estimators), maksimum derinlik (max_depth) vd. gibi hiperparametreler belirlenmiştir. DVM modeli için ise, kernel tipi (linear, rbf vb.), düzenleme parametresi (C), ve kernel parametreleri (gamma) gibi ayarlar Optuna aracılığıyla titizlikle optimize edilmiştir. Bu sayede, her modelin performansı en üst düzeye çıkarılarak sınıflandırma görevinde yüksek doğruluk oranları elde edilmiştir. Sonuç olarak, bu optimizasyon süreci sonunda elde edilen tüm hiperparametreler Tablo 2'de gösterilmiştir.

Tablo 2. Modellerin optuna ile belirlenen optimal hiperparametreleri (Optimal hyperparameters of the models determined by optuna)

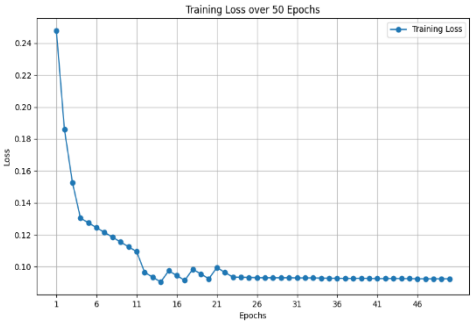
Model	Hiperparametreler	Değer
YSA	n_layers	3
	Units_L0	127
	Activation_L0	ReLU
	Units_L1	55
	Activation_L1	tanh
	Units_L2	61
	Activation_L2	ReLU
	Learning Rate	0.0037
	Batch Size	108
LSTM	lstm_units	76
	n_layers	1
	Units_L0	75
	Learning Rate	2.162e-05
	Batch Size	128
CNN	n_filters	24
	Kernel size	2
	n_layers	2
	Units_L0	119
	Units_L1	57
	Learning Rate	0.0006
	Batch Size	74
TabNet	n_d	16
	n_a	16
	n_steps	5
	gamma	1.5
	Lambda_sparse	0.0001
	Optimizer_fn	Torch.Optim.Adam
	Optimizer_params	"lr": 1e-2
	Mask_type	sparsemax
	n_independent	2
	n_shared	2
	virtual_batch_size	128
	Momentum	0.02
	Clip_value	2.0
	Scheduler_fn	torch.optim.lr_scheduler.StepLR
	Scheduler_params	"step_size": 10, "gamma": 0.1
	Epsilon	1e-15
GBM	n_estimators	120
	Learning Rate	0.1373
	max_depth	9
Rastgele Orman	n_estimators	272
	min_samples_split	5
	max_depth	9
	min_samples_leaf	8
	Bootstrap	True
DVM	Kernel	Linear
	C	0.0959
	Gamma	0.0012

Tablo 3'te yer alan sınıflandırma raporuna göre, kronik hastalığı olan bireyleri (0 Sınıfı) ve olmayan bireyleri (1 Sınıfı) sınıflandırmak üzere kullanılan modellerin test performansları karşılaştırılmıştır. YSA modelinde, 0 sınıfı için F1 puanı 0,92 olarak hesaplanmış olup, duyarlılık değeri 0,93 ve kesinlik değeri 0,92'dir. 1 sınıfı için ise F1 puanı 0,96, duyarlılık 0,96 ve kesinlik 0,95 olarak gözlenmiştir. LSTM modeli benzer sonuçlar vermekle birlikte, 0 sınıfı için F1 puanı 0,92 ve 1 sınıfı için 0,95 olarak hesaplanmıştır. CNN modeli, her iki sınıfta da yüksek performans sergileyerek 0 sınıfında F1 puanı 0,94 ve 1 sınıfında 0,96'ya ulaşmıştır. TabNet modeli ise en yüksek doğruluğa ulaşarak, 0 sınıfında F1 puanını 0,96, 1 sınıfında ise 0,97 olarak vermiştir. GBM ve RF modelleri nispeten daha düşük performans göstermiş, her iki model de 0 sınıfı için 0,90 civarında bir F1 puanı elde etmiş, 1 sınıfı için ise 0,95 seviyesine çıkmıştır. DVM ise genel olarak daha düşük performans sergilemiş olup, 0 sınıfında F1 puanı 0,87, 1 sınıfında ise 0,93 olarak gözlenmiştir. Genel olarak, TabNet ve CNN modelleri tüm sınıflarda daha yüksek doğruluk ve sınıflandırma performansı sergilerken, GB ve RF modelleri daha ortalama bir performans göstermiştir. DVM ise kronik hastalığı olan bireylerin sınıflandırılmasında daha düşük bir performans sergilemiştir.

Şekil 1, en başarılı sınıflandırma performansını gösteren TabNet modelinin eğitim sürecindeki hatanın iterasyon sayısına bağlı olarak nasıl değiştiğini gösteren bir eğriyi göstermektedir. Grafikte gözlemlendiği üzere, eğitim süreci başlangıçta yüksek bir kayıp değeriyle (yaklaşık 0,24) başlamış ve ilk birkaç epoch boyunca kayıpta hızlı bir düşüş gerçekleşmiştir. İlk 10 epoch sonrasında, kayıp değeri önemli ölçüde azalarak 0,10 seviyelerine ulaşmıştır. Bu noktadan itibaren, kayıpta önemli bir değişim görülmemekte ve modelin performansı stabil hale gelmiştir. Özellikle 20. epoch'tan sonra kaybın sabitlenmesi, modelin daha fazla öğrenme sağlayamadığını ve büyük ölçüde yakınsadığını (converge) göstermektedir. Grafik, modelin eğitim süreci boyunca etkin bir şekilde optimize olduğunu ve aşırı öğrenme (overfitting) belirtisi göstermeden kayıp fonksiyonunu minimize ettiğini göstermektedir.

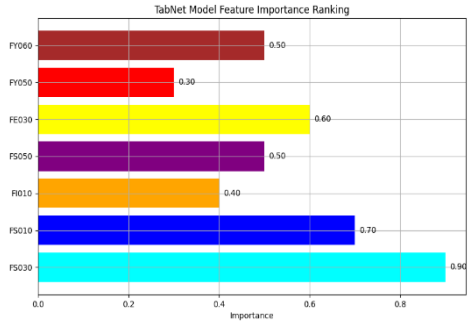
Tablo 3. Modellerin test sonuçlarına ilişkin sınıflandırma raporu (Classification report on the test results of the models)

Model		Kesinlik	Duyarlılık	F1 Puanı	Destek
YSA	0	0,93	0,92	0,92	4.492
	1	0,95	0,96	0,96	8.430
	Doğ.			0,95	12.922
	Genel Ort.	0,94	0,94	0,94	12.922
	Ağır. Ort.	0,95	0,95	0,95	12.922
LSTM	0	0,93	0,91	0,92	4.492
	1	0,95	0,96	0,95	8.430
	Doğ.			0,94	12.922
	Genel Ort.	0,94	0,94	0,94	12.922
	Ağır. Ort.	0,94	0,94	0,94	12.922
CNN	0	0,95	0,93	0,94	4.492
	1	0,96	0,97	0,96	8.430
	Doğ.			0,96	12.922
	Genel Ort.	0,95	0,95	0,95	12.922
	Ağır. Ort.	0,96	0,96	0,96	12.922
TabNet	0	0,96	0,95	0,96	4.492
	1	0,97	0,97	0,97	8.430
	Doğ.			0,97	12.922
	Genel Ort.	0,97	0,97	0,97	12.922
	Ağır. Ort.	0,97	0,97	0,97	12.922
GBM	0	0,90	0,90	0,90	4.492
	1	0,94	0,95	0,95	8.430
	Doğ.			0,93	12.922
	Genel Ort.	0,92	0,92	0,92	12.922
	Ağır. Ort.	0,93	0,93	0,93	12.922
RF	0	0,90	0,89	0,90	4.492
	1	0,94	0,95	0,95	8.430
	Doğ.			0,93	12.922
	Genel Ort.	0,92	0,92	0,92	12.922
	Ağır. Ort.	0,93	0,93	0,93	12.922
DVM	0	0,89	0,85	0,87	4.492
	1	0,91	0,94	0,93	8.430
	Doğ.			0,91	12.922
	Genel Ort.	0,90	0,90	0,90	12.922
	Ağır. Ort.	0,91	0,91	0,91	12.922



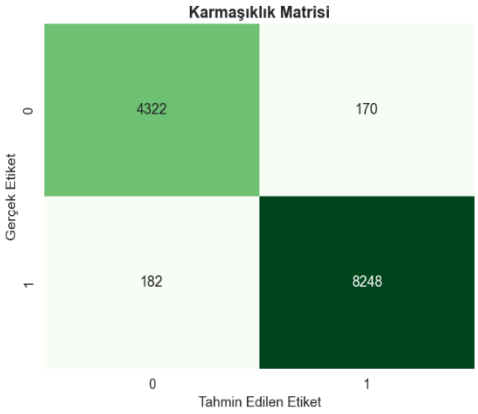
Şekil 1. TabNet modelinin eğitim hatası (Training error of the Tabnet model)

Şekil 2, TabNet modelindeki bağımsız değişkenlerin önem derecelerini gösteren bir sıralama çubuk grafiğidir. Burada, FS030, FS010, FI010, FS050, FE030, FY050 ve FY060 olarak gösterilen değişkenler (Tablo 1'de açıklanmıştır), modelin tahmin performansı üzerindeki etkilerine göre sıralanmıştır. FS030 olarak ifade edilen ve ferdin sağlık probleminden ötürü faaliyetlerde sınırlama olup olmadığını gösteren özellik model için açık ara en önemli değişken olarak belirtilmiş ve en yüksek öneme sahip olduğu gözlemlenmiştir. Bunu 0.70 önem derecesi ile FS010, ardından 0.60 önem derecesine sahip FE030 takip etmektedir. Diğer özellikler arasında FY060 ve FS050'nin her ikisi de 0.50 önem derecesi ile benzer öneme sahipken, FI010'ın önemi 0.40 olarak belirlenmiştir. FY050, 0.30 ile en düşük önem derecesine sahip özellik olarak sıralamada yer almaktadır. Bu sonuçlar, TabNet modelinin dikkat mekanizması aracılığıyla bazı özelliklere daha fazla ağırlık verdiğini ve bu özelliklerin modelin sınıflandırma performansı üzerindeki etkisinin farklı olduğunu göstermektedir. Özellikle FS030 ve FS010 özellikleri, modelin öğrenme sürecinde kritik bir rol oynamaktadır.



Şekil 2. TabNet modeline göre bağımsız değişkenlerin önem sıralaması (Importance ranking of independent variables)

Şekil 3'te, TabNet modelinin test performansını gösteren karmaşıklık matrisi yer almaktadır. "0" etiketi kronik bir hastalığı olan bireyleri, "1" etiketi ise hastalığı olmayan bireyleri temsil etmektedir. Kronik hastalığı olan bireyler için (0), model 4.322 doğru sınıflandırma (Gerçek Pozitif) yaparken, 170 bireyi yanlış bir şekilde kronik hastalığı yok olarak tahmin etmiştir (Yanlış Negatif). Kronik hastalığı olmayan bireyler için (1), model 8.248 bireyi doğru sınıflandırmış (Gerçek Negatif), ancak 182 bireyi yanlış bir şekilde kronik hastalığı var olarak sınıflandırmıştır (Yanlış Pozitif). Modelin genel performansına bakıldığında, hem kronik hastalığı olan bireyler hem de olmayan bireyler için yüksek doğruluk oranları elde ettiği görülmektedir.



Şekil 3. TabNet modelinin test sonuçlarına ilişkin karmaşıklık matrisi (Confusion matrix for the test results of the TabNet model)

Bu aşamadan sonra, modelin genel performansını değerlendirmek ve genelleme yeteneğini ölçmek için kullanılan yaygın bir doğrulama tekniği olan çapraz doğrulama yöntemi uygulanarak, modelin sadece tek bir eğitim ve test veri setiyle değil, farklı veri parçalarıyla eğitilip test edilmesi sağlanmıştır. Bu sayede, modelin aşırı uyum (overfitting) veya eksik uyum (underfitting) yapma olasılığı değerlendirilmiş ve modelin daha genelleştirilebilir sonuçlar üretebilme yeteneği daha doğru bir şekilde ölçülmüştür. Bu süreçte, veri seti 5 eşit parçaya bölünmüştür. Her bir iterasyonda, bu 5 parçadan biri test seti olarak ayrılırken geri kalan 4 parça modelin eğitimi için kullanılmıştır. Bu işlem 5 kez tekrarlanmış ve her seferinde farklı bir parça test seti olarak kullanılmıştır. Sonuç olarak, modelin performansı her bir iterasyonda hesaplanmış ve tüm iterasyonların ortalaması alınarak daha dengeli bir performans ölçümü elde edilmiştir. Bu sürece ilişkin sonuçlar, Tablo 4'te gösterilmiştir.

Tablo 4. TabNet modelinin 5 katlı çapraz doğrulama sonuçları (5-fold cross-validation results of the TabNet model)

Model	Ort.Doğruluk	Kat1	Kat2	Kat3	Kat4	Kat 5
TabNet	0,972	0.973	0,969	0,975	0,973	0,971

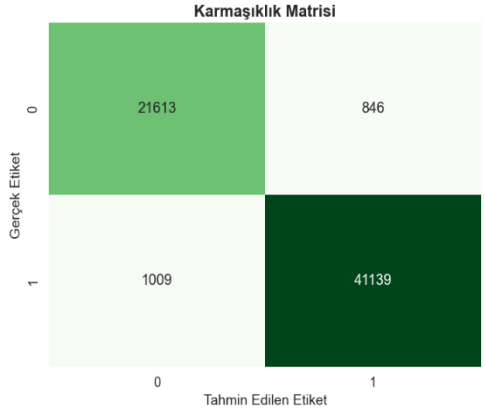
Tablo 4'teki sonuçlar incelendiğinde, her bir katmandaki doğruluk değerleri sırasıyla Kat 1: 0,973, Kat 2: 0,969, Kat 3: 0,975, Kat 4: 0,973 ve Kat 5: 0,971 olarak hesaplanmıştır. Bu değerlerin ortalaması ise 0,972 olarak belirlenmiştir. Sonuçlar, TabNet modelinin sınıflandırma performansının her bir doğrulama setinde oldukça istikrarlı olduğunu ve genelleme yeteneğinin

yüksek olduğunu göstermektedir. Farklı veri bölümlerinde modelin performansındaki küçük farklılıklar (0,969 ile 0,975 arasında değişen doğruluk oranları) modelin tutarlı olduğunu ve aşırı uyuma (overfitting) karşı dirençli olduğunu işaret etmektedir. Ortalama doğruluk oranı olan 0,972, modelin genel olarak veri setinde yüksek bir doğruluk sağladığını ve güvenilir bir performans sergilediğini ortaya koymaktadır. Bu sonuçlar, TabNet modelinin test edilmemiş veriler üzerinde de benzer şekilde başarılı olabileceğini öngörmektedir. TabNet modelinin tüm veri seti üzerinde tahmin performansının değerlendirilmesi, modelin genelleme yeteneğini daha kapsamlı bir şekilde incelemek açısından kritik bir adım olacaktır. 5 katlı çapraz doğrulama sonuçları modelin farklı veri bölümlerinde istikrarlı performans sergilediğini gösterse de, tüm veri seti üzerindeki tahmin performansının değerlendirilmesi, modelin gerçek dünyadaki veri kümeleri üzerindeki etkinliğini daha net ortaya koyacaktır. Bu değerlendirme, modelin yalnızca eğitim ve test verilerindeki başarısını değil, tüm veri seti ile ne kadar iyi genelleme yapabileceğini gösterecektir. Böylece, elde edilen performans ölçütleri (doğruluk, kesinlik, duyarlılık, F1 puanı vb.), modelin çeşitli veri yapıları karşısında sağlam ve güvenilir sonuçlar üretebilme kapasitesini doğrulayacaktır Bu aşama, modelin pratik uygulamalarda kullanılabilirliğini belirleyecek ve potansiyel zayıf yönlerini ortaya çıkararak olası iyileştirmeler için yol gösterecektir. Modelin tüm veri seti üzerindeki tahmin performasına ilişkin sınıflandırma raporu Tablo 5'te sunulmaktadır.

Tablo 5. TabNet modelinin tüm veri seti tahmin performansını gösteren sınıflandırma raporu (Classification report showing the prediction performance of the TabNet model across the entire dataset)

Model		Kesinlik	Duyarlılık	F1 Puanı	Destek
TabNet	0	0,96	0,95	0,95	22.459
	1	0,97	0,97	0,97	42.148
	Doğ.			0,97	64.607
	Genel Ort.	0,96	0,96	0,96	64.607
	Ağır. Ort.	0,97	0,97	0,97	64.607

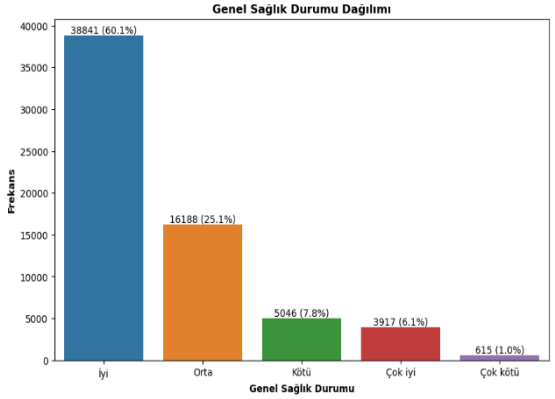
İlgili tablodaki sınıflandırma raporuna göre, kronik hastalığı olan bireyler (0) için modelin kesinlik değeri 0,96, duyarlılık değeri 0,95 ve F1 puanı 0,95 olarak hesaplanmıştır. Kronik hastalığı olmayan bireyler (1) için ise, modelin kesinlik ve duyarlılık değerleri 0,97 olup, F1 puanı da yine 0,97 olarak kaydedilmiştir. Modelin genel doğruluk oranı 0,97 olarak belirlenmiş ve modelin sınıflandırma performansı her iki sınıfta da oldukça yüksektir. Genel ortalamalar incelendiğinde, kesinlik, duyarlılık ve F1 puanlarının tüm veri seti için sırasıyla 0,96, 0,96 ve 0,96 olduğu görülmektedir. Ağırlıklı ortalama değerler de aynı doğrultuda, tüm metriklerde 0,97 seviyesinde olup, TabNet modelinin geniş bir veri seti üzerinde tutarlı ve güçlü bir performans sergilediğini göstermektedir. Bu sonuçlar, modelin sınıflar arasında dengeli bir performans sunduğunu ve hem kronik hastalığı olan hem de olmayan bireyleri yüksek doğrulukla sınıflandırabildiğini ortaya koymaktadır. Bununla birlikte, modelin bu sınıflandırma sürecine ilişkin karmaşıklık matrisi Şekil 4'te sunulmuştur.



Şekil 4. TabNet modelinin tüm veri seti tahmin performansını gösteren karmaşıklık matrisi (Confusion matrix showing the prediction performance of the TabNet model across the entire dataset)

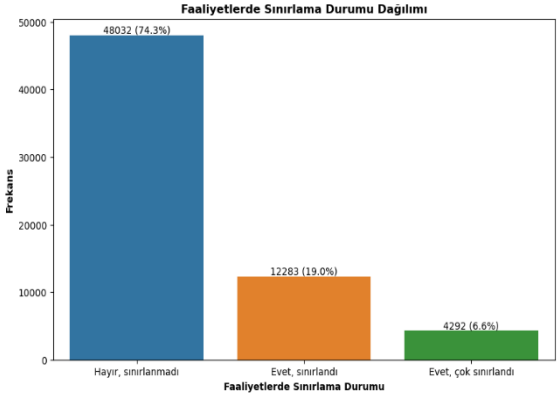
Gerçek etiketlerle karşılaştırıldığında, model kronik hastalığı olan bireyleri (0) 21.613 kez doğru sınıflandırmış (Gerçek Pozitif) ve 846 bireyi yanlış bir şekilde kronik hastalığı yok olarak tahmin etmiştir (Yanlış Negatif). Kronik hastalığı olmayan bireyler (1) için model 41.139 doğru sınıflandırma yapmış (Gerçek Negatif) ve 1.009 bireyi yanlış bir şekilde kronik hastalığı var olarak sınıflandırmıştır (Yanlış Pozitif). Modelin genel doğruluk oranı oldukça yüksek olup, her iki sınıfta da dengeli bir performans sergilediği gözlenmektedir. Ancak, yanlış negatif ve yanlış pozitif sayılarının varlığı, modelin sınıflandırma performansında bazı iyileştirme fırsatlarının olduğunu göstermektedir. Genel olarak, bu matriste TabNet modelinin kronik hastalığa sahip olan ve olmayan bireyleri yüksek doğrulukla sınıflandırdığı ve genelleme yeteneğinin güçlü olduğu anlaşılmaktadır.

Ek olarak, Şekil 5, TÜİK'in 2023 Gelir ve Yaşam Koşulları Araştırması (GYKA) verilerine dayanarak, Türkiye'deki bireylerin genel sağlık durumunun dağılımını göstermektedir. GYKA, Türkiye'de 64607 birey üzerinde uygulanmıştır. Araştırmaya katılan bireylerin büyük bir kısmı (%60.1) kendilerini "iyi" sağlık durumunda olarak tanımlarken, "orta" sağlık durumunda olanların oranı %25.1'dir. "Kötü" sağlık durumuna sahip bireyler %7.8, "çok iyi" durumda olanlar %6.1 ve "çok kötü" sağlık durumunda olanlar ise %1.0 olarak belirlenmiştir. Gradient Boosting sınıflandırma yöntemi kullanılarak yapılan bu analiz, nüfusun sağlık durumları üzerine değerli içgörüler sağlamakta ve sağlık politikalarının şekillendirilmesi için kullanılabilecek önemli bilgiler sunmaktadır.



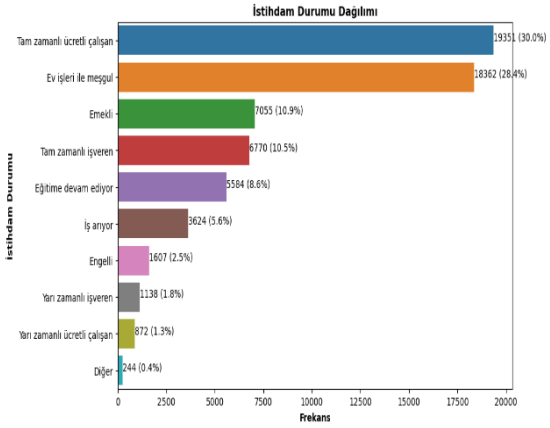
Şekil 5. Genel sağlık durumu dağılımı (Distribution of general health status)

Şekil 6, bireylerin günlük faaliyetlerindeki sınırlamaların dağılımını göstermektedir. Bireylerin büyük bir çoğunluğu (%74.3) günlük faaliyetlerinde herhangi bir sınırlamaya sahip olmadıklarını belirtirken, %19.0'lık bir kesim bazı sınırlamalar yaşadığını, %6.6'lık bir grup ise faaliyetlerinde ciddi sınırlamalar olduğunu ifade etmiştir. Bu veriler, nüfusun sağlıkla ilgili kısıtlamalarının kapsamlı bir resmini çizmekte ve bu konularda müdahale gerektirebilecek alanları belirlemekte kullanılabilir.



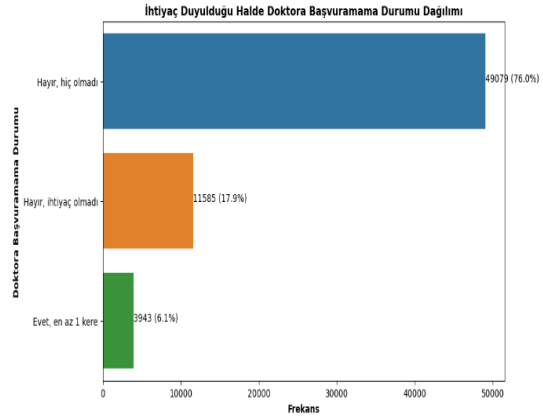
Şekil 6. Faaliyetlerde sınırlama durumu dağılımı (Distribution of restrictions on activities)

Buna ek olarak, Şekil 7'de, bireylerin istihdam durumu dağılımını gösteren bir histogram gösterilmektedir. En yüksek frekanslı kategori, tam zamanlı ücretli çalışanları temsil etmekte (%30.0), ardından ev işleri ile meşgul olanlar (%28.4) gelmektedir. Emekliler %10.9 ile üçüncü en büyük grubu oluştururken, tam zamanlı işverenler ve eğitime devam edenler sırasıyla %10.5 ve %8.6 ile takip etmektedir. İş arayanlar ve engelliler de sırasıyla %5.6 ve %2.5'lik bir orana sahiptir. Yarı zamanlı çalışanlar ve diğer kategoride yer alanlar ise toplamın daha küçük bir yüzdesini oluşturmaktadır. Bu veriler, istihdam piyasasının yapısı ve iş gücüne katılımın çeşitli yönleri hakkında önemli bilgiler sunmaktadır.



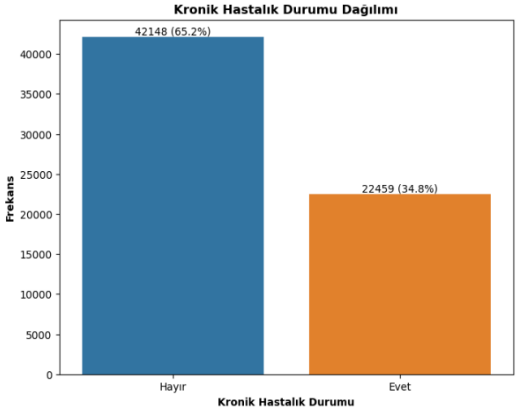
Şekil 7. İstihdam durumu dağılımı (Employment status distribution)

Şekil 8, bireylerin ihtiyaç duydukları halde doktora başvurup başvurmadıklarını gösteren bir dağılımı temsil etmektedir. %76.0'lık büyük bir çoğunluk, ihtiyaç duyduğunda doktora başvurmadığını söylememiş yani herhangi bir engelle karşılaşmadıklarını belirtmiştir. %17.9'luk bir kesim belirli durumlarda doktora başvuramama durumu yaşadığını ifade ederken, sadece %6.1'lik bir grup ihtiyaç duyduklarında en az bir kez doktora başvuramama durumu yaşadığını belirtmiştir. Bu veriler, sağlık hizmetlerine erişim konusunda genel olarak olumlu bir tablo çizerken, azınlıkta olan bir grubun karşılaştığı zorluklara dikkat çekmektedir.



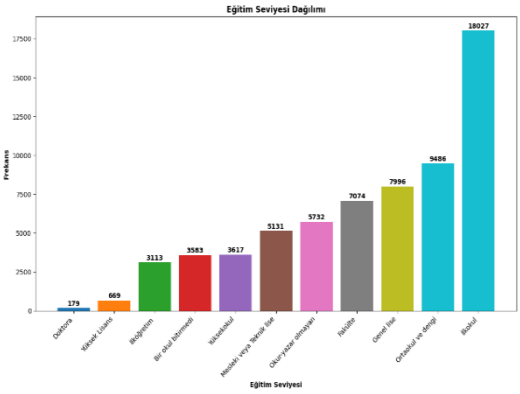
Şekil 8. İhtiyaç duyulduğu halde doktora başvuramama durumu dağılımı (Distribution of inability to consult a doctor when needed)

Şekil 9'da, bireylerin kronik hastalık durum dağılımı gösterilmektedir. Görselde görüldüğü üzere, katılımcıların %65.2'si herhangi bir kronik hastalığa sahip olmadıklarını belirtmiş, buna karşın %34.8'i kronik bir hastalığa sahip olduklarını ifade etmiştir. Bu oranlar, ülkedeki kronik sağlık sorunlarının yaygınlığını ve toplum sağlığına yönelik politika ve kaynak dağılımı için önemli bir veri noktasını temsil etmektedir.



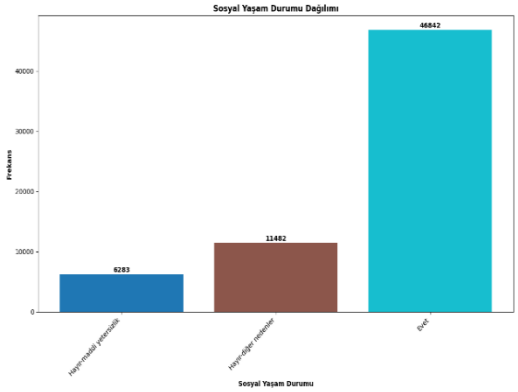
Şekil 9. Kronik hastalık durumu dağılımı (Chronic disease status distribution)

Şekil 10, bireylerin eğitim seviyesi dağılımını göstermektedir. En yüksek frekansa sahip eğitim seviyesi 18.027 birey ile "İlkokul" kategorisi olurken, bu grubu 9.486 birey ile "Ortaokul ve dengi" ve 7.996 birey ile "Genel lise" kategorileri takip etmektedir. "Fakülte" mezunlarının sayısı 7.074 iken, "Okur-yazar olmayan" bireyler 5.732, "Mesleki veya Teknik lise" mezunları ise 5.131 frekansa sahiptir. Daha düşük frekanslar ise "Yüksekokul" (3.617), "Bir okul bitirmedi" (3.583), ve "İlköğretim" (3.113) kategorilerinde gözlemlenmiştir. En az sayıda birey "Yüksek Lisans" (669) ve "Doktora" (179) seviyelerinde bulunmaktadır. Bu dağılım, veri setinde eğitim seviyesi bakımından bir yoğunlaşmanın ilkokul ve ortaokul seviyelerinde olduğunu, yükseköğrenim düzeylerinde ise nispeten düşük bir temsil olduğunu göstermektedir.



Şekil 10. Eğitim seviyesi durumu dağılımı (Distribution of educational status)

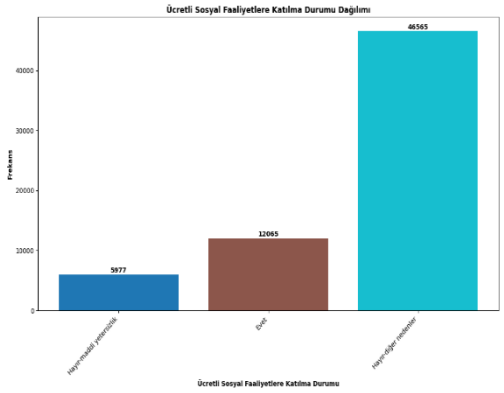
Şekil 11, bireylerin sosyal yaşam durumlarına göre dağılımını göstermektedir. Verilere göre, bireylerin büyük çoğunluğu (46.842) "Evet" kategorisinde yer alarak sosyal yaşam etkinliklerine katılım sağlamaktadır. Buna karşın, "Hayır-diğer nedenler" kategorisinde yer alan bireylerin sayısı 11.482 olup, sosyal etkinliklere katılmayanların önemli bir kısmını oluşturmaktadır. "Hayır-maddi yetersizlik" sebebiyle sosyal etkinliklere katılamayan bireylerin sayısı ise 6.283 olarak tespit edilmiştir. Bu dağılım, sosyal etkinliklere katılımın genellikle yüksek olduğunu, ancak maddi yetersizliklerin katılımı kısıtlayan bir faktör olduğunu göstermektedir. Diğer nedenler ise maddi yetersizlikten daha büyük bir engel olarak gözlemlenmektedir. Bu durum, sosyoekonomik faktörlerin bireylerin sosyal yaşamlarına olan etkisini daha derinlemesine incelemek için önemli bir bulgu olarak değerlendirilebilir.



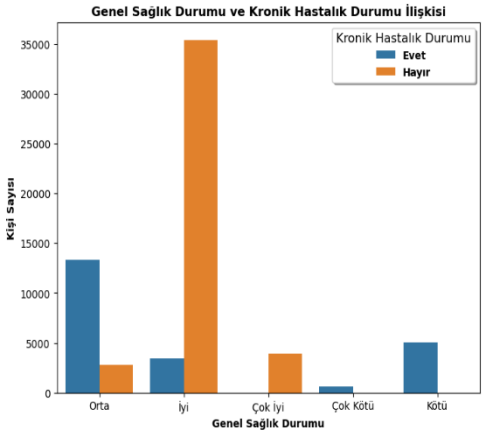
Şekil 11. Sosyal yaşam durumu dağılımı (Distribution of social life status)

Şekil 12, bireylerin ücretli sosyal faaliyetlere katılım durumuna göre dağılımını göstermektedir. Verilere göre, ücretli sosyal faaliyetlere katılmayan bireylerin çoğunluğu "Hayır-diğer nedenler" kategorisinde yer almakta olup, bu grupta 46.565 birey bulunmaktadır. Buna karşın, maddi yetersizlik nedeniyle bu tür faaliyetlere katılamayan birey sayısı 5.977 ile sınırlıdır. Ücretli sosyal faaliyetlere katılan bireylerin sayısı ise 12.065 olarak belirlenmiştir. Bu dağılım, bireylerin ücretli sosyal faaliyetlere katılımında "diğer nedenlerin" maddi yetersizlikten daha büyük bir engel oluşturduğunu göstermektedir. Ücretli faaliyetlere katılımın düşük olmasının, bireylerin zaman yönetimi, ilgi alanları veya sosyal faktörler gibi çeşitli nedenlerden kaynaklanabileceği değerlendirilebilir. Türkiye gibi ülkelerde sosyal aktivitelerin ücretli olması, sosyoekonomik durumun sosyal katılım üzerindeki etkisini de vurgulamakta olup, bu tür analizler sosyoekonomik politikalar ve sosyal programların geliştirilmesi açısından önemli bilgiler sunmaktadır.

Bununla birlikte Şekil 13, bireylerin genel sağlık durumları ile kronik hastalık varlığı arasındaki ilişkiyi gösteren bir sütun grafiği içermektedir. “İyi”, “Çok iyi”, “Orta”, “Kötü” ve “Çok kötü” kategorileri, bireylerin kendilerini sağlık açısından nasıl değerlendirdiklerini temsil ederken, renkler kronik hastalık durumunu (“Evet” ve “Hayır”) göstermektedir. “İyi” kategorisindeki sütunun yüksekliği, bu sağlık durumunu rapor eden bireylerin önemli bir kısmının kronik hastalık rapor etmediğini göstermektedir. Benzer şekilde, ""Orta” kategorisinde de kronik hastalık olmadığını belirten bireylerin sayısı dikkate değerdir. Bu görsel, genel sağlık algısı ile kronik hastalıkların varlığı arasında önemli bir ilişki olduğunu göstermektedir.



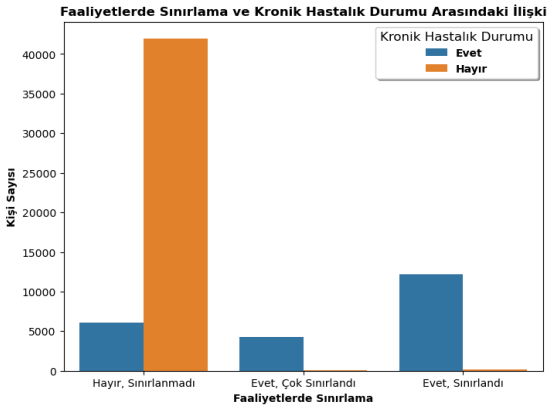
Şekil 12. Ücretli sosyal faaliyetlere katılım durumu dağılımı (Distribution of participation in paid social activities)



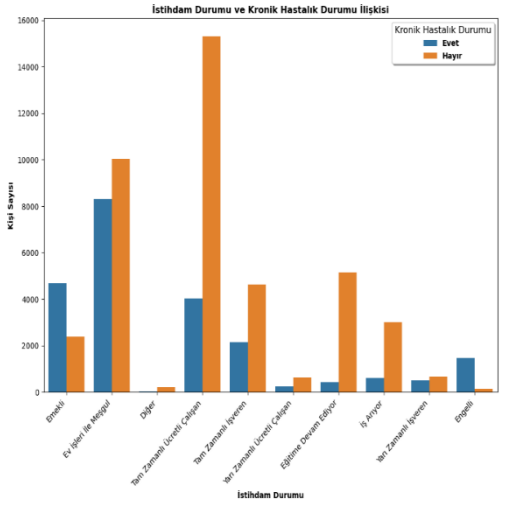
Şekil 13. Genel sağlık durumu ve kronik hastalık durumu arasındaki ilişki (The relationship between general health status and chronic disease status)

Şekil 14, sağlık problemleri nedeniyle günlük faaliyetlerde sınırlama yaşayan bireylerin kronik hastalık durumları ile ilişkisini göstermektedir. Grafikte görüldüğü gibi, sağlık problemlerinden ötürü faaliyetlerinde herhangi bir sınırlama olmayan bireylerin çoğunluğu kronik hastalık bildirmemiştir. Buna karşın, faaliyetlerinde sınırlama yaşayanların büyük bir kısmının aynı zamanda kronik hastalıklara sahip olduğu görülmektedir. Bu durum, kronik hastalıkların bireylerin günlük yaşam aktiviteleri üzerindeki etkisini açıkça ortaya koymaktadır.

Şekil 15, bireylerin istihdam durumları ile kronik hastalık durumları arasındaki ilişkiyi göstermektedir. İstihdam durumu kategorileri arasında, tam zamanlı ücretli çalışanlar arasında kronik hastalık bildirenlerin sayısı, bildirmeyenlere göre daha azken, emekliler arasında kronik hastalık bildirenlerin sayısı daha fazla görünmektedir. Yarı zamanlı çalışanlar ve iş arayanlar arasında da kronik hastalıklar yaygınken, ev işleri ile meşgul olanlar ve eğitime devam edenler arasında daha az yaygındır. Bu dağılım, istihdam türü ve kronik sağlık durumları arasındaki potansiyel korelasyonları gözler önüne sermektedir. Görseldeki veriler, istihdam durumu ve kronik hastalıkların varlığı arasındaki ilişkiyi incelerken, emeklilik ve tam zamanlı çalışma gibi hayatın farklı evrelerinin sağlık üzerinde belirgin bir etkisi olabileceğini işaret etmektedir. Emeklilik durumu, muhtemelen yaşla bağlantılı olarak kronik hastalık prevalansının (belirli bir süre içinde bir hastalığın toplumda görülme sıklığını gösteren ölçüt) yüksekliğini gösterirken, tam zamanlı çalışanlar arasında bu oranın daha düşük olması, çalışan nüfusun genel sağlık durumunun daha iyi olabileceğine işaret edebilir. Yarı zamanlı çalışanlar ve iş arayanlar arasındaki kronik hastalık oranları, bu grupların stres ve yaşam tarzı faktörlerinin sağlık üzerinde etkili olabileceğini düşündürmektedir. Eğitim görenler arasında kronik hastalık oranının düşük olması, genç nüfusun genel sağlık durumunun daha iyi olabileceğini veya eğitimin sağlık üzerinde olumlu bir etkisinin olabileceğini yansıtıyor olabilir. Bu tür veriler, halk sağlığı planlaması ve istihdam politikalarının oluşturulması açısından değerli bilgiler sağlamaktadır.

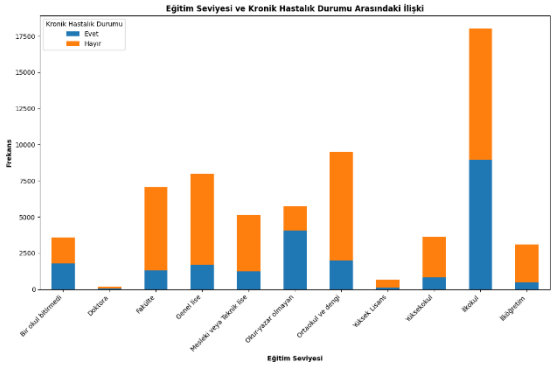


Şekil 14. Faaliyetlerde sınırlama ve kronik hastalık durumu arasındaki ilişki (The relationship between limitation of activities and chronic disease status)



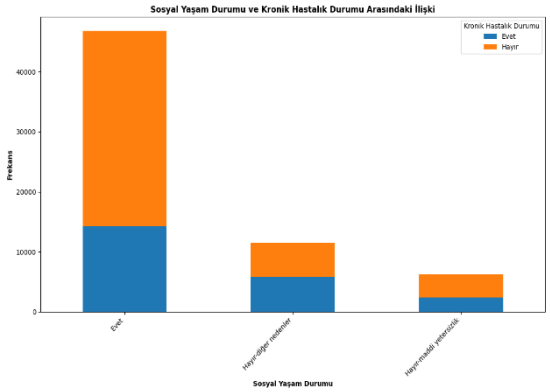
Şekil 15. İstihdam durumu ve kronik hastalık durumu arasındaki ilişki (The relationship between employment status and chronic disease status)

Şekil 16, bireylerin eğitim seviyesi durumları ile kronik hastalık durumları arasındaki ilişkiyi göstermektedir. Grafikte her bir eğitim seviyesi, kronik hastalığı olan bireyler (mavi) ve olmayan bireyler (turuncu) olarak ikiye ayrılmıştır. En yüksek frekansa sahip grup ilkokul mezunları olup, bu grupta kronik hastalığı olan bireylerin oranı diğer eğitim seviyelerine göre oldukça yüksektir. Ortaokul ve dengi okullardan mezun olan bireylerde de kronik hastalığı olan bireylerin oranı dikkat çekicidir. Genel lise ve mesleki/teknik lise mezunlarında ise kronik hastalığı olan bireylerin oranı nispeten daha düşüktür. Doktora ve yüksek lisans mezunları ise, hem toplamda daha düşük sayıda temsil edilmekte hem de kronik hastalık oranı çok düşük seviyelerde kalmaktadır. Genel olarak, eğitim seviyesi arttıkça kronik hastalık oranının azaldığı gözlemlenmektedir. Bu bulgu, düşük eğitim seviyesine sahip bireylerde kronik hastalıkların daha yaygın olduğunu ve eğitimin bireylerin sağlık durumu üzerinde önemli bir etkiye sahip olabileceğini göstermektedir.



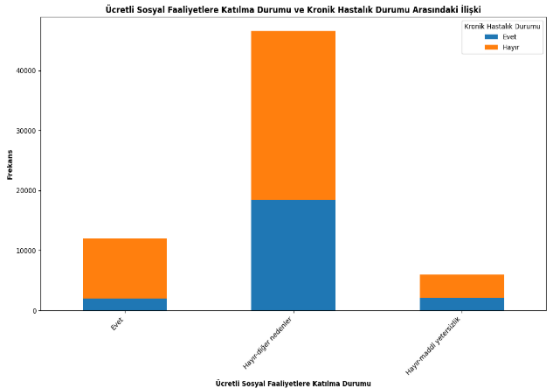
Şekil 16. Eğitim seviyesi ve kronik hastalık durumu arasındaki ilişki (The relationship between education level and chronic disease status)

Şekil 17, bireylerin sosyal yaşam durumları ile kronik hastalık durumları arasındaki ilişkiyi göstermektedir. Grafikte sosyal yaşam durumları "Evet" (sosyal etkinliklere katılanlar), "Hayır-diğer nedenler" (katılmayanlar) ve "Hayır-maddi yetersizlik" (katılamayanlar) olarak üç ana gruba ayrılmıştır. Sosyal etkinliklere katılan bireylerin (Evet) arasında kronik hastalığı olanların sayısı önemli bir orana sahip olup, bu grup içerisindeki bireylerin büyük çoğunluğunu kronik hastalığı olmayanlar oluşturmaktadır. "Hayır-diğer nedenler" kategorisinde, kronik hastalığı olan bireylerin oranı dikkate değer şekilde yüksektir, ancak bu grupta da kronik hastalığı olmayan bireyler çoğunluğu oluşturmaktadır. "Hayır-maddi yetersizlik" kategorisinde ise kronik hastalığı olan bireyler ve olmayan bireyler arasında daha dengeli bir dağılım olduğu gözlemlenmektedir. Genel olarak, sosyal etkinliklere katılma oranı düşük olan bireyler arasında kronik hastalığın daha yaygın olduğu ve sosyal yaşamın sağlık durumu üzerinde belirgin bir etkisinin olabileceği anlaşılmaktadır. Bu bulgu, sosyal etkileşimlerin sağlıkla ilişkili sonuçlar üzerindeki rolünü vurgulamaktadır.



Şekil 17. Sosyal yaşam durumu ve kronik hastalık durumu arasındaki ilişki (The relationship between social life status and chronic disease status)

Şekil 18, ücretli sosyal faaliyetlere katılma durumu ve kronik hastalık durumları arasındaki ilişkiyi göstermektedir. Ücretli sosyal faaliyetlere katılan bireyler ("Evet") arasında kronik hastalığı olan bireylerin (mavi) oranı oldukça düşük olup, büyük çoğunluk kronik hastalığı olmayan bireylerden (turuncu) oluşmaktadır. "Hayır-diğer nedenler" kategorisinde, sosyal faaliyetlere katılmayan bireyler arasında kronik hastalığı olanların oranı daha yüksektir, ancak yine de kronik hastalığı olmayan bireyler bu grupta çoğunluğu oluşturmaktadır. Maddi yetersizlik nedeniyle ücretli sosyal faaliyetlere katılamayan bireyler arasında ise kronik hastalığı olan ve olmayan bireylerin sayısı daha dengeli olup, kronik hastalığı olmayanlar az bir farkla çoğunluktadır. Genel olarak, ücretli sosyal faaliyetlere katılmayan bireyler arasında kronik hastalığın daha yaygın olduğu ve özellikle maddi yetersizlik durumunda bu oranın daha belirgin hale geldiği görülmektedir. Bu bulgular, sosyoekonomik faktörlerin ve sağlık durumunun sosyal etkinliklere katılım üzerindeki etkilerini ortaya koymaktadır.



Şekil 18. Ücretli sosyal faaliyetlere katılma durumu ve kronik hastalık durumu arasındaki ilişki (The relationship between participation in paid social activities and chronic disease status)

5. Sonuçlar (Conclusions)

Bu çalışmada, Türkiye İstatistik Kurumu'nun (TÜİK) 2023 Gelir ve Yaşam Koşulları Araştırması (GYKA) verileri kullanılarak, bireylerin kronik hastalık durumlarının çeşitli makine öğrenimi yöntemleri ile sınıflandırılması amaçlanmıştır. Literatürdeki mevcut boşluk, geniş ve karmaşık veri kümelerinden edinilen bilgiyi etkili bir şekilde kullanarak sağlık alanında karar verme süreçlerini destekleyecek modellerin geliştirilmesine yöneliktir. Bu çalışma, makine öğrenimi yöntemlerinin, özellikle de güçlü tahmin yetenekleri ve karmaşık ilişkileri modelleme kapasitesi ile dikkat çeken TabNet derin öğrenme tekniğinin, sağlık verileri analizinde etkin bir şekilde kullanılmasının önemini vurgulamaktadır. Araştırmanın odak noktası, sağlık hizmetlerine erişimdeki eşitsizlikler ve sosyoekonomik faktörlerin sağlık durumları üzerindeki etkisini anlamak ve değerlendirmektir. Bu çalışmanın bulguları, sağlık politikalarının geliştirilmesi ve sektörel analizler için stratejik kararlar alınmasında kullanılabilecek değerli bilgiler sağlamaktadır.

Bu çalışmanın öne çıkan bulguları arasında, TabNet modelinin %97 doğruluk oranı ile en yüksek performansı gösterdiği tespit edilmiştir. YSA, LSTM, CNN, GBM ve RF gibi modeller de yüksek doğruluk oranları sunmakla birlikte, TabNet modeli özellikle genelleme yeteneği ve sınıflar arasındaki dengeli performansı ile öne çıkmıştır. Optuna kütüphanesi kullanılarak tüm modellerin hiperparametre optimizasyonu yapılmış ve bu sayede modellerin tahmin gücü artırılmıştır. Ayrıca, eğitim seviyesi, sosyal yaşam katılımı, genel sağlık durumu ve ekonomik faktörlerin kronik hastalık riskleri üzerinde önemli bir etkisi olduğu bulgulanmıştır. Sosyal yaşam katılımının ve ekonomik yetersizliklerin kronik hastalık durumu ile güçlü bir ilişkiye sahip olduğu gözlemlenmiştir, bu da sosyal ve ekonomik faktörlerin sağlık üzerindeki rolünün altını çizmektedir.

Bu bulgular doğrultusunda, sağlık politikalarının kronik hastalıkların önlenmesi ve yönetilmesine yönelik olarak sosyal ve ekonomik faktörleri dikkate alacak şekilde yeniden yapılandırılması gerekmektedir. Özellikle, sosyal yaşam katılımının sağlık üzerindeki olumlu etkisi göz önünde bulundurularak, bireylerin sosyal faaliyetlere katılımını teşvik eden programlar geliştirilebilir. Maddi yetersizliklerin kronik hastalıklar üzerindeki olumsuz etkisini azaltmak için, düşük gelirli bireyler ve gruplar için sağlık hizmetlerine erişimi kolaylaştırıcı politikalar hayata geçirilmelidir. Ayrıca, eğitim seviyesinin kronik hastalık riskini azaltıcı etkisi göz önünde bulundurularak, sağlık eğitimi programları yaygınlaştırılmalı ve özellikle risk altındaki gruplara yönelik farkındalık çalışmaları artırılmalıdır. Bu tür bütüncül politikalar, bireylerin hem fiziksel hem de sosyal sağlığını iyileştirmeyi hedeflemeli ve sağlığın sosyal belirleyicilerini göz ardı etmeden daha kapsayıcı bir yaklaşım benimsemelidir. Sektörel bazda alınabilecek stratejik kararlar, sağlık, eğitim ve sosyal hizmetler sektörleri arasında iş birliğini artırmaya yönelik olmalıdır. Sağlık sektöründe, kronik hastalıkların önlenmesi ve yönetimi için daha entegre ve kişiselleştirilmiş tedavi yaklaşımları geliştirilmelidir. Bu doğrultuda, sağlık teknolojilerine ve yapay zekâ destekli erken teşhis sistemlerine yapılan yatırımlar artırılabilir. Eğitim sektöründe, halk sağlığına yönelik bilinçlendirme kampanyaları ve sağlık eğitimi programları geliştirilerek, bireylerin hastalıklar hakkında daha fazla bilgi sahibi olmaları sağlanabilir. Sosyal hizmetler sektöründe ise, sosyal yaşam katılımını destekleyen projeler ve özellikle düşük gelirli gruplara yönelik sosyal destek programları uygulanabilir. Ayrıca, iş dünyasında çalışanların sağlık durumlarının iyileştirilmesi amacıyla, iş yerlerinde sağlık taramaları, fiziksel aktivite programları ve psikososyal destek hizmetleri teşvik edilmelidir. Bu stratejik adımlar, sektörler arası iş birliği ile toplum sağlığının genel olarak iyileştirilmesine katkı sağlayacaktır.

Bu çalışmanın bulgularını literatürde yer alan diğer çalışmalara kıyasladığımızda, TabNet modelinin kronik hastalık sınıflandırmasındaki %97 doğruluk oranıyla öne çıktığını ve özellikle Türkiye İstatistik Kurumu (TÜİK) verileri üzerinde elde edilen sonuçların literatüre önemli bir katkı sağladığını görmekteyiz. Özkan (2019)'un Gradient Boosting algoritmasıyla hepatit hastalığını %98.36 doğruluk, %98.68 kesinlik ve %98.95 duyarlılıkla sınıflandırması, veri ön işleme ve model seçiminin önemini vurgulamakla birlikte, bu çalışmada kullanılan TabNet modeliyle elde edilen sonuçlar, farklı bir sağlık alanında benzer bir performans sergilemiştir. Ahmed vd. (2019)'un CNN ile lösemi alt türlerini sınıflandırma çalışmasında %88.25 doğruluk elde edilmiştir, bu da CNN'in özellikle görüntü tabanlı verilerdeki başarısını göstermektedir. Ancak, TabNet modeli, özellikle tabular veri üzerinde daha yüksek doğruluk sağlamasıyla fark yaratmaktadır. Gaddam ve Pattnaik (2020)'in YSA ile aritmi tespitinde %91 doğruluk elde etmesi de kardiyovasküler hastalıklar için etkili bir sınıflandırma sunarken, TabNet'in kronik hastalıkların sınıflandırılmasındaki başarısı daha geniş bir veri seti üzerinde daha yüksek performans sağlamaktadır. Pacci vd. (2021)'in tüp bebek tedavisinde gebelik tahmini için Destek Vektör Makineleri ile elde ettiği %71.7 doğru pozitif ve %59.4 doğru negatif oranları, TabNet modelinin kronik hastalık sınıflandırmasındaki %97 doğruluk oranına kıyasla daha düşük bir performans sergilemiştir. Tang ve Liu (2021)'in Alzheimer hastalığının ilerleyişini sınıflandırmada Rastgele Orman algoritması ile elde ettikleri %96.14 doğruluk oranı, TabNet modeline yakın bir başarı göstermektedir, ancak TabNet'in daha genel bir veri seti üzerinde bu başarıyı sağlaması dikkat çekicidir. Akcan ve Sertbaş (2021) ile Purwaningsih (2022) tarafından yapılan çalışmalardaki sonuçlarla karşılaştırıldığında, özellikle topluluk öğrenme yöntemleri ve DVM gibi algoritmaların başarılı bir şekilde uygulandığı görülmektedir. Akcan ve Sertbaş'ın göğüs kanseri teşhisinde topluluk öğrenme yöntemlerini

kullanarak Soft Voting, Bagging ve XGBoost ile en yüksek doğruluk oranlarına ulaşması, bu yöntemlerin bireysel sınıflandırma algoritmalarına göre üstün performans sergilediğini göstermektedir. Aynı şekilde, bu çalışmada kullanılan TabNet modeli de kronik hastalıkların sınıflandırılmasında benzer bir şekilde yüksek doğruluk oranı (%97) elde etmiştir. Topluluk öğrenme yöntemleri gibi, TabNet de karmaşık sağlık verileri üzerinde genelleme yeteneği açısından etkili sonuçlar sunmaktadır. Purwaningsih'in çalışmasında ise, DVM modeli ve ileri özellik seçimi (forward selection) kullanılarak kronik böbrek hastalığının %99.75 doğrulukla sınıflandırılması, DVM algoritmasının optimize edilmiş haliyle yüksek performans sağladığını göstermektedir. Bu doğruluk oranı, TabNet'in kronik hastalıklar üzerinde elde ettiği %97 doğruluk oranına kıyasla biraz daha yüksek olsa da, her iki çalışmada da özellik seçimi ve doğru model optimizasyonunun sınıflandırma başarılarını artırdığı vurgulanmaktadır. Ayrıca, Purwaningsih'in çalışmasındaki YSA modelinin %90,5 doğruluk oranı ile TabNet'in elde ettiği %97 doğruluk karşılaştırıldığında, TabNet'in daha yüksek bir performans sergilediği görülmektedir. Sevli (2023)'nin diyabet hastalığını %96.29 doğruluk ile sınıflandırması, Gradient Boosting ve AdaBoost gibi yöntemlerle yüksek başarı elde edilmiştir. Coşkun ve Yüksek (2023) ise Gradient Boosting ile hepatit hastalığının sınıflandırılmasında %98.36 doğruluk oranına ulaşarak başarılı bir sonuç elde etmişlerdir, ancak TabNet'in kronik hastalıklar üzerindeki uygulamaları bu başarıya oldukça yakındır. Son olarak, Kim vd. (2023)'ün LSTM ve CNN tabanlı modelle %94.3 doğruluk elde etmesi, kronik hastalıkların zaman serisi verileri üzerinde incelenmesi açısından önemlidir. Bu çalışmadaki TabNet modeli, tablo verisi üzerinde yüksek doğrulukla çalışarak, kronik hastalık tahmini gibi kritik sağlık problemlerinde başarılı bir performans sergilemektedir. Coşkun ve Yüksek (2023)'in çalışmasında Gradient Boosting algoritması %98.36 doğruluk ile öne çıkmışken, bu çalışmada TabNet %97 doğruluk oranıyla oldukça yakın bir performans sergilemiştir. Bununla birlikte, Gradient Boosting hepatit hastalığında yüksek başarı sağlamış olsa da, TabNet'in geniş bir kronik hastalık yelpazesinde benzer bir başarı elde etmesi, modelin genel uygulanabilirliği açısından önemli bir avantajdır. Kim vd. (2023)'in CNN ve LSTM kombinasyonunu kullanarak kronik hastalık tahmininde %94.3 doğruluk elde etmesi, derin öğrenme modellerinin zaman serisi verilerde etkili olduğunu göstermektedir. Ancak, TabNet modeli %97 doğruluk ile daha yüksek bir performans sergilemiş ve özellikle tabular veri üzerinde güçlü bir alternatif olarak öne çıkmıştır. Zhang vd. (2023)'in LSTM modeliyle %98.82 doğruluk oranı elde etmesi de LSTM'nin ses verileri üzerinde yüksek performans sağladığını gösterirken, TabNet'in sağlık verisi gibi tabular yapılar üzerinde etkili sonuçlar vermesi onu farklı bir bağlamda rekabetçi hale getirmektedir. Özdemir (2023)'in aritmi sınıflandırmasında Gradient Boosting ve Rastgele Orman gibi yöntemlerle yüksek doğruluk elde etmesi, bu çalışmanın bulgularını desteklemekte olup, TabNet'in kronik hastalık tahmininde bu yöntemlerle rekabet edebilecek düzeyde performans sunduğunu göstermektedir. Aynı şekilde, Duyar vd. (2023)'ün TabNet'in boosting yöntemlerine kıyasla daha düşük performans gösterdiğini belirttiği çalışmada, bu modelin farklı veri türlerine ve sağlık koşullarına uygulanabilirliğinin incelenmesi gerektiği ortaya çıkmaktadır. Choubey vd. (2024)'ün PCA ve XGBoost kullanarak kronik hastalık tahmininde %98.8 doğruluk elde etmesi, TabNet ile karşılaştırıldığında biraz daha yüksek bir sonuç sunmaktadır. Ancak, TabNet'in işlem süresi ve model yapısındaki avantajları göz önüne alındığında, bu çalışmanın geniş çaplı sağlık verisi üzerinde sağladığı %97 doğruluk oranı önemli bir başarıdır. Ayrıca, Elkholy vd. (2023)'ün TabNet kullanarak kronik böbrek hastalığını %92.5 doğrulukla tahmin etmesi de, TabNet'in kronik hastalıklar üzerinde etkili olabileceğini vurgulayan diğer bir bulgudur. Sonuç olarak, TabNet modeli, yukarıda bahsi geçen diğer çalışmalarla karşılaştırıldığında, özellikle geniş tabular veri kümeleri üzerinde yüksek doğruluk ve genelleme yeteneği sunarak rekabetçi bir performans sergilemiştir. Gradient Boosting, LSTM ve CNN gibi modellerin spesifik veri setlerinde başarılı sonuçlar elde etmesine rağmen, TabNet'in kronik hastalık sınıflandırmasındaki başarısı, geniş veri setleri ve uygulama alanları için güçlü bir alternatif olduğunu göstermektedir.

Bu çalışma, Türkiye İstatistik Kurumu'nun 2023 Gelir ve Yaşam Koşulları Araştırması verileri kullanılarak makine öğrenimi yöntemleri ile bireylerin kronik hastalık durumlarının sınıflandırılmasına yöneliktir. Veri setinde kullanılan bağımsız değişkenler, kronik hastalıkların sınıflandırılmasında önemli bir rol oynamış ve elde edilen yüksek doğruluk oranlarına katkı sağlamıştır. Özellikle bireylerin eğitim seviyesi, sosyal yaşam katılımı, ekonomik durumu ve genel sağlık durumu gibi demografik ve sosyoekonomik değişkenler, kronik hastalıkların tahmin edilmesinde belirleyici olmuştur. Eğitim seviyesi düşük bireylerde kronik hastalık riskinin daha yüksek olduğu gözlemlenirken, sosyal yaşamdan uzak olan ve maddi yetersizlik çeken bireylerde de hastalık riskinin arttığı bulgulanmıştır. Bu, sosyal ve ekonomik faktörlerin bireylerin sağlık durumu üzerindeki etkisini doğrulamakta ve sağlığın sosyal belirleyicileri kavramının önemini vurgulamaktadır. Ayrıca, veri setinde yer alan diğer değişkenler, bireylerin genel sağlık durumu ve yaşam koşullarıyla ilişkili olup, kronik hastalık riskini etkileyen çok yönlü faktörler olarak karşımıza çıkmaktadır. Bu değişkenlerin modelde doğru şekilde kullanılması, TabNet ve diğer makine öğrenimi algoritmalarının yüksek performans göstermesini sağlamış ve hastalıkların sınıflandırılmasında önemli bir başarı elde edilmiştir.

Bu çalışmanın bazı kısıtlılıkları bulunmaktadır. İlk olarak, kullanılan veri seti Türkiye İstatistik Kurumu'nun (TÜİK) belirli bir zaman dilimine ait verilerine dayandığı için, veriler coğrafi ve zamansal olarak sınırlıdır; bu da modelin farklı popülasyonlar ve zaman dilimlerinde genelleme yeteneğini sınırlayabilir. İkinci olarak, veri setindeki bağımsız değişkenler ağırlıklı olarak sosyoekonomik ve demografik faktörlerden oluşmakta olup, biyomedikal ve klinik verilerin eksikliği, kronik hastalıkların daha derinlemesine ve biyolojik temelli bir analizine olanak sağlamamaktadır. Ayrıca, modelin performansı yüksek olmasına rağmen, kullanılan değişkenlerin doğruluğu ve eksiksizliği modele doğrudan etki etmektedir; eksik veya hatalı veriler modelin tahmin gücünü zayıflatabilir.

Gelecekte yapılacak çalışmalar için birkaç önemli öneri ve tavsiye sunulabilir. İlk olarak, veri setinin coğrafi ve zamansal sınırlarının ötesine geçerek, farklı ülkelerden ve bölgelerden elde edilen daha geniş kapsamlı ve çok merkezli veri setlerinin kullanılması, modelin genelleme yeteneğini artırabilir. Ayrıca, mevcut çalışmada kullanılan sosyoekonomik ve demografik değişkenlerin yanı sıra, biyomedikal verilerin (genetik, klinik test sonuçları ve hastalık geçmişi vs.) entegrasyonu, kronik hastalıkların daha derinlemesine analiz edilmesini sağlayarak modelin tahmin doğruluğunu artırabilir. Ayrıca, gelecekte hibrit model yaklaşımlarının denenmesi, tahmin performansını daha da iyileştirebilir. Ayrıca, modelin açıklanabilirliği ve yorumlanabilirliğini artırmak için SHAP ve LIME gibi model açıklama tekniklerinin kullanılması, sağlık profesyonellerinin ve politika yapıcıların sonuçları daha iyi anlamalarına ve karar süreçlerine dahil etmelerine yardımcı olabilir. Son olarak, zaman serisi analizine dayalı çalışmaların genişletilmesi, kronik hastalıkların ilerleyişi ve uzun vadeli sağlık sonuçları üzerine öngörüler sunarak proaktif ve önleyici sağlık politikalarının geliştirilmesine katkı sağlayabilir.

Kaynaklar (References)

Ahmed, N. A., Yiğit, A., Işık, Z., Alpkoçak, A., 2019. Identification of leukemia subtypes from microscopic images using convolutional neural network. Diagnostics, 9(3), 104. https://doi.org/10.3390/diagnostics9030104

Ahsan, M., Khan, A., Khan, K. R., Sinha, B. B., Sharma, A. 2023. Advancements in medical diagnosis and treatment through machine learning: a review. Expert Systems, 41(3). https://doi.org/10.1111/exsy.13499

Akcan, F., Sertbaş, A., 2021. Topluluk Öğrenmesi Yöntemleri ile Göğüs Kanseri Teşhisi. Electronic Turkish Studies, 16(2).

Albin Ahmed, A., Shaahid, A., Alnasser, F., Alfaddagh, S., Binagag, S., Alqahtani, D., 2023. Android ransomware detection using supervised machine learning techniques based on traffic analysis. Sensors, 24(1), 189. https://doi.org/10.3390/s24010189

Almutairi, M., Chiroma, H., Abubakar, S., 2022. Detecting elderly behaviors based on deep learning for healthcare: recent advances, methods, real-world applications and challenges. IEEE Access, 10, 69802-69821. https://doi.org/10.1109/access.2022.3186701

Al-Shamisi, M. H., Assi, A., Hejase, H., 2013. Artificial neural networks for predicting global solar radiation in al ain city - uae. International Journal of Green Energy, 10(5), 443-456. https://doi.org/10.1080/15435075.2011.641187

Altuntaş, O., Esra, A. K. I., Huri, M., 2015. Kronik hastalıklarda ilaç kullanımının yaşam kalitesi ve sosyal katılıma etkisi üzerine nitel bir çalışma. Ergoterapi ve Rehabilitasyon Dergisi, 3(2), 79-86.

An, W., Liang, M., 2012. A new intrusion detection method based on svm with minimum within-class scatter. Security and Communication Networks, 6(9), 1064-1074. https://doi.org/10.1002/sec.666

Anjum, M. J., Tariq, F., Anjum, K. M., Shaheen, M., Ahmad, F., 2023. Identification of diseases caused by non-synonymous single nucleotide polymorphism using random forest and linear regression algorithms. https://doi.org/10.21203/rs.3.rs-3001745/v1

Arik, S. Ö., Pfister, T., 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, 35(8), 6679-6687. https://doi.org/10.48550/arXiv.1908.07442

Arkin, F. S., Aras, G., Doğu, E., 2020. Comparison of artificial neural networks and logistic regression for 30-days survival prediction of cancer patients. Acta Informatica Medica, 28(2), 108. https://doi.org/10.5455/aim.2020.28.108-113

Bissacco, A., Yang, M.-H., Soatto, S., 2007. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07. (Minneapolis, MN). https://doi.org/10.1109/CVPR.2007.383129

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chen, C., Fan, L., 2022. Cnn-lstm-attention deep learning model for mapping landslide susceptibility in kerala, india. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, X-3/W1-2022, 25-30. https://doi.org/10.5194/isprs-annals-x-3-w1-2022-25-2022

Choubey, S. B., Chitra, T., Hephzipah, J. J., 2024. Big Data Mining for Chronic Disease Prediction using Principal Component Analysis and eXtreme Gradient Boosting. GK International Journal of Advanced Research in Engineering and Technology, 1(1), 1-11.

Coşkun, C., Yüksek, E., 2023. Hepatit hastalığının tespitinde bulanık mantık ve makine öğrenmesi yöntemlerinin karşılaştırılması. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 14(4), 539-546.

Dai, X., Yin, H., Jha, N. K., 2020. Grow and prune compact, fast, and accurate lstms. IEEE Transactions on Computers, 69(3), 441-452. https://doi.org/10.1109/tc.2019.2954495

Dubey, G., Khera, R., Grover, A., Kaur, A., Goyal, A., Rajkumar, R., Srivastava, S., 2023. A hybrid convolutional network and long short-term memory (hbcnls) model for sentiment analysis on movie reviews. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4), 341-348. https://doi.org/10.17762/ijritcc.v11i4.6458

Duyar, C., Senica, S. O., Kalkan, H., 2023. Detection of cardiovascular disease using gut microbiota data. https://doi.org/10.21203/rs.3.rs-2794999/v1

Elkholy, S., Rezk, A., Saleh, A. A., 2023. Enhanced optimized classification model of chronic kidney disease. International Journal of Advanced Computer Science and Applications, 14(2). https://doi.org/10.14569/ijacsa.2023.0140239

El-Shafeiy, E., El-Desouky, A. I., Elghamrawy, S. M., 2024. An optimized artificial neural network approach based on sperm whale optimization algorithm for predicting fertility quality. Studies in Informatics and Control, 27(3), 349-358. https://doi.org/10.24846/v27i3y201810

Ersöz, A. G., 2003. Dünya konferansları belgelerinde aile ve yoksulluk: Saptamalar ve öneriler. Sosyal Politika Çalışmaları Dergisi, 6(6).

Fanelli, G., Dantone, M., Gall, J., Fossati, A., Gool, L., 2012. Random forests for real time 3D face analysis. Int. J. Comput. Vis. 1, 1–22. https://doi.org/10.1007/s11263-012-0549-0

Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Friedman, J., 2001. Greedy boosting approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–407. https://doi.org/10.1214/aos/1016218222

Gaddam, C. M. Pattnaik, S. S., 2020. An ann ensemble based ecg signal classification approach for accurate arrhythmia detection. International Journal of Emerging Technology and Advanced Engineering, 10(8), 57-61. https://doi.org/10.46338/ijetae0820_08

Gao, X., Chen, D., Pan, Q., 2022. An interpretable classification model of breast tumors with tabular mammography data. 2nd International Conference on Signal Image Processing and Communication (ICSIPC 2022). https://doi.org/10.1117/12.2643654

Guhan, T., Bhavishya, S. Kalaiarasan, S., Lalith, K., Dhitchith, O. P., 2024. Chronic Illness Detection using Gradient Boosting Algorithm. Grenze International Journal of Engineering & Technology (GIJET), 10.

Guido, R., Ferrisi, S., Lofaro, D., Conforti, D., 2024. An overview on the advancements of support vector machine models in healthcare applications: a review. Information, 15(4), 235. https://doi.org/10.3390/info15040235

Gündoğdu, S., 2021. Kalp hastalık risk tahmini için Python aracılığıyla sınıflandırıcı algoritmalarının performans değerlendirmesi. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 23(69), 1005-1013.

Hansen, L., Salamon, P., 1990. Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 12, 993–1001. https://doi.org/10.1109/34.58871

Hegde, S. Mundada, M. R., 2020. Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach. International Journal of Pervasive Computing and Communications, 17(1), 20-36. https://doi.org/10.1108/ijpcc-04-2020-0018

Huang, Y., Gao, Z., Zhang, H., 2020. Comparison of common machine learning algorithms trained with multi-zone models for identifying the location and strength of indoor pollutant sources. Indoor and Built Environment, 30(8), 1142-1158. https://doi.org/10.1177/1420326x20931576

Hutchinson, R. A., Liu, L.P., Dietterich, T. G., 2011. "Incorporating boosted regression trees into ecological latent variable models," in AAAI'11, (San Francisco, CA), 1343–1348. Available online at: http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3711

Johnson, R., Zhang, T., 2012. Learning Nonlinear Functions Using Regularized Greedy Forest. Technical Report. arXiv:1109.0887. doi: 10.2172/1052139

Jongjaraunsuk, R., Taparhudee, W., Suwannasing, P., 2024. Comparison of water quality prediction for red tilapia aquaculture in an outdoor recirculation system using deep learning and a hybrid model. Water, 16(6), 907. https://doi.org/10.3390/w16060907

Kim, G., Lim, H., Kim, Y., Kwon, O., Choi, J., 2023. Intra-person multi-task learning method for chronic-disease prediction. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-28383-9

Kim, J. O., Jeong, Y. S., Kim, J. H., Lee, J. W., Park, D., Kim, H. S., 2021. Machine learningbased cardiovascular disease prediction model: A cohort study on the Korean national health ınsurance service health screening database. Diagnostics, 11(6), 943.

Kumsar, A. K., Yılmaz, F. T., 2014. Kronik Hastaliklarda Yaşam Kalitesine Genel Bakiş. ERÜ Sağlık Bilimleri Fakültesi Dergisi, 2(2), 62-70.

Küçükberber, N., Özdilli, K., Yorulmaz, H., 2011. Kalp hastalarında sağlıklı yaşam biçimi davranışları ve yaşam kalitesine etki eden faktörlerin değerlendirilmesi. Anadolu Kardiyol Derg, 11, 619-626.

Lee, E. Y., Fulan, B. M., Wong, G. C. L., Ferguson, A. L., 2016. Mapping membrane activity in undiscovered peptide sequence space using machine learning. Proceedings of the National Academy of Sciences, 113(48), 13588-13593. https://doi.org/10.1073/pnas.1609893113

Liu, Q., Li, S., Li, Y., Yu, L., Zhao, Y., Wu, Z., Zhang, Y., 2023. Identification of urinary volatile organic compounds as a potential non-invasive biomarker for esophageal cancer. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-45989-1

Liu, Y., Wang, Y., Li, Y., Zhang, B., Wu, G., 2004. Earthquake prediction by RBF neural network ensemble, in Advances in Neural Networks - ISNN 2004, eds F.-L. Yin, J. Wang, and C. Guo (Berlin; Heidelberg: Springer), 962–969. https://doi.org/10.1007/978-3-540-28648-6_153

Loey, M., Naman, M. R., Zayed, H. H., 2020. Deep transfer learning in diagnosing leukemia in blood cells. Computers, 9(2), 29. https://doi.org/10.3390/computers9020029

Luo, L., Zhang, F., Yao, Y., Gong, R., Fu, M., Xiao, J., 2018. Machine learning for identification of surgeries with high risks of cancellation. Health Informatics Journal, 26(1), 141-155. https://doi.org/10.1177/1460458218813602

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 55(2), 645-657. https://doi.org/10.1109/tgrs.2016.2612821

McLaughlin, R. T., Asthana, M., Meo, M. D., Ceccarelli, M., Jacob, H. J., Masica, D. L., 2023. Fast, accurate, and racially unbiased pan-cancer tumor-only variant calling with tabular machine learning. NPJ Precision Oncology, 7(1). https://doi.org/10.1038/s41698-022-00340-1

Özdemir, A., 2023. Makine Öğrenmesi Algoritmaları ile Aritmilerin Sınıflandırılması. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, 39(3), 394-402.

Özkan, Y., 2019. Hastalık tanısı verilerinde veri ön işlemenin topluluk öğrenme sınıflandırma algoritmaları üzerindeki etkisinin incelenmesi, Ege Üniversitesi, Sağlık Bilimleri Enstitüsü, Yayınlanmamış Yüksek Lisans Tezi, İzmir.

Pacci, Z., Şengül, Y. A., Attar, R., Alagöz, O., 2021. Yapay Zeka Tabanlı Klinik Karar Destek Sistemi ile Tüp Bebek Tedavisi Gebelik Sonucu Tahmini. EMO Bilimsel Dergi, 11(22), 27-35.

Pittman, S. J., Brown, K. A., 2011. Multi-scale approach for predicting fish species distributions across coral reef seascapes. PLoS ONE 6:e20583. https://doi.org/10.1371/journal.pone.0020583

Qi, Y., 2012. Random forest for bioinformatics, in Ensemble Machine Learning, eds C. Zhang and Y. Ma (New York, NY: Springer), 307. https://doi.org/10.1007/978-1- 4419-9326-7_11

Schapire, R., 2002. The boosting approach to machine learning: an overview. Nonlin. Estimat. Classif. Lect. Notes Stat. 171, 149–171. https://doi.org/10.1007/978-0-387- 21579-2_9

Sevli, O., 2023. Diagnosis of diabetes mellitus using various classifiers. Journal of the Faculty of Engineering and Architecture of Gazi University, 38(2), 989-1001.

Sewell, M., 2011. Ensemble Learning. Technical Report, Department of Computer Science, University College London. Available online at: http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/research/Research_Notes/RN_11_02.pdf Erişim tarihi: 20.02.2024

Shu, C., Burn, D. H., 2004. Artificial neural network ensembles and their application in pooled flood frequency analysis. Water Resour. Res. 40, 1–10. https://doi.org/10.1029/2003WR002816

Sönmez, O., Zengin, K., 2023. Süt Sığırlarının Buzağılama Zamanının Tahmininde Makine Öğrenme Yöntemlerinin Kullanımı Çalışmaları Üzerine Bir Değerlendirme. Journal of New Results in Engineering and Natural Sciences, 2023(18), 27-39.

Tang, X. Liu, J., 2021. Comparing different algorithms for the course of alzheimer's disease using machine learning. Annals of Palliative Medicine, 10(9), 9715-9724. https://doi.org/10.21037/apm-21-2013

Toğaçar, M., Cömert, Z., Ergen, B., 2021. Intelligent skin cancer detection applying autoencoder, MobileNetV2 and spiking neural networks. Chaos, Solitons & Fractals, 144, 110714.

Vidya, G., Hari, V. S., 2023. Lstm network integrated with particle filter for predicting the bus passenger traffic. Journal of Signal Processing Systems, 95(2-3), 161-176. https://doi.org/10.1007/s11265-022-01831-x

Wade, C., Glynn, K., 2020. Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd.

Xi, J., Liang, R., Fei, X., 2017. An algorithm of improving speech emotional perception for hearing aid. Modern Physics Letters B, 31(19-21), 1740094. https://doi.org/10.1142/s0217984917400942

Yangın, G., 2019. XGboost ve Karar Ağacı tabanlı algoritmaların diyabet veri setleri üzerine uygulaması, Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Bilimleri Enstitüsü, Yayınlanmamış Yüksek Lisans Tezi, İstanbul.

Yar, F., 2015. Türkiye'de gelir dağılımı & yoksulluk. Global Analiz, 2, 1-30.

Zhang, P., Swaminathan, A., Uddin, A. A., 2023. Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks. Frontiers in Medicine, 10. https://doi.org/10.3389/fmed.2023.1269784

Research Article
Journal of Intelligent Systems: Theory and Applications 8(1) (2025) 25-34
DOI: 10.38016/jista.1447980

Renewable Energy Forecasting in Turkey: Analytical Approaches

Mehmet Berke Colak[1], Erkan Özhan[2*]

[1, 2] Department of Computer Engineering, Çorlu Faculty of Engineering, Namık Kemal University, Tekirdağ, Türkiye

berkecolak95@gmail.com, eozhan@nku.edu.tr

Abstract

The growing population and industrialization have resulted in an increased demand for energy, which has worsened environmental problems such as pollution and climate change. Renewable energy sources are considered a promising solution due to their environmental benefits and limited potential. This study examines the use of neural networks and time series analysis to predict electricity generation rates from renewable energy sources in Turkey. We use the LSTM, NNAR, and ELM models, all of which utilize the backpropagation algorithm for neural network forecasting. Additionally, we apply ARIMA, Holt's trend, linear regression, mean, and exponential smoothing models for time series analysis. We evaluate the performance using the mean absolute error and root mean square error on the training and test data. The study showed that LSTM models outperformed the ARIMA (1,2,1), ARIMA (2,2,1), ARIMA (3,2,1), and NNAR methods in forecasting accuracy. Although the NNAR model initially had the lowest error, its linear predictions made it less suitable for practical applications. This study highlights the effectiveness of neural networks and time series analysis in predicting renewable energy sources. The ARIMA (1,2,1), LSTM and ARIMA (3,2,1) modeling methods are useful for optimizing the planning and management of Turkey's renewable energy future, contributing to a more sustainable energy landscape.

Keywords: Renewable energy, Turkey, time series, neural networks, climate change, ARIMA, LSTM

Türkiye'de Yenilenebilir Enerji Tahmini: Analitik Yaklaşımlar

Öz

Artan nüfus ve sanayileşme, enerji talebinin artmasına neden olmuş, bu da kirlilik ve iklim değişikliği gibi çevre sorunlarını daha da kötüleştirmiştir. Yenilenebilir enerji kaynakları, çevresel faydaları ve sınırsız potansiyelleri nedeniyle ümit verici bir çözüm olarak değerlendirilmektedir. Bu çalışma, Türkiye'de yenilenebilir enerji kaynaklarından elektrik üretim oranlarını tahmin etmek için sinir ağlarının ve zaman serisi analizinin kullanımını incelemektedir. Sinir ağı tahminleri için her ikisi de geri yayılım algoritmasını temel alan LSTM, NNAR ve ELM modellerini kullanıyoruz. Ayrıca zaman serisi analizi için ARIMA, Holt trendi, doğrusal regresyon, ortalama ve üstel düzeltme modellerini kullanıyoruz. Performansı, eğitim ve test verilerinde ortalama mutlak hata ve kök ortalama kare hata kullanarak değerlendiriyoruz. Çalışma, LSTM modellerinin tahmin doğruluğunda ARIMA (1,2,1), ARIMA (2,2,1), ARIMA (3,2,1) ve NNAR yöntemlerinden daha iyi performans gösterdiğini göstermiştir. NNAR modeli başlangıçta en düşük hataya sahip olmasına rağmen doğrusal tahminleri onu pratik uygulamalar için daha az uygun hale getirdi. Çalışma, yenilenebilir enerji kaynaklarının tahmin edilmesinde sinir ağlarının ve zaman serisi analizinin etkinliğini vurguluyor. ARIMA (1,2,1), LSTM ve ARIMA (3,2,1) modelleme yöntemleri, Türkiye'nin yenilenebilir enerji geleceğinin planlanması ve yönetimini optimize etmek ve daha sürdürülebilir bir enerji ortamına katkıda bulunmak için kullanışlıdır.

Anahtar kelimeler: Yenilenebilir enerji, Türkiye, zaman serileri, sinir ağları, iklim değişikliği, ARIMA, LSTM

1. Introduction

Energy, in its basic form, is a system's ability to perform work or generate heat, while renewable energy refers to naturally replenished sources of energy (Coburn and Farhar, 2004). Renewable energy is a source of energy continually replenished by natural processes (Hersh, 2006). Renewable energy can be obtained from sources with a nearly limitless supply, ensuring sustainability over time ("Renewable energy explained - U.S. Energy Information Administration (EIA)," 2023). Renewable energy sources may vary among countries. Turkey possesses a diverse range of renewable energy sources, including hydropower, wind, and solar energy. According to information from the YTBS website,

*Corresponding Author.
E-mail: eozhan@nku.edu.tr

Recieved : 7 Mar 2024
Revision : 30 Apr 2024
Accepted : 16 Oct 2024

Turkey's renewable energy usage rates in 2023 were as follows: biomass accounted for 2.65%, solar for 5.83%, geothermal for 3.43%, and wind for 10.52%. To increase energy production, it is important to utilize different sources and invest in renewable energy. In this regard, energy prediction studies are of critical importance in accurately forecasting future energy demand and planning necessary investments. Many studies have demonstrated the success of methods such as artificial neural networks (ANNs) in numerical estimation problems. In this study, various prediction models were employed, including neural network autoregression (NNAR), extreme learning machines (ELM), exponential smoothing, Holt's trend, linear regression, mean model, autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM). ANNs can learn dispersed relationships in data (Mossalam and Arafa, 2018). ANNs models possess a structure that mimics the learning functions of the human brain. Information is transmitted through connections between neurons, and the network is trained with preexisting datasets. The goal is to find the network configuration that will produce the most accurate prediction. Various components, such as the learning algorithm and the activation function, contribute to achieving this configuration. Determining the optimal network structure can be considered an optimization problem. Once this structure is established, the developed network model can be used to make future predictions with minimal inaccuracy. Typical ANN models use simplified neuron models similar to human neurons (Nastos et al., 2013). ANN, ARIMA, NNAR, and LSTM models are widely used and effective methods for energy prediction. ANNs have the ability to learn from complex datasets and forecast future trends, enabling the prediction of energy production and consumption levels, price fluctuations, and other factors. Many of these methods have been tested to establish standards, and the selection of the estimation model is based on error rates.

A time series consists of observations produced sequentially over time. A set is considered continuous if it is continuous; otherwise, it is discrete (Baskan, 2008). The primary objective of time series analysis is prediction. The fundamental idea is to utilize past observations to forecast the future, and the model that best describes the data is then employed to predict future outcomes based on historical records (Baccar, 2019).

This study investigated the usability of current methods such as artificial neural networks, ARIMA, ELM, NNAR and LSTM in estimating Turkey's renewable energy production rate.

This study aims to guide Turkey's decision-making process for its renewable energy future and identify the most effective methods. The use of new technologies such as ANNs is important for improving Turkey's energy production capacity and ensuring energy security. Forecasts regarding Turkey's energy production capacity in the future are important for planning investments and ensuring energy security. In this study, the findings of these predictions obtained from the most up-to-date methods are presented.

2. Literature Review

The literature has been surveyed to provide brief summaries of studies conducted chronologically in Turkey and around the world that employ forecasting methods related to renewable energy sources.

Paoli et al. employed neural network (MLP), ARIMA, the k-nearest neighbors algorithm, Bayesian decipherment, and Markov chain estimation methods to predict preprocessed daily solar radiation time series (Paoli et al., 2010a). Hocaoglu and Karanfil utilized Granger causality and impulsive response analysis estimation methods in their study, adopting a time series-based approach to renewable energy modeling (Hocaoglu and Karanfil, 2013a). Golestaneh et al. employed the ELM estimation method (Golestaneh et al., 2016). Jiang et al. utilized the ELM estimation method. Additionally, the investigation incorporated the bacterial-foraging optimization algorithm (BFOA) and empirical mode decomposition (EMD) as estimation methods. This study applied empirical mode decomposition and an advanced ELM optimized with the BFOA to estimate China's renewable energy terminal power consumption (Jiang et al., 2019). Tharani et al. utilized various machine learning techniques to comprehensively examine and project renewable energy trends (Tharani et al., 2020). Goncalves et al. utilized vector autoregression, privacy preservation, and distributed learning prediction methods. Additionally, they conducted a study on confidentiality-preserving distributed learning for renewable energy prediction (Goncalves et al., 2021). In their research, Gullu and Kartal focused on the electricity production objectives derived from renewable energy sources, including solar, wind, and hydroelectric power. They adopted the Box-Jenkins ARIMA methodology to estimate the individual installed capacities of these various renewable energy types. (Güllü and Kartal, 2021). In another study conducted by Cetin et al., future energy production was predicted using real data from a solar power company and employing machine learning algorithms. This study utilized the LSTM method, a type of ANN, to conduct predictions and analyses. The results revealed an error rate ranging from 1% to 15%. Future studies will focus on other renewable sources like wind, geothermal, and hydro energy (Çetin and Işık, 2021).

Erturk et al. developed ANN models using MATLAB for four provinces located in different climatic zones of Turkey (Kayseri, Rize, Hakkari, and Izmir) to accurately determine the amount of solar radiation. The solar radiation predictions made by the model yielded the best results for the province of Hakkari, with an R^2 value of

0.93, followed by Izmir, Kayseri, and Rize. There was a consistent agreement between the values predicted by the ANN models and the measured values for each province (Ertürk et al., 2023).

Kaysal et al. conducted a study using data from a wind farm located in the Mediterranean region, spanning the years 2018 to 2020. They employed convolutional neural network (CNN) and binary long short-term Memory (BLSTM) algorithms for prediction purposes (Kaysal et al., 2023). Çakir (2023), highlighted the increasing challenge of forecasting REG for effective energy management. Various time series models, encompassing physical models, statistical techniques, and artificial intelligence algorithms, have been proposed to address this challenge. Notably, fuzzy time series (FTS) models were applied to forecast Turkey's REG between 2000 and 2020. The results indicate that the proposed integrated model demonstrates high accuracy and serves as a valuable tool not only for REG forecasting but also for addressing other time series forecasting problems. Cakir suggested further exploration of this integrated model (Cakir, 2023).

Rajni et al. examined monthly energy production data spanning from January 1973 to December 2019. They conducted a study on renewable energy production in the United States from January to December 2020. This investigation employed ARIMA time series analysis techniques to forecast ten future time periods (months), specifically considering total renewable energy production (Rajni et al., 2024).

In a 2024 study, Bquet et al. developed an AI-based framework at the Swiss Federal Institute of Technology (EPFL) using a Long Short-Term Memory (LSTM) model to predict solar energy for different time horizons. The dataset consisted of 17,297,280 Global Horizontal Irradiance (GHI) measurements taken every 10 seconds between January 1, 2016, and November 1, 2021. The LSTM model proved highly effective for short-term forecasts, particularly for horizons a few hours ahead (Bouquet et al., 2024).

Solano et al. used Support Vector Regression (SVR), Extreme Gradient Boosting (XGBT), Categorical Boosting (CatBoost) machine learning algorithms for solar radiation prediction and proposed an ensemble feature selection method to select the most relevant input parameters and their past observations. The method called Voting Average (VOA) is an ensemble learning method that includes SVR, XGBT and CatBoost. As a result of the study, they proved that VOA outperformed the other algorithms (Solano et al., 2022).

In this study, in comparison to other studies identified through a literature review, the commonly utilized forecasting methods were ARIMA, employed by Paoli et al., Güllü and Kartal, Rajni et al. Golestaneh et al. and Jiang et al. utilized ELM in their studies. Additionally, LSTM was utilized by Çetin and Işık, Kaysal et al... Furthermore, Çetin and Işık, Erturk et al. and Han et al. employed ANNs.

Previous studies by Çetin and Işık and Ertürk et al. had a narrower focus, examining a single company's production capacity or a localized region, respectively. Güllü and Kartal used the ARIMA methodology to predict solar, wind, and hydroelectric power. In contrast, this study estimates the share of renewable energy in total energy production, excluding hydroelectric power. Unlike previous studies, this research evaluates the most recent data from 1960 to 2023, encompassing all renewable energy sources in Turkey. Additionally, it compares several established algorithms using the most recent data. This study stands out by testing eight different forecasting methods on the most comprehensive dataset and demonstrating the suitability of the ARIMA and LSTM algorithms predictive models by focusing on the top five results.

Upon reviewing the literature, it is evident that studies have focused on forecasting solar energy alone (Çetin and Işık, 2021; Ertürk et al., 2023; Goncalves et al., 2021; Paoli et al., 2010b), wind energy (Kaysal et al., 2023), both solar and wind energy (Cakir, 2023), and a combination of solar, wind, and hydropower (Golestaneh et al., 2016; Güllü and Kartal, 2021; Jiang et al., 2019). Additionally, numerous studies explore renewable energy consumption (Jiang et al., 2019), trends (Tharani et al., 2020), and the relationships between various parameters of renewable energy (Hocaoglu and Karanfil, 2013b).

When examining studies on renewable energy sources in Turkey, Cetin et al. employed the LSTM algorithm, but their work was limited to a specific region within Turkey, and similar to Erturk et al., they only forecast solar energy. Our study, on the other hand, covers the entire country and focuses on predicting energy generation from all renewable sources, excluding hydropower. Hocaoglu et al. investigated the relationships between parameters of renewable energy in their study, while Güllü and Kartal included hydropower among renewable energy sources. Çakır aimed to forecast solar and wind energy generation using a dataset spanning 2000 to 2020. An analysis of Çakır's predictions reveals that the aim was to estimate total production volume. However, our study focuses on predicting the share of renewable energy in Turkey's total energy production as a percentage. When comparing the trends, both studies indicate an upward trend. Moreover, while Çakır's study produced forecasts only up to 2020, our work extends the predictions to 2028, offering continued insights beyond 2020, thus contributing to a more sustained forecasting approach.

3. Method

3.1. Methods Used for Prediction

3.1.1. Extreme Learning Machine (ELM) Method

The single hidden layer feed-forward neural network consists of three neural layers. Its name derives from the nonlinear hidden layer in the model, which processes input layer data features. When the hidden layer performs no computations, the output layer becomes linear and is devoid of any transformation function or bias (Akusok, 2016). The single-layer hidden-layer neural network model with L, a number of training samples in with N, a weight vector between the input and hidden layers with w, an activation function with g, a vias vector with b and an input vector with x. The ELM method is formulated as equation g(x) as shown in Equation 1 (Erdem, 2020), which includes hidden neurons and an activation function.

$$\sum_{j=1}^{L} \beta_j \cdot g\left(\left(w_{j,} x_i\right) + b_j\right) = y_i, \quad i = 1,2,..,N \qquad (1)$$

Extreme Learning Machine (ELM) is a fast algorithm for Single-Layer Feedforward Networks (SLFN), randomly assigning input weights and biases, while analytically determining output weights using the generalized inverse of the hidden layer's output, thus greatly improving learning speed and generalization performance (Guang-Bin Huang et al., 2004).

3.1.2. Neural network autoregression (NNAR) model method

The neural network autoregression (NNAR) model is a three-layer feedforward neural network (Maleki et al., 2018), as illustrated in Figure 1.

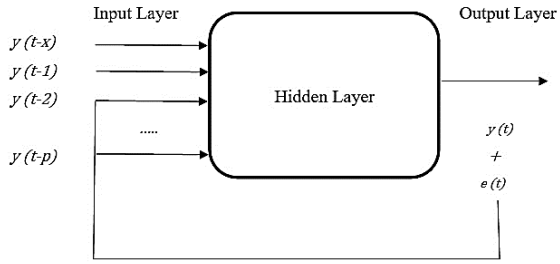


Figure 1. Schematic representation of the NNAR model (Gibson, 2020).

The NNAR (Neural Network Autoregression) model is a data-driven, feedforward neural network that uses the backpropagation algorithm for training and parameter estimation (Sadia et al., 2022). The input layer of the network receives the lagged values of the time series, representing past observations and in the hidden layer, the model learns complex nonlinear relationships between these past values (Daniyal et al., 2022). The backpropagation algorithm is used to minimize the error between the predicted and actual values by adjusting the network's weights. This error is propagated backwards from the output layer through the network, updating the weights to improve future predictions.

3.1.3. Autoregressive Integrated Moving Average (ARIMA) Modeling Method

The ARIMA model was developed according to the methodology described by Box and Jenkins (Box et al., 1994). There are three basic types of ARIMA models: the moving average (MA) model, autoregressive (AR) model, and integrated (I) model (Yang et al., 2020). The nonseasonal model is one of the various ARIMA models. The general equation for the ARIMA (p, d, q) model is formulated in Equation 2 and approximated as shown in Equation 3. In these equations, d represents the number of differences, t denotes the discrete time, $\emptyset_p$ is the autoregressive parameter, $\varepsilon_t$represents the residual and θ_q is the moving average parameter. Additionally, X_t and U_t are both reliable variables. The symbol d represents the difference (Nyatuame and Agodzo, 2018).

$$U_t = \emptyset_1 U_{t-1} + \emptyset_2 U_{t-2} + .. + \emptyset_p U_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - .. - \theta_q \varepsilon_{t-q} \qquad (2)$$

$$U_t = X_t - X_{t-d} \qquad (3)$$

During the tuning of the ARIMA models, autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were constructed to identify the optimal ARIMA parameters based on the lag observations. An automatic configuration was preferred for the lambda parameter. Additionally, the ACF graph is shown in Figure 2, and the PACF graph is shown in Figure 3.

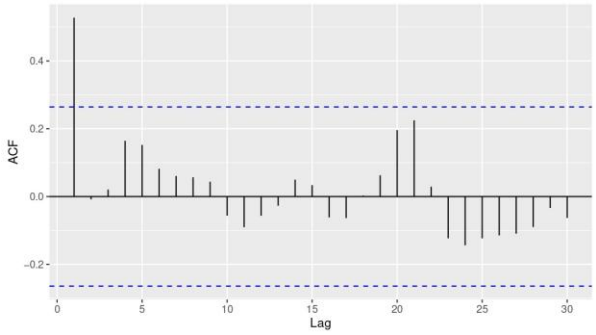


Figure 2. Representation of the ACF Graph

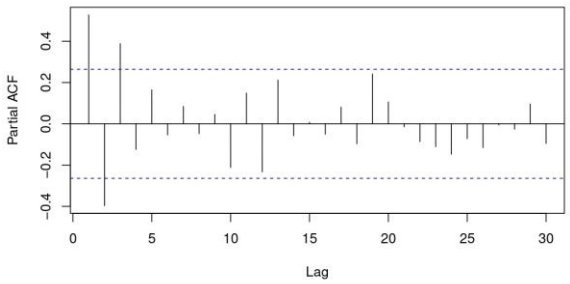


Figure 3. Representation of the PACF Graph

3.1.4. Holt's trend method

Holt's trend method estimates the parameter for trend correction using the Holt-Winters method. The multiplicative model for Holt's trend method is formulated in Equation 4, while the additive model is expressed in Equation 5. In Equation 5, when examining the equation symbols, the symbol S_t represents exponential correction at time t, S_{t-1} signifies exponential correction at time t-1 corresponding to seasonal elements, b_t denotes trend elements at time t and b_{t-1} represents trend elements at time t-1 (Nurhamidah et al., 2020). In addition, β is the smoothing factor for the trend and F is the forecast at steps ahead (Mrutyunjaya, 2020).

$$B_{t-1} = \beta\,(F_{t-1} - F_{t-2}) + (1-\beta)B_{t-2} \quad (4)$$

$$b_t = \beta\,(S_t - S_{t-1}) + (1-\beta)\,b_{t-1} \quad (5)$$

3.1.5. LSTM neural network model

The LSTM model (Hochreiter and Schmidhuber, 1997) is a powerful recurrent neural system specially designed to overcome the exploding/vanishing gradient problems that typically arise when learning long-term dependencies, even when the minimal time lags are very long (Van Houdt et al., 2020).

LSTMs are a type of recurrent neural network (RNN) designed to capture long-term dependencies in time series data, utilizing "cell states" to store important information and "gates" to determine which information should be remembered and which should be forgotten (Olah, 2015)

An LSTM neural network model typically comprises an input sequence layer, one or more LSTM layers arranged sequentially to capture time dependencies in the data, a fully connected layer to transform the output size of preceding layers into the number of classes to be recognized, and a softmax layer to compute the probability of belonging to each class. Additionally, it includes a classification output layer to calculate the cost function (Ghislieri et al., 2021). LSTM networks offer several advantages, including dynamic system modeling capabilities in diverse application domains such as image processing, speech recognition, manufacturing, autonomous systems, communication, and energy consumption (Lindemann et al., 2021).

The adaptive moment estimation (Adam) stochastic gradient descent method, which is based on the adaptive estimation of first and second-order moments, was employed for optimizing the LSTM algorithm.

3.2. Criteria used in the analysis of the results

3.2.1. Mean absolute error (MAE)

The MAE is calculated as the average of the absolute differences, referred to as errors, between the expected and actual observations. As shown in Equation 6, the MAE is calculated as the average of the absolute differences (errors) between the expected and actual observations a_n represents the actual value, is the observed value, n is the number of observations and N is the number of observations.

$$\mathrm{MAE} = \frac{\sum_{n=1}^{N}|o_n - a_n|}{N} \quad (6)$$

3.2.2. Root mean square error (RMSE)

As indicated in Equation 7, the RMSE is employed to calculate the average magnitude of the differences between the predicted and observed values. In this equation, o_n represents the observed value, a_n represents the actual value, n is the number of observations and *N* represents the number of tuples in the test dataset.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(o_n - a_n)^2}{N}} \quad (7)$$

3.3 Dataset

In the experiments, we utilized a dataset encompassing the total rate of electrical energy obtained by Turkey from renewable energy sources, excluding hydroelectricity, for the years 1960–2023. The data for the years 1960 to 2015 were sourced from the World Bank website, while the data for the years 2015 to 2019 were obtained from the Ministry of Energy. Subsequently, data for the years 2019 to 2023 were acquired from the YTBS-TEIAS website ("Yük Tevzi Bilgi Sistemi (YTBS)-Türkiye Elektrik İstatistikleri," 2023).

The dataset contains two attributes: "Date" and "Electricityproduction". Since the dataset was provided as structured Excel and .csv files by the relevant institutions, no additional data cleaning was performed. However, due to the absence of records for the year 1982, the Kalman Filter method was employed to impute the missing data in the time series. Using this method, the data for 1982-1983 was automatically filled in without disrupting the integrity of the series.

The R programming language, with the R-Studio application, was used for the data analysis. In the experimental results section, the outcomes of the estimation processes are discussed in depth. The R programming language is employed for statistical analysis, particularly by data scientists, academics, and health researchers (Lanovaz and Adams, 2019).

4. Experimental Design

The available data were partitioned into training and testing sets, and the dataset underwent time series estimation using the most preferred methods. Among the time series analysis estimation methods, we employed the ARIMA, Holt's trend, linear regression, mean, and exponential smoothing models. On the other hand, artificial neural network estimation methods include the use of the ELM method, the LSTM neural network model and the neural network autoregression (NNAR) method. The methods employed in the study were implemented using the R programming language in the R-Studio application. The 'Keras3' library and the 'TensorFlow' library were utilized within the R software environment to employ the LSTM algorithm. The data used in the analysis from 1960 to 2018 are labeled training data, while the data from 2019 to 2023 are designated test data. The reason for selecting the data from the last few years as the test set is to evaluate the model's predictive ability based on recent trends and variables. This approach may help the model to better predict future trends. We developed several prediction models using training data from 2019 to 2023. The performance of these models was then evaluated using the MAE and the RMSE. This sequence represents a widely accepted approach to solving this type of problem. Furthermore, these values were compared to determine which method yielded better results. Five results were obtained from each estimation method for the years 2019-2023. Finally, future projections were made for the years 2024 to 2028. As a result of this process, 5 outcomes were obtained, contributing to the overall assessment.

5. Experimental Results

The dataset used is displayed in Figure 4. Upon examination of the graph, it becomes evident that certain irregular increases and decreases occurred in the percentage share lines. Consequently, the observed dataset is identified as having a trend component. The absence of a seasonality component is attributed to the uneven distribution of lines in the graph.

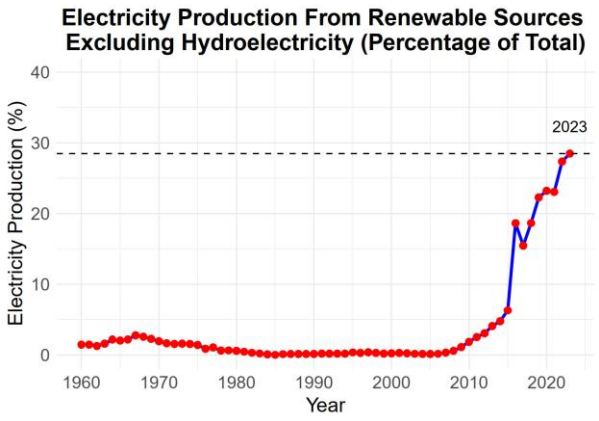


Figure 4. Graphical Representation of the Data Set

Forecasting for the years 2019-2023 was conducted using the training data. First, we tested popular models, and the results are presented in Table 1. Table 2 lists the five estimation methods that yielded the best results. Among the five methods, NNAR demonstrated the best performance when examining the test data segment of the estimation process.

The second-best method was the LSTM modeling method, followed by the ARIMA (1,2,1) modeling method as the third-best estimation approach. The fourth-best estimation method was the exponential smoothing method, followed by the ARIMA (3,2,1) as the fifth best.

Table 1. Performance Values of All Methods

Forecasting Method	MAE	RMSE
NNAR	0.99	1.48
ELM	7.54	8.49
Exponential Smoothing	6.52	7.00
Holt's Trend	7.54	8.50
Linear Regression	20.27	20.42
Mean (constant) Model	22.91	23.05
ARIMA (1,2,1)	1.65	1.84
ARIMA (2,2,1)	7.24	8.28
ARIMA (3,2,1)	6.52	7.18
LSTM	1.40	1.73

Table 2. The Accuracy Values of the Test Data of the Five Methods That Give the Best Results

Forecasting Method	MAE	RMSE
NNAR	0.99	1.48
LSTM	1.40	1.73
Exponential Smoothing	6.52	7.00
ARIMA (1,2,1)	1.65	1.84
ARIMA (3,2,1)	6.52	7.18

Table 3 displays the estimated values and the actual values for the methods that produced the best results.

The table clearly shows that the values obtained using the ARIMA (3,2,1) and ARIMA (2,2,1) modeling methods are closely aligned. Although they exhibit an increase compared to the exponential smoothing method, they do not align well with the actual values. Conversely, it has been revealed that the values obtained through the NNAR method and the LSTM method are more congruent with the actual values than those obtained through other methods.

The actual values in Table 3, along with the estimated values, are depicted in Figure 5. While the estimated values of the ARIMA (3,2,1) and ARIMA (2,2,1) models closely align on the graph, an unrelated pattern is evident in the graph line representing the actual values. The graph shows that the estimation values of the NNAR method, the LSTM method and the ARIMA (1,2,1) modeling method are closer to the graph line of the actual values than those of the other estimation methods.

Table 4 displays the estimated energy percentage values projected for the years 2024–2028, utilizing the five estimation methods that demonstrated the best results on the test data by leveraging the entirety of the training dataset**.**

T**able 3.** Estimated and actual values of the models

Model / Year	2019	2020	2021	2022	2023
NNAR	21.639	24.182	26.143	27.547	28.499
LSTM	22.79	24.466	26.435	28.249	29.393
Exp. Smooth.	18.344	18.344	18.344	18.344	18.344
ARIMA (1,2,1)	20.029	21.551	23.166	24.879	26.697
ARIMA (2,2,1)	23.906	27.488	31.132	36.14	41.859
ARIMA (3,2,1)	24.905	27.29	30.047	34.945	39.765
Actual Values	22.27	23.194	23.057	27.335	28.482

While the percentile values estimated in the NNAR model and the exponential smoothing model, as shown in Table 4, remained relatively stable, those calculated using the LSTM, ARIMA (1,2,1), and ARIMA (3,2,1) modeling methods increased over the years. Notably, the increase observed in the ARIMA (1,2,1) modeling method surpassed that of ARIMA (3,2,1).
In addition, upon examining the estimated values on the graph shown in Figure 6, it is evident that the graph line representing the percentile values obtained in the neural network autoregression (NNAR) model remains constant. In contrast, the values obtained through the LSTM, ARIMA (1,2,1) and ARIMA (3,2,1) modeling methods increase.

Table 4. Forecasts of the Models for the Next 5 Years

Year	NNAR	LSTM	Exp. Smooth.	ARIMA (1,2,1)	ARIMA (3,2,1)
2024	26.072	34.087	28.453	31.875	31.245
2025	25.303	37.958	28.453	35.422	35.306
2026	25.029	40.576	28.453	39.254	38.903
2027	24.928	43.130	28.453	43.377	42.874
2028	24.890	45.017	28.453	47.806	47.412

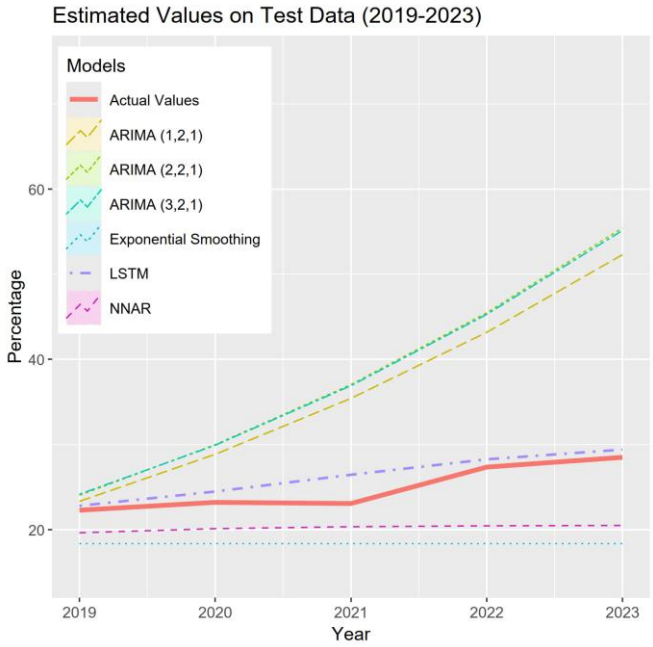


Figure 5. Estimated Values of the Models versus the Actual Values

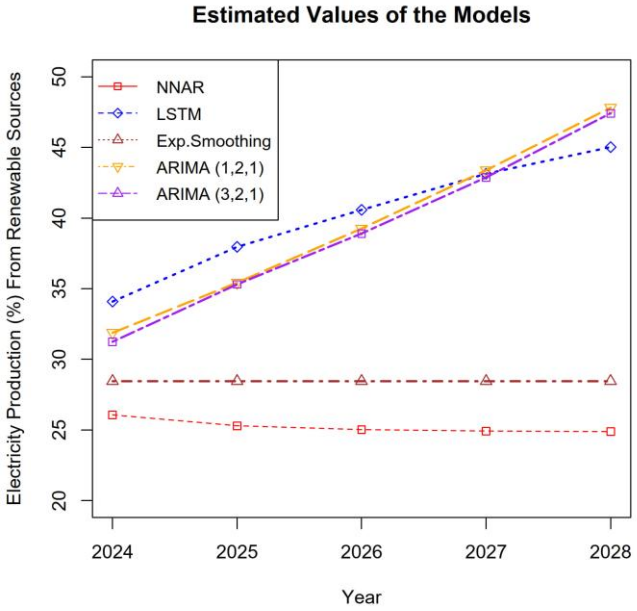


Figure 6. Estimated Values of the Models versus the Actual Values

6. Conclusions and Discussion

We are witnessing the escalating impacts of global warming each day, and it is evident that carbon dioxide emissions, in particular, are the primary cause of this situation. One of the most effective methods for mitigating this effect is to accelerate the adoption of renewable energy resources. Understanding the potential of existing renewable energy sources and assessing the degree to which this potential is already being harnessed will provide valuable insights into the broader landscape. Additionally, for planners and strategists, envisioning the future utilization of renewable energy sources is a crucial consideration. Examining our renewable energy potential and strategizing its future utilization allows us to formulate a comprehensive plan. In this study reviews various estimation approaches for forecasting the future utilization of renewable energy sources and these approaches can be applied to other countries and regions.

The dataset includes information on the electricity generated from renewable sources in Turkey spanning 64 years (1960–2023). For modeling purposes, the initial 59 years of data were used as training data, while the remaining 5 years served as test data. Various time series estimation methods were applied to the dataset, revealing that the NNAR method, a type of artificial neural network method, demonstrated the best performance. However, NNAR consistently iterated a constant value for its future predictions. In contrast, the series exhibited an upward trend, which NNAR failed to capture. The NNAR model makes predictions by establishing a linear relationship from past observations. This may have resulted in an inability to adequately capture the non-linear dynamics inherent in complex and variable processes such as energy production, which is the focus of this study. As a result, NNAR's linear predictions may fail to account for these complexities, making them less suitable for long-term and precise forecasting. Therefore, LSTM algorithm, which had the second-best prediction results, was considered more suitable for forecasting.

Considering the upward trend, the LSTM, ARIMA (1,2,1) and ARIMA (3,2,1) modeling methods produced the best results. According to our findings, it can be predicted that the share of renewable energy in Turkey's total energy production (excluding hydroelectric) from 2024 to 2028 will fall within the range of 34.09% to 45.02%. The techniques employed in this study can be tested for the quantitative estimation of both underground and surface resources. In similar tests, researchers may opt for LSTM and ARIMA modeling methods.

Unlike previous studies that focused on specific regions or types of renewable energy, this study aimed to forecast the share of all renewable energy production in Turkey, excluding hydroelectric power. Furthermore, the study employs eight distinct forecasting methods, thereby offering a more comprehensive understanding of the predictive models applied in the field. Notably, the study validates the effectiveness of well-performing algorithms such as ARIMA and ELM, further contributing to the empirical knowledge base in renewable energy forecasting.

Furthermore, the fact that these forecasting methods can be applied to other resource types and geographical areas implies that comparable approaches can be applied to provide reliable energy projections on a global scale. For their forecasting requirements, researchers and practitioners might investigate the usage of LSTM and ARIMA models. Researchers by modifying the models to take into consideration local characteristics and data accessibility, they can improve future predictions.

In conclusion, by highlighting the advantages of both conventional and modern modeling approaches and offering useful data for future studies and policy formulation, this work seeks to add to the body of knowledge on renewable energy forecasting.

Acknowledgments

This study is part of the Master of Science thesis of Mehmet Berke ÇOLAK, conducted within the Institute of Natural and Applied Science at Tekirdağ Namık Kemal University, under the guidance of thesis advisor Erkan ÖZHAN. The authors would like to express their gratitude to the Institute for its support and valuable contributions. Additionally, the authors extend thanks to the World Bank, the Ministry of Energy of Türkiye, and YTBS-TEIAS for providing the dataset used in this study.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Ethical approval and informed consent

During the preparation process of this study, scientific and ethical principles were followed, and all the studies included in the study were provided in the bibliography.

Author Contributions

The contributions of the authors to this study are equal.

Funding

This research received no external funding.

Availability of data and material

The dataset used in this study is publicly available and can be accessed and downloaded from the following GitHub repository: https://github.com/erkanozhan/renewable_energy. The repository contains the complete dataset in CSV format, along with detailed documentation and scripts for data analysis. Researchers are encouraged to explore, utilize, and contribute to the repository for further studies.

References

Akusok, A., 2016. Extreme Learning Machines : novel extensions and application to Big Data. University of Iowa Iowa Research Online,A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Industrial Engineering in the Graduate College of The University of Iowa.

Baccar, Y.B., 2019. Comparative Study on Time Series Forecasting Models. Master of Science (Data Science)Advisor: Bertrand Lamy, Jacques Doan HUU 1–92. https://doi.org/10.13140/RG.2.2.32241.02408

Baskan, S., 2008. Effect Of Ligand Binding On Protein Dynamics : A Time Series Analysis. Bogazici University 77.

Bouquet, P., Jackson, I., Nick, M., Kaboli, A., 2024. AI-based forecasting for optimised solar energy management and smart grid efficiency. International Journal of Production Research 62, 4623–4644. https://doi.org/10.1080/00207543.2023.2269565

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 1994. Time Series Analysis Forecasting and Control.

Cakir, S., 2023. Renewable energy generation forecasting in Turkey via intuitionistic fuzzy time series approach. Renewable Energy 214, 194–200. https://doi.org/10.1016/j.renene.2023.05.132

Çetin, Ö., Işık, A.H., 2021. Monthly Electricity Generatıon Forecast in Solar Power Plants with LSTM. Düzce Üniversitesi Bilim ve Teknoloji Dergisi 9, 55–64. https://doi.org/10.29130/dubited.1015251

Daniyal, M., Tawiah, K., Muhammadullah, S., Opoku-Ameyaw, K., 2022. Comparison of Conventional Modeling Techniques with the Neural Network Autoregressive Model (NNAR): Application to COVID-19 Data. Journal of Healthcare Engineering 2022, 1–9. https://doi.org/10.1155/2022/4802743

Erdem, K., 2020. Introduction to Extreme Learning Machines | by Kemal Erdem (burnpiro) | Towards Data Science.

Ertürk, S., Kara, H., Akkus, C., Genc, G., 2023. Turkiye'de Farklı İklim Kuşakları İçin Yapay Sinir Ağları Kullanılarak Güneş Isınımının Tahmini. Gazi University Journal of Science Part C: Design and Technology 11, 885–892. https://doi.org/10.29109/gujsc.1331788

Ghislieri, M., Cerone, G.L., Knaflitz, M., Agostini, V., 2021. Long short-term memory (LSTM) recurrent neural network for muscle activity detection. Journal of NeuroEngineering and Rehabilitation 18, 1–15. https://doi.org/10.1186/s12984-021-00945-w

Gibson, K., 2020. The Application Of Machine Learning For Grounwater Level Prediction In The Steenkoppies Compartment Of The Gauteng And North West Dolomite Aquifer , South Africa.

Golestaneh, F., Pinson, P., Gooi, H.B., 2016. Very short-term nonparametric probabilistic forecasting of renewable energy generation - With application to solar energy. IEEE Transactions on Power Systems 31, 3850–3863. https://doi.org/10.1109/TPWRS.2015.2502423

Goncalves, C., Bessa, R.J., Pinson, P., 2021. Privacy-preserving Distributed Learning for Renewable Energy Forecasting. IEEE Transactions on Sustainable Energy 3029, 1–10. https://doi.org/10.1109/TSTE.2021.3065117

Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew, 2004. Extreme learning machine: a new learning scheme of feedforward neural networks, in: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541). Presented at the 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), IEEE, Budapest, Hungary, pp. 985–990. https://doi.org/10.1109/IJCNN.2004.1380068

Güllü, M., Kartal, Z., 2021. Türkiye'nin Yenilenebilir Enerji Kaynaklarının 2030 Yılına Kadar Tahmini. 19 Mayis Sosyal Bilimler Dergisi 2, 288–313. https://doi.org/10.52835/19maysbd.849978

Hersh, M.A., 2006. The Economics and Politics of Energy Generation. IFAC Proceedings Volumes 39, 73–78. https://doi.org/10.1016/S1474-6670(17)30097-6

Hocaoglu, F.O., Karanfil, F., 2013a. A time series-based approach for renewable energy modeling. Renewable and Sustainable Energy Reviews 28, 204–214. https://doi.org/10.1016/j.rser.2013.07.054

Hocaoglu, F.O., Karanfil, F., 2013b. A time series-based approach for renewable energy modeling. Renewable and Sustainable Energy Reviews 28, 204–214. https://doi.org/10.1016/j.rser.2013.07.054

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jiang, P., Dong, J., Huang, H., 2019. Forecasting China's renewable energy terminal power consumption based on empirical mode decomposition and an improved extreme learning machine optimized by a bacterial foraging algorithm. Energies 12. https://doi.org/10.3390/en12071331

Kaysal, K., Yurttakal, A.H., Hocaoğlu, F.O., 2023. Hibrit derin öğrenme yöntemi kullanılarak hiperparametre optimizasyonu ile yenilenebilir elektrik enerjisi tahmini. Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi 12, 770–777. https://doi.org/10.28948/ngumuh.1263782

Lanovaz, M.J., Adams, B., 2019. Comparing the Communication Tone and Responses of Users and Developers in Two R Mailing Lists: Measuring Positive and Negative Emails. IEEE Software 36, 46–50. https://doi.org/10.1109/MS.2019.2922949

Lindemann, B., Müller, T., Vietz, H., Jazdi, N., Weyrich, M., 2021. A survey on long short-term memory networks for time series prediction. Procedia CIRP 99, 650–655. https://doi.org/10.1016/j.procir.2021.03.088

Maleki, A., Nasseri, S., Aminabad, M.S., Hadi, M., 2018. Comparison of ARIMA and NNAR Models for Forecasting Water Treatment Plant's Influent Characteristics. KSCE Journal of Civil Engineering 22, 3233–3245. https://doi.org/10.1007/s12205-018-1195-z

Mossalam, A., Arafa, M., 2018. Using artificial neural networks (ANN) in projects monitoring dashboards' formulation. HBRC Journal 14, 385–392. https://doi.org/10.1016/j.hbrcj.2017.11.002

Mrutyunjaya, P., 2020. Application of ARIMA and Holt-Winters forecasting model to predict the spreading of COVID-19 for India and its states. Department of Computer and Applications, Utkal University, Vani Vihar, India 14, 1–4.

Nastos, P.T., Moustris, K.P., Larissi, I.K., Paliatsos, A.G., 2013. Rain intensity forecast using Artificial Neural Networks in Athens, Greece. Atmospheric Research 119, 153–160. https://doi.org/10.1016/j.atmosres.2011.07.020

Nurhamidah, N., Nusyirwan, N., Faisol, A., 2020. Forecasting Seasonal Time Series Data Using the Holt-Winters Exponential Smoothing Method of Additive Models. Jurnal Matematika Integratif 16, 151. https://doi.org/10.24198/jmi.v16.n2.29293.151-157

Nyatuame, M., Agodzo, S.K., 2018. Stochastic ARIMA model for annual rainfall and maximum temperature forecasting over Tordzie watershed in Ghana. Journal of Water and Land Development 37, 127–140. https://doi.org/10.2478/jwld-2018-0032

Olah, C., 2015. Understanding LSTM Networks.

Paoli, C., Voyant, C., Muselli, M., Nivet, M.L., 2010a. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy 84, 2146–2160. https://doi.org/10.1016/j.solener.2010.08.011

Paoli, C., Voyant, C., Muselli, M., Nivet, M.L., 2010b. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy 84, 2146–2160. https://doi.org/10.1016/j.solener.2010.08.011

Rajni, Banerjee, T., Kumar, P., 2024. Forecasting of renewable energy production in United States: An ARIMA based time series analysis. AIP Conference Proceedings 3010, 030014. https://doi.org/10.1063/5.0193938

Renewable energy explained - U.S. Energy Information Administration (EIA) [WWW Document], 2023. . EIA. URL https://www.eia.gov/energyexplained/renewable-sources/ (accessed 1.15.24).

Sadia, I., Mahmood, A., Binti Mat Kiah, L., Azzuhri, S.R., 2022. Analysis and Forecasting of Blockchain-based Cryptocurrencies and Performance Evaluation of TBATS, NNAR and ARIMA, in: 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). Presented at the 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), IEEE, Kota Kinabalu, Malaysia, pp. 1–6. https://doi.org/10.1109/IICAIET55139.2022.9936798

Solano, E.S., Dehghanian, P., Affonso, C.M., 2022. Solar Radiation Forecasting Using Machine Learning and Ensemble Feature Selection. Energies 15, 7049. https://doi.org/10.3390/en15197049

Tharani, K., Kumar, N., Srivastava, V., Mishra, S., Pratyush Jayachandran, M., 2020. Machine learning models for renewable energy forecasting. Journal of Statistics and Management Systems 23, 171–180. https://doi.org/10.1080/09720510.2020.1721636

Van Houdt, G., Mosquera, C., Nápoles, G., 2020. A review on the long short-term memory model. Artif Intell Rev 53, 5929–5955. https://doi.org/10.1007/s10462-020-09838-1

Yang, Q., Wang, J., Ma, H., Wang, X., 2020. Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy. Journal of Infection and Public Health 13, 1415–1418. https://doi.org/10.1016/j.jiph.2020.06.019

Yük Tevzi Bilgi Sistemi (YTBS)-Türkiye Elektrik İstatistikleri [WWW Document], 2023. . Yük Tevzi Bilgi Sistemi (YTBS). URL https://ytbsbilgi.teias.gov.tr/ytbsbilgi/frm_istatistikler.jsf (accessed 1.15.24).

Clinical Research Article
Journal of Intelligent Systems: Theory and Applications 8(1) (2025) 35-46
DOI: 10.38016/jista.1501164

Modified Hard Voting Classifier Implementation on MEFV Gene Variants Increases in Silico Tool Performance: A Novel Approach for Small Sample Size

Tarık Alay[1*], İbrahim Demir[2], Murat Kirisci[3]

[1*] Ankara Etlik Integrated Healthcare Campus, Ankara, Türkiye

[2] Turkish Statistical Institute (TUIK), Ankara, Türkiye

[3] Department of Biostatistics and Medical Informatics, Istanbul University-Cerrahpaşa, Istanbul, Türkiye
mtarikalay@gmail.com, idemir@gmail.com, mkirisci@hotmail.com

Abstract
Objective: There are a limited number of pathogenic variants known in the MEFV gene. In silico tools fail to classify many MEFV gene variants. Therefore, it is essential to implement novel approaches. Our goal is to develop a new strategy to solve the even number classification problem while improving MEFV gene variant prediction accuracy using small datasets.
Material - methods: First, we determined the optimal number of computational tools for the model. We then applied eight distinct ML algorithms on the training dataset containing MEFV gene variants using the determined tools. We initiated the application of modified hard voting machine learning algorithms, using a training and validation dataset. Subsequently, we implemented a comparative analysis between the prediction results and existing algorithms and studies. Finally, we evaluated the gene and protein level ascertainment to identify hotspot regions.
Results: The ensemble classifier scored an average ROCAUC of 88%. The modified hard voting method correctly classified all known variants with 82% accuracy, outperforming both the soft voting (75%) and hard voting (70%) methods. The results showed that the prevalence of LP variants was approximately 2.5 times higher in domains compared to LB variants(χ2: 13.574, $p < 0.001$, OR: 2.509 [1.532-4.132]).
Conclusion: Considering the limited understanding of the clinical implications associated with MEFV gene mutations, employing a modified hard voting classifier approach may improve the classification accuracy of computational tools.

Keywords: Classification, FMF, Machine learning, MEFV, Voting Classifier

MEFY Gen Varyantlarında Modifiye Edilmiş Sert Oylama Sınıflandırıcısı Uygulaması, In-Silico Araç Performansını Artırıyor: Küçük Örneklem Boyutu İçin Yeni Bir Yaklaşım

Öz
Amaç: MEFV geninde bilinen sınırlı sayıda patojenik varyant bulunmaktadır. İn siliko araçlar, birçok MEFV gen varyantını sınıflandıramamaktadır. Bu nedenle, yeni yaklaşımların uygulanması gerekmektedir. Sert oylama sınıflandırıcıları ve sağlam doğrulama teknikleri sınıflandırma için kullanılabilir; ancak çift sayı sınıflandırması doğru bir şekilde yapılamamaktadır. Amacımız, hem çift sayı sınıflandırma sorununu çözmek hem de küçük veri setleri kullanarak MEFV gen varyantı tahmin doğruluğunu artırmak için yeni bir strateji geliştirmektir.
Yöntem: İlk olarak model için optimal sayıda hesaplama aracını belirledik. Daha sonra, belirlenen araçlar kullanılarak MEFV gen varyantlarını içeren eğitim veri setinde sekiz farklı makine öğrenme algoritması uygulandı. Eğitim ve doğrulama veri setinin kullanımıyla, modifiye edilmiş sert oylama makine öğrenme algoritmalarının uygulanmasına başlandı. Bundan sonra, tahmin sonuçları ile mevcut algoritmalar ve çalışmalar arasında karşılaştırmalı bir analiz gerçekleştirildi. Son olarak, gen ve protein düzeyinde değerlendirme yapılarak hotspot bölgeler belirlendi.
Bulgular: Topluluk sınıflandırıcısı, ortalama ROC AUC puanlarının %88 olduğunu gösterdi ve modifiye edilmiş sert oylama sınıflandırıcı yöntemi ile bilinen tüm varyantları %82 doğrulukla sınıflandırdı. Bu oran, hem yumuşak (%75) hem de sert oylama sınıflandırıcı (%70) yöntemlerinden daha yüksektir. Tüm varyantların kolektif değerlendirilmesi, LP varyantlarının, LB varyantlarına göre alanlarda yaklaşık 2,5 kat daha yaygın olduğunu ortaya koymuştur (χ2:13.574, $p < 0.001$, OR: 2.509 [1.532-4.132]).

* Corresponding Author.
E-mail: mtarikalay@gmail.com

Recieved : 14 June 2024
Revision : 22 Nov 2024
Accepted : 22 Jan 2025

Sonuç: MEFV gen mutasyonlarının klinik sonuçlarıyla ilgili bilgi yetersizliği göz önüne alındığında, modifiye edilmiş sert oylama sınıflandırıcı yaklaşımını kullanmak, hesaplama araçlarının sınıflandırma doğruluğunu artırmak için küçük örneklemlerde makul bir yöntem olabilir.
Anahtar Kelimeler: Sınıflandırma, Ailevi Akdeniz Ateşi, Makine Öğrenmesi, *MEFV*, Oylama Sınıflandırıcısı

1. Introduction

The widespread utilization of next-generation sequencing technology enhances the probability of diagnosing familial Mediterranean fever (FMF), ascertains the carrier rates within the population, and forecasts the likelihood of disease recurrence. Although the widespread utilization of Next-Generation Sequencing (NGS) assays has led to the discovery of a multitude of novel variants within the MEFV gene(Kırnaz, Gezgin and Berdeli, 2022)[1], The International Study Group on Systemic Autoinflammatory Disorders (INSAID) consensus criteria found that the clinical outcomes of more than half of the MEFV gene variants are categorized as variation of unknown significance fort the American College of Medical Genetics (ACMG)(Van Gijn *et al.*, 2018) [2].

Physicians and patients face difficulties in comprehending and interpreting the clinical implications of variants of uncertain significance (VOUS). In order to ascertain the clinical implications of the VOUS variant, it is necessary to conduct well-executed functional and hereditary investigations. However, these studies are associated with substantial costs and time requirements. Consequently, there is a need for innovative approaches that are both rapid and cost-effective, while also posing minimal risk, to predict the consequences of MEFV variants(Richards *et al.*, 2015; Nykamp *et al.*, 2017) [3, 4]. The utilization of existing variant prediction tools was considered as the second option. Nevertheless, there is a divergence of viewpoints regarding the selection and utilization of protein prediction methods and meta-predictors for the purpose of clinical variant evaluation, as highlighted by Richards et al. (Richards *et al.*, 2015)[3]. The ACMG and Clingen organizations have recommended conducting extensive evaluations at the gene level(Pyeritz and for the Professional Practice and Guidelines Committee, 2012; Stewart *et al.*, 2018; Harrison, Biesecker and Rehm, 2019; Burdon *et al.*, 2022; Lai *et al.*, 2022)5–9]. Although despite these efforts, it is still insufficient to accurately predict the clinical implications of most genes, including MEFV.

Our research endeavours focused on the exploration of a novel approach that incorporates an optimal selection of tools and machine learning algorithms, aiming to achieve a level of accuracy that is close to perfection. The accuracy of predicting outcomes is dependent on the training data exhibiting high levels of responsiveness. Therefore, the implementation of novel machine learning selection methods is expected to mitigate uncertainties. Nevertheless, numerous machine learning algorithms are currently employed in various amino acid prediction scores, meta scores, and ensemble algorithms. However, conventional machine learning (ML) algorithms are developed by choosing the classification method that yields the highest level of accuracy. This process fails to adequately acknowledge the success achieved by other machine learning algorithms. Many ML algorithms work well in large datasets(Song *et al.*, 2021). However, some datasets contain many uncertainties, making it impossible to achieve larger sample sizes(Accetturo, Bartolomeo and Stella, 2020; Alay, 2024). Therefore, in these situations, it is imperative to develop novel methodologies. Hard and soft voting classifiers are employed to enhance the performance of in silico tools and to evaluate the contribution of multiple scores to the classification process. However, hard voting classifiers perform binary classification (1 or 0), making accurate assignments in cases involving an even number of classifiers challenging. Many previous studies have reported difficulties in achieving consensus with an even number of algorithms(Awe *et al.*, 2024). To address this specific limitation, it may be beneficial to develop a method that incorporates only the most effective algorithms into the prediction process.In this study, we propose and evaluate a method called the "modified hard voting classifier" designed to overcome this issue.

This study aims to present a novel methodology for improving the accuracy of *MEFV* gene variant classification by utilizing optimal amino acid prediction scores and machine-learning algorithms. Our objective is to establish a more precise categorization of *MEFV* variants while minimizing uncertainty through the development of a new voting classifier. The findings of this study will provide valuable insights for clinicians in interpreting the clinical significance of variants with ambiguous effects on health outcomes and contribute to the development of gene-specific interpretation guidelines.

2. Material-Methods

2.1. Machine Learning Analyses

Libraries

Python were utilized for machine learning analysis step. The following libraries were utilized: sklearn for machine learning analysis, seaborn and matplotlib for data visualization, statsmodel for statistical models, and pandas and numpy for data manipulation. All versions of libraries were compatible with Python 3.7.1. Evaluation.

2.2. Data Retrieval Process

We obtained 389 MEFV variants from the Infevers database (https://infevers.umai-montpellier. fr/web, last access date:05/04/2022), focusing solely on single

nucleotide variants within the coding region such as missense and silent variants. Variants such as frameshift/inframe deletions, termination gain, termination loss, insertions, , and indels were omitted for analysis. In line with Clingen and ACMG guidelines, only clinically validated SNV predictors endorsed or evaluated by the Clingen group or ACMG guidelines were considered(Richards *et al.*, 2015; Ioannidis *et al.*, 2016; Tian *et al.*, 2019; Savige *et al.*, 2021; Pejaver *et al.*, 2022; Cheng *et al.*, 2023)

2.3.Feature Determination

During the process of selecting in-silico tools, we evaluated a number of conditions. First, we chose *in-silico* tools because they were up-to-date, validated, and recommended by ACMG guidelines(Richards *et al.*, 2015) and Clingen Group (Savige *et al.*, 2021; Waring *et al.*, 2021; Pejaver *et al.*, 2022; Wilcox *et al.*, 2022). Second, we meticulously determined that missing values for relevant variant scores in the entire dataset should not be included(Palanivinayagam and Damaševičius, 2023). Third, we compared all scores multicollinearity by using Spearman correlation. According to these rules, four in silico tools (Revel,MetaLR,SIFT, FATHM) were detected compatible with our algorithms. Other details of the selection *in silico tool* process are indicated in Supplementary File 1.

2.3. Feature Engineering

Data Preprocessing

For encoding dummy variables, the "Label Encoder" and "Ordinal Encoder" methods of sklearn.preprocessing are utilized. The "standard scaler" method was implemented for data standardization. The standard scaler method implemented after dataset split into training validation and prediction.

$\frac{X-u}{z}$ **(1)** Standardization(Z-score normalization)

Checking for Normality, Data Transformation and Dimension Reduction

We examined the distribution patterns for four distinct scoring metrics and determined that three of them were right-skewed, while the remaining one was left-skewed. In response to these findings, we applied square root and logarithmic transformations to normalize our dataset. Given the non-normal distribution of all four scores, we employed the Kruskal-Wallis H test as the appropriate non-parametric statistical method.

After conducting a thorough investigation, we found a total of 266 distinct MEFV gene mutations in our dataset. The breakdown can be outlined as follows: The recorded values are as follows: The distribution of the classifications of the variations is as follows: Benign (B): 3, Likely Benign (LB): 46, Likely Pathogenic (LP): 44, Pathogenic (P): 5, Variations of Unknown Significance (VOUS): 110, Not Categorized (NC): 26, Unsolved (US): 32. Given the diverse attributes of this dataset and the challenges associated with the seven-tier classification system, we recognized the need to decrease the number of dimensions to achieve a fairer and more understandable analysis. Several prior research employed the same methodology. (Accetturo *et al.*, 2020; Accetturo, Bartolomeo and Stella, 2020; Mighton *et al.*, 2022) Figure 1 provides a visual depiction and comparative examination of the seven-tier and three-tier classification systems. By employing dimensionality reduction techniques, we have circumvented the "curse of dimensionality," thereby defining boundaries that facilitate more accurate discrimination between damaging and benign genetic variants.

Upon conducting an extensive review, we identified 98 clinically recognized variants, evenly split between 49 likely benign and 49 likely pathogenic. These variants were selected to form a balanced training dataset. The dataset was subsequently partitioned, allocating 80% (n=78) for training purposes and 20% (n=20) for validation. Utilizing this set of clinically

Figure 1. MEFV variants distributions according to seven-tier and three-tier categories. For this step, we verified whether REVEL, SIFT, MetaLR, and FATHMM scores for the classes "LB" vs. "B" and "LP" vs. "P" did not show a statistically significant difference. In no cases, the medians of the two benign and two pathogenic classification groups demonstrated statistically significant differences (Kruskal-Wallis test not significant for non-parametric ANOVA of "LP" vs. "LP/P vs. "P" and "B" vs. "LB"). Therefore, we merged into LB and B as LB, LP and P as LP, and NC, VOUS, and US as VOUS. a1,b1) The frequency of variants according to infevers seven-tier classification system and our three-tier classification system,respectively. a2,a3,a4,a5) The box and plot distribution of infevers seven-tier classification according to Revel, MetaLR, SIFT, and FATHMM classification systems. b2,b3,b4,b5) The box and plot distribution of new three-tier classification according to Revel, MetaLR, SIFT, and FATHMM classification systems c1,d1) Comparison of variants according to exonic placements betwen seven-tier Infevers Classification and three-tier new classification system. Most of the pathogenic variants were placed in exon 10. c2,c3,c4,c5) Variant distribution by cDNA position according to Infevers seven-tier classification system. No certain pattern of clustering detected d2,d3,d4,d5) Variant distribution by cDNA position according to new three-tier classification system. Higher than 0.9 Revel scores most likely associated with variant pathogenicity similar to Clingen PP3

classification evaluation. No clear distinguished threshold is evident for other scores e1,f1) Variant distribution according to pyrin protein domains. PF02758: PAAD/DAPIN/pyrin domain, PF00643: Domain b-Box Zinc Finger domain, PF13765:SPRY-associated domain, PF00622: SPRY domain Most of the pathogenic variants placed in SPRY domain of pyrin protein. e2,e3,e4,e5) Variant distribution by aminoacid position according to infevers seven-tier classification system. f2,f3,f4,f5) Variant distribution by aminoacid position according to new three-tier classification system

Corroborated variants, our aim was to ascertain the optimal number of features necessary for reliable predictions. A review of existing literature, coupled with sample size determinations, revealed that a quartet of in-silico tools yielded the most favorable performance (Ogundimu, Altman and Collins, 2016; Riley *et al.*, 2019; Accetturo *et al.*, 2020; Acharjee *et al.*, 2020; Luan *et al.*, 2020).

2.4. Feature Selection

2.4.1. Selection of Machine learning Methods

We utilized seven machine learning techniques—K-nearest neighbor (KNN), Decision Tree(DT), Random Forest (RF), Multilayer perceptron Logistic regression (LR), Linear Support Vector Machine (SVM-linear), and Radial basis function Support Vector Machine (SVM-RBF)—to analyze four scores (SIFT, FATHMM, Revel, and MetaLR).

2.4.2. Dataset Evaluation

We trained RF, DT, KNN, LR, LSVM, KSVM, and PSVM on four scores (REVEL, MetaLR, SIFT, FATHMM). We did k-fold crossvalidation, leave one out of crossvalidation, leave p out of crossvalidation, and validation dataset techniques for a model validation and generalizability techniques. As compatible with our dataset nature k-fold cross-validation put forward best results with 10 values. As our training dataset was balanced, so we determined our threshold value according to the accuracy score. We used other paramaters such as precision, recall and F1 metrics for dataset evaluation explained at Supplementary File S2.

2.4.3. Modified Hard Voting Classifier

Problem

A Hard Voting Classifier cannot make an assignment in the case of a tie. In such instances, weighting is necessary; however, applying weights requires prior assumptions about these characteristics. This approach neglects the individual performance metrics of the algorithms. A new classification method that considers performance metrics for binary classification could resolve this issue for the Hard Voting Classifier.

Formulas

Given a set of n algorithms $\{A_1, A_2, \ldots, A_n\}$, where n is an even number and $\frac{n}{2}$ is an odd number, we aim to select machine learning algorithms with the highest ROCAUC scores and use hard voting to combine their predictions. Let the ROCAUC scores of algorithms be $\{ROC_{A1}, ROC_{A2}, \ldots, ROC_{An}\}$.

1. **Select the algorithms with ROC AUC scores greater than 0.80:**
   Successful_algorithms = $\{ A_i \mid ROC_{A_i} > 0.80 \}$
2. **Iterative Reduction to an Odd Number**
   While the number of successful algorithms is even and greater than 1, reduce it by half: While $|Successful_{algorithms}| \% 2 = 0$ and $|Successful_{algorithms}| > 1$:
   Successful_algorithms = $\{ Ai \mid Accuracy_{A_i} \text{ in top half of } Accuracy \text{ scores of } Successful\_algorithms \}$
3. **Final Set of Algorithms**
   After the iterative reduction, let $\{A_{f1}, A_{f2}, \ldots, A_{fm}\}$ be the final set of algorithms, where m is an odd number.
4. **Hard Voting Classifier**
   Combine the predictions of the final set of algorithms using a hard voting mechanism: $\hat{y} = argmax_k \sum_{j=1}^{m} \chi (C_{A_{fj}}(x) = k)$
   $C_{A_{fj}}(x)$ is the prediction of algorithm for A_{fj} instance x, and χ is the indicator function that equals 1 if the condition is true and 0 otherwise.

Application

We conducted assessments of machine learning algorithms with a focus on those that exceeded a pre-established accuracy threshold. Our analysis techniques were built on ensemble models, specifically using a voting prediction approach. This method did not assign weighted scores for predictive accuracy; instead, our classification system was binary, labeling outcomes as either "classified" or "not classified." For instance, should all three machine learning algorithms concur in identifying variant "X" as LP, it would receive a score of “3” and be categorized accordingly as LP. Conversely, if only two algorithms determined variant "Y" to be LB, it would garner a score of “2” and be categorized as LB. Thus, our method operates under stringent criteria without utilizing weighted scoring, leading us to describe it as a “modified hard voting classifier.”

Consequently, we predicted our variants similar to the hard voting classifier algorithm. However, four approaches were different from the hard voting classifier: (1) Voting classifier did not solve even numbers classification problem. (2)We did not only implemented hard voting classifier on not only training and validation dataset, but also prediction scores. (3) Different from classic hard voting classifier we did not calculate all scores, we only included voting a showed outstanding area under curve scores which was accepted as higher than 80%. (4) We selected each algorithms

best paramaters not only combination of best paramaters of scores [Figure 2].

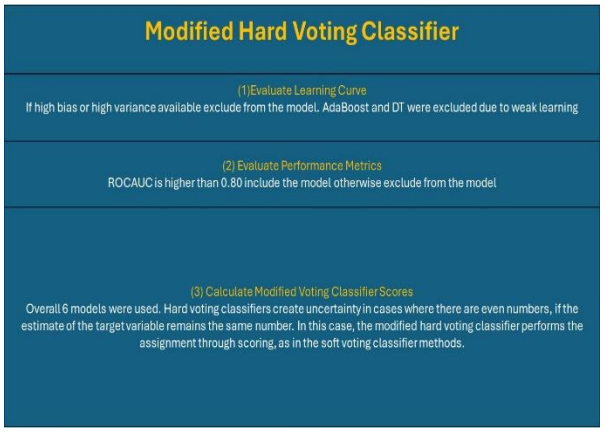


Figure 2. Establishing a modified Hard Voting Classifier

2.5. Functional and Clinical Level Evaluation

We evaluated each variant in two categories for functional-level ascertainment: gene-level and protein-level. While we established gene-level evaluation based on exonic position, we implemented protein-level evaluation by comparison of pyrin protein domain distributions. We evaluated *MEFV* domains initiation and termination location according to protein databank (https://www.rcsb.org/), Ensemble, Prosite (https://www.expasy.org/resources/prosite), conserved domain databases (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), InterPro (https://www.ebi.ac.uk/interpro/), and existing literature(Grandemange *et al.,* 2011). *MEFV* (NM_000243.3)

transcripts were based on variants distribution on pyrin protein.

2.6. Sample Size Calculation

Before implementing machine learning analysis, we had to conduct sample size calculations. Our sample size calculation was implemented on the basis of study of Accetturo et al. (Accetturo *et al.*, 2020). Furthermore, the metapredictors and aminoacid prediction tools that we implemented in the study have already been trained on a larger dataset(Waring *et al.*, 2021; Pejaver *et al.*, 2022; Sallah *et al.*, 2022). Considering the number features and sample size it is sufficient to implement machine learning methods(Ogundimu, Altman and Collins, 2016; Riley *et al.*, 2019; Luan *et al.*, 2020; Rajput, Wang and Chen, 2023).

2.7. Statistical Analysis

Our statistical analyses were performed utilizing Python version 3.7.1 alongside SPSS version 25.0 for Windows (IBM, Chicago, IL). We established a 95% confidence interval for the entirety of our statistical tests. The significance levels, denoted as alpha (α) and beta (β), were set at 0.05 and 0.20, respectively. A p-value threshold was determined to be 0.05, with values falling below this cutoff being considered statistically significant.

To evaluate the distribution of both discrete and continuous numerical variables, normality was probed using a suite of graphical and analytical techniques. Conformity with normal distribution assumptions allowed the use of means and standard deviations; in their absence, medians and interquartile ranges were employed. Categorical variables, either nominal or ordinal, were quantified and expressed as frequencies and percentages, with ordinal variables arranged according to their inherent hierarchy.

Graphical methods such as Q-Q plots, detrended plots, boxplots, histograms, and stem-and-leaf plots, alongside the analytical Kolmogorov-Smirnov test, were utilized to assess the normality of the data. The range for skewness and kurtosis was considered acceptable between -1 and +1, while skewness and kurtosis indices – calculated by dividing the respective values by their standard errors – were deemed to reflect normality when falling within the -2 to +2 range.

For variables that adhered to normal distribution, we applied the Analysis of Variance test. This was followed by post hoc analysis using the Tukey test to identify significant pairwise differences. In the case of non-normally distributed data, the non-parametric Kruskal-Wallis H-test was administered, succeeded by the Dunn-Bonferroni test for post hoc comparisons.

3. Results

3.1. Evaluation of Training dataset

It is highly recommended that, when developing novel algorithms, one should not only concentrate on novelty but also identify a good feature dataset(Khalid and Sezerman, 2018). Therefore, we examined our training dataset and determined the threshold of 80% ROCAUC required for machine learning algorithms to succeed. All algorithms exceeded the threshold value. However, Adaboost and DT were excluded from the model due to their overfitting[Supplementary Figure 3 and 4]. Revel score is detected as the most important feature when classifying datasets according to the most accurate classifier algorithm, RF [Figure 3]

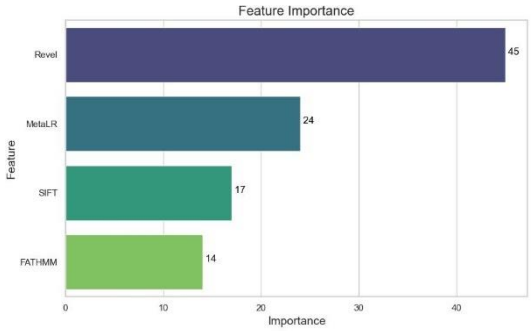


Figure 3. Feature Importance Metrics according to Random Forest Classifier. Revel is the most important feature which is contributed roughly about 50% to classifier.

3.2. Validation and hyperparamater tuning

Overall 98 known variants analyzed under two dataset: training dataset (n=78) and validation dataset (n=20). All. After implementing the machine learning algorithm, we conducted hyperparameter tuning for our accurate machine learning classifier algorithms [Table 1]. The learning curve demonstrates low bias and variance, even when trained on a small dataset, indicating robust and reliable model performance [Figure 4]. K-fold Crossvalidation (CV) and nested CV methods were used for validation methods which were more robust to sample size(Vabalas *et al.*, 2019; Larracy, Phinyomark and Scheme, 2021; Dalmaijer, Nord and Astle, 2022).

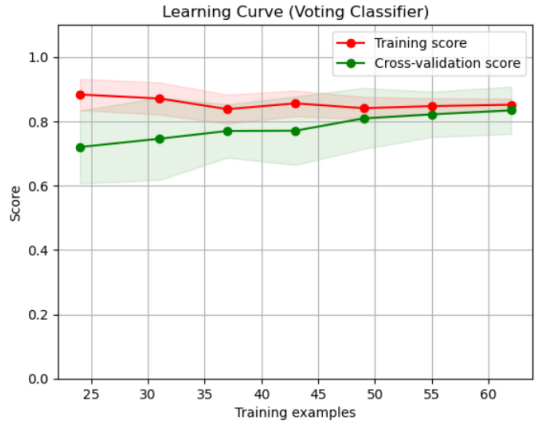


Figure 4. Modified Hard Voting Classifier Learning Curve. Both the training and validation curves were generated from a small dataset, exhibiting relatively low bias and variance, thus indicating a robust and reliable model performance.

Table 1. Cross validation results in validation dataset(n=78 for training dataset, and n=20 for validation dataset)

ML Methods	Precision	Recall	Accuracy	ROC AUC
LR	0.79	0.75	0.76	**0.9**
SVM-RBF	**0.81**	0.77	0.77	0.89
SVM-Linear	0.77	**0.86**	**0.81**	**0.9**
Gaussian NB	**0.89**	0.75	**0.81**	0.89
kNN	**0.82**	0.77	0.79	0.85
RF	**0.86**	**0.79**	**0.82**	**0.91**

Stratified K-fold CV implemented (The best results obtained in 10-fold CV. The 10-fold CV results were represented above). The results were checked with nested cv which is more robust to sample size. The similar results were obtained. The best parameters obtained for each classifier are as follows: **RF classifier**: - 'bootstrap': False - 'max_depth': None - 'min_samples_leaf': 3 - 'min_samples_split': 10 - 'n_estimators': 10 max_features=3 **KNN classifier**: - 'algorithm': 'auto' - 'n_neighbors': 3 - 'p': 2 - 'weights': 'uniform'. **DT classifier:** - 'criterion': 'entropy' - 'max_depth': None - 'min_samples_leaf': 2 - 'min_samples_split': 2 **NB Classifier** n_jobs=-1, cv=5, verbose=5, var_smoothing= 1e-6 , **LR Classifier** penalty=L2, C:1000, **SVM Classifier** 'C': 1000, 'gamma': 0.01, 'kernel': 'rbf' **SVM-linear Classifier** {'C': 1000, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False} **SVM-RBF classifier** {'C': 1000, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}. For hyperparameter optimization, the GridSearchCV algorithm was implemented with a 10-fold CV. However, due to the overfitting observed in DT and AdaBoost Classifier, both of the algorithms were excluded from the analysis.

3.3. Comparison of the Training Dataset Results with Existing Literature

The next step we compared our results with existing literature and scores we used in our dataset. According to this comparison, modified hard voting classifier has most accurate classifying known variants [Figure 5]. The mean ROCAUC of six remaining ML methods was detected as 88% in both training and validation dataset. Interestingly, modified hard voting classifier classified more than 82% of known variants correctly in overall (training and validation) dataset. In the literature, the second most accurate classifier was Linear Discriminant Analysis conducted by Accetturo et al. classified variants with 75 % accuracy(Accetturo *et al.*, 2020). Most of the predictors classified LB variants with higher ROCUAC scores than 80%; however, LP classification showed a wide range of variety in accuracy scores between 2% - 62.5(Ioannidis *et al.*, 2016; Liu *et al.*, 2016; Knecht *et al.*, 2017; Tian *et al.*, 2019; Accetturo *et al.*, 2020).

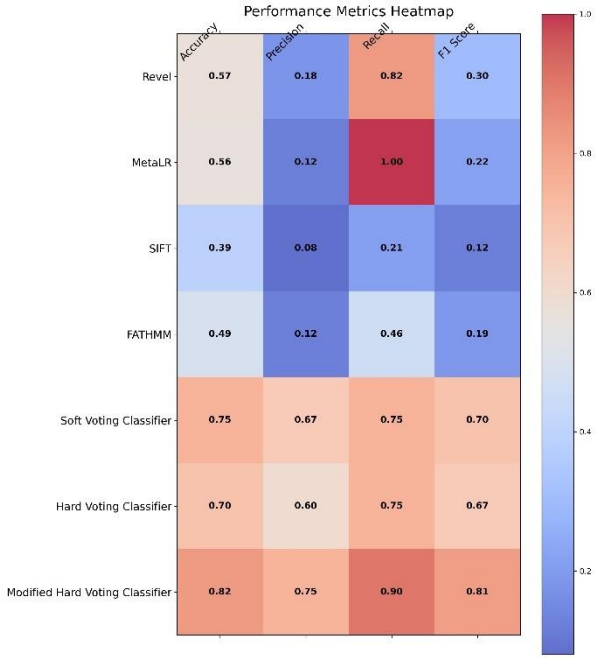


Figure 5. Comparison of modified hard voting classifier with existing algorithms by evaluating their success in classifying known variants. This figure illustrates the improved classification metrics achieved by a modified hard voting classifier for the prediction of MEFV gene variants. Traditional in silico predictors have struggled to distinguish MEFV gene variants, often performing at levels comparable to random chance. The modified hard voting classifier, however, demonstrates enhanced accuracy, sensitivity, and specificity, showcasing its superior discriminatory power in the analysis of MEFV gene variants. Additionally, this classifier has improved the classification performance of the existing hard voting classifier. As a result, it has outperformed the soft voting classifier. The modified hard voting classifier, especially for small sample sizes, can be combined with well-tuned k-fold cross-validation or nested CV methods, which are not significantly affected by the sample size.

3.4. Prediction Outcomes and Evaluation of Machine Learning Algorithms on VOUS variants

After the voting classification of training (n=78) and validation (n=20) dataset, overall 94 out of 98 (95.91%) variants were classified accurately in our dataset. The same prediction implemented for VOUS variants. Overall, we found 85 LP variants and 83 LB variants. As a result, we discovered 134 LP variants and 132 LB variants in the overall dataset. New distribution of all *MEFV* gene variants indicated in Figure 6.

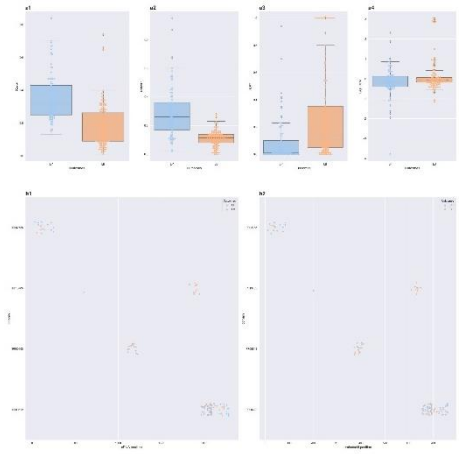


Figure 6. Visualization of Prediction Algorithm Results. a) While, Revel, MetaLR, and SIFT algorithms contributed statistically significant effect on model ($p<0.05$), FATHMM algorithm plays a supporting role on it. b1) Domain distribution of MEFV gene variants prediction results according to cDNA position b2) Domain distribution of MEFV gene variants prediction results according to aminoacid position. While most of the pathogenic variants distributed into PF00622 domain (2.595[1.525-4.425], $p<0.001$) , most of the benign variants distributed into PF00643 domain. PF02758: PAAD/DAPIN/pyrin domain, PF00643: Domain b-Box Zinc Finger domain, PF13765:SPRY-associated domain, PF00622: SPRY domain Most of the pathogenic variants placed in SPRY domain of pyrin protein.

3.5. Functional Evaluation

3.5.1. Gene-level (Exonic) Ascertainment

In our initial assessment of variants of uncertain significance, we ascertained that exon 10 harbored 37.6% (32/85) of variants predicted as likely pathogenic, whereas exon 2 contained 41.0% (34/83) of those predicted to be likely benign. Through our prediction methodology, it was concluded that 61.5% (32/52) of exon 10 variants and 58.6% (34/58) of exon 2 variants were classified as LP and LB, respectively. A disproportionate distribution was observed, with exons 7, 9, and 10 presenting a greater prevalence of LP variants in contrast to the preponderance of LB variants in other exons. In particular in exon 10, 42.5% (57/134) of the variants were categorized as likely pathogenic (LP), and in exon 2, 43.2% (57/132) were classified as likely benign (LB). Statistical analysis demonstrated a significant discrepancy in the distribution between LP and LB variants in these exons. As a result of our gene-level analyses revealed that exon 10 variants were 2.6 times more prone to be classified as LP than LB ($\chi2$: 12.858, $p < 0.001$, odds ratio [OR]: 2.629; 95% CI: 1.539-4.493). In contrast, exon 2 variants had a higher likelihood of being labeled as LB compared to LP ($\chi2$: 12.693, $p < 0.001$, OR: 2.595; 95% CI: 1.532-4.132). Afterwards, we combined our prediction outcomes with the training datasets (LB and LP) and assessed them according to exonic positions [Table 2].

Table 2. Distribution of all variants by exons and variant prediction outcomes

Exons	Variant prediction outcomes		p values	95% Interval		
	LB (n=132)	LP (n=134)		Odds ratios	Lower	Upper
1	4	11	0.118[a]	2.862	0.887	9.228
2	57	30	**<0.001**[b]	0.380	0.223	0.646
3	20	14	0.334[a]	0.653	0.315	1.356
4	3	2	0.683[c]	0.652	0.107	3.963
5	14	10	0.463[a]	0.680	0.291	1.590
6	-	-	*	*	*	*
7	-	4	0.122[c]	*	*	*
8	1	4	0.370[c]	4.031	0.445	36.549
9	4	2	0.445[c]	0.485	0.087	2.693
10	29	57	**<0.01**[b]	2.629	1.539	4.493

a Chi-square(Yates correction), b. Pearson chi-square test, c. Fisher Exact test,*not calculated. Evaluation of each exon is based on the LP to LB ratio.

3.5.2. Protein-Level (Domain-based) Evaluation

Within the domains, we properly identified 47% (40/85) predicted LP variants and 28.92% (24/83) predicted LB variants. After assessing the anticipated variations (n = 168), it was discovered that variants located within the domain were 2.766 times more likely to be classified as LP compared to LB (2:10.566, p:0.002, OR: 2.766 [1.462-5.233]). Subsequently, we combined the training dataset with the predicted VOUS variants. After collectively evaluating all variants, we found that LP variants were approximately 2.5 times more common in domains compared to LB variants (χ2:13.574, $p < 0.001$, OR: 2.509 [1.532-4.132]). On the other hand, B30.2 domain variants had a 2.5-fold higher likelihood of being LP compared to LB. This difference was statistically significant (χ2:12.693, $p < 0.001$, OR: 2.595 [1.532-4.132]). Nevertheless, the likelihood of variants that were not found in any domains being LB was 2.6 times higher compared to LP (χ2:14.508, $p < 0.001$, OR: 0.386 [0.235–0.633]). Upon identifying this statistically significant disparity, we assessed all variations within their respective domains [Table 3].

Table 3. Distribution of all variants by domains and variant prediction outcomes

Domains	Variant prediction outcomes		p values	95% Interval		
	LB (n=132)	LP (n=134)		Odds ratios	Lower	Upper
PYD	4	11	0.118[a]	2.862	0.887	9.228
bZIP	6	1	0.065[b]	0.158	0.019	1.330
B	4	8	0.390[a]	2.032	0.597	6.918
CC	4	2	0.445[b]	0.485	0.087	2.693
B30.2	31	58	**<0.001**[c]	2.595	1.525	4.425
Not identified	83	54	**<0.001**[c]	0.386	0.235	0.633

a Chi-square (Yates correction), b Fisher Exact test c. Chi-square test

4. Discussion

Many novel ML algorithms are designed to predict outcomes for larger datasets. However, few strategies are available for small datasets(Liu *et al.*, 2013; Vabalas *et al.*, 2019; Albaradei *et al.*, 2021; El-Sofany, Bouallegue and El-Latif, 2024). In this context, the modified hard voting classifier demonstrates superior performance, surpassing traditional hard voting and soft voting methods while effectively addressing challenges such as odd-number classification, which refers to scenarios where standard voting methods struggle to

make definitive decisions in cases with an uneven distribution of votes. By optimizing predictions, this approach enhances the accuracy of in silico tools and offers a reliable solution for analyses involving limited sample sizes.

New applications and Implementation Steps

This study includes a number of enhanced methods and new technologies. We base our new approach on a three-fold framework. The first step involves using big data analysis and comparing the datasets with existing algorithms and previous research findings. The second and the third steps include functional and protein-level evaluations, respectively.

Evaluation of Results

Initially, we applied seven machine learning algorithms on the training set, specifically the LP and LB variants. For the prediction of VOUS variants, we selected three out of the six machine learning techniques that had a minimum ROCUAC of 80%. We obtained 88% mean ROCAUC results for all 6 algorithms: LR, SVM-RBF, SVM-linear, Gaussian NB, KNN, RF. According to our sample, our voting classifier model correctly classified LB and LP variants. Subsequently, we assessed our training dataset by comparing it to established variant prediction tools and previous research. Based on the comparison results, the modified hard voting classifier method demonstrated superior performance in classifying MEFV variants compared to existing *in silico* algorithms and previous studies(Accetturo *et al.*, 2020). In the second and third steps, we conducted a comprehensive functional level analysis, evaluating all variants from both gene-level and protein-level perspectives. Our analysis at the functional level revealed that the SPRY domain(Papin *et al.*, 2007), which corresponds to exon 10 (Dundar *et al.*, 2022)and accounts for a significant portion of predicted damaging MEFV gene variants, exhibited a statistically significant increase in LP variants in non-evolutionarily conserved regions. However, this increase was nearly equivalent to that observed in other evolutionarily conserved regions, and the difference was not statistically significant when compared to these conserved regions.

Modified hard voting classifier

The modified hard voting classifier introduces several novelties in the literature. First of all, the modified hard voting classifier approach incorporates an optimal quantity of protein prediction tools(Ng and Henikoff, 2003) or meta-predictors(Ioannidis *et al.*, 2016). Additionally, this method assesses the influence of all effective machine-learning techniques. To the best of our knowledge, we have made the initial modification to a hard voting classifier for the purpose of variant classification and two distinct classification methods. Rather than relying on the traditional hard voting classifier, which explicitly votes on a single target variable, our approach utilizes both LB and LP variations, establishing a precise threshold for decision-making. Models that exhibited overfitting or underfitting were systematically eliminated, and the voting process was repeated until an optimized model was identified for predicting classification outcomes. Each model was evaluated with its own optimal parameters, ensuring rigorous performance testing. The modified hard voting classifier incorporates voting mechanisms to provide a rigorous classification procedure. Our high training data accuracy score stems from an optimum number of tools(Megantara and Ahmad, 2021; Hu *et al.*, 2024). In contrast to the first study on *MEFV* gene unknown variant prediction conducted by Accetturo et al.(Accetturo *et al.*, 2020), and existing tools , our prediction was derived from an ensemble method rather than relying on the most effective sole machine learning algorithm.

Literature review

The existing study provides a significant contribution to the literature by offering innovative solutions to three issues that previous *in-silico* tools have failed to address. The first issue is that current methods fail to successfully classify MEFV gene variants using numerous variant prediction algorithms (Accetturo *et al.*, 2020). Therefore, it is difficult to interpret variants according to current in silico tools (Ioannidis *et al.*, 2016). However, the modified hard voting classifier does not rely solely on a single *in-silico* tool or one ML method. The selection criteria of ML methods and *in-silico* tools are based on strict criteria, and only include most accurate methods or best features. Second, a significant issue is that during the variant classification process, many predictors correctly classify benign variants; however, many tools often fail to detect pathogenic variants accurately at the desired level (Adzhubei, Jordan and Sunyaev, 2013; Knecht *et al.*, 2017; Fortuno *et al.*, 2018; Pejaver *et al.*, 2022; Wilcox *et al.*, 2022). The comparative analysis revealed that our innovative methodology, the modified hard voting classifier, outperformed current in silico algorithms in classifying MEFV variants. This outcome arises from the modified hard voting classifier, which depends on a consensus of multiple machine learning techniques. Third, significant novel tools present better results day after day; unfortunately, still many variants remain unresolved. Even the newly developed in silico tool, Alphamissense, cannot classify 20% of all gene variants (Cheng *et al.*, 2023). The modified hard voting classifier effectively resolves uncertainties in variant interpretation.

Limitations of the study

Although this method produces very high classification rates, it has some drawbacks when applied to our dataset. First, identifying the optimal classifiers from among hundreds of in silico tools remains a challenging task. (Gunning *et al.*, 2021; Cheng *et al.*, 2023). Therefore, we only applied ClinGen- and ACMG-recommended tools(Waring *et al.*, 2021; Pejaver *et al.*, 2022; Wilcox *et al.*, 2022). This approach enabled us to use more reliable tools in our study. Second, due to the lack of research on MEFV gene

classification, we had to base our sample size calculations on the study by Accetturo et al. (Accetturo *et al.*, 2020). However, we also confirmed this 'optimum number of features' by looking at the literature and their classification accuracy (Vu and Braga-Neto, 2009; Accetturo *et al.*, 2020).

The main drawback of our study is the absence of validation via clinical or functional studies. While integrating our model with the ClinVar dataset could provide an avenue for external validation, ClinVar currently reports only 33 missense variants (Accessed: 12/5/2024). As we have already integrated all these variants into our training dataset, we could not utilize them as an external validation dataset. However, the explicit methodology of the modified hard voting classifier facilitates its straightforward application to diverse datasets. Further studies are necessary to fully understand the efficacy of the modified hard voting classifier.

5. Conclusion

Brief Summary of Findings and Evaluation of the study

This study holds significance for both machine learning applications and routine clinical practice. In this work, an algorithm was developed to enhance the performance of hard voting classifiers, demonstrating optimal results even with small sample sizes. However, the primary limitation of the study is that it has not been validated on an external dataset.

Consequently, this approach addresses the three previously identified gaps in in silico tools, reduces existing prediction errors of other in silico tools by offering gene-specific optimization, and, most importantly, provides an alternative method for bioinformaticians working on in silico tool optimization while also serving as a helpful tool for clinicians. Given that 60% of the clinical implications associated with MEFV gene variants are still incompletely understood, it would be advantageous to apply a modified hard voting vote classifier approach to enhance the classification accuracy of machine learning techniques. However, more testing of the improved modified hard voting approach is required on other gene variations.

Future Implications

The impact of this modified hard voting classifier on other datasets also needs to be evaluated to better understand its significance compared to the standard hard voting classifier. Additionally, from a clinical perspective, functional studies specifically designed for the *MEFV* gene are required to fully comprehend the true success of these classifications.

6. References

Accetturo, M. *et al.* (2020) 'Improvement of MEFV gene variants classification to aid treatment decision making in familial Mediterranean fever.', *Rheumatology (Oxford, England)*, 59(4), pp. 754–761. Available at: https://doi.org/10.1093/rheumatology/kez332.

Accetturo, M., Bartolomeo, N. and Stella, A. (2020) 'In-silico Analysis of NF1 Missense Variants in ClinVar: Translating Variant Predictions into Variant Interpretation and Classification.', *International journal of molecular sciences*, 21(3). Available at: https://doi.org/10.3390/ijms21030721.

Acharjee, A. *et al.* (2020) 'A random forest based biomarker discovery and power analysis framework for diagnostics research', *BMC Medical Genomics*, 13(1), p. 178. Available at: https://doi.org/10.1186/s12920-020-00826-6.

Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013) 'Predicting functional effect of human missense mutations using PolyPhen-2.', *Current protocols in human genetics*, Chapter 7, p. Unit7.20. Available at: https://doi.org/10.1002/0471142905.hg0720s76.

Alay, M.T. (2024) 'An Ensemble Model Based on Combining BayesDel and Revel Scores Indicates Outstanding Performance: Importance of Outlier Detection and Comparison of Models', *Cerrahpasa Medical Journal*, 48(2), pp. 179–184.

Albaradei, S. *et al.* (2021) 'Machine learning and deep learning methods that use omics data for metastasis prediction.', *Computational and structural biotechnology journal*, 19, pp. 5008–5018. Available at: https://doi.org/10.1016/j.csbj.2021.09.001.

Awe, O.O. *et al.* (2024) 'Weighted hard and soft voting ensemble machine learning classifiers: Application to anaemia diagnosis', in *Sustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network, Ghana, 2022*. Springer, pp. 351–374.

Burdon, K.P. *et al.* (2022) 'Specifications of the ACMG/AMP variant curation guidelines for myocilin: Recommendations from the clingen glaucoma expert panel.', *Human mutation*, 43(12), pp. 2170–2186. Available at: https://doi.org/10.1002/humu.24482.

Cheng, J. *et al.* (2023) 'Accurate proteome-wide missense variant effect prediction with AlphaMissense.', *Science (New York, N.Y.)*, 381(6664), p. eadg7492. Available at: https://doi.org/10.1126/science.adg7492.

Dalmaijer, E.S., Nord, C.L. and Astle, D.E. (2022) 'Statistical power for cluster analysis', *BMC Bioinformatics*, 23(1), pp. 1–28. Available at: https://doi.org/10.1186/s12859-022-04675-1.

Dundar, M. *et al.* (2022) 'Clinical and molecular evaluation of MEFV gene variants in the Turkish population: a study by the National Genetics Consortium.', *Functional & integrative genomics*, 22(3), pp. 291–315. Available at: https://doi.org/10.1007/s10142-021-00819-3.

El-Sofany, H., Bouallegue, B. and El-Latif, Y.M.A. (2024) 'A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method.',

Scientific reports, 14(1), p. 23277. Available at: https://doi.org/10.1038/s41598-024-74656-2.

Fortuno, C. *et al.* (2018) 'Improved, ACMG-compliant, in silico prediction of pathogenicity for missense substitutions encoded by TP53 variants.', *Human mutation*, 39(8), pp. 1061–1069. Available at: https://doi.org/10.1002/humu.23553.

Van Gijn, M.E. *et al.* (2018) 'New workflow for classification of genetic variants' pathogenicity applied to hereditary recurrent fevers by the International Study Group for Systemic Autoinflammatory Diseases (INSAID).', *Journal of medical genetics*, 55(8), pp. 530–537. Available at: https://doi.org/10.1136/jmedgenet-2017-105216.

Grandemange, S. *et al.* (2011) 'The regulation of MEFV expression and its role in health and familial Mediterranean fever', *Genes & Immunity*, 12(7), pp. 497–503. Available at: https://doi.org/10.1038/gene.2011.53.

Gunning, A.C. *et al.* (2021) 'Assessing performance of pathogenicity predictors using clinically relevant variant datasets', *Journal of Medical Genetics*, 58(8), pp. 547–555. Available at: https://doi.org/10.1136/jmedgenet-2020-107003.

Harrison, S.M., Biesecker, L.G. and Rehm, H.L. (2019) 'Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines.', *Current protocols in human genetics*, 103(1), p. e93. Available at: https://doi.org/10.1002/cphg.93.

Hu, Y.-H. *et al.* (2024) 'A novel MissForest-based missing values imputation approach with recursive feature elimination in medical applications', *BMC Medical Research Methodology*, 24(1), p. 269. Available at: https://doi.org/10.1186/s12874-024-02392-2.

Ioannidis, N.M. *et al.* (2016) 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.', *American journal of human genetics*, 99(4), pp. 877–885. Available at: https://doi.org/10.1016/j.ajhg.2016.08.016.

Khalid, Z. and Sezerman, O.U. (2018) 'Computational drug repurposing to predict approved and novel drug-disease associations', *Journal of Molecular Graphics and Modelling*, 85, pp. 91–96. Available at: https://doi.org/https://doi.org/10.1016/j.jmgm.2018.08.005.

Kırnaz, B., Gezgin, Y. and Berdeli, A. (2022) 'MEFV gene allele frequency and genotype distribution in 3230 patients' analyses by next generation sequencing methods.', *Gene*, 827, p. 146447. Available at: https://doi.org/10.1016/j.gene.2022.146447.

Knecht, C. *et al.* (2017) 'IMHOTEP-a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants.', *Nucleic acids research*, 45(3), p. e13. Available at: https://doi.org/10.1093/nar/gkw886.

Lai, A. *et al.* (2022) 'The ClinGen Brain Malformation Variant Curation Expert Panel: Rules for somatic variants in AKT3, MTOR, PIK3CA, and PIK3R2.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 24(11), pp. 2240–2248. Available at: https://doi.org/10.1016/j.gim.2022.07.020.

Larracy, R., Phinyomark, A. and Scheme, E. (2021) 'Machine learning model validation for early stage studies with small sample sizes', in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, pp. 2314–2319.

Liu, C. *et al.* (2013) 'Applications of machine learning in genomics and systems biology.', *Computational and mathematical methods in medicine*, p. 587492. Available at: https://doi.org/10.1155/2013/587492.

Liu, X. *et al.* (2016) 'dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.', *Human mutation*, 37(3), pp. 235–241. Available at: https://doi.org/10.1002/humu.22932.

Luan, J. *et al.* (2020) 'The predictive performances of random forest models with limited sample size and different species traits', *Fisheries Research*, 227, p. 105534. Available at: https://doi.org/https://doi.org/10.1016/j.fishres.2020.105534.

Megantara, A.A. and Ahmad, T. (2021) 'A hybrid machine learning method for increasing the performance of network intrusion detection systems', *Journal of Big Data*, 8(1). Available at: https://doi.org/10.1186/s40537-021-00531-w.

Mighton, C. *et al.* (2022) 'Data sharing to improve concordance in variant interpretation across laboratories: results from the Canadian Open Genetics Repository', *Journal of Medical Genetics*, 59(6), pp. 571 LP – 578. Available at: https://doi.org/10.1136/jmedgenet-2021-107738.

Ng, P.C. and Henikoff, S. (2003) 'SIFT: Predicting amino acid changes that affect protein function.', *Nucleic acids research*, 31(13), pp. 3812–3814. Available at: https://doi.org/10.1093/nar/gkg509.

Nykamp, K. *et al.* (2017) 'Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 19(10), pp. 1105–1117. Available at: https://doi.org/10.1038/gim.2017.37.

Ogundimu, E.O., Altman, D.G. and Collins, G.S. (2016) 'Adequate sample size for developing prediction models is not simply related to events per variable.', *Journal of clinical epidemiology*, 76, pp. 175–182. Available at: https://doi.org/10.1016/j.jclinepi.2016.02.031.

Palanivinayagam, A. and Damaševičius, R. (2023) 'Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods', *Information*, 14(2), p. 92.

Papin, S. *et al.* (2007) 'The SPRY domain of Pyrin, mutated in familial Mediterranean fever patients, interacts with inflammasome components and inhibits proIL-1beta processing.', *Cell death and differentiation*, 14(8), pp. 1457–1466. Available at: https://doi.org/10.1038/sj.cdd.4402142.

Pejaver, V. *et al.* (2022) 'Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria.', *American journal of human genetics*, 109(12), pp. 2163–2177. Available at: https://doi.org/10.1016/j.ajhg.2022.10.013.

Pyeritz, R.E. and for the Professional Practice and

Guidelines Committee, A. (2012) 'Evaluation of the adolescent or adult with some features of Marfan syndrome', *Genetics in Medicine*, 14(1), pp. 171–177. Available at: https://doi.org/10.1038/gim.2011.48.

Rajput, D., Wang, W.-J. and Chen, C.-C. (2023) 'Evaluation of a decided sample size in machine learning applications', *BMC Bioinformatics*, 24(1), p. 48. Available at: https://doi.org/10.1186/s12859-023-05156-9.

Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 17(5), pp. 405–424. Available at: https://doi.org/10.1038/gim.2015.30.

Riley, R.D. *et al.* (2019) 'Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes', *Statistics in medicine*, 38(7), pp. 1276–1296.

Sallah, S.R. *et al.* (2022) 'Improving the clinical interpretation of missense variants in X linked genes using structural analysis.', *Journal of medical genetics*, 59(4), pp. 385–392. Available at: https://doi.org/10.1136/jmedgenet-2020-107404.

Savige, J. *et al.* (2021) 'Consensus statement on standards and guidelines for the molecular diagnostics of Alport syndrome: refining the ACMG criteria.', *European journal of human genetics : EJHG*, 29(8), pp. 1186–1197. Available at: https://doi.org/10.1038/s41431-021-00858-1.

Song, X. *et al.* (2021) 'Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis.', *International journal of medical informatics*, 151, p. 104484. Available at: https://doi.org/10.1016/j.ijmedinf.2021.104484.

Stewart, D.R. *et al.* (2018) 'Care of adults with neurofibromatosis type 1: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG)', *Genetics in Medicine*, 20(7), pp. 671–682. Available at: https://doi.org/10.1038/gim.2018.28.

Tian, Y. *et al.* (2019) 'REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification', *Scientific Reports*, 9(1), p. 12752. Available at: https://doi.org/10.1038/s41598-019-49224-8.

Vabalas, A. *et al.* (2019) 'Machine learning algorithm validation with a limited sample size.', *PloS one*, 14(11), p. e0224365. Available at: https://doi.org/10.1371/journal.pone.0224365.

Vu, T.T. and Braga-Neto, U.M. (2009) 'Is bagging effective in the classification of small-sample genomic and proteomic data?', *EURASIP journal on bioinformatics & systems biology*, 2009(1), p. 158368. Available at: https://doi.org/10.1155/2009/158368.

Waring, A. *et al.* (2021) 'Data-driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomyopathy.', *Journal of medical genetics*, 58(8), pp. 556–564. Available at: https://doi.org/10.1136/jmedgenet-2020-106922.

Wilcox, E.H. *et al.* (2022) 'Evaluating the impact of in silico predictors on clinical variant classification.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 24(4), pp. 924–930. Available at: https://doi.org/10.1016/j.gim.2021.11.018.

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 8(1) (2025) 47-62
DOI: 10.38016/jista.1543187

Restoran Müşteri Yorumlarının Duygu Analizi: Sıfır-Atış Metin Sınıflandırma Yaklaşımı

Kutan Koruyan[1*]

[1] Yönetim Bilişim Sistemleri, İktisadi ve İdari Bilimler Fakültesi, Dokuz Eylül Üniversitesi, İzmir, Türkiye

kutan.koruyan@deu.edu.tr

Öz

Bu makale, restoranlara yapılan çevrimiçi müşteri yorumlarından yararlanarak müşteri memnuniyetini değerlendirmek ve artırmak amacıyla makine öğrenmesi ve doğal dil işleme temelli bir yöntem önermektedir. Araştırma, çoğunluğu İzmir Körfezi çevresinde yer alan ilçelerdeki 89 balık restoranına odaklanmakta olup, veri seti 2013-2023 yılları arasında yapılan, 43 farklı dili içeren yaklaşık 15.000 müşteri yorumundan oluşmaktadır. Bu kapsamda, çalışmada hedef tabanlı duygu analizi kullanılarak, yemek kalitesi, servis kalitesi, fiziksel çevre ve adil fiyat restoran kalite boyutları temel alınarak sıfır-atış metin sınıflandırma yöntemiyle müşteri yorumlarının analiz edilmesi amaçlanmaktadır. Model değerlendirme metrikleri ümit verici sonuçlar vermekte olup, her sınıf için %75-%88 arası doğruluk ve %72-%88 arası F1 puanı elde edilmiştir. Önerilen yöntem, restoran yöneticilerinin müşteri yorumlarını otomatik olarak farklı kalite boyutlarında değerlendirmesine, restoranın güçlü ve zayıf yönlerini belirlemesine, zaman içinde müşteri memnuniyetindeki değişimleri izlemesine, rakip restoranlarla performans karşılaştırması yapmasına ve Türkçe ile yabancı dildeki müşteri yorumlarını birlikte veya ayrı ayrı analiz etmesine olanak tanımaktadır. Çalışmada önerilen bu yaklaşım, restoran yöneticilerine müşteri beklentilerini daha derinlemesine anlama ve restoran kalitesini iyileştirme konusunda veri analizi odaklı bir yol haritası sunmaktadır.

Anahtar kelimeler: Sıfır-Atış Metin Sınıflandırma, Hedef Tabanlı Duygu Analizi, Müşteri Yorumları, Müşteri Memnuniyeti, Çok Dilli Veri Seti

Sentiment Analysis of Restaurant Customer Reviews: A Zero-Shot Text Classification Approach

Abstract

This paper proposes a machine learning and natural language processing-based method to evaluate and increase customer satisfaction by using online customer reviews of restaurants. The research focuses on 89 fish restaurants, mostly located in the districts around Gulf of Izmir, and the dataset consists of approximately 15,000 customer reviews written between 2013 and 2023, covering 43 different languages. In this context, the study aims to analyse customer reviews using target-based sentiment analysis using zero-shot text classification method based on restaurant quality dimensions of food quality, service quality, physical environment, and fair price. Model evaluation metrics give promising results, with accuracy between 75% and 88% and F1 score between 72% and 88% for each class. The proposed method allows restaurant managers to automatically evaluate customer reviews on different quality dimensions, identify restaurant strengths and weaknesses, monitor changes in customer satisfaction over time, compare performance with competitor restaurants, and analyse Turkish and foreign language customer reviews together or separately. This approach proposed in the study provides restaurant managers with a data analysis-focused roadmap to understand customer expectations more deeply and improve restaurant quality.

Keywords: Zero-Shot Text Classification, Target-Based Sentiment Analysis, Customer Reviews, Customer Satisfaction, Multilingual Data Set

1. Giriş (Introduction)

Küreselleşme ve şehirleşmenin etkisiyle bireylerin yaşam tarzları, gelir düzeyleri, tüketim ve yeme-içme alışkanlıkları hızla değişmektedir. Yeme-içme alışkanlıklarındaki değişimin bir sonucu olarak ise restoran işletmeciliği giderek daha fazla önem

* Sorumlu yazar.
E-posta adresi: kutan.koruyan@deu.edu.tr

Alındı : 3 Eylül 2024
Revizyon : 27 Ekim 2024
Kabul : 4 Şubat 2025

kazanmakta; bireylerin evleri dışında yemek yeme eğilimi, etkin hizmet sunan restoranların sayısının artmasına neden olmaktadır (Yüksekbilgili, 2014). Günümüzde, restoranlar insanların sadece beslenme ihtiyacını karşılamakla kalmayıp, aynı zamanda sosyal etkileşim merkezleri haline dönüşmüştür. Bunun sonucu olarak da dışarıda yemek yeme alışkanlığı günlük bir aktivite haline gelmiş, bu da birçok yeni restoranın açılması sonucunda rekabetin artmasına sebep olmuştur (Bengül ve Dinç, 2023). Ayrıca, yerel halkın dışında, turistlerin yeni yiyecek ve içecek deneyimleri kazanmasına imkân tanıyan gastronomi turizmi, geçmişten günümüze yaygınlaşarak, önemli bir pazar haline gelmiştir (Akkurt Kurnaz, 2024; Küçükkömürler vd., 2018). Gastronomi turizmi bölge turizminin tanıtımına yardımcı olarak, bölgenin ekonomik gelişimine katkı sağlamaktadır (Hall vd., 2003).

Rekabetin yüksek olduğu restoran sektöründe, müşterilerin beklentilerini karşılamak ve hizmet sonrası memnun kalmalarını sağlamak tüm işletmelerin kritik hedefidir (Kukanja vd., 2017). Başarılı bir restoran yönetimi, hedef pazarlarının ana özelliklerini, yeni müşterileri nasıl çekeceğini, mevcut müşterileri elde tutmak için hangi rekabet faktörlerinin önemli olduğunu bilmek zorundadır (Yi vd., 2018).

Son zamanlarda, sosyal ağlar gibi birçok çevrimiçi kanal, her sektörde olduğu gibi restoranlar için de daha fazla performans izleme fırsatı sunmaktadır (Lepkowska-White ve Parsons, 2019). Çünkü, restoran yöneticileri neyi doğru veya neyi yanlış yaptıklarını, başka bir deyişle, müşteri memnuniyetini (veya memnuniyetsizliğini) müşteri yorumları vasıtasıyla çevrimiçi mecralardan takip edebilmektedir. Bunun yanı sıra işletmeler, bu yorumları zamanında değerlendirip analiz ettiğinde iş fırsatlarını geliştirmek ve iyileştirmek için kullanabilmektedirler (Revathi vd., 2023). Tüketiciler ise bir restorana gitmeden önce çevrimiçi kanallardan restoranı daha önce ziyaret edenlerin tavsiye ve yorumlarını dikkate almakta; böylece, kendi algıları şekillenmektedir (Jurafsky vd., 2014; Kumar vd., 2020).

Tüm bunlar göz önünde bulundurulduğunda, çevrimiçi müşteri yorumları restoran yöneticileri için önemli bir veri kaynağıdır. Diğer taraftan, bazen müşteri yorumları teker teker okunup analiz edilemeyecek kadar büyük boyutta veya devamlı akan yapıda olabilmektedir. Dolayısıyla, müşterilerin işletme hakkında yazdıkları yorumlardaki memnuniyeti belirlemek, başka bir deyişle, metinlerdeki pozitif veya negatif duygu durumunu otomatik olarak nitelendirmek için literatürde duygu analizi olarak adlandırılan yöntem kullanılmaktadır. Duygu analizi, bir metin parçasındaki duyguyu belirlemek amacıyla, belirli kelimelerin varlığı veya belirli konuların ele alınma düzeyi gibi özelliklere dayalı olarak, doğal dil işleme ve makine öğrenmesi teknikleri kullanılarak belgelerin sınıflandırılması olarak tanımlanmaktadır (Henrickson vd., 2019).

Geleneksel duygu analizi metindeki genel duyguyu ortaya çıkarmaktadır. Fakat, bazen müşteri yorumları birçok boyuta odaklanabilmektedir. Örneğin müşteri, tek bir yorumda bir şeyi çok sevebildiği gibi aynı anda başka bir şeyden memnun kalmadığını belirtebilir. Böyle durumlarda, geleneksel duygu analizinin kullanımı restoranın hangi özelliğinin iyi veya kötü olduğunu ortaya çıkarmayıp, sadece genel bir sonuç verecektir. Literatürde hedef tabanlı duygu analizi olarak adlandırılan yaklaşımla ise metnin belirli parçaları veya özellikleri üzerine odaklanılmakta ve bu farklı özellikler üzerine duygu analizi gerçekleştirilmektedir. Böylece, bu yöntem sayesinde işletmelerde iyileştirilmesi gereken spesifik boyutlar belirlenebilip, bu boyutlara yönelik gerekli düzenlemeler yapılabilecektir.

Bu çalışmada, Google Maps haritalama servisi bünyesindeki Google Yerel Rehberler'de yer alan çoğunluğu İzmir Körfezi çevresindeki ilçelerde konumlanmış 89 adet balık restoranına yapılan müşteri yorumları veri seti olarak kullanılarak, sıfır-atış metin sınıflandırma yöntemi ile hedef tabanlı duygu analizi gerçekleştirilmiştir. Çalışmada metinler, Gagić vd.'nin (2013) önerdiği restoran müşterilerinin memnuniyetini etkileyen en önemli boyutlar olan yemek kalitesi, servis kalitesi, fiziksel çevre ve adil fiyat, olumlu ve olumsuz olmak üzere sekiz kategoride sınıflandırılmıştır.

Bu çalışmanın amacı; restoranların müşteri memnuniyetini etkileyen kalite boyutlarının analiz edilerek, restoranların güçlü ve zayıf yönlerinin belirlenmesi ve bu doğrultuda iyileştirilmesi gereken alanların tespit edilmesine yönelik bir yöntem ortaya koymaktır. Buna yönelik olarak da müşteri yorumları kullanılarak restoranların zaman içindeki kalite boyutlarındaki değişimler incelenmiştir. Ayrıca, çok dilli yapıya sahip olan müşteri yorumları üzerinden Türkçe ve yabancı dilde yazan kullanıcıların görüşleri de karşılaştırılmıştır. Çalışmada önerilen yöntem sayesinde, restoran yöneticilerinin stratejik karar alma süreçlerinde veri odaklı bir yaklaşımla zaman içindeki müşteri memnuniyetindeki değişimleri izleyebilmeleri, müşteri beklentilerini daha iyi anlayabilmeleri ve hizmet kalitesinin iyileştirilmesine yönelik somut adımlar atmalarının desteklenmesi hedeflenmiştir. Bu çerçevede, işletmelerin restoran sektöründe artan rekabet koşullarında sürdürülebilir müşteri memnuniyetini sağlama kapasitelerinin arttırılması amaçlamaktadır.

Restoran işletmelerine yönelik olarak literatürde çoğunlukla metnin genel duygusunu belirleyen çalışmalar bulunurken, bu çalışmada farklı boyutların aynı anda değerlendirmesini sağlayan bir yöntem önerilmektedir. Bunun yanında, genellikle metnin sınıflandırılması ve duygu analizi şeklinde iki aşamada uygulanan geleneksel hedef tabanlı duygu analizi bu çalışmada tek adımda gerçekleştirilebilmektedir. Ayrıca, önerilen yöntem ile farklı dillerde yazılan yorumlardaki konular ve bunlara ait duygular dilden bağımsız olarak analiz edilmektedir.

2. Literatür Taraması (Literature Review)

Literatürde çeşitli sektörlere yönelik makine öğrenmesi ve doğal dil işleme yöntemleri kullanılarak müşteri yorumlarının duygu analizi ile değerlendirildiği birçok çalışma mevcuttur. Bu bölümde, duygu analizi, sıfır-atış öğrenme ve sıfır-atış metin sınıflandırma tanıtılacak ve literatürde müşteri yorumları kullanılarak restoran işletmelerine yönelik gerçekleştirilen duygu analizi çalışmaları incelenecektir.

2.1. Duygu Analizi (Sentiment Analysis)

Duygu analizi çeşitli makine öğrenmesi ve doğal dil işleme teknikleri kullanılarak metinsel verideki duygu durumunun ortaya çıkarılmasıdır. Duygu durumu, metinlerin pozitif ve negatif (ve bazen nötr) olarak sınıflandırılmasıyla belirlenmektedir.

Duygu analizinde veri kaynağı olarak kelime veya kelime öbekleri (cümleler); yani, blog yazıları, SMS ve sohbet mesajları gibi birçok metin kullanılmaktadır (Mohammad, 2017). Son yıllarda, insanların çeşitli konularda fikirlerini belirttiği, ürün ve hizmetler hakkında yorumlar yaptıkları sosyal ağlar (X, Facebook, vb.), çevrimiçi alışveriş siteleri (Amazon, Hepsiburada, vb.), çevrimiçi seyahat bilgi ve rezervasyon platformları (TripAdvisor, Etstur) ve yerel işletme rehberleri (Yelp, Google Maps: Google Yerel İşletmeler) gibi platformlar giderek daha popüler hale gelmiştir. Bunun sonucu olarak da bu platformlarda yapılan yorumlar, duygu analizi için önemli bir veri kaynağı olarak öne çıkmaktadır (Ahmed vd., 2020; Soleymani vd., 2017). Duygu analizi; ürün veya hizmetler için müşteri eğilimlerini incelemek, devletlerin düşmanca veya olumsuz tavırlarını belirlemek ve siyasi parti seçmenlerinin düşüncelerini anlamak gibi farklı alanlarda kullanılmaktadır (Pang ve Lee, 2008).

Duygu analizi çok kullanılan bir yöntem olsa da zorlukları da bulunmaktadır. Chifu ve Fournier (2023) eş sesli kelimeler, dildeki muğlaklık, alaycılık ve kültürel farklılıklar, veri kalitesi ve miktarı, eğitim verisinin yanlılığı gibi faktörlerin duygu analizi başarısını etkilediğini belirtmiştir. Bunun yanında, dillerin kendine özgü farklı gramer yapıları ve alfabeleri de buna eklenebilir. Nankani vd. (2020) metinlerdeki yazım yanlışları, kelimelerin farklı alanlarda farklı duygular içerebilmesi veya sahte içerikler gibi olgulara ek olarak, çok dilli (multilingual) metinlerde dillerin gramer ve morfolojik yönden farklılıklarını doğal dil işleme uygulamalarında karşılaşılan zorluklar olarak sıralamıştır.

Geleneksel duygu analizi, duygunun ne hakkında olduğunu nitelendirmeden genel duyguyu sınıflandırmaya odaklanır. Fakat, metin aynı anda farklı konular veya varlıklarla ilgiliyse ve muhtemelen farklı konulara yönelik farklı duygular ifade ediyorsa, burada geleneksel duygu analizi yeterli olmayacaktır (Hoang vd., 2019). Bu tür problemlerde metin içindeki birden çok konu hakkındaki duyguları belirlemek için hedef tabanlı duygu analizi kullanılmaktadır. Hedef tabanlı duygu analizi, bir cümledeki belirli bir özelliğe yönelik duygu kutupluluğunu belirlemeyi amaçlamakta ve hedef özellik, bir varlığın bir özelliğini tanımlayan kelime veya ifadeye atıfta bulunmaktadır (Jiang vd., 2019).

Örneğin, "*Hizmet çok iyiydi. Mehmet Bey'e teşekkür ederim.*" cümlesi pozitif bir anlamdadır ve cümlede tek bir konuya (yani hizmete) ait duygu durumu vardır. Başka bir örnekte ise "*Yemekler harika! Özellikle kalamar. Fakat restoran çok gürültülüydü ve ışıklandırma rahatsız ediciydi*." cümlesinde müşteri restoranda yediği yemeği çok beğenmesine rağmen, ambiyanstan rahatsız olmuştur. Bu tür cümlelerde aynı anda farklı konular hakkında farklı duygular yer almaktadır. Geleneksel duygu analizi yerine, böyle durumlarda hedef tabanlı duygu analizi kullanılarak her bir konu için duygu durumu belirlenebilmektedir.

2.2. Restoran Yorumlarının Duygu Analizi (Sentiment Analysis of Restaurant Reviews)

Restoranları ziyaret eden kişilerin çevrimiçi yaptıkları müşteri yorumlarını duygu analizi ile inceleyen bazı çalışmalarda, geleneksel denetimli makine öğrenmesi algoritmaları ve kelime gömme (Word Embedding) yaklaşımları kullanılarak müşterilerin restoranlar hakkındaki yorumları hiçbir kategoriye ayrılmadan analiz edilmiştir. Krishna vd. (2019) İngilizce restoran yorumlarının duygu analizini gerçekleştirmek için Naïve Bayes, Destek Vektör Makinesi, Karar Ağacı ve Rastgele Orman sınıflandırma algoritmalarını kullanmışlardır. Çalışmanın sonuçlarına göre, Destek Vektör Makinesi diğer algoritmalara kıyasla %94,56 gibi bir doğruluk puanına ulaşmıştır. Hossain vd. (2020) Bangladeş'teki restoranlara ait müşteri yorumlarını Evrişimli Sinir Ağı ve Uzun-Kısa Vadeli Bellek algoritmalarını birlikte kullanarak olumlu ve olumsuz olarak sınıflandırmışlardır. Hossain vd. (2021) farklı kaynaklardan edindikleri Bengalce restoran yorumlarını Çift Yönlü Uzun-Kısa Vadeli Bellek, Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Naïve Bayes, Destek Vektör Makinesi modellerini kullanarak duygu analizini gerçekleştirmişlerdir. Çift Yönlü Uzun-Kısa Vadeli Bellek modeli %95,35 ile en yüksek doğruluk puanına ulaşmıştır. Patil vd. (2022) olumlu ve olumsuz restoran yorumlarını sınıflandırmak için K-En Yakın Komşu Sınıflandırıcı, Lojistik Regresyon, Destek Vektör Sınıflandırıcı ve Naïve Bayes algoritmaları kullanmış ve en iyi doğruluk puanını %78 ile Destek Vektör Sınıflandırıcısı ile elde etmişlerdir. Abdullah vd. (2023) Arapça çevrimiçi restoran yorumlarının duygu analizini gerçekleştirmek için Naïve Bayes, Karar Ağacı, Destek Vektör Sınıflandırıcısı, K-En Yakın Komşu, Rastgele Orman ve Lojistik Regresyon algoritmalarını kullanmışlardır. Destek Vektör Sınıflandırıcısı %97,6 ile en yüksek doğruluk puanına ulaşarak yorumları pozitif, negatif ve nötr olarak sınıflandırmada en iyi performansı göstermiştir. Gedif vd. (2023) Etiyopya'nın bir dili olan Amharca yazılan restoran yorumlarını

Destek Vektör Makinesi, K-En Yakın Komşu ve Naïve Bayes sınıflandırıcı kullanarak duygu analizini gerçekleştirmiştir. Değerlendirme sonucunda Destek Vektör Makinesi modelinin en yüksek doğruluğu verdiği görülmüştür. Lavanya vd. (2023) çalışmalarında restoran incelemeleri için Yelp veri setini kullanarak kelime çantası (Bag of Words, BoW), Terim Frekansı-Ters Doküman Frekansı (Term Frequency-Inverse Document Frequency, TF-IDF), GloVe, Word2Vec ve Doc2Vec gibi farklı kelime gömme yaklaşımlarını test etmişler, Lojistik Regresyon ve Destek Vektör Makinesi gibi denetimli makine öğrenmesi algoritmalarını kullanarak, sonuçları doğruluk, kesinlik, duyarlılık ve F1 puanı gibi performans metriklerine göre değerlendirmişlerdir. Karşılaştırmalı bulgular, Destek Vektör Makinesi ve TF-IDF'nin birlikte kullanımının %98 doğruluk puanı ile sonuçlar ürettiğini göstermiştir. Bozkurt ve Yalçın (2024) Amazon yemek yorumları üzerinde topluluk öğrenmesi algoritmalarını kullanarak duygu analizi yapmayı amaçlamıştır. Çalışmada, Rastgele Orman, CatBoost ve XGBoost algoritmaları kullanılarak olumlu, olumsuz ve nötr duygu sınıflandırması yapılmışlardır. Çalışmada, farklı vektörleştirme tekniklerinin başarısı da karşılaştırılmış, çeşitli değerlendirme metrikleri kullanılarak en yüksek %90,22 test doğruluk değeri, Rastgele Orman ve CountVectorizer tekniği ile elde edilmiştir. Ayrıca, web kazıma ile yeni bir veri seti oluşturulmuş ve modeller bu veri seti üzerinde de test edilmiştir.

Bazı çalışmalarda ise duygu analizi yanında bir de lezzet, hizmet, ambiyans ve fiyatlandırma gibi kategoriler ele alınıp, müşterileri yorumları her kategoride hedef tabanlı duygu analizi yapılarak incelenmiştir. Suciati ve Budi (2019) çalışmalarında Endonezya'daki restoranların Endonezyaca ve İngilizce müşteri yorumlarını yemek, fiyat, hizmet ve ambiyans boyutlarını kullanarak olumlu, olumsuz ve nötr olarak sınıflandırmışlardır. Rastgele Orman, Multinomial Naïve Bayes, Lojistik Regresyon, Karar Ağacı ve Ekstra Ağaç sınıflandırıcısı algoritmaları kullanılan çalışmada Lojistik Regresyon en yüksek skoru yemek (%81,76) ve ambiyans (%77,29) boyutlarıyla, en yüksek skor fiyat (%78,71) ve hizmet (%85,07) boyutları için Karar Ağacı algoritması ile elde edilmiştir. Zahoor vd. (2020) Karachi Pakistan'daki restoran yorumlarını Facebook'dan temin ederek, Naïve Bayes sınıflandırıcı, Lojistik Regresyon, Destek Vektör Makinesi ve Rastgele Orman algoritmalarından yararlanarak, yorumları pozitif ve negatif olarak sınıflandırmışlardır. En yüksek doğruluk oranı %95 Rastgele Orman modeli ile elde edilmiştir. Daha sonra, aynı veriyi lezzet, ambiyans, servis ve fiyatlandırma olacak şekilde yine aynı dört algoritmayı kullanarak kategorilere ayırmışlardır. Bunun sonucunda, lezzet kategorisi %97 doğruluk oranı ile en yüksek performansı göstermiştir. Diğer kategoriler için ise doğruluk, fiyat %84, ambiyans %86,49 ve hizmet %89 oranlarında bulunmuştur. Ara vd. (2020) bir restoranın web portalı üzerinden edinmiş oldukları yemek kalitesi, hizmet, ortam, fiyat, online yemek siparişi gibi konular hakkındaki müşteri yorumları vasıtasıyla, SentiStrength sınıflandırıcısı kullanılarak görüşlerde ifade edilen kelimelerin duygu gücünü bulmuşlardır. Daha sonra, yorumlar standart sapma tekniği kullanılarak pozitif, negatif ve nötr olarak sınıflandırılmış ve %85,71'lik bir doğruluk elde etmişlerdir. Zhang vd. (2022) Yelp'deki restoran incelemelerini veri kaynağı olarak kullanarak ve fiyat, zaman, yemek, hizmet ve konum konularını temel alarak Gizli Dirichlet Ayrımı yöntemi ile konu modelleme gerçekleştirmişlerdir. Daha sonra, TextBlob ile her konu için duygu analizi yapmışlardır.

Bazı çalışmalarda ise dönüştürücü (transformer) mimarisini kullanan BERT'den (Bidirectional Encoder Representations from Transformers, Dönüştürücülerden Çift Yönlü Kodlayıcı Temsilleri) yararlanılmıştır. Tuna vd. (2023) yaptıkları çalışmada, Word2vec, Glove, fastText ve BERT gibi çeşitli modeller kullanmışlardır. SemEval'15 ABSA yarışmasında sunulan restoran müşterilerine ait yorumlardan oluşan veri seti kullanılarak hedef terim, hedef kategori ve duygu sınıfları belirlenmiştir. fastText yöntemi hedef terim ve kategori tespitinde en yüksek başarıyı göstermiş, BERT yöntemi ile duygu sınıflandırmasında en iyi sonuç elde edilmiştir. Bu sonuçlar, fastText ve BERT yöntemlerinin Türkçe metinlerde başarılı olduğunu göstermiştir. Branco vd. (2024) Portekizce restoran yorumlarının duygu analizi için BERT ve RoBERTa (Robustly Optimized BERT Approach, Güçlü Bir Şekilde Optimize Edilmiş BERT Yaklaşımı) modellerini kullanılarak, yorumların duygu sınıflandırmasını gerçekleştirmiş ve önceden eğitilmiş derin öğrenme modellerinin uygunluğunu değerlendirmişlerdir. Modelin performansının tespiti için doğruluk ve Alıcı Çalışma Karakteristiği (Receiver Operating Characteristic, ROC) eğrisi altındaki alan metrikleri kullanılmış ve çalışmada %80'in üzerinde bir sonuç elde edilmiştir. Bazı çalışmalarda BERT yöntemine ek olarak bir de OpenAI GPT'den (Generative Pre-trained Transformer, Üretici Önceden Eğitilmiş Dönüştürücü) yararlanılmıştır. Carrasco ve Dias (2023) Portekiz'in Algarve bölgesinde yer alan restoranlara yapılan yorumları TripAdvisor'dan edinmiş, müşteri memnuniyetinin beş temel özelliği olan gıda kalitesi, hizmet, ortam, fiyat ve restoranın konumu ile ilgili konuları BERT, USE (Universal Sentence Encoding, Evrensel Cümle Kodlama) ve OpenAI GPT ile kategorize etmişlerdir. Daha sonra, önceden belirlenen ve duygular için referans olan cümleler temel alınarak kosinüs benzerliği kullanılmış, pozitif ve negatif duygulara yönelik bir sınıflandırma gerçekleştirilmiştir. Değerlendirme için üç model karşılaştırıldığında doğruluk %93, kesinlik %86, duyarlılık %83 ve F1 puanı %85 ile OpenAI GPT en yüksek puana ulaşmıştır. Benzer şekilde, Carrasco ve Dias (2024) yine Algerve bölgesindeki restoran yorumlarını kullanarak bu sefer müşteri memnuniyetinin beş temel özelliği ayrıntılandırıp, BART (Bidirectional and Auto-Regressive

Transformers, Çift Yönlü ve Oto-Regresif Dönüştürücüler), DeBERTa (Decoding-enhanced BERT with disentangled attention, Ayrıştırılmış dikkat ile kod çözümü geliştirilmiş BERT), ChatGPT 3.5 ve ChatGPT 4.0'ı kullanmışlardır. Çalışmada ChatGPT 4.0 modelinde en yüksek F1 skoru elde edilmiştir.

2.3. Sıfır-Atış Öğrenme (Zero-Shot Learning)

Bilindiği üzere, veri sınıflandırma görevlerinde kullanılan denetimli makine öğrenmesinde, veri içindeki sınıflar önceden belirlenmekte (veya bilindiği varsayılmakta) ve her bir örnek için eğitim verisinin hangi sınıfa ait olduğuna dair etiketleme yapılarak, sonrasında eğitim ve test işlemi gerçekleştirilmektedir. Bunun sonucu olarak da bazen bilinmeyen sınıfların gözden kaçma olasılığı bulunmaktadır. Yani, sınıflandırıcı -eğer iyi bir eğitim için her sınıfta yeterli sayıda etiketli veri bulunursa- sadece eğitim verilerinin kapsadığı sınıflara ait örnekleri sınıflandırabilmektedir (Wang vd., 2019). Bununla birlikte, projelerde öğrenme aşamasından sonra yeni sınıflar da ortaya çıkabilmektedir (Romera-Paredes ve Torr, 2015). Dahası, etiketlemenin zaman alıcı ve maliyetli bir süreç olduğu da göz ardı edilmemelidir (Liu vd., 2004). Bu noktada, Chang vd. (2008) geleneksel makine öğrenmesi çalışmalarının aksine, etiketli örneklere ihtiyaç duyulmadan verinin sınıflandırma işlemini gerçekleştiren ve başta "Verisiz Sınıflandırma" (Dataless Classification) olarak adlandırılan bir model önermişlerdir. Daha sonra sıfır-atış öğrenme olarak adlandırılan bu model, geleneksel denetimli makine öğrenmesi modelleri ile karşılaştırıldığında, görülen ve görülmeyen sınıflar arasındaki boşluğu doldurmaktadır. Bir başka deyişle sıfır-atış öğrenme, anlamsal bilgi yardımıyla görülen diğer sınıflardan elde edilen bilgiyi aktararak, görülmeyen sınıflardaki nesneleri sınıflandırabilen bir model eğitmeyi amaçlamaktadır (Pourpanah vd., 2022). Bu yüzden, eğitim ve çıkarım olmak üzere iki aşamalı bir sürece sahip olan sıfır-atış öğrenmede, niteliklerle ilgili bilgiler eğitim aşamasında yakalanırken, bu bilgiler çıkarım aşamasında örnekleri yeni bir sınıf kümesi arasında kategorize etmek için kullanılmaktadır (Çelik ve Dalyan, 2023).

Sıfır-atış öğrenme geçmişten günümüze daha çok bilgisayarlı görü ve doğal dil işleme projelerinde uygulama alanı bulmuştur (Rezaei ve Shahidi, 2020). Sıfır-atış öğrenme, nesne veya hayvanların algılanması (Bansal vd., 2018; Lampert vd. 2014), tıbbi görüntüleme ve görüntü sınıflandırma (Mahapatra vd., 2021; Vétil vd., 2022), arazi örtüsü haritalanması (Li vd., 2021; Pradhan vd., 2020), video sınıflandırma ve eylem tanımlama (Brattoli vd., 2020; Estevam vd., 2021), fotoğraf iyileştirme (Kar vd., 2021; Zheng ve Gupta, 2022), araç rotası belirleme (Yu vd., 2020), metinden görüntü oluşturma (Sanghi vd., 2022) ve insan duygularının tanımlanması (Zhan vd., 2019) gibi bilgisayarlı görü çalışmalarında başarı ile kullanılmıştır.

Sıfır-atış öğrenme doğal dil işleme görevlerinde birçok araştırmada kullanılmıştır. Örneğin, bir sohbet robotunun kullanıcının ne istediğini anlayabilmesini sağlayan niyet sınıflandırması (intent classification) (Liu vd., 2019), metnin anlamının daha iyi anlaşılması için metinden nesneleri ve kavramları çıkarma işlemi olan öznitelik çıkarımı (feature extraction) (McInerney vd., 2023), metin içindeki varlıkları (kişiler, yerler, kuruluşlar vb.) tanımlama ve onları bir bilgi kaynağındaki ilgili varlıklara bağlama işlemi olan varlık bağlama (entity linking) (Logeswaran vd., 2019), bir dilden başka bir dile çeviri için kullanılan makine çevirisi (Johnson vdl., 2017; Thompson ve Post, 2020; Zhang vd., 2020), bir metnin yazarının verilen hedeften yana mı yoksa ona karşı mı olduğunun algılanması sağlayan duruş tespiti (stance detection) (Choi ve Ko, 2023; Jiang vd., 2023), otomatik özetleme (Goodwin vd., 2020) ve devam eden bölümde anlatılacak metin sınıflandırma ve duygu analizi olmak üzere birçok alanda kullanılmaktadır.

2.4. Sıfır-Atış Metin Sınıflandırma Kullanılarak Duygu Analizi (Sentiment Analysis Using Zero-Shot Text Classification)

Doğal dil işlemenin bir alt dalı olan metin sınıflandırma, metin belgelerini içeriklerine göre otomatik olarak kategorilere ayırma işlemini ifade etmektedir. (Quazi ve Musa, 2022). Uzun yıllardır doğal dil işleme alanında bir araştırma konusu olan metin sınıflandırma, duygu analizi, konu modelleme ve bilgi çıkarımı gibi çeşitli görevlerde kullanılmaktadır (Jiao, 2023).

Sıfır-atış metin sınıflandırmada, geleneksel olarak belgelerin belirli bir sınıftaki diğer belgelerle karşılaştırılarak sınıflandırılması yerine, sınıflar arasındaki bilinen ilişkilerden yararlanılmaktadır. Böylece, bir taksonominin tüm sınıfları için eğitim verisi gerekmemekte ve eğitim sırasında görülmeyen sınıfların tahmin edilmesi mümkün kılınmaktadır (Hoppe vd., 2021). Vaswani vd.'nin (2017) dönüştürücü mimarisini tanıtması ile birçok farklı doğal dil anlama görevinde kullanım alanı bulan sıfır-atış metin sınıflandırma, önceden eğitilmiş dil modellerinden (Pre-trained language models) yararlanmaktadır. Önceden eğitilmiş dil modellerinde transfer öğrenme kullanılmakta, önceden eğitilmiş bir model alınıp, modelin ilgili ancak farklı bir görev üzerinde çalışması için ince ayar yapılmaktadır (Azunre, 2021). Bu modeller genel olarak Wikipedia, X, haber siteleri gibi büyük ölçekli metinlerden eğitilmektedir. Ayrıca, kelime dağarcığı daha spesifik olan bilim, tıp veya hukuk gibi alanlara özgü modeller de geliştirilmektedir (Pérez vd., 2021). Önceden eğitilmiş dil modellerinin bir özelliği de çok dilli bir yapıya sahip olmasıdır. Diller arası dil anlama (Cross-lingual language understanding) olarak adlandırılan bu özellik, bir makine öğrenimi modelinin her dil için ayrı bir eğitim gerektirmeden, birden fazla dilde doğal dil metnini anlama ve işleme

yeteneğini ifade etmektedir (Conneau vd., 2019). Böylece bu vasıf, özellikle uluslararası düzeyde, birçok farklı dildeki sosyal ağ kullanıcı gönderileri veya müşteri yorumları gibi metne dayalı verilerin sınıflandırılmasında büyük kolaylık sağlamaktadır.

Sıfır-atış metin sınıflandırmaya yönelik olarak yapılan çalışmalarda, örneğin, Lin ve Wen (2022) semantik benzerlik (Sentence-BERT) ve sıfır-atış sınıflandırma olmak üzere iki modeli kullanarak müşteri geri bildirimlerini sınıflandırmak için bir prototip tanıtmıştır. Çalışma sonucunda sıfır-atış metin sınıflandırmada en yüksek doğruluk elde edilmiştir. Cherapanukorn ve Sugunnasil (2022) görüş madenciliği tekniği kullanarak ve sıfır-atış metin sınıflandırma yöntemi ile turistik cazibe merkezleri için turist memnuniyeti bileşenlerini belirlemişlerdir. Buna yönelik olarak da 2010 ile 2021 yılları arasında TripAdvisor'da yayınlanan Tayland'daki 40 turistik merkeze ait toplam 40.000 çevrimiçi turist yorumu analiz edilmiştir. Rey-Moreno vd. (2023) 2018 ile 2021 yılları arasında Airbnb ve otel konaklamalarına ait toplamda 24.436 cümleyi kullanarak, misafir memnuniyeti ve güveniyle ilgili hizmet özelliklerini tanımlamak için BERTopic, sıfır-atış sınıflandırma ve temel bileşen analizi yöntemlerinin kombinasyonunu kullanmışlardır. Das vd. (2023) on altı popüler Google Play Store uygulamasına ait yorumları veri kaynağı olarak kullanarak uygulama incelemelerinin otomatik olarak analiz edilmesi ve uygulama geliştiricilerinin kullanıcı geri bildirimlerini daha verimli şekilde yönetmelerine yardımcı olmak için bir çerçeve sunmuştur. Çalışmada; duygu analizi için BERT, sınıflandırılan metnin temalarını belirlemek için sıfır-atış sınıflandırma, metin özetlerinin oluşturulması için ise GPT-3 kullanılmıştır.

Sıfır-atış metin sınıflandırma kullanılarak yapılan duygu analizi çalışmalarında ise örneğin Masarifoglu vd. (2021) bankacılık alanındaki Türkçe müşteri yorumlarının duygu analizini gerçekleştirmiş, bunun için önceden eğitilmiş çok dilli BERT, BERTurk, sıfır-atış metin sınıflandırmada kullanılan XLM-Roberta-Large-XNLI modeli ve geleneksel makine öğrenmesi yöntemlerini (Destek Vektör Makineleri, Naïve Bayes ve Lojistik Regresyon) karşılaştırmışlardır. Çalışmada farklı senaryolar test edilmiş ve etiketli eğitim verisinin olmadığı durumda XLM-Roberta-Large-XNLI modeli %83 ağırlıklı F1 skoru ile en yüksek başarıyı elde etmiştir. Kumar ve Albuquerque (2021) Hintçe üzerine duygu analizi yapmak için XLM-RoBERTa çapraz dilli modeli ve sıfır-atış çapraz-dilli transfer öğrenme (Zero-shot Cross-lingual Transfer Learning) yöntemini kullanmışlardır. İngilizce veri setiyle eğitilen model, Hintçe cümle seviyesinde duygu analizinde test edilmiş ve çalışmada hibrit derin öğrenme yönteminden (Evrişimli Sinir Ağır, CNN ve Destek Vektör Makinesi) daha yüksek bir doğruluk oranı elde edilmiştir. Modelin performansını değerlendirmek için ortalama duyarlılık, F1 puanı ve makro F-puanı doğrulama metrikleri kullanılmıştır. XLM-R modeliyle ortalama %60,93 doğruluk elde edilmiştir. Sahar vd. (2022) BERT ve sıfır-atış öğrenme yardımıyla Amazon'dan edindikleri kozmetik ürün incelemelerini kullanarak duygu analizini gerçekleştirmişlerdir. Manias vd'nin. (2023) çalışmasında ise Twitter verileri ile dört farklı BERT tabanlı çok dilli sınıflandırıcı (mBERT, XLM-R, DistilBERT, ve BERT-m) ve sıfır-atış metin sınıflandırma yöntemini doğruluk ve uygulanabilirlik açısından karşılaştırmışlardır. Sonuçlar, çok dilli BERT tabanlı modellerin yüksek performans sağladığını ve sıfır-atış öğrenme yaklaşımının daha hızlı, verimli ve ölçeklenebilir çözümler sunduğunu göstermiştir.

2.5. Literatürden Elde Edilen Temel Bulgular (Key Findings from Literature)

Literatürde restoran müşterilerinin yorumlarının duygu analizi ile değerlendirilmesi çalışmalarında çoğunlukla metinlerdeki genel duygu ortaya çıkarılmış ve geleneksel makine öğrenmesi ve derin öğrenme algoritmaları kullanılmıştır. Fakat, metinde farklı konulara ait farklı duygular araştırılmak istenirse bu yaklaşımlar yetersiz kalmaktadır. Buna yönelik olarak da lezzet, hizmet, ambiyans ve fiyatlandırma gibi spesifik boyutlar ele alınarak hedef tabanlı duygu analizi gerçekleştirilmiştir. Ayrıca son dönemlerde hedef tabanlı duygu analizi için BERT ve OpenAI GPT gibi dönüştürücü tabanlı modeller de kullanılmaya başlanmıştır.

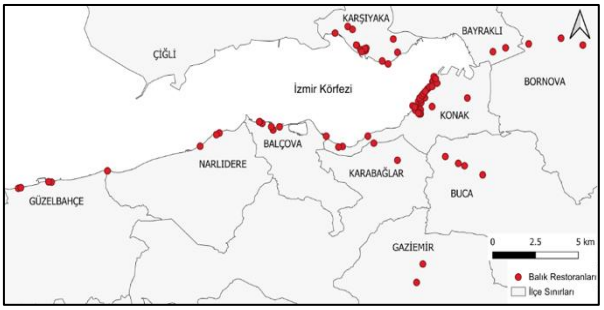
Sıfır-Atış öğrenme ile etiketli verilere ihtiyaç duyulmadan verilerin sınıflandırılması gerçekleştirilmektedir. Ayrıca, yaklaşımın özellikle metin sınıflandırma ve duygu analizi görevlerinde önceden eğitilmiş dil modelleri ile kullanılması farklı dillerden oluşan veri setlerinin sınıflandırılması zorluğunu ortadan kaldırmıştır. Önceden yapılan çalışmalar, bu yöntemin kullanımında yüksek doğruluk oranlarına ulaşıldığını ve diğer modellere kıyasla hızlı ve etkili olduğunu göstermektedir.

3. Metodoloji (Methodology)

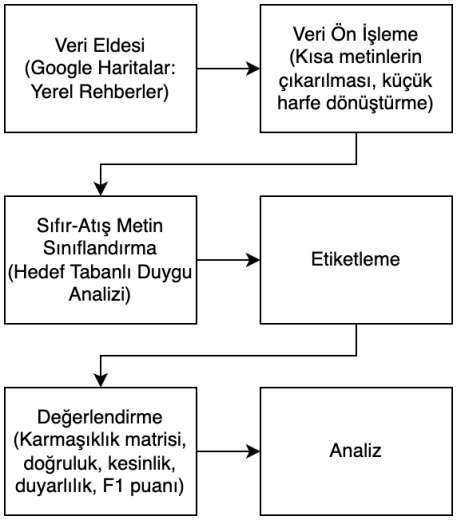
İzmir, yeme-içme sektörü bakımından zengin, aynı zamanda potansiyeli olan ve geliştirilmesi gereken bir şehirdir (Akgündüz vd., 2024; Altıntaş ve Hazarhun, 2020). İzmir yüzyıllardır farklı kültürlere ev sahipliği yapmış, bu da İzmir mutfağına yansımıştır. Özellikle İzmir'e özgü çeşitli bitkiler, otlar, sebzeler, zeytinyağlılar, balık ve deniz ürünleri İzmir'e özgü yemekler olarak sıralanmaktadır (Erdoğan ve Özdemir, 2018). Bilhassa, İzmir'in Ege Denizi'ne kıyısı ve uzun sahil şeridine sahip olması, deniz ürünleri mutfağının da gelişimine yol açmıştır. Böylelikle İzmir, kıyı şeridi balık restoranları ve balık pazarları açısından zenginlik göstermektedir (Yentür ve Demir, 2022).

Çalışmada, çoğunluğu İzmir Körfezi çevresindeki ilçelerde yer alan balık restoranlarına yönelik olarak sıfır-atış metin sınıflandırma yaklaşımı kullanılarak hedef tabanlı duygu analizi gerçekleştirilmiştir (Şekil 1). Veri kaynağı olarak Google Haritalar: Google Yerel

Rehberler'de yer alan balık restoranlarına yapılan müşteri yorumları kullanılmıştır. Veri eldesi, Selenium Python kütüphanesinden yararlanılarak web kazıma metodu ile Şubat-Mart 2023 aralığında gerçekleştirilmiştir ve 89 adet balık restoranı çalışma kapsamında yer almıştır. Çalışmanın akış şeması Şekil 2'de verilmektedir.



Şekil 1. Çalışmada incelenen balık restoranları konumları (Locations of fish restaurants analysed in the study)



Şekil 2. Akış şeması (Flow chart)

Çalışmada işlenen veri setinde tüm restoranlar için toplam yorum sayısı 15.305'tir. En eski yorum 2013, en yeni yorum ise 2023 yılına aittir. Çok dilli bir yapıya sahip olan ve 43 farklı dili barındıran veri setinde yorumların büyük bir çoğunluğu Türkçe (89%) olup, diğer diller ise İngilizce (%7), Almanca (%1), Rusça (%0,9), Fransızca (%0,5), Arapça (%0,4), Korece (%0,2) ve diğer dillerdedir (%1). Veri setinde kullanıcıların yaptıkları restoran yorumları, zaman (yıl), restoran adı, kullanıcı adı ve puan verileri yer almakta, bu çalışma kapsamında sadece yorum ve zaman verisi kullanılmıştır.

3.1. Veri Önişleme (Data Preprocessing)

Veri önişleme aşamasında, sadece puan verilip ve boş olan girdiler ve üç kelimenin altında olan "*Harika!*", "*Muhteşem restoran*" veya "*Berbat!*" gibi anlamsal olarak çok genel olan yorumlar göz ardı edilmiş ve veriden çıkarılmıştır. Ayrıca, tüm yorumlar küçük harfe dönüştürülmüştür. Buna ek olarak, Liu vd. (2021), Leburu-Dingalo vd. (2022) ve Manias vd.'nin (2023) çalışmalarında değindikleri üzere, mikro-metinler kullanılarak yapılan çok dilli duygu analizinde noktalama işaretleri, emojiler ve yüz ifadelerini temsil eden karakterler (emoticon) metindeki duyguyu ve anlamı pekiştirdiği için metinden çıkartılmamıştır.

3.2. Veri Analizi (Data Analysis)

Çalışmada, Gallego'nun (2023) huggingface.co'da sunduğu, "XLM-RoBERTa-large" modelinin birkaç doğal dil çıkarımı (Natural Language Inference, NLI) veri kümesi üzerinde ince ayarlanmış "vicgalle/xlm-roberta-large-xnli-anli" modeli kullanılmıştır. Model çok dilli olup, "XLM-RoBERTa-large" temel modeli 100 farklı dilde önceden eğitilmiştir (Conneau vd., 2020).

Kodlama ve veri işleme, Google Colaboratory'de Python programlama dili kullanılarak ve GPU (Graphics Processing Unit, Grafik İşlemci Birimi) hizmetinden yararlanılarak yapılmıştır. BERT, GPT, RoBERTa gibi birçok farklı dil modelini ve mimarisini içeren transformers kütüphanesindeki pipeline fonksiyonu yardımıyla, "vicgalle/xlm-roberta-large-xnli-anli" modeli ile sıfır-atış sınıflandırma işlemi için bir hattın (pipeline) oluşturulması sağlanmıştır. Hat, basit bir uygulama programlama arayüzü (Application Programming Interface, API) sunarak, ham metni girdi olarak almakta, onu kelimelere veya alt kelimelere ayırmakta ve ardından belirli görevler için modele beslemektir (Hugging Face, 2024).

Kategoriler, Gagić vd.'nin (2013) restoran kalitesinin ölçülmesi için önerdiği yemek kalitesi, servis kalitesi, fiziksel çevre ve adil fiyat boyutları temel alınarak oluşturulmuştur. Buna yönelik olarak da kategoriler hem olumlu hem de olumsuz yorumları kapsayacak şekilde kaliteli yemek (KY), kalitesiz yemek (KsizY), kaliteli servis (KS), kalitesiz servis (KsizS), güzel ambiyans (GA), kötü ambiyans (KA), düşük fiyat (DF) ve yüksek fiyat (YF) olmak üzere sekiz sınıf olarak belirlenmiştir. Müşteriler yorumlarda restoranın birkaç özelliğinden aynı metin içinde bahsedebildikleri için model çok-etiketli sınıflandırma (multi-label classification) şeklinde ayarlanmış, böylece her sınıf için puanlar ayrı ayrı hesaplanmıştır. Tablo 1'de çeşitli restoranlara yönelik örnek yorum ve tahminlenmiş kategorilerin puanları verilmektedir. Yüksek puanlar yorumun hangi sınıfa ait olabileceğini göstermektedir.

Tablo 1. *Örnek restoran yorum ve tahmin puanları (Examples of restaurant reviews and prediction scores)*

Müşteri Yorumları	KY	KsizY	KS	KsizS	GA	KA	YF	DF
Lezzetleri mükemmel, özellikle mezeler harika ve leziz, ayrıca işletme ve çalışanları gayet güler yüzlü ve sıcakkanlılar. Hizmet olarak kaliteli ve keyifli bir mekân.	0,9751	0,0002	0,9996	0,0002	0,5721	0,0002	0,0150	0,0002
Потрясающая кухня, обслуживание на высоте, один из самых лучших ресторанов города Рекомендую! (Harika mutfak, mükemmel hizmet, şehrin en iyi restoranlarından biri. Tavsiye ederim!)	0,9989	0,0002	0,9994	0,0002	0,0010	0,0003	0,0228	0,0008
الأكل عادي و أسعارو غالية (Yemek sıradan ve fiyatları pahalı.)	0,0003	0,0005	0,0007	0,0008	0,0002	0,0080	0,0020	0,9115
Pricy but excellent service and delicious food. (Pahalı ama mükemmel servis ve lezzetli yemekler.)	0,9992	0,0002	0,9997	0,0002	0,0014	0,0003	0,6844	0,0007
Böyle yüksek faturalı bir yer için balıklar rezaletti. Ayrıca servis de kötüydü. Hiç tavsiye etmiyorum.	0,0002	0,9993	0,0004	0,9371	0,0003	0,3136	0,0004	0,9988

Kaliteli Yemek (KY), Kalitesiz Yemek (KsizY), Kaliteli Servis (KS), Kalitesiz Servis (KsizS), Güzel Ambiyans (GA), Kötü Ambiyans (KA), Düşük Fiyat (DF), Yüksek Fiyat (YF)

3.3. Model Performans Değerlendirmesi (Model Performance Evaluation)

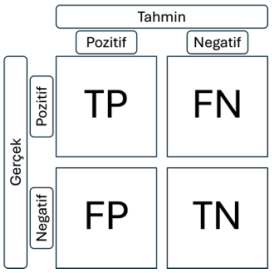
Çalışmada kullanılan modelin tahmin performansını ölçmek için tüm veriden yaklaşık %11'lik bir kısım gelişigüzel örnek olarak alınmış, 0 ve 1 ikili yapı (binary) şeklinde etiketlenmiştir. Tahminlenmiş veri setinden de aynı satırların seçilimi yapılıp, veri 0,5 eşik değerine göre yine ikili yapıya dönüştürülmüştür. Örnek alınan veri seti içindeki Türkçe haricindeki müşteri yorumlarının oranı yaklaşık %10'dur.

Veri setinin çok-etiketli yapısından dolayı her bir sınıf içerisinde yer alan pozitif ve negatif örneklerin eşit sayıda olması neredeyse mümkün değildir ve sınıflar dengesiz bir yapıya sahiptir (Tablo 2). Bu tür durumlarda doğruluk, kesinlik, duyarlılık gibi metriklerin hesaplanmasında verilerin bu haliyle doğrudan kullanılması performans değerlendirmesinde yanıltıcı olabilmektedir (Iram vd., 2016). Bu yüzden, her sınıfı dengeli hale getirmek amacıyla azınlık sınıfı sayısını çoğunluk sınıfı sayısına eşitlemek için aşırı örnekleme (over-sampling) veya çoğunluk sınıfı sayısını azınlıklık sınıfı sayısına eşitlemek için alt örnekleme (under-sampling) gibi yöntemler kullanılmaktadır. Aşırı örnekleme metodu için yeni metin üretimi veya azınlık sınıfındaki verileri çoğunluk sınıfı sayısına eşitlemek için azınlık sınıfındaki bazı satıların aynılarını ekleyerek veriyi çoğaltmak gibi teknikler bulunmaktadır. Bu yöntem, azınlık sınıfındaki verilerin çoğaltılmasıyla sınıf dengesizliklerini azaltmasına rağmen, yukarıda bahsedilen tekniklerin, sırasıyla, yeni veri üretiminin başka bir araştırma konusu olduğundan veya aynı verilerin tekrar kullanılması ile yeni bilgi sağlanamadığından tercih edilmemiştir. Bu yüzden, çalışmada alt örnekleme metodu kullanılarak pozitif ve negatiflerde en küçük azınlık sınıfı temel alınarak, tüm sınıflarda pozitif ve negatif sayıları eşitlenmiştir. Bir başka deyişle, en küçük eleman sayısına sahip KA sınıfının pozitif sayısı 101 olduğu için diğer tüm sınıf sayıları bu değere indirgenmiştir (Bknz. Tablo 2, KA sınıfı, Pozitif: 101). Bu işlemden sonra, karmaşıklık matrisi ve doğruluk, kesinlik, duyarlılık, F1 puanı metrikleri hesaplanmıştır.

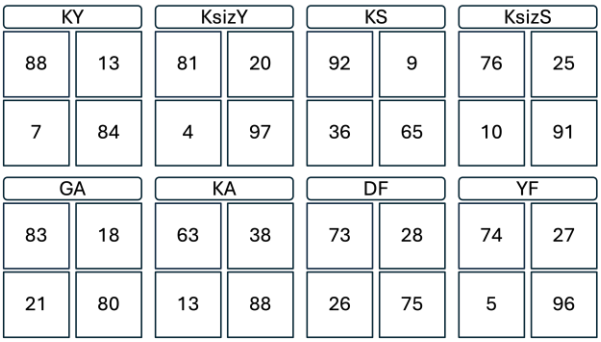
Karmaşıklık matrisi, makine öğrenimi modellerinin performansını değerlendirmek için kullanılan bir tablo yapısıdır. Bu tablo, bir sınıflandırma modelinin tahmin sonuçlarını dört temel kategoride özetler: doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) (Khakhar ve Dubey, 2022). TP, doğru sınıflandırılan pozitif örneklerin sayısını, TN ise doğru sınıflandırılan negatif örneklerin sayısını gösterir. Benzer şekilde, FP, pozitif ve yanlış olarak sınıflandırılan örneklerin sayısını, FN ise negatif ve yanlış olarak sınıflandırılan örneklerin sayısı ifade etmektedir (Kulkarni vd., 2020). Şekil 3'te karmaşıklık matrisi örneği ve her sınıf için karmaşıklık matrisleri Şekil 4'te verilmektedir.

Tablo 2. Örnek alınan verilerde sınıfların pozitif ve negatif dağılımları (Positive and negative distributions of classes in the sampled data)

		KY	KsizY	KS	KsizS	GA	KA	YF	DF
Pozitifler (1'ler)	Frekans	1005	171	652	229	510	101	243	305
	Frekans (%)	58,13	9,89	37,71	13,24	29,50	5,84	14,05	17,64
Negatifler (0'lar)	Frekans	724	1558	1077	1500	1219	1628	1486	1424
	Frekans (%)	41,87	90,11	62,29	86,76	70,50	94,16	85,95	82,36



Şekil 3. Karmaşıklık Matrisi (Confusion matrix)



Şekil 4. Her sınıfa göre karmaşıklık matrisleri (Confusion matrices for each class)

Çalışmada kullanılan değerlendirme metriklerinden ilki doğruluktur. Bu metrik, bir modelin ne kadar doğru tahmin yaptığını gösteren bir performans ölçütüdür. Doğru sınıflandırılanların tüm sınıflandırılanlara oranıdır ve

$$\text{Doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

şeklinde hesaplanır. Kesinlik, modelin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekten pozitif olduğunu ölçmektedir. Yani, TP'lerin toplam pozitif tahminlere oranıdır ve

$$\text{Kesinlik} = \frac{TP}{TP + FP} \tag{2}$$

formülü ile hesaplanır. Duyarlılık, modelin gerçek pozitif örnekleri ne kadar iyi tespit ettiğini ölçer. TP'nin toplam gerçek pozitiflere oranıdır ve

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \tag{3}$$

ile hesaplanır. F1 puanı ise bir modelin kesinlik ve duyarlılığı birlikte değerlendiren bir metriktir. F1 puanı, kesinlik ve duyarlılığın harmonik ortalamasıdır. Kesinlik ve duyarlılığın dengeli olduğu durumlarda en yüksek değere ulaşır. F1 puanı, aşağıdaki formül ile hesaplanır:

$$\text{F1} = \frac{2 \cdot \text{Kesinlik} \cdot \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \tag{4}$$

Çalışmada alt örnekleme metodu kullanıldığından, yani azınlık sınıfına göre her sınıftaki pozitif ve negatiflerden rastgele örnek alındığından değerlendirme metriklerinin hesaplaması on kere tekrarlanmış ve her bir metriğin aritmetik ortalaması alınmıştır. Sonuçlar Tablo 3'te verilmektedir.

Tablo 3. Her sınıf için değerlendirme metrikleri ortalamaları (Evaluation metrics averages for each grade)

Kalite Boyutları	Doğruluk	Kesinlik	Duyarlılık	F1 Puanı
KY	0,8807	0,8867	0,8743	0,8800
KsizY	0,8673	0,9514	0,7743	0,8533
KS	0,7609	0,6986	0,9188	0,7936
KsizS	0,8332	0,9030	0,7475	0,8175
GA	0,8262	0,8297	0,8228	0,8256
KA	0,7604	0,8613	0,6238	0,7229
DF	0,7490	0,7643	0,7218	0,7415
YF	0,8312	0,9519	0,6980	0,8051

Tablo 3'ten görüldüğü üzere, modelin performansı her sınıfa göre değişiklik göstermekte, değerlendirme sonucu mükemmel olmasa da iyi bir performans elde edildiği söylenebilmektedir. Buna göre doğruluk, tüm sınıflar için genelde %75 ile %88 arasında değişmekte, çoğu sınıfta %80'in üstünde değer almaktadır. Kesinlik, özellikle KsizY ve YF sınıflarında oldukça yüksektir ve bu da modelin bu sınıflar için tahmin ettiği pozitif örneklerin çoğunlukla doğru olduğunu göstermektedir. KS sınıfında %70 ile en düşük kesinlik değeri gözlenmiştir. Bu da modelin pozitif tahminlerinin bir kısmının hatalı olduğunun işaretidir. Duyarlılıkta KS %91 ile en yüksek seviyededir. Bu, modelin bu sınıftaki gerçek pozitifleri iyi bir şekilde yakaladığını göstermektedir. Ancak, KA ve YF sınıflarında duyarlılık değerleri sırasıyla %62 ve %70 olarak hesaplanmıştır. Bu da modelin bu sınıflarda daha fazla hatalı negatif tahmin yaptığını işaret eder. F1 puanı incelendiğinde en yüksek F1 puanı KY sınıfında (%88) elde edilmiş, en düşük F1 puanı ise KA sınıfında (%72) gözlemlenmiştir.

4. Bulgular (Findings)

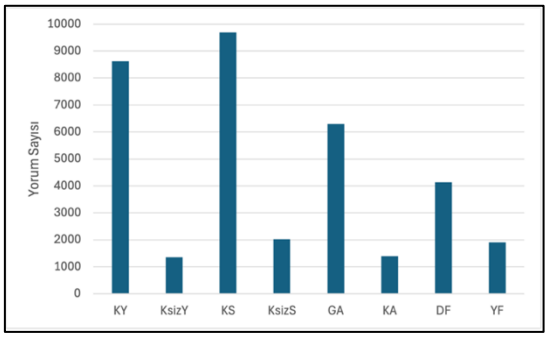
Çalışmada, 89 balık restorana yapılan müşteri yorumları kullanılmıştır. En fazla yorum alan restoranın yorum sayısı 609, en düşük ise 8'dir. Restoran başına ortalama yorum sayısı ise 171,97'dir.

Müşteri memnuniyetini etkileyen farklı boyutların karşılaştırılarak, restoranların güçlü ve zayıf yönlerinin belirlenmesi ve bu alanlara yönelik iyileştirmeler yapılması restoranlar için kritik önem taşımaktadır. Bunun yanında, yorumların farklı kategorilerdeki pozitif ve negatif duygu dağılımlarının zaman içindeki değişimlerinin incelenmesi, restoranların müşteri memnuniyeti eğilimlerini anlamalarına ve bu doğrultuda gelecek için stratejik kararlar almalarına olanak tanımaktadır. Ayrıca, müşteri yorumları içinde farklı

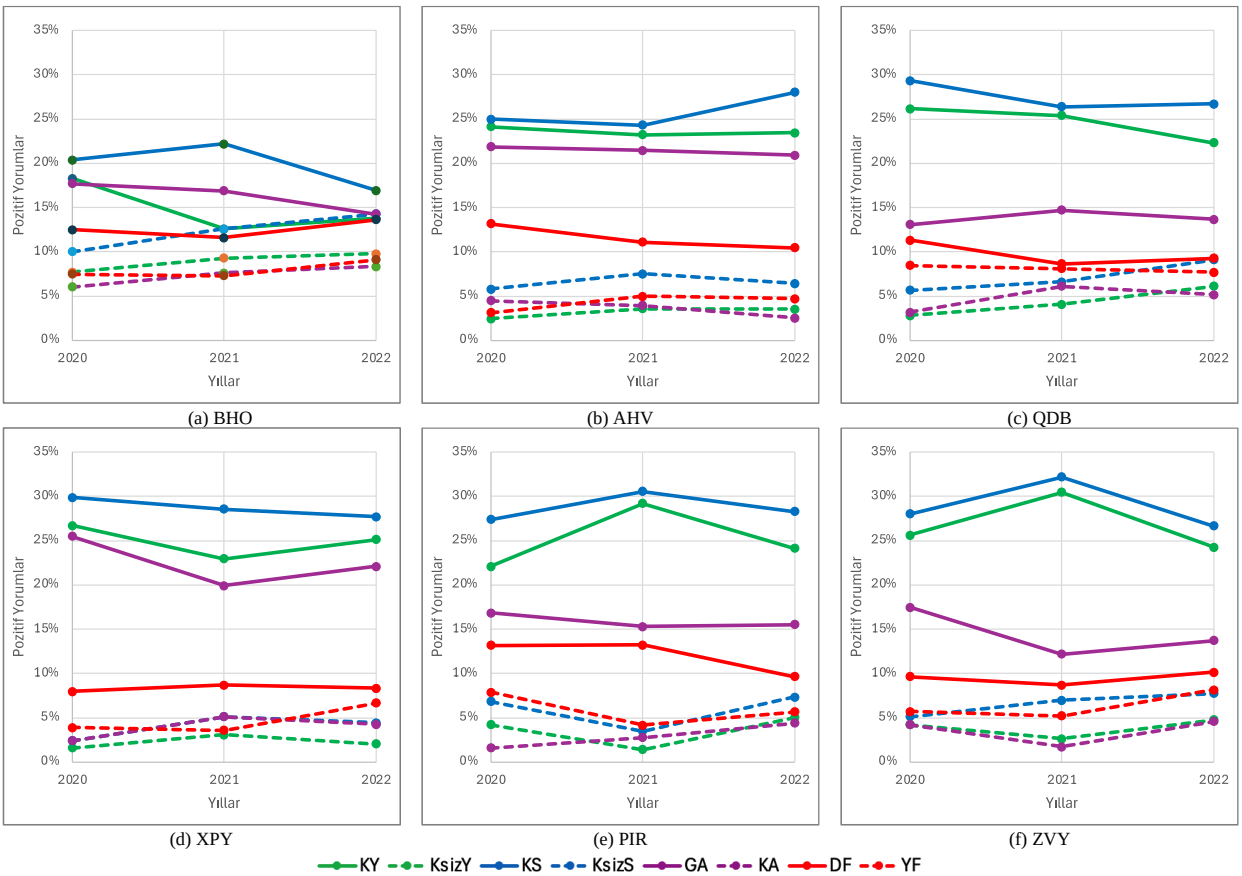
dillerde yapılan yorumlar da bulunmaktadır. İzmir'in turistik bir yer olması ve balık restoranları ile ilgili ünü göz önünde bulundurulduğunda, yabancı müşterilerin algılarının da araştırılması önem taşımaktadır.

Buna yönelik olarak, başta 2013-2023 yılları arasında İzmir genelindeki balık restoranları için müşterilerin memnuniyetini etkileyen en önemli boyutlar araştırılmıştır. Şekil 5'ten de görüleceği üzere 89 restoran için en çok KS üzerine yorum yapılmıştır. En az yorum ise KsizY üzerinedir. Ayrıca, müşteriler çoğunlukla servis boyutu (KS + KsizS) üzerine yorum yapmışlar ve en az yorumlanan ise fiyat boyutudur (DF + YF). Ayrıca, olumlu yorumlar olumsuz yorumlara göre çok daha fazla sayıdadır.

Tüm yorumların %20'lik kısmını kapsayan en çok yorum yapılan ilk altı restoran seçilmiş ve yıllara göre müşteri yorumlarındaki değişim incelenmiştir (Şekil 6). Veriler en son Mart 2023'te alındığından ve 2023 yılında çok az veri olduğundan bu yıla ait veriler incelemeye dahil edilmemiştir. Her yıl için farklı yorum sayıları bulunduğundan, bu değerler o yılda kategorilerde kaç yorum bulunuyorsa toplam yorum sayısına göre yüzde olarak hesaplanmıştır. Bu analiz sayesinde zamansal olarak restoran işletmelerinin kalite boyutlarında olumlu veya olumsuz değişimler gözlenebilmekte, böylece restoranların hangi kalite boyutunda iyileştirme yapmaları gerektiği belirlenebilmektedir. Ayrıca, rakip restoranlar karşılaştırılarak, birbirlerine yönelik eksi veya artı yönler gösterilebilmektedir.



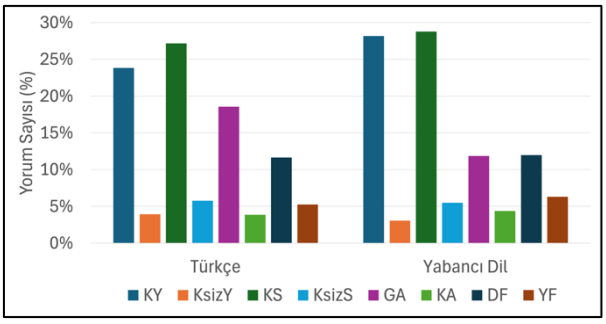
Şekil 5. İzmir balık restoranları kalite boyutu sınıfları dağılımları (Distribution of quality dimension classes of İzmir fish restaurants)



Şekil 6. En çok yorum alan ilk altı restoran için sınıfların yıllara göre değişimi (Yearly change in classes for the top six most reviewed restaurants)

Örneğin, BHO kodlu restoran incelendiğinde her yıl GA üzerine yapılan pozitif yorum sayısı azalmış, aynı şekilde, KA boyutuna yönelik pozitif yorum sayısı artmıştır. Bu durum, restoranın ambiyansa yönelik iyileştirme yapması gerektiğini göstermektedir. Aynı zamanda, GA ve KA sınıfları arasındaki artış ve azalıştaki denge de göz önüne serilmektedir. Bu durum, örnek verilen diğer restoranların çoğunda da farklı boyutlarda gözlenmektedir. XPY ve ZVY restoranları incelendiğinde, ikisinde de DF'a yönelik pozitif olarak pek bir değişim olmadığı görülürken, YF'da özellikle 2021 yılından sonra az da olsa küçük bir artış gözlenmektedir. Bu da bu iki restoranın nispeten fiyatlandırma stratejilerini doğru uyguladıklarını açıklamaktadır. AHV restoranında 2021-2022 yılları arasında KS'de gözle görülür bir iyileşme olduğu görülmektedir. Bu da restoran açısından müşteri memnuniyetini arttırıcı bir özelliktir. Rakip restoranlar karşılaştırıldığında örneğin, XPY 2021'den sonra rakiplerine göre özellikle GA sınıfında müşterilerden gelen pozitif yorumlar neticesinde daha başarılı görülürken, diğer restoranlar bu sınıfta pek de bir başarı gösterememiştir. Buna yönelik olarak da restoranların iyileştirme yapmaları gerekmektedir.

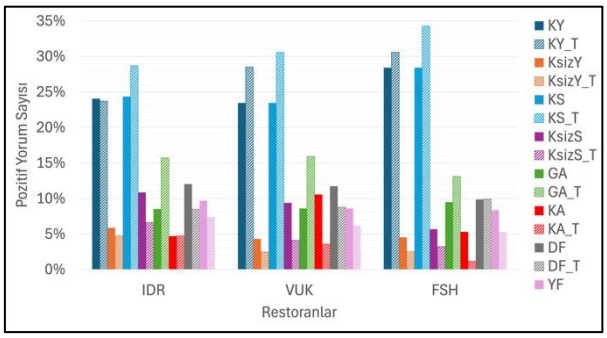
Çalışmada ayrıca Türkçe haricindeki dillerde yapılan yorumlar da incelenmiştir. 15305 yorum için yabancı dilde yapılan yorum sayısı yaklaşık %11'lik paya sahiptir. Müşteri yorumlarındaki Türkçe harici dillerin tespiti için Google Sheets'de yer alan DETECTLANGUAGE() fonksiyonu kullanılmıştır. Şekil 7'de İzmir'deki tüm restoranlar için Türkçe ve yabancı dilde yapılan yorumlar kalite boyutlarına göre karşılaştırılmıştır. 31.287 Türkçe pozitif ve 4137 yabancı dil pozitif olan veriler yüzdeliğe dönüştürülmüştür.



Şekil 7. İzmir'deki balık restoranlarına yapılan yorumların sınıflar ve dillere göre karşılaştırılması (Comparison of comments on fish restaurants in İzmir by classes and languages)

Şekil 7'den görüldüğü üzere, KY ve KS sınıflarında yabancı dilde yazan müşteriler bu sınıflarda Türkçe yazanlara göre daha fazla oranda pozitif yönde görüşlerini belirtmişlerdir. Türkçe yorumda bulunanlarda ise GA sınıfı öne çıkmaktadır.

En çok yabancı dilde yorumun yapıldığı ilk üç restoran incelendiğinde, KS ve GA sınıfları hakkında Türkçe yorumlar, yabancı dildeki yorumlardan oranca daha fazladır (Şekil 8). VUK ve FSH restoranlarında KY sınıfı Türkçe yorum yapanlarda daha yüksek orandadır. Yabancı dilde yorum yapan müşteriler GA'a pek önem vermezken, Türkçe yorum yapan müşterilerde bu oran daha yüksektir. Fiyatlandırma incelendiğinde ise IDR ve VUK restoranlarında yabancı dilde yazan müşteriler Türkçe yazan müşterilere göre DF sınıfı üzerine pozitif yorumlarda bulunmasının yanında, her üç restoranda da fiyatların yüksekliğinden bahsedilmektedir.



Şekil 8. En çok yabancı dilde yorum yapılan üç restorana ait müşteri yorumlarının Türkçe yorumlarla karşılaştırılması (Comparison of customer reviews of the three restaurants with the most foreign language reviews with Turkish reviews)

5. Sonuçlar ve Tartışmalar (Conclusions and Discussions)

Bu makalede, İzmir'deki balık restoranlarına ait müşteri yorumları üzerinde sıfır-atış metin sınıflandırma yöntemi kullanılarak hedef tabanlı duygu analizi gerçekleştirilmiştir. Çalışmada kullanılan yöntem ile restoranların yemek kalitesi, servis kalitesi, fiziksel çevre ve adil fiyat kalite boyutlarında güçlü ve zayıf yönleri belirlenebilmektedir.

Çalışmada kullanılan yöntem, çeşitli açılardan önemli avantajlar sunmaktadır. İlk olarak, sıfır-atış metin sınıflandırma yönteminde kullanılan önceden eğitilmiş dil modelleri, özellikle büyük veri setlerinin hızlı ve etkin bir biçimde sınıflandırılmasını ve kullanıcıların uzun süren eğitim ve test gibi geleneksel makine öğrenmesi aşamalarını atlamasını sağlamaktadır. Bu durum, ayrıca, sürekli artan müşteri yorumlarını gerçek zamanlı olarak analiz etme ihtiyacı olan işletmeler için büyük faydalar sağlayacaktır. İkinci olarak, geleneksel doğal dil işleme projelerinde genellikle tek bir dil üzerinde çalışılırken, diller arasındaki farklılıklar çok dilli veri setleri ile çalışmaya engel olmakta, çalışmada önerilen yöntem ise farklı dillerdeki metinlerden oluşan veri setlerinin dilden bağımsız şekilde analiz edilmesini mümkün kılmaktadır. Üçüncü olarak, sıfır-atış metin sınıflandırma geleneksel hedef tabanlı duygu analizine kıyasla önemli bir avantaj sağlamaktadır. Çünkü, geleneksel yöntem, iki safhada gerçekleştirilmektedir. Yani, başta konular belirlenmekte (örneğin konu modelleme) ve daha sonra belirlenen konular üzerine duygu analizi gerçekleştirilmektedir. Çalışmada önerilen yöntem, bu adımları birleştirerek tek seferde

verilen konu üzerinden duyguyu belirlemekte, böylece, adımlar basitleşerek analiz süresi kısalmaktadır.

Modelin performansının değerlendirilmesi aşamasında, veri setindeki farklı kalite boyutları için pozitif ve negatif örnek sayısındaki dengesizlik, performans değerlendirmesinde zorluk yaratmıştır. Bu durumu azaltmak için alt örnekleme yöntemi kullanılmış, bazı bilgilerin kaybolması dezavantajını en aza indirmek için azınlıklık sınıfı sayısına göre örnekleme tekrarlanmış ve çıkan sonuçların aritmetik ortalaması alınmıştır.

Değerlendirme metrikleri incelendiğinde, modelin farklı sınıflarda değişken bir performans sergilediği gözlemlenmektedir. Doğruluk incelendiğinde, tüm sınıflarda %75 ve üzerinde çıkmıştır. Bu durum, modelin çoğu durumda doğru tahminler yaptığını ve güvenilir olduğunun göstergesidir. KY, KsizY, KsizS, GA ve YF sınıflarında %80 ve üzerindeki F1 puanları modelin bu sınıflarda güvenilir bir şekilde çalıştığını göstermektedir. Bununla birlikte, KS, KA ve DF sınıflarında F1 puanları (sırasıyla, 0,79, 0,72 ve 0,74) modelin bu sınıflardaki performansında dengesizlikler olabileceğini işaret etmektedir. Özellikle KS sınıfında yüksek duyarlılığa (0,92) rağmen nispeten düşük kesinlik (0,70), bu sınıfta hatalı pozitif oranının varlığını göstermektedir. Benzer şekilde, KA sınıfında yüksek kesinlik (0,86) ancak düşük duyarlılık (0,62), modelin doğru pozitifleri belirlemede zorluk yaşadığını ortaya koymaktadır. Fakat diğer sınıflardaki performans, modelin etkili tahminler yapma kapasitesini açıkça gözler önüne sermektedir.

Hataların sebeplerinin birçok nedene bağlı olduğu düşünülmektedir. Örneğin model, "*Çok nezih ve balığı lezzetli bir restoran. Fiyatlar biraz yüksek*." veya "*Yeri çok merkezi. Yiyecekler ve mezeler çok lezzetli*." yorumlarında KS'den bahsedilmemesine rağmen eşik değerinin üstünde bir puanı KS sınıfına vermiştir. Bu tür hatalar kesinlik puanının KS sınıfında düşük çıkmasına yol açmıştır. Aynı şekilde, KA sınıfında, örneğin, "*El değiştirmiş ve gayet bozmuş...*" veya "*Mezgit şiş enfes.*" yorumlarında model KA'ya yüksek puan vermiştir. Yukarıda bahsedilen bu tür hataların sebebinin restoran hakkında yapılan bazı yorumların genel bir değerlendirme olduğundan kaynaklandığı ve böylece modelin yanlış yorumladığı kanısına varılmıştır. Bununla birlikte, her ne kadar veri ön işleme aşamasında üç kelimenin altındaki yorumlar veri setinden çıkarılmış olsa da yukarıda örneği verilen nispeten kısa cümlelerden ötürü, modelin hatalar yapmış olduğu olasıdır. Örneğin, "*Yemek kalitesi iyiydi, fiyatlarda bu kalite için çok pahalı değil. Fakat restorandaki yaklaşım maalesef iyi değildi. Mekân genel anlamda boş olduğu halde talep ettiğimiz masayı bize vermediler, eşimle beni ısrarla kuytu köşe bir yere oturttular. 600 TL'lik bir hesaptan sonra meyve ikram edilmesi usuldendir, yapamayacaklarını söylediler, zorunda değiller tabi. Ama çay ikram ettiler.*" yorumu gayet başarılı bir şekilde sınıflandırılmıştır.

Bu çalışmada kullanılan yöntem, yöneticilerin karar alma süreçlerinde restoranların güçlü ve zayıf yönlerini hızlı ve etkin bir biçimde belirlemelerine yardımcı olacak ve hizmet kalitesini artırmak için gerekli iyileştirme alanlarını tespit etmelerini sağlayacaktır. Bu sayede yöneticiler, stratejik kararlar alabilecek ve operasyonel iyileştirme süreçlerini başlatabileceklerdir. Örneğin, yemek kalitesinin iyi olduğu, ancak servis kalitesinin veya fiyatlandırmanın olumsuz geri bildirim aldığı tespit edildiğinde, bu bilgi, yöneticilerin ilgili alanlara yönelik iyileştirme ve yatırımlar yapma kararı almasına olanak sağlayacaktır.

Çalışmada önerilen yöntem, yöneticilerin, farklı müşteri gruplarının restoran hizmetleri konusundaki ihtiyaç ve beklentilerini, bunların ne ölçüde karşılandığını daha iyi değerlendirmelerini sağlayacaktır. Örneğin, yabancı müşterilerin belirli bir boyuta verdikleri önem, Türk müşteriler için farklı olabilmektedir. Bu bilgiler, müşteri segmentasyonuna göre özelleştirilmiş hizmet sunum stratejilerinin geliştirilmesine yardımcı olacaktır.

Gelecekteki çalışmalar için model performansının daha etkin bir şekilde değerlendirilmesinin önemli olduğu düşünülmektedir. Örneğin, KY boyutunda 1005 olan pozitif yorum sayısı, alt örnekleme yöntemiyle KA boyutundaki pozitif yorum sayısına, yani 101'e düşürülmüştür. Rastgele on kez örnek alınarak alt örnekleme uygulanıp sonuçların aritmetik ortalaması alınsa da azınlık sınıflarının temsil edilebilirliğini artırmak ve modelin düşük sayıya sahip sınıflarda daha tutarlı performans göstermesini sağlamak için bu sınıfların veri sayısının artırılması gerekmektedir. Bu amaçla, sentetik veri üretim yöntemleri kullanılabilirliği olasıdır. Ayrıca, çalışmada kullanılan önceden eğitilmiş dil modeli yerine farklı modellerin denenmesiyle hataların azaltılması ve model performansının iyileştirilmesi de mümkündür. Buna ek olarak, veri setine özgü bir ince ayar yapılmasıyla modelin hedef boyutlar üzerindeki duyarlılığını ve genel performansını artırabileceği öngörülmektedir.

İleriki çalışmalarda, kullanılan yöntemin GPT gibi yeni nesil dönüştürücü tabanlı modeller ile karşılaştırılması hedeflenmektedir. Bu sayede, karmaşık dil yapılarının işlenmesi konusunda yeni açılımlar sağlanacaktır. Ayrıca, metinlerdeki duygu yoğunluğunun ölçülmesi (örneğin, müşterinin bir şeyi çok, az veya orta seviye sevmesi), müşteri taleplerinin ve beklentilerinin daha derinlemesine anlaşılmasını sağlayacaktır. Bu sayede, müşterilerin belirli bir hizmet veya ürüne dair farklı unsurlara yönelik duygu yoğunlukları arasındaki farklar net bir şekilde ortaya konabilecektir.

Kaynaklar (References)

Abdullah, M., Waheed, S., Hossain, S., 2023. Sentiment analysis of restaurant reviews using machine learning. Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022, 17–18 December 2022, Tangail, Bangladesh, pp. 419–428.

Ahmed, M., Chen, Q., Li, Z., 2020. Constructing domain-dependent sentiment dictionary for sentiment analysis. Neural Computing & Applications, 32(18), 14719–14732.

Akgündüz, Y., Koba, Y., Alkan, C., 2017. Yerel halkın bakış açısıyla İzmir'in gastronomi turizmi değerleri ve gelişme potansiyeli. Aydın gastronomy, 8(1), 169–186.

Akyurt Kurnaz, H., 2024. Gastronomi rehberliği. Anatolia: Turizm araştırmaları dergisi, 35(1), 129–132.

Altıntaş, V., Hazarhun, E., 2020. İzmir'in gastronomi turizmi potansiyeline turist rehberlerinin bakış açıları. International journal of applied economic and finance studies, 5(2), 13-36.

Ara, J., Hasan, Md. T., Al Omar, A., Bhuiyan, H., 2020. Understanding customer sentiment: Lexical analysis of restaurant reviews. 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh, pp. 295-299.

Azunre, P., 2021. Transfer learning for natural language processing. Manning, New York.

Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A., 2018. Zero-shot object detection. ECCV 2018, 15th European Conference, 8-14 September 2018, Munich, Germany, pp. 397–414.

Bengül, S., Dinç, T. E., 2023. Marka deneyimi ile müşteri sadakati arasındaki ilişkinin kanonik korelasyon analizi ile incelenmesi, yiyecek içecek işletmeleri üzerinde bir uygulama. Pazarlama ve pazarlama araştırmaları dergisi, 16(2), 421-450.

Bozkurt, A. H., Yalçın, N., 2024. Topluluk öğrenmesi algoritmaları kullanarak Amazon yemek yorumları üzerine duygu analizi. Bilecik Şeyh Edebali Üniversitesi fen bilimleri dergisi, 11(1), 128–139.

Branco, A., Parada, D., Silva, M., Mendonça, F., Mostafa, S. S., Morgado-Dias, F., 2024. Sentiment analysis in Portuguese restaurant reviews: Application of transformer models in edge computing. Electronics, 13(3), 589.

Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., Chalupka, K., 2020. Rethinking zero-shot video classification: End-to-end training for realistic applications. Conference on Computer Vision and Pattern Recognition (CVPR), 13-19 June 2020, Seattle, WA, USA, pp. 4612-4622.

Carrasco, P., Dias, S., 2023. Exploring natural language processing and sentence embeddings for sentiment analysis of online restaurant reviews. Atas da 23ª Conferência da Associação Portuguesa de Sistemas de Informação, 19-21 October 2023, Peja, Portugal.

Carrasco, P., Dias, S., 2024. Enhancing restaurant management through aspect-based sentiment analysis and NLP techniques. Procedia computer science, 237, 129–137.

Chang, M-W., Ratinov, L., Roth, D., Srikumar, V., 2008. Importance of semantic representation: dataless classification. 23rd National Conference on Artificial Intelligence – Volume 2, 13–17 July 2008, Chicago Illinois, pp. 830–835.

Cherapanukorn, V., Sugunnasil, P., 2022. Tourist attraction satisfaction factors from online reviews. A case study of tourist attractions in Thailand. Journal of environmental management and tourism, 13(2), 379.

Chifu, A.-G., Fournier, S., 2023. Sentiment difficulty in aspect-based sentiment analysis. Mathematics, 11(22), 4647.

Choi, K., Ko, Y., 2023. Meta-learning with topic-agnostic representations for zero-shot stance detection. Pattern recognition letters, 171, 15–20.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale. 58th Annual Meeting of the Association for Computational Linguistics, July 2020, Online, pp. 8440-8451.

Çelik, E., Dalyan, T., 2023. Unified benchmark for zero-shot Turkish text classification. Information processing & management, 60(3), 103298.

Das, S., Deb, N., Chaki, N., Cortesi, A., 2023. Driving the technology value stream by analyzing app reviews. IEEE transactions on software engineering, 49(7), 3753–3770.

Erdoğan, S., Özdemir, G., 2018. İzmir destinasyonunda gastronomi turizmi üzerine bir araştırma. Journal of tourism and gastronomy studies, 10(3), 249–272.

Estevam, V., Pedrini, H., Menotti, D., 2021. Zero-shot action recognition in videos: A survey. Neurocomputing, 439, 159–175.

Gagić, S., Tešanović, D., Jovičić, A., 2013. The vital components of restaurant quality that affect guest satisfaction. Turizam, 17(4), 166–176.

Gallego, V., 2023. (5 Haziran 2024) XLM-RoBERTa-large-XNLI-ANLI. https://huggingface.co/vicgalle/xlm-roberta-large-xnli-anli

Gedif, B., Alemu, A., Assefa, Y., Nibret, S., 2023. Design amharic text sentiment analysis model using machine learning techniques. In case of restaurant reviews. 2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), 26-28 October 2023, Bahir Dar, Ethiopia, pp. 150–154.

Goodwin, T. R., Savery, M. E., Demner-Fushman, D., 2020. Towards zero-shot conditional summarization with adaptive multi-task fine-tuning. Findings of the Association for Computational Linguistics: EMNLP 2020, 16-20 November 2020, Online, pp. 3215–3226.

Hall, C.M., Sharples, L., Mitchell, R., Macionis, N., Cambourne, B. (Eds.), 2003. Food tourism around the world. Butterworth-Heinemann, Oxford.

Henrickson, K., Rodrigues, F., Pereira, F. C., 2019. Data preparation. C. Antoniou, L. Dimitriou, F. Pereira (Eds.), Mobility patterns, big data and transport analytics pp. 73–106, Elsevier, Amsterdam.

Hoang, M., Bihorac, O. A., Rouces, J., 2019. Aspect-based sentiment analysis using BERT. 22nd Nordic Conference on Computational Linguistics, 30 September – 2 October 2019, Turku, Finland, pp. 187–196.

Hoppe, F., Dessi, D., Sack, H., 2021. Understanding class representations: An intrinsic evaluation of zero-shot text classification. Workshop on Deep Learning for Knowledge Graphs (DL4KG@ISWC2021), 25 October, Online.

Hossain, E., Sharif, O., Hoque, M. M., Sarker, I. H., 2021. SentiLSTM: A deep learning approach for sentiment analysis of restaurant reviews. 20th International

Conference on Hybrid Intelligent Systems (HIS 2020), 14-16 December, Online, pp. 193–203.

Hossain, N., Bhuiyan, M. R., Tumpa, Z. N., Hossain, S. A., 2020. Sentiment analysis of restaurant reviews using combined CNN-LSTM. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-3 July 2020, Kharagpur, India.

Hugging Face, 2024. Pipelines. (10 Temmuz 2024) https://huggingface.co/docs/transformers/main_classes/pipelines

Iram, S., Vialatte, F.-B., Qamar, M. I., 2016. Early diagnosis of neurodegenerative diseases from gait discrimination to neural synchronization. D. Al-Jumeily, A. Hussain, C. Mallucci, C. Oliver (Eds.), Applied computing in medicine and health (pp. 1–26). Elsevier, Waltham.

Jiang, Q., Chen, L., Xu, R., Ao, X., Yang, M., 2019. A challenge dataset and effective models for aspect-based sentiment analysis. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3-7 November 2019, Hong Kong, China, pp. 6280–6285.

Jiang, Y., Gao, J., Shen, H., Cheng, X., 2023. Zero-shot stance detection via multi-perspective contrastive learning with unlabeled data. Information processing & management, 60(4), 103361.

Jiao, Q., 2023. A brief survey of text classification methods. 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), 26-28 May 2023, Chongqing, China, pp. 1384-1389.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J., 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5, 339–351.

Jurafsky, D., Chahuneau, V., Routledge, B. R., Smith, N. A., 2014. Narrative framing of consumer sentiment in online restaurant reviews. First Monday, 19(4).

Kar, A., Dhara, S. K., Sen, D., Biswas, P. K., 2021. Zero-shot single image restoration through controlled perturbation of Koschmieder's model. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20-25 June 2021, Nashville, TN, USA, pp. 16200-16210.

Khakhar, P., Dubey, R. K., 2022. The integrity of machine learning algorithms against software defect prediction. R. Pandey, S. K. Khatri, N. K. Singh, P. Verma (Eds.), Artificial Intelligence and Machine Learning for EDGE Computing, (pp. 65–74). Elsevier, London.

Krishna, A., Akhilesh, V., Aich, A., Hegde, C., 2019. Sentiment analysis of restaurant reviews using machine learning techniques. International Conference, ICERECT 2018, 23-24 August 2018, Mandya, India, pp. 687–696.

Kukanja, M., Gomezelj Omerzel, D., Kodrič, B., 2017. Ensuring restaurant quality and guests' loyalty: An integrative model based on marketing (7P) approach. Total quality management & business excellence, 28(13–14), 1509–1525.

Kulkarni, A., Chong, D., Batarseh, F. A., 2020. Foundations of data imbalance and solutions for a data democracy. F. A. Batarseh, R. Yang (Eds.), Data Democracy (pp. 83–106). Elsevier, London.

Kumar, A., Albuquerque, V. H. C., 2021. Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor Indian language. ACM transactions on Asian and low-resource language information processing, 20(5), 1–13.

Kumar, J., Konar, R., Balasubramanian, K., 2020. The impact of social media on consumers' purchasing behaviour in Malaysian restaurants. Journal of tourism, sustainability and well-being, 8(3), 197-216.

Küçükkömürler, S., Şırvan, N. B., Sezgin, A. C., 2019. Dünyada ve Türkiye'de gastronomi turizmi. Uluslararası turizm ekonomi ve işletme bilimleri dergisi, 2(2), 78-85.

Lampert, C. H., Nickisch, H., Harmeling, S., 2014. Attribute-based classification for zero-shot visual object categorization. IEEE transactions on pattern analysis and machine intelligence, 36(3), 453-465.

Lavanya, B. N., Shenoy, P. D., Venugopal, K. R., 2023. Sentiment analysis of social media reviews using machine learning and word embedding techniques. 2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON), 5-7 August 2023, Mysore, India.

Leburu-Dingalo, T., Ntwaagae, K. J., Motlogelwa, N. P., Thuma, E., Mudongo, M., 2022. Application of XLM-RoBERTa for multi-class classification of conversational hate speech. FIRE 2022 Working Notes, 9-13 December 2022, Kolkata, India, pp 590-595.

Lepkowska-White, E., Parsons, A., 2019. Strategies for monitoring social media for small restaurants. Journal of Foodservice Business research, 22(4), 351–374.

Li, Y., Zhu, Z., Yu, J.-G., Zhang, Y., 2021. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification. IEEE transactions on geoscience and remote sensing, 59(12), 10590–10603.

Liu, B., Li, X., Lee, W S., Yu, P. S., 2004. Text classification by labeling words. 19th national conference on artificial intelligence (AAAI'04), 25-29 July 2004, San Jose, California, USA, pp. 425–430.

Liu, C., Fang, F., Lin, X., Cai, T., Tan, X., Liu, J., Lu, X., 2021. Improving sentiment analysis accuracy with emoji embedding. Journal of safety science and resilience, 2(4), 246–252.

Liu, H., Zhang, X., Fan, L., Fu, X., Li, Q., Wu, X.-M., Lam, A. Y. S., 2019. Reconstructing capsule networks for zero-shot intent classification. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3-7 November 2019, Hong Kong, China, pp. 4799–4809.

Logeswaran, L., Chang, M.-W., Lee, K., Toutanova, K., Devlin, J., Lee, H., 2019. Zero-shot entity linking by reading entity descriptions. 57th Annual Meeting of the Association for Computational Linguistics, 28 July – 2 August 2019, Florence, Italy, pp. 3449–3460.

Mahapatra, D., Bozorgtabar, B., Ge, Z., 2021. Medical image classification using generalized zero shot learning. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 11-17 October 2021, Montreal, BC, Canada, pp. 3337-3346.

Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., Kyriazis, D., 2023. Multilingual text categorization and sentiment analysis: A comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural computing & applications, 35, 21415–21431.

Masarifoglu, M., Tigrak, U., Hakyemez, S., Gul, G., Bozan, E., Buyuklu, A. H., Özgür, A., 2021. Bankacılık alanında müşteri yorumlarının BERT tabanlı yaklaşımlar ile duygu analizi. 2021 29th Signal Processing and Communications Applications Conference (SIU), 9-11 June, Istanbul, Turkey.

McInerney, D. J., Young, G., van de Meent, J.-W., Wallace, B. C., 2023. (15 Temmuz 2024). CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. https://arxiv.org/abs/2302.12343

Mohammad, S. M., 2017. Challenges in sentiment analysis. E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco (Eds.), A practical guide to sentiment analysis (pp. 61–83), Springer, Cham.

Nankani, H., Dutta, H., Shrivastava, H., Rama Krishna, P. V. N. S., Mahata, D., Shah, R. R., 2020. Multilingual sentiment analysis, B. Agarwal, R. Nayak, N. Mittal, S. Patnaik (Eds.), Deep learning-based approaches for sentiment analysis (pp. 193–236). Springer, Singapore.

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and trends® in information retrieval, 2(1–2), 1–135.

Patil, D. R., Shukla, D., Kumar, A., Rajanak, Y., Pratap Singh, Y., 2022. Machine learning for sentiment analysis and classification of restaurant reviews. 2022 3rd International Conference on Computing, Analytics and Networks (ICAN), 18-19 November 2022, Rajpura, Punjab, India.

Pérez, J. M., Furman, D. A., Alemany, L. A., Luque, F. M., 2022. RoBERTuito: A pre-trained language model for social media text in Spanish. Thirteenth Language Resources and Evaluation Conference, 20-25 June 2022, Marseille, France, pp. 7235–7243.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Lim, C. P., X-Z., Wang, Wu, Q. M. J., 2023). A review of generalized zero-shot learning methods. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4051–4070.

Pradhan, B., Al-Najjar, H. A. H., Sameen, M. I., Tsang, I., Alamri, A. M., 2020. Unseen land cover classification from high-resolution orthophotos using integration of zero-shot learning and convolutional neural networks. Remote Sensing, 12(10), 1676.

Quazi, S., Musa, S. M., 2022. Text classification and categorization through deep learning. 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), 4-6 December 2022, Al-Khobar, Saudi Arabia, pp. 513-519.

Revathi, K. L., Satish, A. R., Rao, P. S., 2023. Feature level fine grained sentiment analysis for classifying online restaurant reviews. 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 5-7 April 2023, Trichirappalli, India.

Rey-Moreno, M., Sánchez-Franco, M. J., Rey-Tienda, M. D. la S., 2023. Examining transaction-specific satisfaction and trust in Airbnb and hotels. An application of BERTopic and zero-shot text classification. Tourism & management studies, 19(2), 21-37.

Rezaei, M., Shahidi, M., 2020. Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review. Intelligence-based medicine, 3(100005), 100005.

Romera-Paredes, B., Torr, P. H. S., 2017. An embarrassingly simple approach to zero-shot learning. R. S. Feris, R. S., C. Lampert, D. Parikh (Eds.), Visual attributes (pp. 11–30). Springer, Cham.

Sahar, A., Ayoub, M., Hussain, S., Yu, Y., Khan, A., 2022. Transfer learning-based framework for sentiment classification of cosmetics products reviews. Pakistan journal of engineering and technology, 5(3), 38–43.

Sanghi, A., Chu, H., Lambourne, J. G., Wang, Y., Cheng, C.-Y., Fumero, M., Malekshan, K. R., 2022. CLIP-forge: Towards zero-shot text-to-shape generation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18-24 June 2022, New Orleans, LA, USA, pp. 18582-18592.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., Pantic, M., 2017. A survey of multimodal sentiment analysis. Image and vision computing, 65, 3–14.

Suciati, A., Budi, I., 2019. Aspect-based opinion mining for code-mixed restaurant reviews in Indonesia. 2019 International Conference on Asian Language Processing (IALP), 15-17 November 2019, Shanghai, China, pp. 59–64.

Thompson, B., Post, M., 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 16-20 November 2020, Online, pp. 90–121.

Tuna, M. F., Polatgil, M., Kaynar, O., 2023. Restoran müşterilerinin geri bildirimleri üzerinde hedef kategorinin tespiti ve hedef tabanlı duygu analizi. Süleyman Demirel Üniversitesi vizyoner dergisi, 14(40), 1205–1221.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. 31st International Conference on Neural Information Processing Systems (NIPS 2017), 4-9 December 2017, Long Beach, CA, USA.

Vétil, R., Abi-Nader, C., Bône, A., Vullierme, M.-P., Rohé, M.-M., Gori, P., Bloch, I., 2022. Learning shape distributions from large databases of healthy organs: Applications to zero-shot and few-shot abnormal pancreas detection. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, 18-22 September 2022, Singapore, Singapore, pp. 464–473.

Wang, W., Zheng, V. W., Yu, H., Miao, C., 2019. A survey of zero-shot learning. ACM transactions on intelligent systems and technology, 10(2), 1–37.

Yentür, F., Demir, C., 2022). The current perceptions of travel agencies in Izmir about gastronomy tourism and their actual gastronomic tourism offers. Journal of gastronomy hospitality and travel, 5(1), 238-249.

Yi, S., Zhao, J., Joung, H.-W. (david), 2018. Influence of price and brand image on restaurant customers' restaurant selection attribute. Journal of foodservice business research, 21(2), 200–217.

Yu, L., Feng, Q., Qian, Y., Liu, W., Hauptmann, A. G., 2020, June 14-19). Zero-VIRUS: Zero-shot vehicle route understanding system for intelligent transportation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 14-19 June, 2020, Seattle, WA, USA, pp. 2534-2543.

Yüksekbilgili, Z., 2015. Restoran seçim ölçütleri üzerine bir araştırma. Journal of Yaşar University, 9(36), 6453-6360.

Zahoor, K., Bawany, N. Z., Hamid, S., 2020. Sentiment analysis and classification of restaurant reviews using machine learning. 2020 21st International Arab Conference on Information Technology (ACIT), 28-30 November 2020, Giza. Egypt.

Zhan, C., She, D., Zhao, S., Cheng, M.-M., Yang, J., 2019. Zero-shot emotion recognition via affective structural embedding. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 October-2 November 2019, Seoul, Korea (South), pp. 1151-1160.

Zhang, B., Williams, P., Titov, I., Sennrich, R., 2020. Improving massively multilingual neural machine translation and zero-shot translation. 58th Annual Meeting of the Association for Computational Linguistics, 5-10 July 2020, Online, pp. 1628–1639.

Zhang, S., Ly, L., Mach, N., Amaya, C., 2022. Topic modeling and sentiment analysis of Yelp restaurant reviews. International journal of information systems in the service sector, 14(1), 1–16.

Zheng, S., Gupta, G., 2022. Semantic-guided zero-shot learning for low-light image/video enhancement. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 4-8 January 2022, Waikoloa, HI, USA, pp. 581-590.

Research Article
Journal of Intelligent Systems: Theory and Applications 8(1) (2025) 63-83
DOI: 10.38016/jista.1566965

Outliers Treatment for Improved Prediction of CO and NOx Emissions from Gas Turbines Using Ensemble Regressor Approaches

Vahid Sinap[1*]

[1] Department of Management Information Systems, Ufuk University, Ankara, Türkiye

vahidsinap@gmail.com

Abstract
Gas turbines are widely used in power generation plants due to their high efficiency, but they also emit pollutants such as CO and NOx. This study focuses on developing predictive models for predicting CO and NOx emissions from gas turbines using machine learning algorithms. The dataset used includes pollutant emission data from a combined cycle gas turbine (CCGT) in Türkiye, collected hourly between 2011 and 2015. Various outlier treatment methods such as Z-Score, Interquartile Range (IQR), and Mahalanobis Distance (MD) are applied to the dataset. Machine learning algorithms including Random Forest, Extra Trees, Linear Regression, Support Vector Regression, Decision Tree, and K-Nearest Neighbors are used to build the predictive models, and their performances are compared. Additionally, Voting Ensemble Regressor (VR) and Stacking Ensemble Regressor (SR) methods are employed, using Gradient Boosting, LightGBM, and CatBoost as base learners and XGBoost as a meta-learner. The results demonstrate that the SR model, when applied to the dataset processed using the IQR method, achieves the highest prediction accuracy for both NOx and CO emissions, with R^2 values of 0.9194 and 0.8556, and RMSE values of 2.7669 and 0.4619, respectively. These findings highlight the significant role of the IQR method in enhancing model accuracy by effectively handling outliers and reducing data noise. The improved data quality achieved through this method contributes to the superior performance of the SR model, making it a reliable approach for predicting NOx and CO emissions with high precision.

Keywords: Gas turbine emissions, Machine learning, Outlier processing, Combined cycle power generation, Interquartile range, Mahalanobis distance

Gaz Türbinlerinden Kaynaklanan CO ve NOx Emisyonlarının Tahmininde Aykırı Değer İşleme ve Topluluk Regresyon Yaklaşımlarının Kullanımı

Öz
Gaz türbinleri, yüksek verimlilikleri nedeniyle enerji üretim tesislerinde yaygın olarak kullanılmaktadır; ancak, aynı zamanda CO ve NOx gibi zararlı gaz emisyonlarına da neden olmaktadırlar. Bu çalışma, gaz türbinlerinden kaynaklanan CO ve NOx emisyonlarını tahmin etmek için makine öğrenmesi algoritmalarını kullanarak tahmin modelleri geliştirmeye odaklanmaktadır. Kullanılan veri seti, Türkiye'deki bir kombine çevrim gaz türbininden (CCGT) 2011 ve 2015 yılları arasında saatlik olarak toplanan emisyon verilerini içermektedir. Veri setine Z-Skoru, Çeyrekler Arası Aralık (IQR) ve Mahalanobis Mesafesi (MD) gibi çeşitli aykırı değer işleme yöntemleri uygulanarak modellerin performansına etkisine incelenmiştir. Modeller oluşturulurken Rastgele Orman, Ekstra Ağaçlar, Doğrusal Regresyon, Destek Vektör Regresyonu, Karar Ağacı ve K-En Yakın Komşu gibi makine öğrenmesi algoritmaları kullanılmış ve performansları karşılaştırılmıştır. Ayrıca, Gradient Boosting, LightGBM ve CatBoost algoritmalarını temel temel öğrenici ve XGBoost'u meta-öğrenici olarak kullanan Oylama Topluluk Regresyonu (VR) ve İstifleme Topluluk Regresyonu (SR) yöntemlerinin de performansları incelenmiştir. Sonuçlar, IQR yöntemiyle işlenen veri seti üzerinde uygulanan SR modelinin hem NOx hem de CO emisyonları için en yüksek tahmin doğruluğunu sağladığını göstermektedir. Modelin R^2 değeri NOx için 0.9194, CO için 0.8556 olarak bulunmuş; RMSE ise sırasıyla 2.7669 ve 0.4619 olarak elde edilmiştir. IQR yöntemiyle elde edilen iyileştirilmiş veri kalitesi, SR modelinin üstün performans göstermesine katkı sağlamakta ve modelin NOx ve CO emisyonlarını yüksek hassasiyetle tahmin edebilmesi açısından güvenilir bir yaklaşım olduğunu ortaya koymaktadır.

Anahtar Kelimeler: Gaz türbini emisyonları, Makine öğrenmesi, Aykırı değer işleme, Kombine çevrim enerji üretimi, Çeyrekler arası aralık, Mahalanobis uzaklığı

* Corresponding Author.
E-mail: vahidsinap@gmail.com

Recieved : 14 Oct 2024
Revision : 24 Jan 2025
Accepted : 5 Feb 2025

1. Introduction

Gas turbines (GT) are a widely preferred energy conversion technology in power generation plants due to their high efficiency and reliability. Simply put, GT consist of a series of turbines and compressors on a shaft rotating at high speed. GT take in air from outside and compress it using a compressor. The compressed air is then mixed with a fuel (usually natural gas or oil) and sent to a combustion chamber. In the combustion chamber, the fuel-air mixture is ignited and burns under high temperature and pressure. The high-pressure and high-temperature gases generated by this combustion expand and pass through the turbine, causing the turbine blades to rotate. This rotational movement turns the turbine shaft, which is connected to a generator, thus generating electrical energy. The open cycle of a GT is illustrated in Figure 1.

While the high efficiency and reliability of GT make them an ideal choice for electricity generation, harmful gases such as carbon monoxide (CO) and nitrogen oxides (NOx) released during the combustion process cause environmental impacts. CO is produced as an incomplete combustion product when there is not enough oxygen, or the combustion process is incomplete. CO is a colorless, odorless gas and can be dangerous to humans. When inhaled at dangerous levels, CO can cause severe poisoning and even death (Liu et al. 2021). NOx is the general term for compounds formed because of the reaction of nitrogen and oxygen in the atmosphere under high temperature and pressure, including nitrogen monoxide (NO) and nitrogen dioxide (NO_2). NOx can contribute to acid rain, ozone formation, and air pollution, causing respiratory diseases and environmental damage (Pandey and Chandrashekhar, 2014).

Various methods and technologies are used to reduce the release of harmful gases such as CO and NOx from GT into the environment. These include exhaust gas treatment systems such as selective catalytic reduction (SCR) and selective non-catalytic reduction (SNCR). The SCR system lowers NOx emissions by injecting ammonia (NH_3) or urea into the exhaust gas stream. In the presence of a catalyst, these substances react with NOx, converting it into harmless nitrogen (N_2) and water (H_2O) (Wardana and Lim, 2022). The SNCR system reduces NOx by injecting ammonia or urea into the exhaust gas without the use of a catalyst, relying on high temperatures to facilitate the reaction (Mahmoudi et al. 2010). Additionally, improving combustion efficiency and optimizing the air-fuel ratio can significantly reduce the formation of CO and NOx emissions (Tian et al. 2024). For example, low NOx combustion techniques can be used to provide higher combustion efficiency while minimizing NOx formation. Techniques such as water or steam injection can be used to lower combustion temperatures, thereby reducing NOx emissions. In addition, regular maintenance and cleaning of GT can help reduce CO and NOx emissions. Other effective strategies for reducing CO and NOx emissions include utilizing cleaner fuels, enhancing combustion chamber design, and implementing exhaust gas recirculation (EGR) systems (Kumar et al. 2022). However, traditional emission control methods often face disadvantages such as high costs, complexity, and efficiency issues. Implementing and operating large-scale exhaust gas treatment systems often involve significant costs (Lott et al. 2024). Additionally, the environmental impact of some technologies must be considered. For example, certain exhaust gas treatment systems can produce harmful by-products that may be released into the environment (Lopes et al. 2015). This necessitates a broad and comprehensive evaluation of emission control processes on an ongoing basis to ensure environmental sustainability and regulatory compliance.

In recent years, machine learning (ML) has become a prominent technology for evaluating emission control processes by predicting emissions from GT. ML algorithms analyze large amounts of data and take into account various variables such as operating conditions of GT, fuel composition, air temperature and other environmental factors. By identifying complex relationships between these variables, they create models to predict emissions from GT. These models can quickly respond to changes in the operating conditions of GT and keep emission predictions up to date. For instance, if there is a sudden change in the operating conditions of a gas turbine and its impact on emissions is immediately identified and assessed, control measures can be taken automatically if necessary. Furthermore, predictive models allow proactive measures to be taken, considering operating conditions. ML models can predict future operating conditions by analyzing historical performance data and environmental conditions of the plant. In a scenario where the developed model predicts an impending air temperature increase and identifies its potential impacts on CO and NOx emissions, power plant operators can adjust plant operating parameters based on this information or take proactive measures to minimize emissions by making a specific process change. In this way, predictive models can guide power plants to develop and implement strategies to reduce environmental impacts.

The promising advantages of ML in predicting emissions from GT have been evaluated by some important studies in the literature. Aslan (2024) evaluates the performance of machine learning models, including AdaBoost, XGBoost, and Random Forest (RF), in predicting gas turbine emissions. Using random search optimization, the study finds that AdaBoost achieves the highest accuracy (99.97%) and lowest mean square error (MSE = 2.17). Dirik (2022) conducted a study using the Adaptive Neural Fuzzy Inference System (ANFIS) method to model and predict NOx emissions from a natural gas-fired combined cycle power plant (CCPP). The results demonstrated that the ANFIS models achieved high accuracy in predicting

NOx emissions. Pachauri (2024) discusses the importance of monitoring harmful gas emissions from GT in CCPPs, particularly CO and NOx, to ensure compliance with emission standards. The study proposes a stacked ensemble machine learning (SEM) model for predicting CO and NOx emissions from a CCPP gas turbine. The model uses neural network for regression (NNR), generalized additive model (GAM), and bagging of regression trees (BT) as base learners, with a generalized regression neural network (GRNN) as a meta-learner. The hyperparameters of SEM are optimized using a Bayesian optimization algorithm. The performance of SEM is compared with support vector regression (SVR), decision tree (DT), and linear regression (LR). Simulation results show that SEM significantly reduces the root mean square error (RMSE) for NOx and CO compared to other ML techniques, demonstrating its higher predictive accuracy. The study by Kochueva and Nikolskii (2021) investigates the utility of predictive emission monitoring systems (PEMS) as software solutions to validate and complement costly continuous emission monitoring systems for natural gas electrical generation turbines. The research focuses on building a model for predicting CO and NOx emissions based on ambient variables and technological process parameters using various ML methods. The developed models achieve coefficients of determination of $R^2 = 0.83$ for NOx emissions and $R^2 = 0.89$ for CO emissions. In their study, Kaya et al. (2019) introduce a novel PEMS dataset collected over a period of five years from a gas turbine, specifically for the predictive modeling of CO and NOx emissions. The data is analyzed using a contemporary ML approach, providing valuable insights into emission predictions. It is noted in the study that the most successful algorithm model for the exhaust gas emission prediction is Extreme Learning Machines (ELM). In the study conducted by Dalal et al. (2023), commonly used ML regression models such as Multiple Linear Regression (MLR), DT, RF, Adaboost Regressor, Gradient Boosting Regressor (GB), and XGBoost Regressor were compared using the same dataset for predicting emissions like CO and NOx. According to the results of the research, the RF Model showed the best performance with the highest accuracy of 0.60 for NOx prediction and 0.65 for CO prediction. Coelho et al. (2024) conducted a study to estimate CO and NOx emissions from a gas turbine using the PEMS dataset. They employed four feature generation methods: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), and Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE). Various regression models, including Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), K-Nearest Neighbors (KNN), Cubist Regression, RF, Light Gradient Boosting Machine (LGBM), Categorical Boosting, and Deep Forest Regression (DFR), were evaluated with all the generated features. The DFR model achieved the best results, with an R^2 value of 0.53 for CO emissions in the validation dataset. For NOx emissions, the DFR model achieved an R^2 value of 0.47 for the validation dataset. The study by Yousif et al. (2024) aims to predict gas emissions from natural gas power plants. A hybrid model combining Feed Forward Neural Network (FFNN) and Particle Swarm Optimization (PSO) was developed for this purpose. The FFNN predicts NOx and CO emissions, while the PSO optimizes the FFNN weights to enhance prediction accuracy. The PSO employs a unique random number selection strategy using the KNN algorithm. Neighbor Component Analysis (NCA) is used to select parameters most correlated with emissions. The model was tested with publicly available datasets and evaluated using MSE, mean absolute error (MAE), and RMSE metrics. Results show that the PSO significantly improves FFNN training, increasing CO and NOx prediction accuracy by 99.18% and 82.11%, respectively. Naghibi (2024) develops an advanced gas turbine forecasting model using ensemble decision trees and robust preprocessing. The bagging structure outperforms boosted trees, achieving a lower RMSE (1.4176) with fewer estimators. While effective overall, the model has limitations in specific operating ranges. The study offers insights for optimizing gas turbine efficiency and improving electricity supply reliability.

The aim of this study is to develop predictive models for predicting CO and NOx emissions from GT using algorithms such as RF, Extra Trees (ET), LR, SVR, DT, and KNN, and to compare their performances with VR and SR methods. In VR and SR methods, GB, LightGBM and CatBoost algorithms are used as base learners and XGBoost algorithm is used as meta learner. The models are trained on a dataset containing emission data collected over a five-year period (01/01/2011–31/12/2015). Unlike previous studies that used the same dataset, this research focuses on the problem of outliers in the dataset. By examining and comparing the effects of various outlier treatment methods such as Z-Score, IQR, and MD on the developed models, this study aims to provide a novel contribution to literature. Additionally, the findings of this study are expected to contribute to the field of emission prediction from GT and provide valuable insights for environmental management in the energy sector. By enhancing the accuracy of the developed models, the study aims to play a critical role in environmental impact assessments and sustainable energy policies.

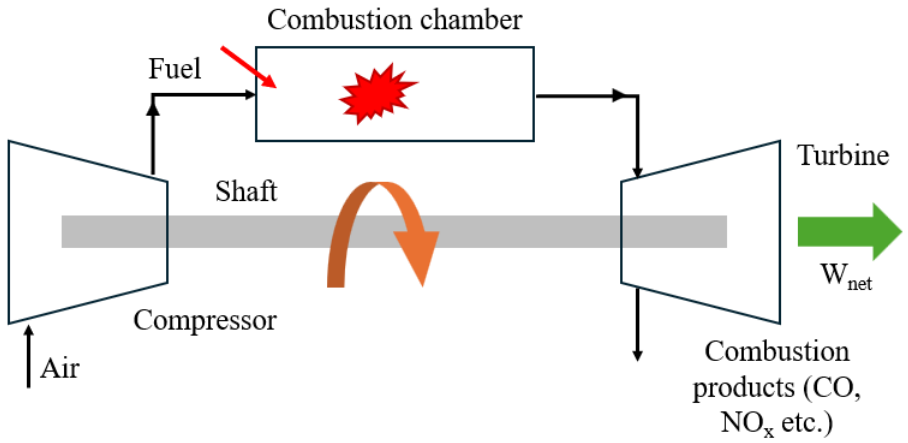


Figure 1. Open cycle of a GT

2. Methodology

In this section, the characteristics of the dataset used in the study, descriptions of the ML algorithms, performance criteria used in the comparison of the algorithms and information about the data preparation process are given. With this information, it is aimed to establish the methodological and analytical foundations of the research, to increase the scientific contribution of the study and to ensure its reproducibility.

2.1. Dataset

The dataset used in this study includes pollutant emission data from a CCGT in Türkiye. The dataset consists of sensor data collected hourly between 2011 and 2015 and is openly available through the UCI repository (Kaya et al. 2019). This set, which includes 36,733 data records in total, belongs to the periods when the power plant operated between 75% and 100% load factors. A graphical representation of the output features is given in Figure 2.

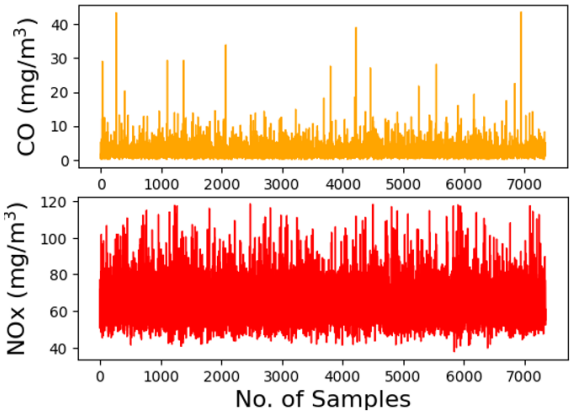


Figure 2. Visual depiction of the output features NOx and CO.

The dataset includes various environmental and operational parameters that affect the performance of the gas turbine. These parameters include ambient temperature (AT), ambient pressure (AP), ambient humidity (AH), air flow differential pressure (AFDP), gas turbine exhaust pressure (GTEP), compressor discharge pressure (CDP), turbine energy yield (TEY), and turbine inlet temperature (TIT) and turbine afterburner temperature (TAT). The main pollutants produced by GT are CO and NOx, while sulfur oxides (SOx) and other pollutants vary depending on the type of fuel used. Table 1 presents the statistical analysis of two output variables (CO and NOx) and nine input variables (AT, AP, AH, AFDP, GTEP, TIT, TAT, CDP and TEY) in the dataset. Especially the atmospheric parameters such as AT, AP, and AH play an important role in predicting CO and NOx emissions (Farzaneh-Gord, and Deymi-Dashtebayaz, 2011). For CCGT, AFDP, GTEP, TIT, TAT, and CDP parameters are very influential, and sensor locations and measurement methods of these variables are of great importance (Wood, 2023). AFDP sensors are usually placed before and after the air filter, GTEP sensors in the exhaust duct, TIT sensors at the turbine inlet, TAT sensors at the turbine outlet, and CDP sensors at the compressor outlet. Correct positioning and regular calibration of these sensors ensures efficient and safe operation of CCGT systems.

Understanding the relationships between input and output variables is critical for improving the accuracy of predictive models. These relationships are analyzed using the correlation coefficient (CC), which is calculated with Pearson correlation. CC values indicate the level of dependence between variables, with positive values indicating a direct relationship and negative values indicating an inverse relationship. Figure 3 shows the correlations between CO, NOx, and other input variables. The concentration of CO demonstrates negative correlations with several operational parameters, including AT, AFDP, GTEP, TIT, TEY, and

CDP, with correlation coefficient values of -0.17, -0.45, -0.52, -0.71, 0.57, and -0.55, respectively. This implies that as the turbine's inlet temperature and the compressor's discharge pressure decrease, the emission of CO increases. Conversely, NOx emissions exhibit a higher level of correlation with a decrease in AT (-0.56). During the winter season, it is recommended to operate the gas turbine at higher temperatures to mitigate NOx emissions. However, GT operation is also negatively correlated with AFDP (-0.19), GTEP (-0.20), TIT (-0.21), TAT (-0.09), TEY (-0.12), and CDP (-0.17), respectively. Moreover, CO demonstrates positive correlations with AP (0.07), AH (0.11), and TAT (0.06), while NOx is positively correlated with AP and AH, with correlation coefficient values of 0.19 and 0.16, respectively. Figure 4 illustrates the schematic diagram of the CCPP, encompassing all input and output features.

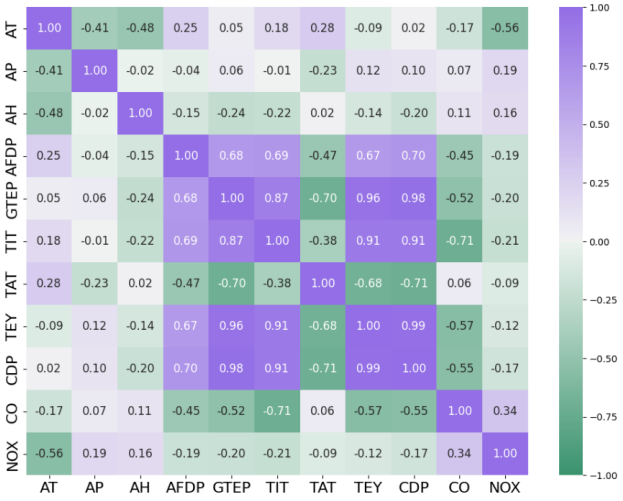


Figure 3. Matrix of Pearson correlation coefficients among the features

Table 1. Statistical overview of the dataset

Features	Unit	Average	Min	Max	Skewness	Kurtosis	Standard Dev.
AT	°C	17.71	-6.23	37.10	-0.0435	-0.8266	7.4474
AP	mbar	1013.07	985.85	1036.60	0.1941	0.4419	6.4633
AH	%	77.86	24.08	100.20	-0.6280	-0.2745	14.4613
AFDP	mbar	3.92	2.08	7.61	0.3810	0.2246	0.7739
GTEP	mbar	25.56	17.69	40.71	0.3290	-0.6538	4.1959
TIT	°C	1081.42	1000.80	1100.90	-0.8882	-0.0457	17.5363
TAT	°C	546.15	511.04	550.61	-1.7559	2.0167	6.8423
CDP	mbar	12.06	9.85	15.15	0.2367	-0.6315	1.0887
TEY	MWh	133.50	100.02	179.50	0.1165	-0.5001	15.6186
CO	mg/m^3	2.37	0.0003	44.10	4.8381	49.0817	2.2626
NOx	mg/m^3	65.29	25.90	119.91	1.0267	2.0375	11.6783

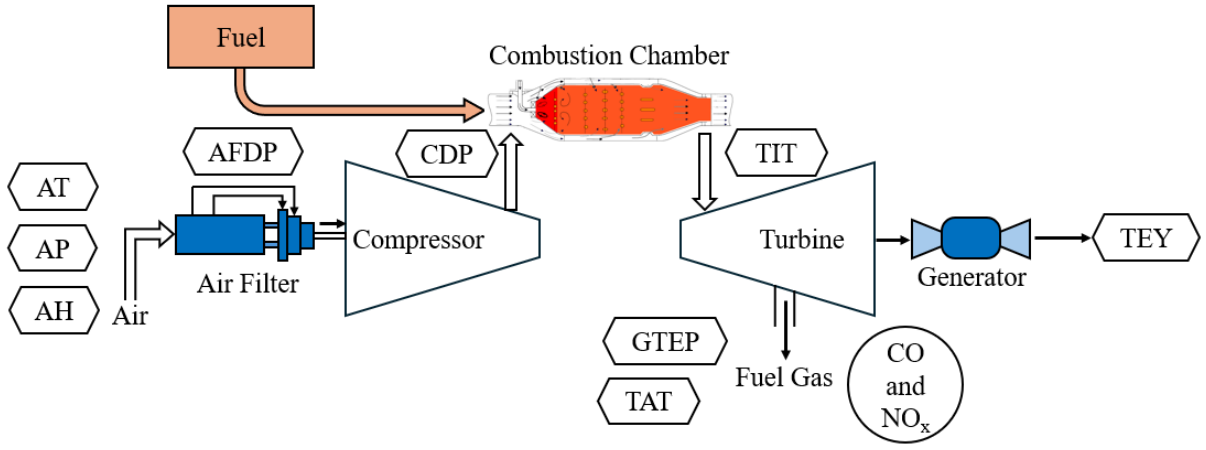


Figure 4. The schematic diagram illustrating the CCPP includes all input and output features

2.2. Data preparation

In any ML or data analysis project, data preparation is a critical step that significantly influences the accuracy and performance of the resulting models. Proper data preparation involves cleaning, transforming, and organizing the raw data into a format suitable for analysis. This process helps to ensure that the models are trained on high-quality data, which is essential for achieving reliable and meaningful predictions. It also involves handling missing values, removing outliers, and normalizing data, all of which contribute to the robustness of the analysis. In this study, careful attention has been paid to the data preparation phase to maximize the predictive performance of the models for CO and NOx emissions from a gas turbine. The dataset was carefully inspected, and no null or missing values were

found, ensuring the dataset remained complete and representative. To test the data's compliance with the normal distribution assumption, the Shapiro-Wilk test was applied.

Figure 5 presents box plots for the various input and output features used in this study, providing a visual summary of their distributions. The box plots illustrate the central tendency and variability of each variable, as well as the presence of any potential outliers. For instance, variables like AP and AH exhibit a relatively tight IQR, indicating low variability, whereas variables such as CO and NOx show a wider IQR, signifying higher variability and the presence of numerous outliers.

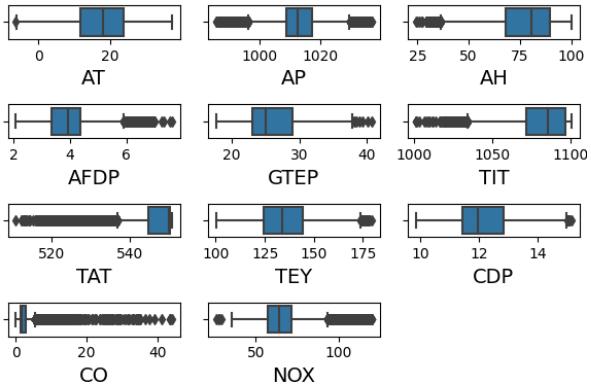


Figure 5. Box plots of input and output features showing distribution and potential outliers

The study focuses on the impact of different outlier treatment methods on the performance of the developed forecasting models. The outlier handling methods examined include Z-Score, IQR and MD. Each method offers a unique approach to identifying and handling outliers, which can significantly impact model accuracy and reliability.

- **Z-Score**

The Z-Score method identifies outliers based on the number of standard deviations a data point is from the mean. The Z-Score for a data point x is calculated using the formula in Equation 1.

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the mean of the data and σ is the standard deviation. Data points with Z-Scores greater than a specified threshold (commonly ± 3) are considered outliers. This method assumes that the data follows a normal distribution, and it is particularly useful for detecting extreme values in symmetric distributions. However, the Z-Score method is sensitive to the assumption of normality. If the data is not normally distributed, the Z-Score method may incorrectly identify outliers. Additionally, it is less effective for small datasets or datasets with high variability, as the mean and standard deviation can be heavily influenced by extreme values.

- **Interquartile Range**

IQR method identifies outliers based on the spread of the middle 50% of the data. The IQR is calculated as the difference between the third quartile $(Q3)$ and the first quartile $(Q1)$ given in Equation 2.

$$IQR = Q3 - Q1 \tag{2}$$

Outliers are typically defined as data points that fall below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$. This method is robust to non-normal distributions and is effective in handling skewed data. The IQR method is less sensitive to extreme values compared to the Z-Score method, but it may not be as effective for datasets with a small number of observations, as the quartiles may not accurately represent the data distribution. The choice of the multiplier (e.g., 1.5 or 3) can also affect the number of outliers detected, requiring careful tuning.

- **Mahalanobis Distance**

The MD method identifies outliers by considering the distance of a data point from the mean of the distribution, taking into account the correlations between variables. The MD for a data point x is given as in Equation 3.

$$D^2 = (x - \mu)^{T\,\Sigma^{-1}} \quad (x - \mu) \tag{3}$$

where μ is the mean vector of the data, and Σ is the covariance matrix. Data points with a MD exceeding a certain threshold (determined by the chi-square distribution with degrees of freedom equal to the number of variables) are considered outliers. This method is particularly effective for multivariate data and can identify outliers that may not be evident when considering variables individually. However, the MD method is sensitive to the distribution of the dataset. If the dataset is small or homogeneous, the MD method may incorrectly identify outliers, leading to overfitting (Caicedo et al. 2017). The method assumes that the data follows a multivariate normal distribution, and if this assumption is violated, the MD may not accurately identify outliers. In datasets with high dimensionality, the MD method can be computationally expensive and may struggle with the "curse of dimensionality," where the distance metric becomes less meaningful. For small and homogeneous datasets, the MD method may overfit the model by identifying extreme data points as outliers, which can lead to a model that performs well on the training data but poorly on new or unseen data. In

datasets with high variability or noise, the MD method may misinterpret the variance and flag some data points as outliers, which can weaken the generalization ability of the model.

2.3. Regression algorithms

In this study, RF, ET, LR, SVR, DT and KNN algorithms were used to predict CO and NOx emissions from GT. RF is an ensemble learning algorithm consisting of many decision trees. Each decision tree is trained on subsets of randomly selected features and data samples. This increases the generalization ability of the model and makes it more resistant to overfitting. Final predictions are usually made by averaging the predictions of these trees. The basic idea of RF is based on the idea that many different and random trees can come together to form a more powerful model (Biau, 2012). In this way, the errors within each tree compensate for each other and a better prediction can be made overall. The formula for RF for regression is given in Equation 4.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x) \tag{4}$$

Here, $\hat{y}$ is the predicted value, N is the total number of decision trees and $h_i(x)$ is the prediction of the i-th decision tree. With this formula, the final prediction is calculated by averaging the predictions of all trees.

ET is an ensemble learning algorithm similar to RF, but with certain differences. ET aims to increase diversity by generating decision trees in a more randomized way. When building decision trees, the best split point for each node is randomly selected. This allows the trees to be more diverse from each other, which helps the ensemble model to become more generalizable. It has been observed that ET can have faster training times compared to RF (Ahmad et al. 2018).

LR is a basic regression algorithm used to model the relationship between dependent and independent variables. This algorithm attempts to capture the linear relationship between the values of the independent variables in the dataset and the dependent variable (Maulud and Abdulazeez, 2020). The model determines the coefficients of the features in the dataset and a constant (cut-off point). These coefficients and constant represent the linear relationship that will best explain the observations in the dataset. LR is particularly effective when the dependent variable is continuous and there is a linear relationship between the variables. The LR formula is presented in Equation 5.

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{5}$$

When Equation 5 is analyzed, $\hat{y}$ represents the predicted value, β_0 represents the cut-off point, β_j represents the coefficient of the j-th independent variable and x_j represents the value of the j-th independent variable.

SVR is an adaptation of the Support Vector Machines (SVM) algorithm for regression analysis. While SVM was originally developed for classification problems, SVR modifies it to address regression tasks. SVR employs the concept of a hyperplane used in SVM for classifying data points, but in this case, the hyperplane is determined to ensure that the data points lie within a specified margin (Valkenborg et al. 2023). The basic idea of SVR is to fit the data points, i.e., the training data, around a hyperplane in such a way that the hyperplane is positioned to provide the widest margin possible. This margin is defined as the distance between the hyperplane and the closest data points on either side. However, unlike classification where data points are expected to be separated by the hyperplane, in regression, it is unrealistic to expect all data points to lie exactly on the hyperplane. Instead, SVR aims to find a balance between fitting the data points closely while maintaining a margin of tolerance, allowing for some deviation from the hyperplane within a defined threshold (Yu and Kim, 2012). Therefore, a tolerance (ε) margin is defined and the hyperplane tries to classify data points within this margin. Other data points may fall outside the tolerance margin. SVR determines the regression line by minimizing a cost function, which can also be controlled by hyperparameters "C" and "ε". The parameter "C" controls the model's resistance to overfitting, while the parameter "ε" determines the tolerance margin. The regression formula for SVR is presented in Equation 6.

$$\underset{\mathbf{w},\, b}{min}\ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\varepsilon_i + \varepsilon_i^*) \tag{6}$$

Here, $\mathbf{w}$ represents the weight vector, b the bias term, C the regularization parameter and $\varepsilon_i,\ \varepsilon_i^*$ the error terms. This formulation determines the hyperplane according to the maximum margin principle while minimizing the error terms for data points that lie within the specified tolerance margin.

DT is a regression analysis technique that uses features and outcomes from the dataset. It works based on decision trees and each tree is used to predict outcomes based on the values of features in the data set. The algorithm starts by creating a decision node for each feature in the dataset. These nodes test the feature values in the dataset and create smaller subsets by partitioning the data according to a specific rule. Each node tries to choose the best feature and threshold value to best partition the dataset. The tree-building process continues until the dataset is divided by a certain criterion (for example, until a certain depth or minimum number of samples is reached). As a result, each leaf node makes a

prediction, and the average or weighted average of these predictions is used as the tree's prediction based on the values of the features in the dataset (Quinlan, 1996). DT provides a flexible modeling method to capture complex relationships. In Equation 7, the basic estimation formula for DT is given.

$$\hat{y} = \frac{1}{N_t} \sum_{x_i \in R_t} y_i \quad (7)$$

In this formula, the predicted value $\hat{y}$ is calculated as the average of the actual values of the samples in a leaf node. During this calculation, the number of samples in the leaf node N_t, the leaf node R_t and the actual value y_i for each sample are used.

KNN is a fundamental classification and regression algorithm. In classification, the class of a data point is determined by the majority vote of its k nearest neighbors. In regression, the output value of a data point is predicted by taking the average of the values of its k nearest neighbors. KNN relies on the similarity between data points and is generally considered a simple and effective method. However, it can be sensitive to noise and high-dimensional data issues and is often computationally expensive because it requires comparing the target data point against the entire training dataset to make predictions (Song et al. 2017). Equation 8 defines the regression formula for the KNN algorithm.

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \quad (8)$$

In the formula, the estimate $\hat{y}$ is calculated as the average of the true values y_i of the k nearest neighbors. In this calculation, k neighbors (k) and the true value for each neighbor (y_i) are used.

2.4. Ensemble learning

In this study, the ensemble learning methods VR and SR are used. In these methods, GB, LightGBM and CatBoost algorithms are used as base learners, while XGBoost is chosen as the meta-learner to combine their predictions. While selecting the base learners and the meta-learner, the diversity, performance, and compatibility of the base learners were taken into account. Also, the overall impact of the meta-learner on the ensemble was considered. Ensemble methods aim to improve the overall performance of the model by combining the predictions of more than one base learner.

The VR combines the predictions of different base learners to generate the final prediction. The number of base learners is set to N and y^{ij} denotes the prediction of base learner i about sample j. In this case, the final prediction $y^{ensemble}$ of the VR is calculated as in Equation 9.

$$y^{ensemble} = \frac{1}{N} \sum_{i=1}^{N} y^{ij} \quad (9)$$

In the formula in Equation 9, $y^{ensemble}$ is defined as the final prediction and y^{ij} is defined as the prediction of the base learner i about the sample j.

A SR creates the final prediction by training a meta-learner on the predictions of base learners. The meta-learner is trained on a new dataset formed by the predictions of the base learners, and this meta-learner then takes the base learners' predictions as inputs to produce the final prediction (Divina et al. 2018). For example, if there are three different base learners, the predictions from these base learners form a new dataset. This dataset consists of the predictions of each base learner for each sample. Then, the meta-learner is trained on this new dataset along with the true values. The meta-learner uses the predictions of the base learners and the true values to make a more accurate prediction. The final prediction of the SR is the prediction made by the meta-learner. This method enhances the performance of the ensemble model by transforming the base learners' predictions into a structure that can model more complex relationships. The steps involved in the SR process are as follows:

1. Base learners' predictions: The predictions of the base learners are denoted as y^{ij}, where i represents the i-th base learner and j represents the index of the j-th sample.
2. Creating a new dataset: A new dataset is created using the predictions of the base learners. This new dataset is used to train the meta-learner. For each sample, the new dataset contains the predictions of all the base learners. Thus, the new representation of a sample's predictions can be denoted as $X^j = [y_{1j}, y_{2j}, \ldots, y_{nj}]$.
3. Meta-learner prediction: The meta-learner makes predictions using this new dataset and the true values. The prediction of the meta-learner is denoted as $\hat{y}_{\text{meta}}$.
4. Final prediction: The final prediction of the SR is achieved using the meta-learner's prediction, which is $\hat{y}_{\text{meta}}$.

Once this process is formalized, the final estimate of the SR is calculated as in Equation 10.

$$y^{ensemble} = \hat{y}_{\text{meta}} \quad (10)$$

2.5. Validation method

Many ML models rely on splitting datasets into training and testing to measure their performance. However, this method can lose reliability in terms of

accuracy as the size of the dataset used to test a small portion of the dataset decreases. In this study, the Stratified K-Fold Cross-Validation (SKCV) method is used to evaluate the generalization ability of the model. In the SKCV method, the dataset is divided into k equal parts. Each part is respectively selected as the test set, while the remaining k-1 parts are used as the training set. This process is repeated k times, and the model is trained and tested in each iteration. The overall performance of the model is evaluated by averaging the obtained performance metrics. This method can be more robust to noisy datasets and scatter of data points, which can better reflect the performance of the model on real-world data (Prusty et al. 2022). The basic formula of the SKCV method is given in Equation 11.

$$\text{SKCV} = \frac{1}{k}\sum_{i=1}^{k} L\left(y_i, \hat{y}_{-i}\right) \tag{11}$$

In the formula, k represents the number of layers ($k = 10$ in this study), y_i represents the true values in each layer, and $\hat{y}_{-i}$ represents the predicted values of the model trained outside that layer. The function L is used to measure the error between the true and predicted values. Usually, the mean squared error is used for regression problems or zero-one loss for classification problems.

2.6. Performance metrics

Evaluating research and validating its results requires the use of specific measurements and metrics. These metrics are used to evaluate the success of the model or algorithm and to understand its performance. This study examines performance metrics commonly used in regression problems such as R^2, RMSE and MSE. R^2 is a measure of a model's ability to explain variance in observed values. Its formula is given in Equation 12.

$$\text{R}^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{12}$$

In the formula in Equation 12, SS_{res} represents the residual sum of squares and SS_{tot} represents the total sum of squares. SS_{res} is defined by the formula in Equation 13 and SS_{tot} is defined by the formula in Equation 14.

$$SS_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{14}$$

In the formulas, y_i represents the actual values, $\hat{y}_i$ the predicted values and $\bar{y}$ the mean of the observed values. R^2 takes values between 0 and 1 and the higher it is, the better the model explains the observed data.

RMSE is a metric that measures how far the model's predictions are from the true values. By taking the square root of the prediction errors, it shows the magnitude of the errors on average. Its formula is expressed in Equation 15.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{15}$$

In the RMSE formula, y_i represents the actual values, $\hat{y}_i$ represents the values predicted by the model, and n represents the total number of data points. The steps to calculate the RMSE are as follows:

1. For each observation, the difference (error) between the predicted value and the actual value is calculated: $e_i = y_i - \hat{y}_i$
2. The squares of these errors are taken: ${e_i}^2 = (y_i - \hat{y}_i)^2$
3. The mean of all the squared errors is computed: $\frac{1}{n}\sum_{i=1}^{n} {e_i}^2$
4. The square root of this mean is taken to find the RMSE.

MSE is the square of RMSE and represents the mean squared error of the model. The formula for MSE is given in Equation 16.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{16}$$

In the formula, y_i is the actual values, $\hat{y}_i$ is the values predicted by the model, and n is the total number of data points.

2.7. Model setups

For the development of the models, the data set was divided into two parts, 80% training and 20% testing. The randomness factor was set to 42 in all algorithms. The best hyperparameter settings for the models were determined using Bayesian Optimization. This method aims to discover the optimal parameter combinations using knowledge in the search domain. Bayesian Optimization is optimized to improve the performance of the model, working particularly effectively in complex and high-dimensional hyperparameter search spaces (Yang and Shami, 2020). This method offers a more efficient alternative to the classical Grid Search and Random Search techniques.

Table 2 lists the best hyperparameter settings determined using Bayesian Optimization for the ML models used in the study. The RF model was configured with 100 estimators, a maximum depth of 10, and

minimum samples of 2 to split a node and 1 to be at a leaf node, balancing complexity and generalization to prevent overfitting. The ET model, which randomizes tree generation, used 150 estimators, a maximum depth of 12, and minimum samples of 4 to split and 2 at a leaf, enhancing diversity and capturing complex patterns. For LR, fit_intercept was set to True to calculate the intercept, while normalize was set to False, as the data was preprocessed, ensuring the model captures linear relationships without unnecessary normalization. The SVR model used C = 1.0, ε = 0.1, and an 'rbf' kernel to handle non-linear relationships, with C controlling complexity and ε defining the error tolerance margin. The DT model was set with a maximum depth of 12 and minimum samples of 4 to split and 3 at a leaf, controlling tree growth to avoid overfitting. The KNN model used 10 neighbors, a 'distance' weighting function, and the 'ball_tree' algorithm to efficiently handle high-dimensional data, prioritizing closer neighbors for predictions. The VR combined Gradient Boosting, LightGBM, and CatBoost with equal weights, leveraging multiple algorithms for robust predictions. Finally, the SR used the same base learners as VR, with XGBoost as the meta-learner, creating a hierarchical model that captures complex relationships and achieves higher predictive accuracy by combining the strengths of multiple algorithms.

Table 2. Hyperparameter settings for ML models

Model	Hyperparameter	Settings
RF	n_estimators, max_depth, min_samples_split, min_samples_leaf	100, 10, 2, 1
ET	n_estimators, max_depth, min_samples_split, min_samples_leaf	150, 12, 4, 2
LR	fit_intercept, normalize	True, False
SVR	C, epsilon, kernel	1.0, 0.1, 'rbf'
DT	max_depth, min_samples_split, min_samples_leaf	12, 4, 3
KNN	n_neighbors, weights, algorithm	10, 'distance', 'ball_tree'
VR	estimators, weights	[('gb', GradientBoostingRegressor()), ('lgbm', LGBMRegressor()), ('cat', CatBoostRegressor())], [1, 1, 1]
SR	estimators, final_estimator	[('gb', GradientBoostingRegressor()), ('lgbm', LGBMRegressor()), ('cat', CatBoostRegressor())], XGBRegressor()

In the study, the hyperparameter settings of Z-score, IQR and MD methods, which are used to detect outliers and solve this problem, were determined, and analyzed. For the Z-score method, the "threshold" hyperparameter, which determines how many standard deviations away a data is from the standard deviation, is set to 3.0. This setting means that data that are more than 3 standard deviations away will be considered abnormal. For the IQR method, the "k" hyperparameter, which determines the distance between the upper and lower quartiles, is set to 1.5. For the MD method, the "threshold" hyperparameter, which determines whether the data are anomalous according to the MD distribution, is set to 95.0 percentile. This setting means that data with MD greater than a certain percentile will be considered outliers. These hyperparameter settings had a significant impact on model performance in anomaly detection and data preprocessing. The hyperparameter settings are presented in Table 3.

Table 3. Hyperparameter settings for outliers' treatment methods

Method	Hyperparameter	Settings
Z-score	threshold	3.0
IQR	k	1.5
MD	threshold	95.0 percentile

Python programming language was used for data analysis and model testing. In the data analysis process, pandas and NumPy libraries were used for data processing and manipulation. For model building and testing, the scikit-learn library was preferred. This library provides various ML algorithms and model evaluation tools. All processes were carried out in the Jupyter Notebook development environment, where code, text and visuals are presented together. A PC running on a Ryzen 7800x3D processor with a processor speed of 4.2 GHz was used for training the models. In addition, the PC has NVIDIA 4070 Ti GPU and 32 gigabyte 6000 MHz DDR5 RAM. Windows 11 was used as the operating system. The overview of the CO and NOx emission prediction system realized in the study is given in Figure 6.

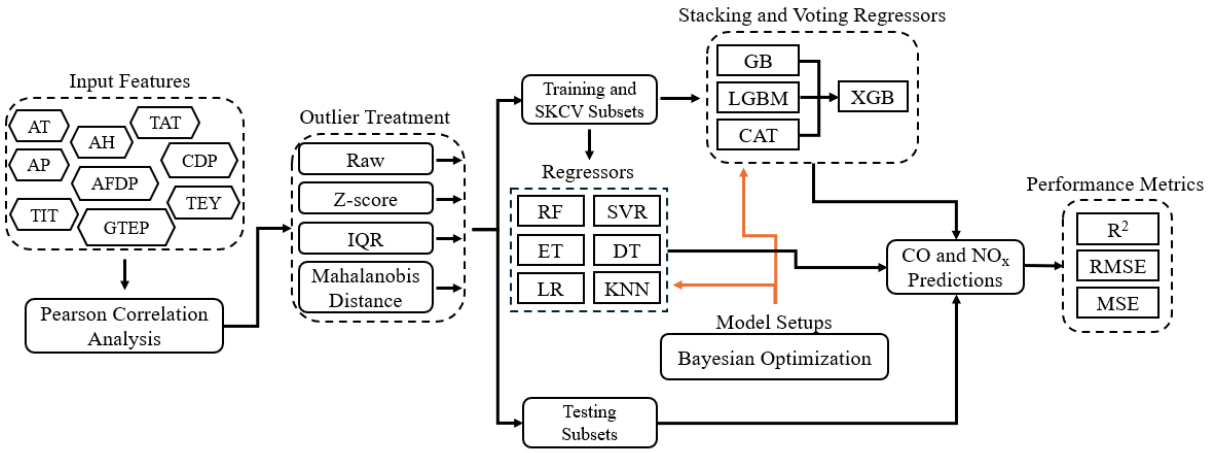


Figure 6. Overview of the CO and NOx emissions prediction system

3. Experimental Study and Results

In this section, the performance results of the regression models are presented and discussed. Table 4 shows the NOx emission prediction performance of the regression models and Table 5 shows the CO emission prediction performance of the models. The results are presented using three different metrics (R^2, RMSE, MSE) and four different outliers treatment methods (Raw, Z-Score, IQR, MD). R^2 indicates the explanatory power of the model, while RMSE and MSE indicate the error rates. The values of the best performing models for each outlier's treatment method are expressed in bold font. Table 4 shows that when the outliers in the dataset are treated with the MD method, ML models and ensemble methods reach the highest performance values in NOx emission prediction. In fact, the Stacking Regressor (SR) method reached the highest determination coefficient with $R^2 = 0.9974$ (RMSE = 0.5906, MSE = 0.3488). The predictions made by the SR model for NOx emissions closely match the observed (actual) NOX values, suggesting a strong agreement between the model's predictions and the real-world data. Figure 7 presents the R^2 scores of regression models for predicting NOx emissions, providing a clear comparison of their performance across different outlier treatment methods.

Upon reviewing the Table 4, it appears that the ET and DT models achieved an $R^2 = 1.0$, suggesting perfect performance compared to the SR. However, the respective RMSE values of 9.4325 and 1.5509 indicate that these models exhibit significantly larger prediction errors than expected. This discrepancy suggests that the models may have overfitted the training data, demonstrating excellent fit to the training set while lacking the ability to generalize to new data. Additionally, the MD method's sensitivity to data distribution means that outlier values in the dataset could negatively impact model performance, contributing to these observed errors. There is another very important point to be considered here. Figure 9 displays the comparison between predicted and actual NOx and CO emissions using the MD method and the SR model. The left panel plots the predicted NOx values against the actual NOx values. When the NOx predictions in Figure 9 are examined, it is seen that the points are ideally concentrated on the y = x line. This shows that the model predicts NOx values almost perfectly. These near-perfect predictions for NOx suggest that the model might be overfitting. A model that fits the training data exceptionally well may not maintain this performance when faced with new or noisier data. Analyzing the graphs and performance indicators reveals that the models trained on the dataset created using the MD method exhibit overfitting, indicating that the MD method does not yield accurate results for NOx prediction in the context of this study.

When Table 4 is further analyzed, the model created with SR in the dataset processed for outliers using the IQR method achieved the highest coefficient of determination ($R^2 = 0.9194$) and the lowest error values (RMSE = 2.7669, MSE = 7.6562). When the scatter plot in Figure 10 is examined, it is understood that the performance values obtained by the SR method seem to be suitable for real world data and the possibility of overfitting the model is low. In Figure 11, Hydrographs are given for each model trained on the dataset created with the IQR method and the prediction performances of the models are revealed. The success of the IQR method and the model built with SR in predicting NOx emissions at GT compared to other methods and models is clearly seen in the graphs. In addition to all this, it is found that the models trained on the dataset where outliers are processed with the Z-score method perform lower than the models trained on raw data.

Table 4. NOx emission prediction performance of regression models with outliers' treatment methods

	Raw			Z-Score		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.8765	4.0443	16.3566	0.8716	4.1235	17.0035
ET	0.8908	3.8037	14.4683	0.8875	3.8607	14.9055
LR	0.4946	8.1832	66.9648	0.4911	8.2112	67.4251
SVR	0.7574	5.6698	32.1475	0.0739	11.0776	122.714
DT	0.7367	5.9059	34.8805	0.7383	5.8882	34.6718
KNN	0.8549	4.3846	19.2255	0.7893	5.2839	27.9197
VR	0.8575	4.3447	18.8766	0.8159	4.9384	24.3883
SR	**0.8942**	**3.7430**	**14.0101**	**0.8889**	**3.8353**	**14.7100**
	IQR			MD		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.8992	3.0947	9.5776	0.9822	1.5602	2.4343
ET	0.9162	2.8218	7.9629	1.0	9.4325	8.8972
LR	0.7259	5.1046	26.0579	0.5233	8.0913	65.4694
SVR	0.0748	9.3789	87.9644	0.0788	11.2477	126.5129
DT	0.7672	4.7047	22.1351	1.0	1.5509	2.4052
KNN	0.8191	4.1473	17.2002	0.8723	4.1864	17.5262
VR	0.8511	3.7624	14.1559	0.8836	3.9972	15.9782
SR	**0.9194**	**2.7669**	**7.6562**	**0.9974**	**0.5906**	**0.3488**

Table 5 shows that ML models and ensemble methods achieve the highest performance values in CO emission prediction when outliers in the dataset are processed using the MD method. Similarly, examining the CO predictions in Figure 9, it is observed that the points are concentrated around the y = x line, but there is a more pronounced scatter and deviations. Although the CO predictions are more modest, there is a high probability that the model may have overfitted. Therefore, when Table 5 is further analyzed using other methods, it is seen that the model created with SR reaches the highest coefficient of determination (R^2 = 0.8556) and the lowest error values (RMSE = 0.4619, MSE = 0.2133) in the dataset processed for outliers using the IQR method. The scatter plot in Figure 10 also supports these results obtained by SR with the IQR method. In addition, the Hydrographs given for each model in Figure 12 reveal the high success of the IQR method and the model built with SR in predicting CO emissions in GT compared to other methods and models. Also, similar to the results for NOx emission prediction, the CO prediction performance of the models trained on raw data is lower than the models trained on the dataset generated by the IQR method, but higher than the scenario using the Z-score method. Figure 8 displays the R^2 scores of regression models for CO emission prediction, offering a concise comparison of their performance across various outlier treatment methods.

Table 5. CO emission prediction performance of regression models with outliers' treatment methods

	Raw			Z-Score		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.7649	1.1306	1.2783	0.7509	1.1637	1.3543
ET	0.7990	1.0453	1.0926	**0.7682**	**1.1225**	**1.2602**
LR	0.5566	1.5526	2.4107	0.5511	1.5621	2.4404
SVR	0.6755	1.3282	1.7641	0.4405	1.7441	3.0419
DT	0.5081	1.6354	2.6747	0.4055	1.7978	3.2321
KNN	0.7630	1.1350	1.2883	0.6626	1.3543	1.8343
VR	0.7631	1.1349	1.2880	0.6220	1.4335	2.0550
SR	**0.8026**	**1.0359**	**1.0731**	0.6705	1.3384	1.7914
	IQR			MD		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.8426	0.4822	0.2325	0.9682	0.4002	0.1602
ET	0.8542	0.4640	0.2153	1.0	4.4572	1.9867
LR	0.6150	0.7542	0.5689	0.5651	1.4803	2.1915
SVR	0.5491	0.8163	0.6664	0.4658	1.6408	2.6922
DT	0.6658	0.7027	0.4939	1.0	3.1728	1.0066
KNN	0.7752	0.5764	0.3322	0.7983	1.0080	1.0162
VR	0.8060	0.5353	0.2866	0.9286	0.5998	0.3597
SR	**0.8556**	**0.4619**	**0.2133**	**0.9974**	**0.1122**	**0.0126**

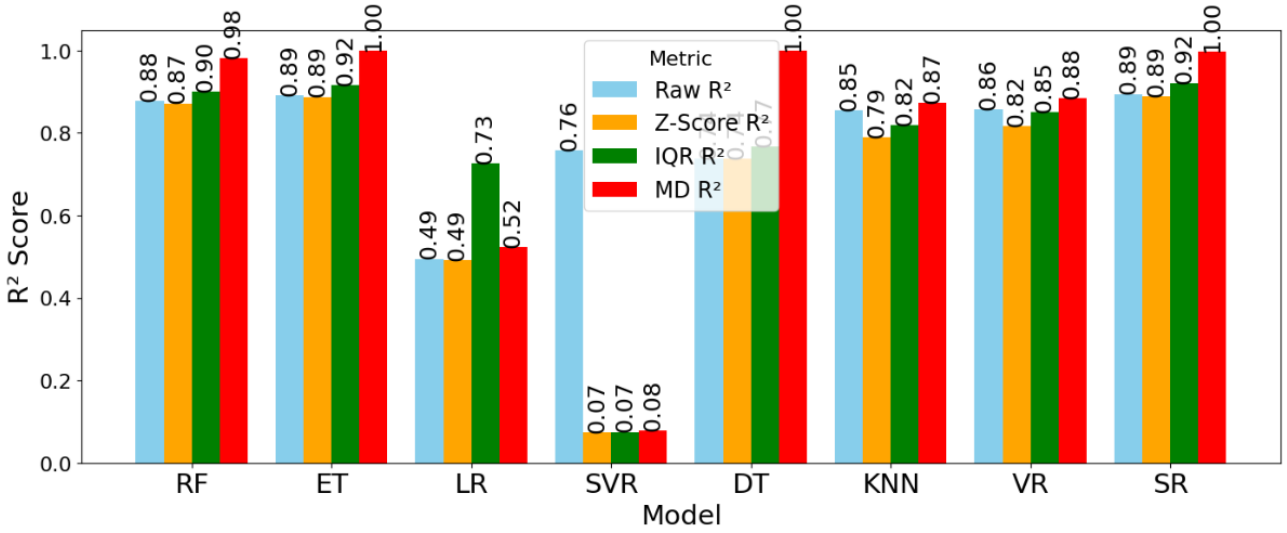


Figure 7. Comparison of R^2 scores for NOx emission prediction models with different outlier treatment methods

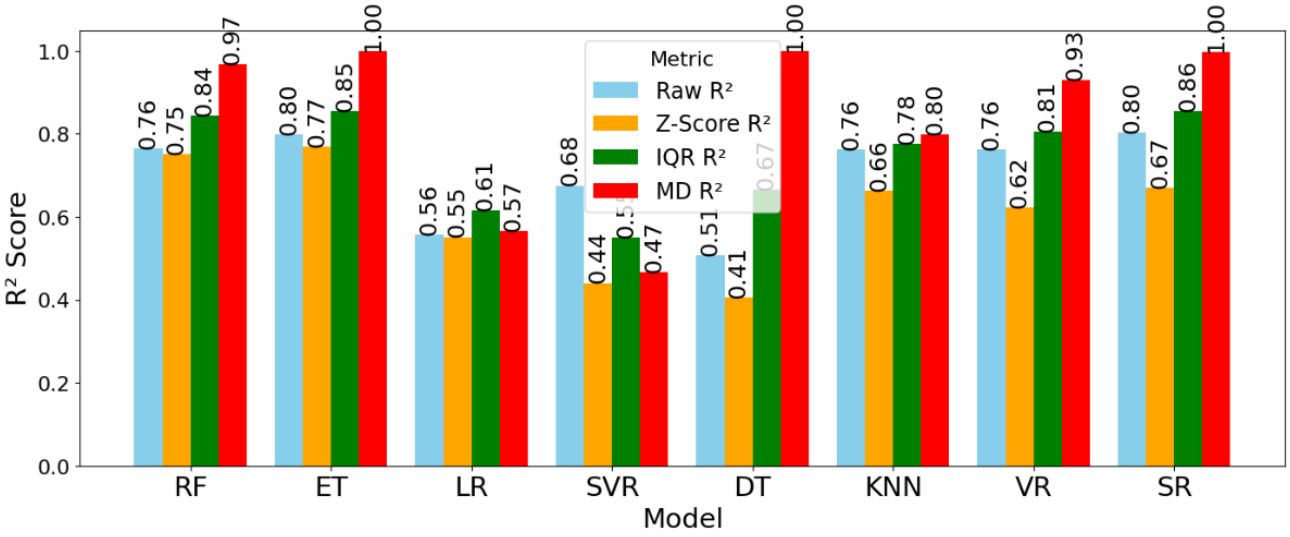


Figure 8. Comparison of R^2 scores for CO emission prediction models with different outlier treatment methods

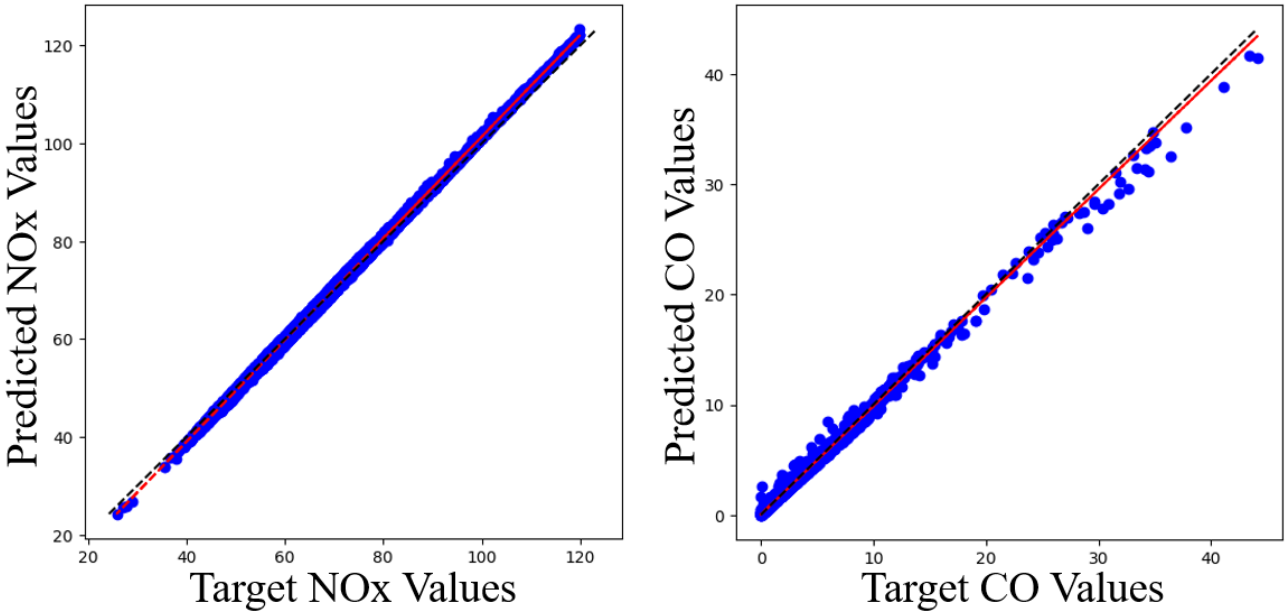


Figure 9. Scatter plots generated on the MD method for the SR prediction model

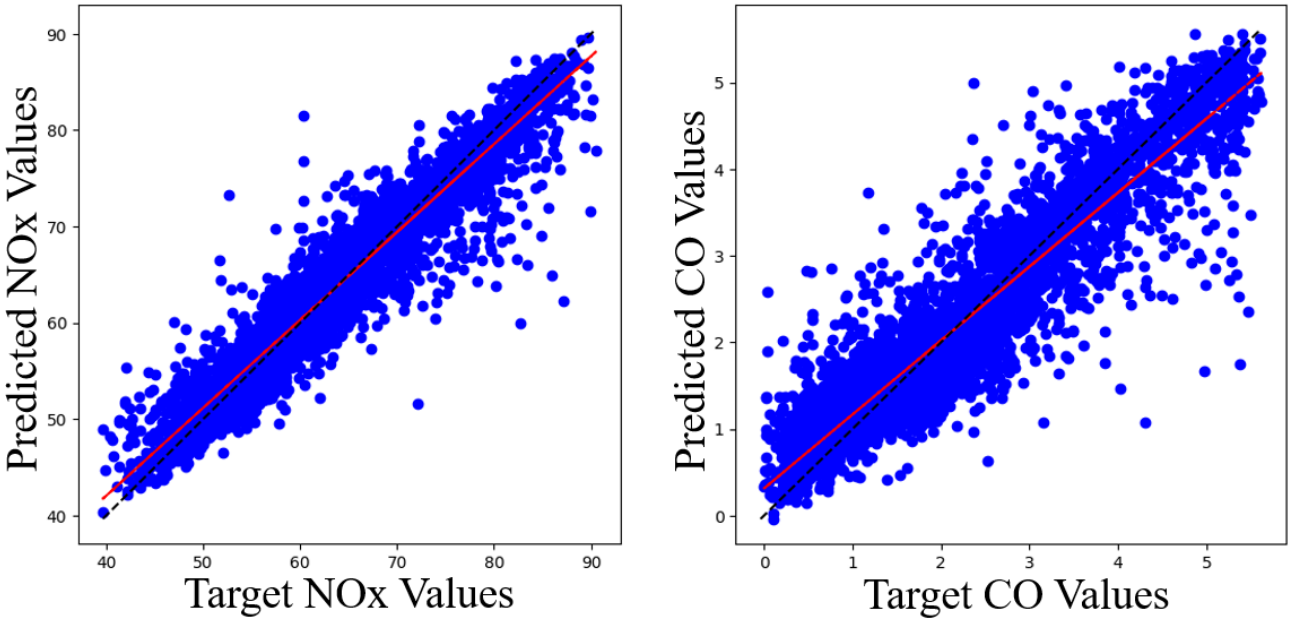


Figure 10. Scatter plots generated on the IQR method for the SR prediction model

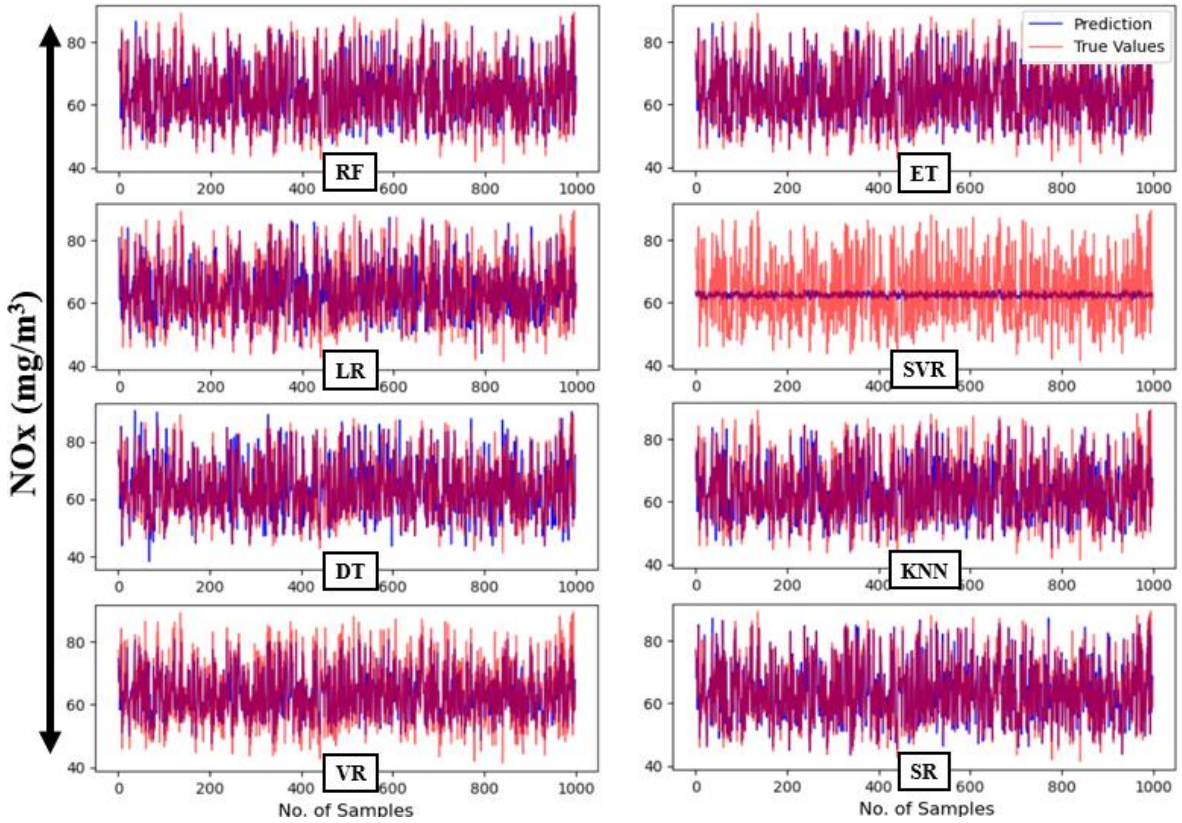


Figure 11. Hydrographs for all designed ML models for NOx in the IQR-processed dataset

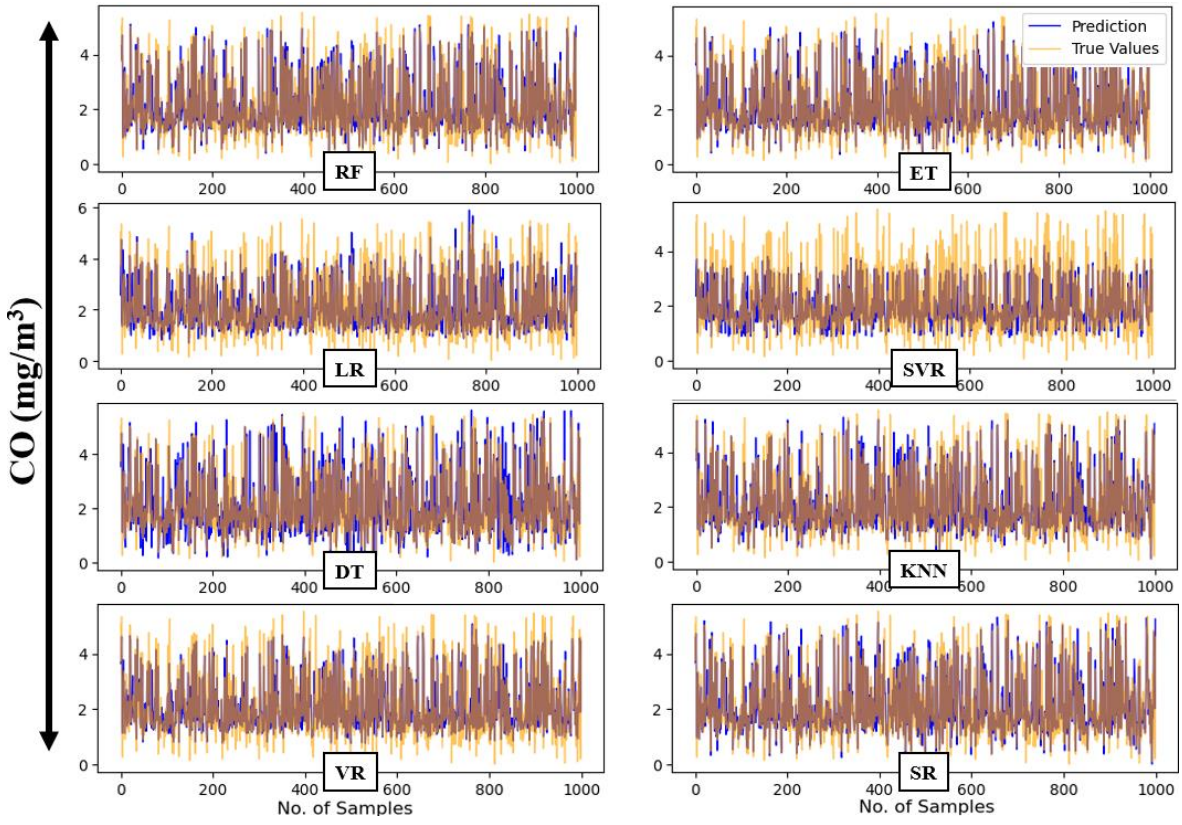


Figure 12. Hydrographs for all designed ML models for CO in the IQR-processed dataset

Figure 13 shows the results of the sensitivity analysis for CO and NOx emissions. The graph reflects the impact of each feature on the predictive performance of the model. Blue bars represent CO sensitivity and red bars represent NOx sensitivity. The sensitivity analysis for each feature is performed as follows:

1. Baseline score: First, the error of the model's predictions (MSE) with the available features is calculated (Equation 16).
2. Perturbed scores: For each feature, the error of the model's predictions is recalculated by randomly permuting the values of this feature. This process is repeated 10 times and the MSE values obtained each time are recorded (Equation 17). Here $\hat{y}_j^{(m)}$, denotes the values predicted by the model after the m-th permutation and m represents the number of permutations (10).

$$\text{Perturbed\_Score}^{(m)} = \frac{1}{n}\sum_{j=1}^{n}\left(y_i - \hat{y}_j^{(m)}\right)^2 \quad (17)$$

3. Sensitivity value: The sensitivity value for each feature is calculated by subtracting the average error from the permuted scores from the baseline score (Equation 18).

$$\text{Sensitivity}(i) = \left(\frac{1}{10}\sum_{m=1}^{10}\text{Perturbed\_Score}^{(m)}\right) - \text{Baseline\_Score} \quad (18)$$

When examining Figure 13, it is found that AT has a significant impact on CO emissions, with a sensitivity value of 0.6234. It also shows a notable effect on NOx emissions (2.5019). Higher temperatures generally enhance combustion, promoting NOx formation while potentially reducing CO emissions. AP has a low-level effect on CO and a moderate impact on NOx emissions. The impact of AP on CO is measured at 0.0587, while its impact on NOx is 0.7457. TIT has the highest impact on CO emissions, with a sensitivity value of 10.0492. Its effect on NOx is negative, with a sensitivity value of -0.6029. The high sensitivity of TIT on CO emissions indicates that TIT plays a critical role in combustion efficiency and, consequently, CO formation. The negative effect on NOx emissions may indicate that high inlet temperatures can reduce NOx formation. TAT has a significant impact on CO emissions, with a sensitivity value of 6.819. Its effect on NOx is quite low and positive, measured at 0.0694. The high impact of TAT on CO emissions suggests that the turbine outlet temperature can affect the composition of post-combustion gases. Overall, it is understood that TIT and TAT parameters are particularly critical in predicting CO emissions, while AT has significant effects on NOx emissions.

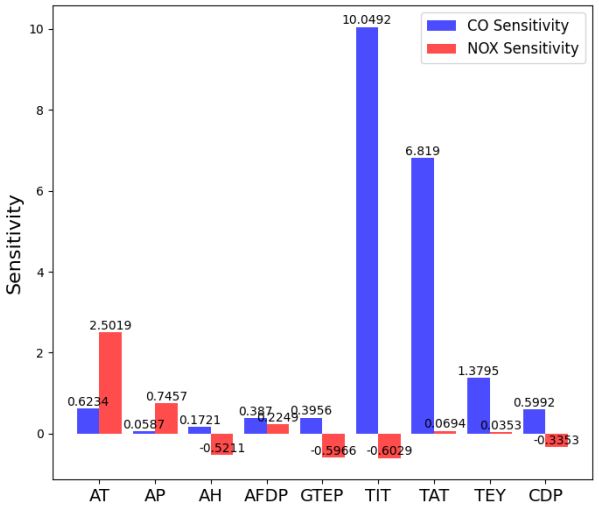


Figure 13. Parameter sensitivity analysis results for CO and NOx emissions

The error deviation graphs presented in Figure 14 show the differences between the predicted and actual CO and NOx emission rates. Table 6 provides the minimum and maximum error values obtained for each predictive ML model created. For the SR model trained on the dataset processed with the IQR method for outlier treatment, the minimum and maximum error values for NOx are -21.5874 and 24.5953, respectively, while for CO they are -2.6273 and 3.5575, respectively. Compared to other models, it can be concluded that the SR model has the lowest error deviation. This indicates that the SR model is more successful in predicting CO and NOx emissions and that these predictions are closer to the actual values.

Table 6. Error deviation for all designed predictive models in the IQR-processed dataset

Error deviation	RF	ET	LR	SVR	DT	KNN	VR	SR
NOx								
Min deviation	-27.1536	-21.9554	-27.3576	-21.9011	-34.0320	-25.9326	-18.3893	-21.5874
Max deviation	30.1443	24.4523	33.1878	28.0568	37.6850	31.1432	26.4236	24.5953
CO								
Min deviation	-2.7295	-2.6710	-3.3075	-2.8513	-4.7361	-2.9963	-2.9103	-2.6273
Max deviation	3.7549	3.6115	4.0626	3.8812	4.6908	3.5781	3.4712	3.5575

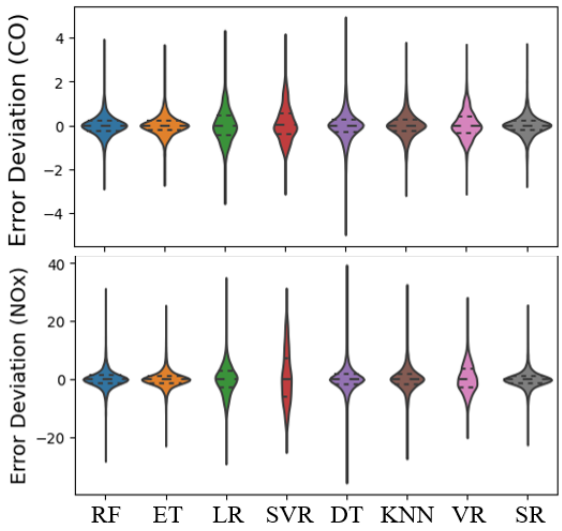


Figure 14. Error deviation comparison of all predictive models for NOx and CO in the IQR-processed dataset

4. Discussion

This study aims to develop predictive models for predicting CO and NOx emissions from GT using various ML algorithms and compare their performance with VR and SR methods. A comprehensive evaluation of these models is conducted to provide insights into their practical applicability in industrial settings. The findings show the importance of hyperparameter tuning and outlier processing in improving the prediction accuracy of these models.

Tree-based methods using a combination of models provided better performance compared to other regression models for both CO and NOx predictions. SR achieved the highest determination coefficient values when outliers in the dataset were treated using the MD method. However, despite the high R^2 values, high RMSE values indicate the possibility of overfitting, mainly in NOx predictions. While the MD method effectively detects outliers, it can lead to overfitting if the dataset does not follow a normal distribution. Overfitting occurs when the model performs well on training data but poorly on new data (Karthikeyan et al. 2023). Furthermore, the risk of overfitting associated with the MD method, as observed in this study, has also been noted in prior research. For instance, Ghorbani (2019) highlighted that the MD method, while effective for multivariate outlier detection, can be overly sensitive to data distribution and may lead to overfitting in small or homogeneous datasets. This corroborates our findings, where the MD method achieved high R^2 values but showed poor generalization in real-world scenarios.

In contrast, the IQR method's robustness and consistency make it a more suitable choice for practical applications in emission monitoring and control. In addition, the MD method considers the distance of each data point in the data set to other data points in multiple dimensions (Leys et al. 2018). This method can identify extreme data points as outliers and remove them from the model or cause the model to overfit these points. Another disadvantage of the MD method is that it is very sensitive to the distribution of the dataset (Todeschini et al. 2013). If there is heterogeneity or too much noise in the dataset, this method may detect false positive outliers. In this case, the model may focus on false positive outliers instead of true outliers, which weakens the generalization ability of the model. Furthermore, if there is too much variance in the dataset, the MD method may misinterpret this variance and flag some data points as outliers (Wu et al. 1997). This can cause the model to be trained on an incorrect subset of data and lead to overfitting.

On the other hand, the IQR method for outlier treatment produced more robust models with lower error rates. The SR model trained on the IQR-processed dataset achieved the highest determination coefficient ($R^2 = 0.9194$) and the lowest RMSE and MSE values for NOx predictions. The scatter plots in Figure 10 confirm the model's performance close to the true values, indicating a low risk of overfitting. This aligns with previous research by Yaro et al. (2023), who demonstrated that the IQR method is highly effective in reducing the impact of outliers in datasets with non-normal distributions, leading to more robust and generalizable models. Similarly, Mishra et al. (2019) emphasized that the IQR method outperforms Z-Score and MD in scenarios where data variability is high, as it is less sensitive to outlier values and provides a balanced approach to outlier detection. These studies collectively support the conclusion that the IQR method is a reliable choice for emission prediction tasks, especially in datasets with complex operational variability, such as those from gas turbines.

Similar patterns were observed for CO predictions. Although the MD method initially seemed to give the best results, further analysis revealed significant spread and biases, indicating the possibility of overfitting. The SR model trained on the IQR processed dataset outperformed the other models, again achieving the highest R^2 (0.8556) and the lowest RMSE and MSE values. This observation supports that the IQR method offers a balanced approach to dealing with outliers and ensures the robustness and accuracy of the model.

The worse performance of models trained on raw data compared to models trained on the processed dataset emphasizes the necessity of outlier handling methods in improving the reliability of emission prediction models. This finding is in line with previous research (Osborne and Overbay, 2019), which indicates that untreated outliers can significantly skew model training and prediction results. However, in the dataset where outliers were treated with the Z-score method, the performance of the models decreased compared to the raw dataset. The Z-score method is a common method for detecting and treating outliers, but it is based on the assumption of a normal distribution. If the dataset is not normally distributed or is multidimensional, this method may detect false positive outliers and over-smooth the

variance in the dataset (Mare et al. 2017). This reduces the model's ability to capture true patterns and leads to performance degradation. The degradation of the performance of the models in the dataset processed with the Z-score method compared to the raw dataset indicates that this method is not always appropriate and should be applied carefully according to the characteristics of the dataset.

In the study, sensitivity analysis was performed to improve the performance of emission prediction models and to identify the parameters that have the most impact on emissions. The findings show that turbine inlet and outlet temperatures (TIT and TAT) have a significant effect on CO emissions. TIT has the highest impact on CO emissions, indicating that TIT plays a critical role on combustion efficiency and hence CO formation. The negative effect on NOx emissions indicates that higher inlet temperatures can reduce NOx formation. Similarly, AT has a significant effect on NOx emissions, but less on CO emissions. This finding suggests that higher temperatures can reduce CO emissions while promoting NOx formation.

In Table 7, the results obtained by using the SR model proposed in this study and the dataset processed with the IQR method are compared with other studies in literature. The table shows that the model proposed in this study outperforms other studies in the literature in the prediction of NOx and CO emissions. The RMSE value of the proposed model in NOx prediction is 2.76 and the RMSE value in CO prediction is 0.46. These error values are significantly lower than the results of other models in the literature. In addition, higher R^2 values were obtained in both NOx and CO emissions prediction compared to other models in the literature. Possible reasons for this high performance include the effective treatment of outliers with the IQR method, which reduces the noise in the dataset. The IQR method provides a robust approach to handling non-normally distributed data by eliminating extreme values without over-penalizing potential influential data points. Moreover, the structure of the SR model and hyperparameter optimization are other factors that positively affect the prediction performance. Although the MD method initially appeared to provide competitive results in terms of R^2 values, its over-performance can be attributed to its sensitivity to multivariate relationships and its assumption of normal data distribution. However, this sensitivity may lead to overfitting, as the MD method can classify influential but valid data points as outliers, thereby reducing model generalization. This explains why, despite high R^2 values, the RMSE values remained relatively high, indicating possible model overfitting and reduced performance on new data.

Table 7. Comparison of NOx and CO emission prediction models from the literature with the proposed SR

Model and Reference	NOx			CO		
	RMSE	MSE	R^2	RMSE	MSE	R^2
ANN with feature normalization (Nino-Adan et al. 2021)	7.06	-	0.57	-	-	0.43
Symbolic regression (Kochueva and Nikolskii, 2021)	-	-	0.83	-	-	0.89
KNN (Rezazadeh, 2021)	-	-	0.89	-	-	-
DFR (Coelho et al. 2024)	5.54	30.69	-	1.35	1.84	-
KNN (Wood, 2023)	5.12	-	-	-	-	-
ANFIS (Dirik, 2022)	4.98	24.8	-	-	-	-
SR (Pachauri, 2024)	3.83	14.70	0.87	0.61	0.37	0.77
This study (SR with IQR Treatment)	2.76	7.65	0.92	0.46	0.21	0.85

5. Conclusion

This study aims to evaluate the performance of various ML algorithms for predicting CO and NOx emissions from GT. The algorithms used in the study include RF, ET, LR, SVR, DT, KNN, as well as VR and SR methods. In the ensemble methods, GB, LightGBM and CatBoost algorithms were used as base learners and XGBoost was determined as the meta-learner.

The study examined the effects of processing the outliers in the dataset with various methods (Z-Score, IQR, MD) on model performance. The findings show that the MD method provides high performance of the models, especially in NOx emission prediction, but it also brings the risk of overfitting. Although the SR model provides high R^2 and low error values in NOx predictions, scatter plots and hydrographs reveal that the data processed with the MD method do not show a distribution suitable for real world data. Due to the sensitivity of the MD method to outlier values, some models tend to overfit, indicating that caution should be exercised in the use of this method.

The treatment of outliers with the IQR method provided more balanced and generalizable results in CO and NOx emission predictions. In the dataset processed with the IQR method, the SR model achieved R^2 values of 0.9194 in NOx emission predictions and 0.8556 in CO emission predictions and showed low error rates in other metrics. Scatter plots and hydrographs also showed a consistent distribution. These results show that the IQR method is an effective approach to deal with outliers and has an impact on improving the performance of ML models. In the dataset processed with the Z-Score method, the performance of the models was lower compared to the models trained with raw data. This finding suggests that the Z-Score method may be less effective in identifying and processing outliers,

especially in this dataset, and that each outlier processing method may yield different results depending on the characteristics of the dataset and the problem definition.

The results show that the methods used to handle outliers have a significant impact on the performance of ML models and that the right choice of method can improve model accuracy and generalizability. While the IQR method gives balanced results in predicting CO and NOx emissions, the MD method can provide high performance in some cases, although it increases the risk of overfitting. ML algorithms and outlier treatment methods need to be carefully selected for gas turbine emission prediction.

The sensitivity analysis highlighted the significant impact of TIT, TAT, and AT on CO and NOx emissions in the prediction models. In particular, the high impact of TIT on CO emissions shows how important a role TIT plays on combustion efficiency and CO formation. This highlights the need for careful control of TIT to optimize combustion processes and reduce CO emissions. Furthermore, the negative effect of TIT on NOx emissions indicates that high inlet temperatures can reduce NOx formation. This finding suggests that high temperature processes in power generation have the potential to control NOx emissions. The significant effect of AT on NOx emissions suggests that environmental temperature conditions should also be considered in emission control strategies.

Based on the results, some recommendations can be made to improve gas turbine emission predictions and develop environmental management strategies in the energy sector. First of all, the quality of sensors and data collection systems in power plants should be improved and data accuracy should be ensured through regular calibrations, as accurate detection and processing of outliers directly affects model performance. The use of integrated model approaches should be encouraged as combinations of different ML algorithms can provide more balanced and accurate results in emission predictions. In addition, it is understood that temperature parameters are critical for improving the accuracy of emission prediction models and obtaining reliable predictions. To improve emissions management in the power sector, turbine and environmental temperatures need to be optimized. These approaches will enable more effective implementation of emission reduction strategies and contribute to reducing environmental impacts. To increase the generalizability of the results, future studies should test these methods on different datasets or systems, ensuring their applicability across various operational contexts and conditions.

The findings of this study hold significant potential to support sustainable energy policies and environmental management practices. For instance, the high accuracy of the SR model can help Turkish industries comply with stringent emission regulations, such as the EU Industrial Emissions Directive, while aligning with global climate goals like the Paris Agreement. By enabling precise real-time emission monitoring, these models can empower policymakers to design data-driven regulations and incentivize the adoption of low-emission technologies. Furthermore, optimizing TIT based on sensitivity analysis insights could reduce CO emissions by up to 10–15% in practical scenarios, contributing to cleaner air and improved public health in countries. Integrating ML-driven outlier detection and ensemble methods into existing emission control systems can also reduce operational costs by minimizing fuel waste and avoiding non-compliance penalties. These advancements not only enhance the sustainability of gas turbine operations but also position the countries' energy sector to meet evolving environmental standards while maintaining energy security and supporting its transition to a greener economy.

References

Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production, 203*, 810-821. https://doi.org/10.1016/j.jclepro.2018.08.207

Aslan, E. (2024). Prediction and Comparative Analysis of Emissions from Gas Turbines Using Random Search Optimization and Different MachineL earning Based Algorithms. *Bulletin of the Polish Academy of Sciences Technical Sciences*, e151956-e151956. https://doi.org/10.24425/bpasts.2024.151956

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., ... & Carpenter, A. E. (2017). Data-analysis strategies for image-based cell profiling. *Nature Methods*, *14*(9), 849-863.

Coelho, D. S. L., Ayala, H. V. H., & Mariani, V. C. (2024). CO and NOx emissions prediction in gas turbine using a novel modeling pipeline based on the combination of deep forest regressor and feature engineering. *Fuel*, *355*, 129366. https://doi.org/10.1016/j.fuel.2023.129366

Dalal, A. S., Sultanova, N., Jayabalan, M., & Mustafina, J. (2023, December). Gas turbine–CO & NOx emission data analysis with predictive modelling using ML/AI approaches. In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 100-104). IEEE. https://doi.org/10.1109/DeSE60595.2023.10469322

Dirik, M. (2022). Prediction of NOx emissions from gas turbines of a combined cycle power plant using an ANFIS model optimized by GA. *Fuel*, *321*, 124037. https://doi.org/10.1016/j.fuel.2022.124037

Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., & Torres, J. F. (2018). Stacking ensemble learning for short-term electricity consumption forecasting. *Energies*, *11*(4), 949. https://doi.org/10.3390/en11040949

Farzaneh-Gord, M., & Deymi-Dashtebayaz, M. (2011). Effect of various inlet air cooling methods on gas turbine performance. *Energy*, *36*(2), 1196-1205. https://doi.org/10.1016/j.energy.2010.11.027

Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*, 583-595. https://doi.org/10.22190/FUMI1903583G

Karthikeyan, S., Kathirvalavakumar, T., & Prasath, R. (2023, June). Classification of the Class Imbalanced Data Using Mahalanobis Distance with Feature Filtering. In *International Conference on Mining Intelligence and Knowledge Exploration* (pp. 45-53). Cham: Springer Nature Switzerland.

Kaya, H., Tüfekci, P., & Uzun, E. (2019). Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS. *Turkish Journal of Electrical Engineering and Computer Sciences*, *27*(6), 4783-4796. https://doi.org/10.3906/elk-1807-87

Kochueva, O., & Nikolskii, K. (2021). Data analysis and symbolic regression models for predicting CO and NOx emissions from gas turbines. *Computation*, *9*(12), 139. https://doi.org/10.3390/computation9120139

Kumar, M. V., Babu, A. V., Reddy, C. R., Pandian, A., Bajaj, M., Zawbaa, H. M., & Kamel, S. (2022). Investigation of the combustion of exhaust gas recirculation in diesel engines with a particulate filter and selective catalytic reactor technologies for environmental gas reduction. *Case Studies in Thermal Engineering*, *40*, 102557.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150-156. https://doi.org/10.1016/j.jesp.2017.09.011

Liu, Z., Meng, H., Huang, J., Kwangwari, P., Ma, K., Xiao, B., & Li, L. (2021). Acute carbon monoxide poisoning with low saturation of carboxyhaemoglobin: a forensic retrospective study in Shanghai, China. *Scientific Reports*, *11*(1), 18554.

Lopes, C., Antelo, L. T., Franco-Uría, A., Alonso, A. A., & Pérez-Martín, R. (2015). Valorisation of fish by-products against waste management treatments–Comparison of environmental impacts. *Waste Management*, *46*, 103-112. https://doi.org/10.1016/j.wasman.2015.08.017

Lott, P., Casapu, M., Grunwaldt, J. D., & Deutschmann, O. (2024). A review on exhaust gas after-treatment of lean-burn natural gas engines–From fundamentals to application. *Applied Catalysis B: Environmental*, *340*, 123241.

Mahmoudi, S., Baeyens, J., & Seville, J. P. (2010). NOx formation and selective non-catalytic reduction (SNCR) in a fluidized bed combustor of biomass. *Biomass and bioenergy*, *34*(9), 1393-1409. https://doi.org/10.1016/j.biombioe.2010.04.013

Mare, D. S., Moreira, F., & Rossi, R. (2017). Nonstationary Z-score measures. *European Journal of Operational Research*, *260*(1), 348-358. https://doi.org/10.1016/j.ejor.2016.12.001

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140-147. https://doi.org/10.38094/jastt1457

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, *22*(1), 67-72. https://doi.org/10.4103/aca.ACA_157_18

Naghibi, A. (2024). Exploring explainable ensemble machine learning methods for long-term performance prediction of industrial gas turbines: A comparative analysis. *Engineering Applications of Artificial Intelligence*, *138*, 109318. https://doi.org/10.1016/j.engappai.2024.109318

Nino-Adan, I., Portillo, E., Landa-Torres, I., & Manjarres, D. (2021). Normalization influence on ANN-based models performance: A new proposal for Features' contribution analysis. *IEEE Access*, *9*, 125462-125477. https://doi.org/10.1109/ACCESS.2021.3110647

Osborne, J. W., & Overbay, A. (2019). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6. https://doi.org/10.7275/qf69-7k43

Pachauri, N. (2024). An emission predictive system for CO and NOx from gas turbine based on ensemble machine learning approach. *Fuel*, *366*, 131421. https://doi.org/10.1016/j.fuel.2024.131421

Pandey, R. A., & Chandrashekhar, B. (2014). Physicochemical and biochemical approaches for treatment of gaseous emissions containing NOx. *Critical Reviews in Environmental Science and Technology*, *44*(1), 34-96. https://doi.org/10.1080/10643389.2012.710430

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*, 972421. https://doi.org/10.3389/fnano.2022.972421

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71-72.

Rezazadeh, A. (2021). Environmental pollution prediction of NOx by predictive modelling and process analysis in natural gas turbine power plants. *Pollution*, *7*(2), 481-494. https://doi.org/10.22059/poll.2021.316327.977

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, *251*, 26-34. https://doi.org/10.1016/j.neucom.2017.04.018

Tian, J., Wang, L., Xiong, Y., Wang, Y., Yin, W., Tian, G., ... & Ji, S. (2024). Enhancing combustion efficiency and reducing nitrogen oxide emissions from ammonia combustion: A comprehensive review. *Process Safety and Environmental Protection*, 183, 514-543. https://doi.org/10.1016/j.psep.2024.01.020

Todeschini, R., Ballabio, D., Consonni, V., Sahigara, F., & Filzmoser, P. (2013). Locally centred Mahalanobis distance: a new distance measure with salient features towards outlier detection. *Analytica Chimica Acta*, *787*, 1-9. https://doi.org/10.1016/j.aca.2013.04.034

Valkenborg, D., Rousseau, A. J., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(5), 754-757.

Wardana, M. K. A., & Lim, O. (2022). Review of improving the NOx conversion efficiency in various diesel engines fitted with SCR system technology. *Catalysts*, *13*(1), 67.

Wood, D. A. (2023). Long-term atmospheric pollutant emissions from a combined cycle gas turbine: Trend monitoring and prediction applying machine

learning. *Fuel*, *343*, 127722. https://doi.org/10.1016/j.fuel.2023.127722

Wu, T. J., Burke, J. P., & Davison, D. B. (1997). A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, 1431-1439. https://doi.org/10.2307/2533509

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316. https://doi.org/10.1016/j.neucom.2020.07.061

Yaro, A. S., Maly, F., & Prazak, P. (2023). Outlier detection in time-series receive signal strength observation using z-score method with s n scale estimator for indoor localization. *Applied Sciences*, *13*(6), 3900. https://doi.org/10.3390/app13063900

Yousif, S. T., Ismail, F. B., & Al-Bazi, A. (2024). A hybrid neural network-based improved PSO algorithm for gas turbine emissions prediction. *Advanced Theory and Simulations*, 2301222. https://doi.org/10.1002/adts.202301222

Yu, H., & Kim, S. (2012). SVM Tutorial-Classification, Regression and Ranking. *Handbook of Natural Computing*, *1*, 479-506.

Research Article
Journal of Intelligent Systems: Theory and Applications 8(1) (2025) 84-94
DOI: 10.38016/jista.1620633

Due Date Determination in Dynamic Job Shop Scheduling with Artificial Neural Network

Mümtaz İpek[1*], İsmail Hakkı Cedimoğlu[2]

[1] Industrial Engineering Department, Faculty of Engineering, Sakarya University, Sakarya, Türkiye

[2] Department of Information Systems Engineering,Faculty of Computer and Information Sciences, Sakarya University, Sakarya, Türkiye

ipek@sakarya.edu.tr, cedim@sakarya.edu.tr

Abstract
In this study, an artificial neural network approach that is thought to produce better results as an alternative to due date determination methods in dynamic job shop scheduling environment is presented and its feasibility is demonstrated. The performance of the neural network model is compared with five different regression models. An event oriented simulation software is developed for the determination of the coefficients of the regression models and for the generation of data to be used in the training of the neural network model. Back-propagation artificial neural network was used as an artificial neural network model and a software was developed. After the regression models were created and the neural network was trained, the simulation software was run for the shortest processing time and earliest due date priority rules for comparison purposes. In order to compare the models, average absolute deviation from the due date, mean square of absolute deviation from the due date, average tardiness, number of tardy jobs, average earliness and number of early jobs were used as performance metrics. As a result of the study, the artificial neural network model was found to be effective in due date determination. Both the shortest processing time first and the earliest due date first priority rules gave good results in terms of several performance metrics. It was observed that the neural network gave better results in the shortest processing time priority rule in general.

Keywords: Dynamic Job Shop, Due Date Determination, Artificial Neural Networks

1. Introduction

In a production system, activities requiring decision-making occur hierarchically at three levels. These are strategic, tactical, and control level. At the strategic level, production plans are required to meet market demands. At the tactical level, the planned production schedule is coordinated with some shop floor constraints such as inventory, machine capacity, maintenance plan and labour productivity. At the control level, the flow of work is continuously regulated to realise the execution of the planned production schedules and schedules disturbed by unexpected events are immediately updated.

1.1. Dynamic scheduling

The dynamic problem causes difficulties in determining a finite schedule. Unlike make-to-stock production, there is no master production schedule to help predict future workload in a manufacturing facility. Unknown future work fluctuations make it difficult to develop efficient scheduling algorithms. Furthermore, finite scheduling techniques that attempt to detail the future state of jobs within the shop floor may not be appropriate if there is significant uncertainty about processing times.

1.2. Due date determination

The importance of assigning accurate due dates for the delivery of jobs in a production system is well recognised by academic researchers and managers in practice. Due to developments in manufacturing systems and idealised concepts in inventory systems, due dates-based research has attracted attention and a rich literature has been reported in this area (Cheng et al., 1989). In a manufacturing system, each job is assigned a due date before it is released for processing on the shop floor. The literature analysis shows that various decision rules have been proposed for due date assignment. The literature on due date assignment emphasises simple, regression-based approaches to deadline setting for the dynamic multi-machine case (Philipoom, 1994).

* Corresponding Author.
E-mail: ipek@sakarya.edu.tr

Recieved : 15 Jan 2025
Revision : 31 Jan 2025
Accepted : 10 Mar 2025

1.3. Regression analysis

Regression analysis is a statistical analysis technique that is frequently used to determine the relationship between two or more variables that have a cause-effect relationship between them with a mathematical function and to make estimation or prediction about the dependent variable using this relationship (Tarı, 1999, Orhunbilge, 2000). Regression analysis also reveals the structural relationships between variables. It is possible to find cause-effect relationships in most economic, social, and natural events.

After fitting the regression model, checking the adequacy of the model is the most important part of regression analysis. It is necessary to ensure that the applied model is close enough to the correct model and to check whether it meets all the assumptions of the least squares regression analysis. In regression analysis, analysis of variance and multiple coefficient of determination (R^2) are usually used for model adequacy. It is not enough to demonstrate the adequacy of the model by analysis of variance. In addition, the statistical significance of regression parameters should also be investigated by t-tests.

1.4. Artificial neural networks

An artificial neural network (ANN) can be defined as an inference mechanism based on the human brain (Negnevitsky, 2002). In other words, artificial neural networks are planned hierarchical structures with simple elements connected in parallel to each other and interacting with real world objects in the same way as biological nervous systems do (Kohonen, 1987). A general neural network model is characterised by processing elements. A processing element consists of five components:

- Inputs bring information to the processing element. This information is provided by other processing elements or external sources. Sometimes the processing element can provide information itself.
- Weights determine the effect of a certain input on a processing element. It is the weight values that need to be optimised during the training process in order for the network to produce the correct outputs.
- The summation function sums the weighted inputs of the process element. There are various summation functions (Neuralware Inc, 1990). The most common one is to find the weighted sum. Here, each input value is multiplied by its weight and summed (Oztemel, 2003).
- The transfer function determines the output of the processing element by modifying the result of the addition function. Again, there are various transfer functions (Neuralware Inc, 1990). Some of the popular ones are sigmoid function, linear function, and step function.
- The output sends the result of the transfer function to the connected processing elements or to external sources.

The topology of the network is the second feature that characterizes the network. A group of processing elements forms a structure called a layer. A typical neural network contains three interconnected layers. These are the input layer that accepts input from the outside world, the hidden layer (or hidden layers) that processes the information from the input layer and sends it to the output layer, and the output layer that informs the outside world about the decision of the network. Information flows between or within layers of the network.

1.5. Priority rules

Priority rules are used when there is more than one job in the queue in front of a machine and the job needs to be assigned when the machine becomes available. In this study, the shortest processing time and the earliest due date are used as priority rules. Since both priority rules lead to schedules with different flow times, therefore different completion times, separate simulations of the dynamic job shop under study were performed for each rule.

1.6. Performance metrics

Performance measures help comment on the success of a schedule. Therefore, it can be said whether the schedule is good or not according to a performance metric. In this study, an artificial neural network model is proposed for due date determination. The proposed model is compared with the regression models in terms of performance. Since the due dates are in question, the average absolute deviations of the due dates from the actual completion times, the squares of the average absolute deviations, the tardiness and earliness from the actual completion times were selected as performance measures. Comparisons were made on the basis of these metrics. In addition, the number of tardy and early jobs were also calculated for both types of models to give an idea.

1.7. Assumptions

Both real and hypothetical systems are available in the literature for dynamic workshop simulation. Hypothetical systems typically contain a small number of machines (usually less than 10) and some assumptions are made in simulation studies (Baker, 1974, French, 1982).

1.8. Aim

The aim of this study is to demonstrate the feasibility of artificial neural network for deadline determination in dynamic job shop scheduling. For this purpose, an event oriented simulation software using dynamic variables

has been developed. A back-propagation artificial neural network software using dynamic variables is also presented for modelling the artificial neural network. The event oriented simulation and the artificial neural network softwares are written in C. The due dates were determined by regression models and artificial neural network and compared according to the selected performance metrics. When determining the due date with regression models, missing data or a possible error may cause the due date information to be calculated very differently from what it should be. Artificial neural network can give appropriate answers in such a situation due to its ability to work with missing data.

2. Literature Review

The production control system for a shop floor can be analysed in a structure consisting of three sequential stages (Philipoom, 1994): Order stage, order release and shop floor stage. In the first stage the customer's work arrives and a due date is assigned. Assigning a due date is the first important task of shop floor control. Due date-related performance is characterised by the quality of the due date assignment rules. Due date assignment and products delivered to the customer on time will provide customer satisfaction and competitive advantage (Sha and Hsu, 2004). In this section, studies on date date assignment in the literature are examined in two groups: numerical and heuristic.

2.1. Numerical methods

In these methods, the problem is defined by mathematical models. These models are solved by mathematical programming techniques and the optimum solution is sought. Mosheiov (2001) studied the due date assignment problem and job shop scheduling in parallel similar machines. He proposed that the cost of a schedule is a function of maximum earliness cost, maximum tardiness cost, and due date cost. The aim of the study is to develop a due date scheduling algorithm that minimizes these three cost functions. Biskup and Jahnke (2001) considered a general due date assignment to jobs and scheduling of jobs on a single machine. They considered that the processing times are controllable. However, in contrast to previous approaches, they emphasised the case where all processing times can be reduced at the same rate. They concentrated on minimising the number of early, tardy, and late jobs as well as due date assignment. They found algorithms that can be solved polynomially. Veral (2001) tried to find out that it is possible to set static due date with flow time analysis. The proposed model has been compared with the TWK (Total Work) model. The comparison was carried out considering light, medium, and heavy shop floor loads. The author's model outperformed the TWK at all three different shop floor loads and performed better in workflow time prediction accuracy. Gupta et al. (2002) studied the permutation flow type problem. In this study, each job centre consists of parallel similar machines. Each job has different release dates and is processed in the same order on machines in different job centres. In the study, 20 jobs and 10 work centres were considered. In addition, the cost of due date assignment was added to the objective function. Wang and Uzsoy (2002) investigated the feasibility of job due dates in batch processing machines used in the metalworking and microelectronics industries when jobs are dynamically sent to the workshop. A genetic algorithm technique was used together with a dynamic programming algorithm. It was concluded that the study showed excellent average performance. Sabuncuoglu and Comlekci (2002) proposed a new flow time estimation method that uses route information about the operations of jobs, such as detailed job, shop floor and machine imbalances. They state that such information is now available in computer integrated manufacturing systems. They measured the performance of their proposed method by simulation under various experimental conditions. They compared with existing flow time estimation methods in terms of various performance measures. They found that the performance of manufacturing systems can be improved by information intensive methods rather than simple methods (TWK). They claimed that the use of detailed information in flow time prediction provides significant improvements over methods that use more integrated information to improve system performance. They stated that the results of the study showed that predicting flow time for each operation is a better approach than traditional job-dependent prediction. Song et al. (2002) stated that the determination of product due dates is an important part of production planning and studied the determination of product due dates in complex multi-stage assembly operations. Product lead times were used to minimise earliness and tardiness. Sha and Liu (2005) argue that although just-in-time production philosophy is gaining importance, the ability to deliver orders on time will increase customer satisfaction and provide a competitive advantage to the organisation. In this study, in order to improve the performance of TWK, which is one of the due date assignment rules, they represented the dynamic workshop conditions with IF-THEN rules and the k coefficient was determined by evaluating the situation in the workshop at the arrival of the job, thus reducing the due date error of the TWK method. As a result, the rule-based TWK method gave better results compared to static and dynamic TWK methods. Shabtay and Steiner (2006) argued that on-time delivery of orders is one of the most important issues in scheduling and supply chain management. The authors aimed to minimise the weighted earliness, tardiness and due date assignment penalties and to minimise the weighted number of late jobs and due date assignment costs for the single machine problem. Zhao (2016) examines a single-machine scheduling problem where jobs have specific release times. The research aims to determine an optimal common due date and scheduling sequence to minimize a cost function that includes the weighted

number of tardy jobs and due date assignment costs. The problem is proven to be NP-hard, and the authors propose a dynamic programming algorithm and a fully polynomial-time approximation scheme as solutions. Teymourifar and Ozturk (2018) designed new due date assignment models and dispatching rules for dynamic job shop scheduling problems, developed dispatching rules based on modified and composite characteristics of jobs, and obtained competitive results compared to existing models. Vinod et al. (2019) investigated the interaction between dynamic due date assignment methods and scheduling decision rules in a dynamic job shop with queue-dependent preparations. They developed analytical models based on regression using simulation results and found that the proposed scheduling rules improve the performance with respect to average lateness. Kianpour et al. (2021) introduced an automated model that develops job shop scheduling by integrating Industry 4.0 and project management principles. The model adapts to real-time information about processing times and due dates, aiming to minimise early and late costs by considering rescheduling expenses. Wang et al. (2022) investigates a single-machine scheduling problem that considers due date assignment alongside past-sequence-dependent setup times. Under common, slack, and different due date assignment methods, the objective is to find the optimal sequence and due dates that minimize the weighted sum of lateness, the number of early and delayed jobs, and due date costs, where weights depend on job positions in the sequence. The authors provide optimal properties and propose a polynomial-time algorithm to obtain the optimal solution. Mosheiov and Sarig (2024) addresses a single-machine scheduling and due-date assignment problem incorporating acceptable lead-times. The study combines elements of common and different due-date models, aiming to determine job-dependent due dates. The objective function, of a minmax type, consists of four cost components: job earliness, job tardiness, due-date cost, and due-date tardiness cost. The authors present a simple procedure for identifying different job types and introduce a polynomial-time solution.

2.2. Heuristic methods

In heuristic approaches, the solution is found by narrowing the area to be searched based on the findings obtained from experimental studies. It is called heuristic screening and is based on advanced searching algorithms (Tasgetiren, 1996). Philipoom (2000), who stated that due date determination is a difficult situation for manufacturing managers, examined the trends in the choice of priority rule in a job shop where due dates are determined depending on lead time and tardiness penalties. As a result of his study, he stated that the first rule with the shortest processing time works well for lean tardiness penalties. As the penalty for tardiness increased, he found that priority rules such as first-in-first-out worked well. He stated that the earliest due first rule does not work well due to the interaction between the earliest due date first rule and the parameters of the due date determination rule. Yang and Wang (2001) presented a new adaptive artificial neural network and heuristic hybrid approach for job shop scheduling. The neural network has the ability to adapt the connection weights and bias values of the processing elements during the feasible solution. They presented two heuristics that can be combined with the neural network. One of them was used to speed up the solution of the neural network and to guarantee the approximation of the network. The other one is used to obtain delay-free schedules from the feasible solutions provided by the neural network. Computer simulations showed that the proposed hybrid approach is fast and efficient. Cheng et al. (2002) studied the single machine problem in their study. In their problem, a common due date is assigned to all jobs. The objective is to determine the due date and schedule that will minimise the earliness, tardiness and the total penalty associated with the dur date. The authors claim that they have developed an algorithm that obtains the optimal due date and schedule if the job order is predetermined or if all jobs have the same processing time. Xiao and Li (2002) considered the problem of assigning a general due date to jobs and scheduling jobs on parallel machines by minimising the weighted sums of due date, total earliness, total tardiness and an absolute performance ratio for this heuristic. They presented a better worst-case bounded heuristic for the case with zero earliness penalty. They also developed an approximation scheme that is completely polynomial. They claimed that their heuristic contributes to job shop scheduling and general due date assignment algorithm development. Birman and Mosheiov (2004) studied the due date and scheduling problem in a two-machine flow-type production. They stated that due date scheduling problems have attracted a lot of attention in recent years. The objective of their study was to minimise the maximum earliness, tardiness and due date determination costs. As a result, they claimed that they found a more effective solution with Johnson's algorithm. Min and Cheng (2006) proposed a type of genetic algorithm based on regional coding that determines the optimal scheduling policy to determine the optimal overall due date, the processing sequence and the number of jobs on each machine, and minimises the costs of due date assignment, earliness, and tardiness. For the genetic algorithm, they also added a simulated annealing mechanism and an iterative heuristic fine-tuning operator to construct 3 types of hybrid genetic algorithms with good performance. Focusing on similar parallel machine scheduling and general parallel machine scheduling problem, the numerical computational results show that these algorithms outperform heuristic algorithms and are suitable for large-scale parallel machine earliness, tardiness scheduling problem. In their study, Mosheiv and Oron (2006) wanted to determine the sequence of

work, the overall due date and the placement of rapid maintenance activity. Jobs scheduled after or before the due date are penalised according to their early or tardy finish time. The processing time of a job scheduled after the maintenance activity is reduced by a job-dependent factor. The objective is the minimum total earliness, tardiness, and deadline costs. They proposed a polynomial solution for this problem. In this first work, where maintenance scheduling and due date assignment are performed simultaneously, they state that the problem is solvable in polynomial time. In his study, Chen (2007) focused on output time estimation, which is a critical task in a biscuit factory. He proposed an intelligent hybrid system to improve the accuracy of output time estimation. Firstly, he applied the concept of input classification to the Chen fuzzy backpropagation network approach by pre-classifying batches of biscuits with a k-means classifier before estimating their output time with a fuzzy backpropagation network. Examples belonging to different categories were taught by networks with different but identical topology. Secondly, the factory future shipment plan was also included in the intelligent hybrid request. In order to evaluate the effectiveness of the proposed methodology, production simulation was performed to create test cases. According to the experimental results, the prediction accuracy of the intelligent hybrid system is significantly better than other approaches. Baykasoğlu and Gökçen (2009) propose a due date assignment approach for a multi-stage job shop using Gene expression Programming (GEP), a genetic programming technique. Simulation experiments showed that the GEP-based method outperformed several conventional due date assignment models under various test conditions. Yang et al. (2012) considered a job shop scheduling problem involving due dates, aiming to minimise the sum of weighted earliness and tardiness. They proposed an improved genetic algorithm that uses an operation-based scheme to represent schedules as chromosomes. The effectiveness of the algorithm is demonstrated through tests on various job shop scheduling problems of different sizes. Inal et al. (2023) to solve the dynamic scheduling problem, propose a multi-agent system with reinforcement learning aimed at the minimization of tardiness and flow time to improve the dynamic scheduling techniques. The performance of the proposed multi-agent system is compared with the first-in-first-out, shortest processing time, and earliest due date dispatching rules in terms of the minimization of tardy jobs, mean tardiness, maximum tardiness, mean earliness, maximum earliness, mean flow time, maximum flow time, work in process, and makespan. Under a heavy workload, the proposed multi-agent system gives the best results for five performance criteria, which are the proportion of tardy jobs, mean tardiness, maximum tardiness, mean flow time, and maximum flow time.

In addition to the traditional due date setting rules, the use of artificial neural network for prediction is quite common in the literature. The ability of artificial neural network to learn from examples or to reach a conclusion by considering different values related to the workshop can also be used for due date setting. Considering the ability of artificial intelligence to learn from examples and applications in dynamic workshop scheduling, it was deemed appropriate to carry out such a research for due date determination.

3. Due Date Determination

3.1. Regression models

In this study, 5 regression models were used to determine the due date (Philipoom et al., 1994):

1. Total work (TWK):

$$F_i = \mathrm{k}\mathrm{P}_i \quad (1)$$

The predicted flow time (F_i) of job i is a function of the total processing time P_i. k is a coefficient and is calculated by regression.

2. Number of operations (NOP):

$$F_i = \mathrm{k}\mathrm{N}_i \quad (2)$$

Here, the predicted flow time of the job is a function of the number of operations (N_i) of job i. Again k is a coefficient and is calculated by regression.

3. Total work and Number of operations (TWK+NOP):

$$F_i = k_1 P_i + k_2 N_i \quad (3)$$

In this model, both the number of operations and the total processing time are used. k_1 and k_2 coefficients are calculated by regression.

4. Number of jobs in queue (JIQ):

$$F_i = k_1 P_i + k_2 (JIQ_i) \quad (4)$$

When job i arrives at the job shop, the number of waiting jobs in the queues is summed (JIQi). This job shop data is combined with the job characteristic P_i. The coefficients k_1 and k_2 are calculated by regression.

5. Work in queue (WIQ):

$$F_i = k_1 P_i + k_2 (WIQ_i) \quad (5)$$

This model differs from model 4 in that it does not use the number of jobs in the queues but uses the total processing time in the job shop. Again, the coefficients k_1 and k_2 are calculated by regression.

3.2. Due date determination with regression models

In these models, the flow time is estimated (Figure 1). When the ready time (r_i), which is the time when the jobs arrive at the workshop, is added to the estimated flow times, the due date (d_i) can be estimated.

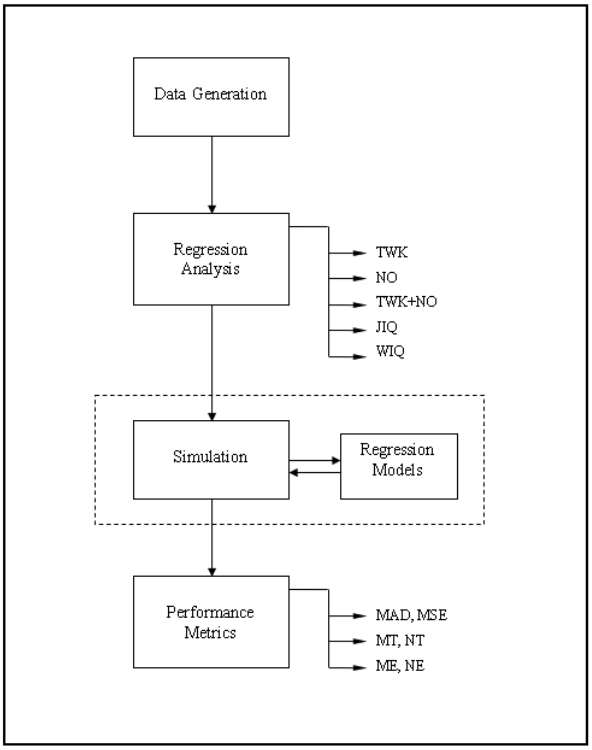


Figure 1. Due date determination with regression models

3.2.1. Data generation for regression analysis

The data required for determining the coefficients of the total processing time of job i, the number of operations of job i, the sum of the number of jobs in the queues when job i arrives at the job shop and the sum of the total processing times of the waiting jobs in the workshop (workload) used in the estimation of the flow time in the 5 regression models mentioned above were obtained by simulation. For SPT (Shortest Processing Time) and EDD (Earliest Due Date) priority rules, 10 simulations were performed, each starting with a different random number. In each simulation run, after a warm-up period of 5000 jobs, the data to be used in the regression models for 10000 jobs were recorded. This resulted in 10 data files with 10000 data in each file. Then, a single data set with 10000 jobs data was obtained by taking the data of one of the 10 jobs from each data file.

3.2.2. Creation of regression models

The k coefficients required for 5 regression models were obtained by linear regression using this data of 10000 jobs (Table 1, Table 2). Regression analysis were carried out for each model, and sample outputs for the SPT first priority rule and the TWK model are given in the appendix.

3.2.3. Due date determination with regression models

After the regression model coefficients were obtained, 10 simulations were performed using SPT and EDD priority rules. In these 10 simulations, different initial random numbers were used from each other and also from the simulations performed to obtain the data set required for the determination of the regression model coefficients. These initial random numbers will be used as the same for the proposed neural network model in the future.

Table 1. Coefficients of regression models (SPT)

Model	k_1	Std. Dev.	Sig.	k_2	Std. Dev.	Sig.
TWK	6,963	0,080	0,000	-	-	-
NOP	50,062	0,867	0,000	-	-	-
TWK+NOP	21,511	0,226	0,000	-143,238	2,124	0,000
JIQ	14,294	0,193	0,000	-17,614	0,428	0,000
WIQ	12,213	0,175	0,000	-0,048	0,001	0,000

Table 2. Coefficients of regression models (EDD)

Model	k_1	Std. Dev.	Sig.	k_2	Std. Dev.	Sig.
TWK	7,336	0,019	0,000	-	-	-
NOP	68,550	0,196	0,000	-	-	-
TWK+NOP	4,184	0,058	0,000	30,960	0,542	0,000
JIQ	3,170	0,025	0,000	6,190	0,038	0,000
WIQ	4,254	0,034	0,000	0,027	0,000	0,000

In these 10 simulations, the due dates are determined when the jobs arrive at the job shop and the due date is determined by adding the flow time estimated using the regression model to the arrival time of the jobs. In each simulation, data were recorded for 10000 jobs after the warm-up period of 5000 jobs.

The mean absolute deviation (MAD) was calculated by taking the absolute differences of the estimated due dates of 10000 jobs from the actual completion times (Ci) and the mean square errors (MSE) were calculated. In addition, the number of tardy jobs (NT), mean tardiness (MT), number of early jobs (NE) and mean earliness (ME) values were calculated. These calculations were made separately for 10 data files. By taking the arithmetic averages of these 10 calculated values, MAD, MSE, NT, MT, NE, and ME values are calculated for the priority rule and regression model to be compared with the proposed model values.

3.2. Artificial neural network model

In this study, an artificial neural network model is proposed for due date determination in job shop type

production (Figure 2). The artificial neural network used is a back-propagation neural network. In the previously mentioned regression models, one or more information about the job, job shop or both is used in the equation of that regression model. In the proposed model, in addition to the information used in the regression models, other job and job shop information is used to predict the work flow time of the artificial neural network.

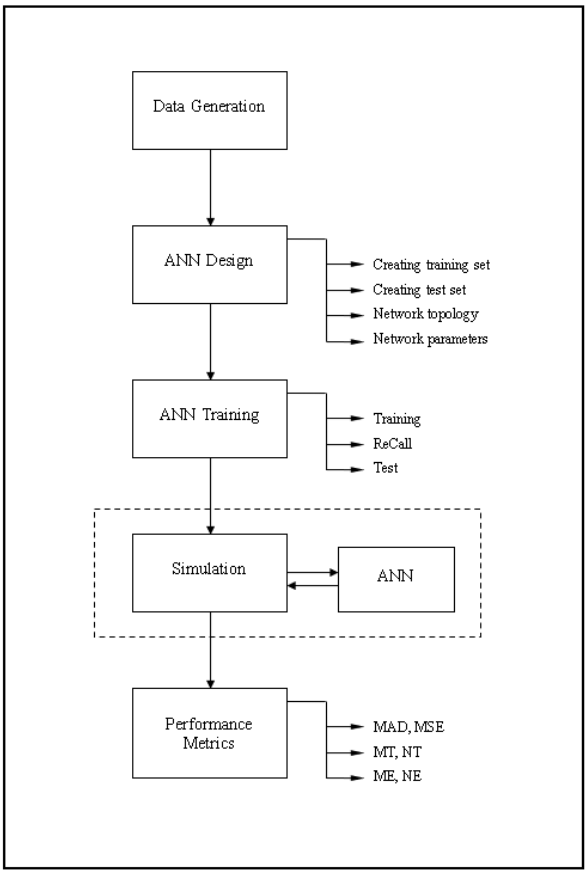


Figure 2. Articial neural network model

The information to be used in the prediction of the flow time and therefore the due date using the artificial neural network is shown in Table 3. This information is the input information of the artificial neural network and its number is 15. Therefore, the number of inputs of the artificial neural network is 15, there are 15 processing elements in the input layer. In the output layer, there is only one processing element, the flow time information.

Table 3. Inputs of artificial neural network

Input	Information
1	Maximum operation time of the work
2	Sum of the operation times of the work
3	Total number of jobs in the workshop
4	Total number of operation of jobs waiting in queues
5	Sum of average lateness times of works
6	Sum of operation times of jobs waiting in queues
7..15	1st, 2nd, ..., 9th operation times of the work

3.2.1. Creation of training set

The samples (input/output) to be used as a training set for the artificial neural network are the same as the dataset produced to determine the coefficients of the regression models. After the warm-up period of 5000 jobs in the regression model, a single dataset of 10000 jobs created by taking one of every 10 jobs from 10 data files obtained as a result of 10 simulations of 10000 jobs was also used for artificial neural network training. There are 2 datasets using SPT and EDD priority rules. Different neural networks were trained for both priority rules. This information in the dataset is not suitable for neural network training. Since sigmoid function will be used as activation function in the neural network, the results will be in the range of 0-1. In other words, it is not possible for the network to produce a value greater than 1 or less than 0 for the flow time (Öztemel, 2003). For this reason, the input and output values of the network were normalised and a scaled dataset was obtained. The output values of the network after training were also reverse normalised to obtain normal flow time values.

The normalisation process was performed according to the following equation:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (6)$$

The inverse normalisation process is obtained from the normalisation equation as follows:

$$x = x'(x_{max} - x_{min}) + x_{min} \qquad (7)$$

In order for the dataset to be ready for the training of the artificial neural network, the dataset should be divided into two parts. Because two sets, training and test, are required for training an artificial neural network. The network already sees all the examples in the training set during training. Until the desired error level is achieved or the desired number of iterations or epochs (the network sees all the examples once) is completed, it calculates output values from the inputs and changes the weight values by looking at the difference between the desired and actual output. Therefore, it cannot be said that the network learnt with a single set. Therefore, the normalised dataset is divided into two sets: training and test. The splitting was done randomly so that 80% of the first dataset was in the training set and 20% in the test set (Philipoom et al., 1994).

3.2.2. Creation of neural network model

Since 15 values will be used as input in the flow time estimation with the artificial neural network, 15 input and one output processing element estimating the flow time are used. For the data obtained using the SPT and EDD priority rule, the artificial neural network was

designed as a single hidden layer with 7 and 9 processing elements, respectively. The training process with the data of both priority rules was performed with the same parameters (Table 4).

Table 4. Parameters of articial neural network (SPT, EDD)

Parameter	Value
Learning coefficient	0.2
Momentum coefficient	0.8
Initial weights values	Random between -0.1 and 0.1
Example presentation	Sequential
Number of epochs	2000

3.2.3. Training of artificial neural network

The neural network training was performed in 2000 epochs for both SPT and EDD. Since a training set of 8000 jobs was used during the training, each training was performed as 16000000 iterations. During the training process, absolute deviation and squared error values were calculated and recorded for the output processing element at the end of each epoch. Figure 3 and Figure 4 shows the graphs of the mean deviation squared values of the neural networks.

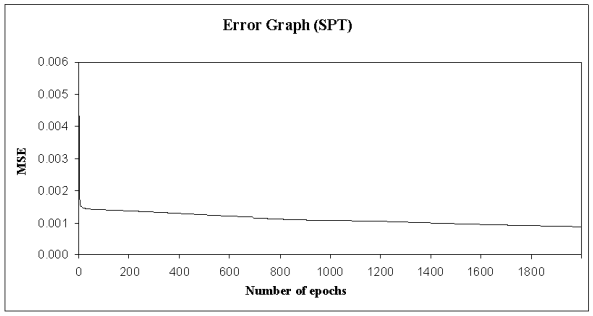


Figure 3. Articial neural network error graph (SPT)

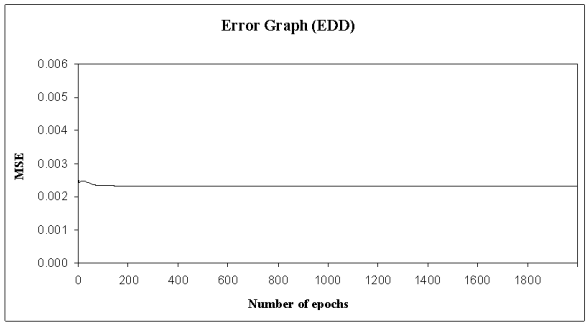


Figure 4. Articial neural network error graph (EDD)

In artificial neural network training, in order to determine whether the network has learnt or not, the training set and the test set that the network has never seen are given as input to the network and the outputs are compared with the actual output values. If the difference is below the acceptable margin of error, the network is said to respond correctly to that input sample, and if it is below, it is said to respond incorrectly. Table 5 shows the training success percentages of both networks.

In artificial neural network training, to determine whether the network has learned, the training set and the test set, which the network has never seen before, are given as input to the network and the outputs are compared with the actual output values. If the difference is below the acceptable error margin, it is said that the network responded correctly to that input sample, if not, it is said to have responded incorrectly. Table 5 shows the training success percentages of both networks.

Table 5. Training performance of artificial neural network

Epoch	SPT		EDD	
	Training set (%)	Test set (%)	Training set (%)	Test set (%)
500	96	96	94	94
1000	97	97	95	95
1500	97	97	95	95
2000	97	97	95	95

3.2.4. Due date determination with artificial neural network

After the training of the artificial neural networks was completed, 10 simulations were performed using SPT and EDD priority rules. In these 10 simulations, different initial random numbers were used from each other and from the simulations performed to obtain the dataset required for the training of the artificial neural networks and also for the determination of the regression model coefficients. These initial random numbers were the same as those used in the simulations in which the due date was decided by using regression models.

In these 10 simulations, the due dates are determined by adding the flow time estimated using the trained artificial neural network to the arrival time of the jobs when the jobs arrive at the workshop. In each simulation, data were recorded for 10,000 jobs after a warm-up period of 5,000 jobs.

The mean absolute deviation was calculated by the absolute deviations of the estimated due dates of the 10000 jobs from the actual completion times (C_i) and the mean absolute deviation squares were calculated. In addition, the number of jobs tardy, mean tardiness, number of jobs early mean earliness values were calculated. These calculations were made separately for 10 data files. By taking the arithmetic averages of these 10 calculated values, MAD, MSE, NT, MT, MT, NE, and ME values are calculated for the priority rules and artificial neural network model to be compared with the values of the previously mentioned regression models.

3.6. Performance metrics of models

In order to compare these simulation results of the proposed artificial neural network model and 5 regression models, arithmetic averages of the

performance metric values obtained from 10 simulations were taken and evaluations were made based on these average values. The mean values of the performance metrics obtained for the SPT and EDD priority rules for the regression models and the neural network model are given in Tables 6 and 7.

Based on the statistical analyses (statistical test outputs of the models for the MAD performance metric with the SPT priority rule are given in the appendix as an example), in Table 6, it is seen that the ANN model gives the best result with a value of 211 when the MAD performance metric is considered. This model is followed by the models of TWK+NOP, JIQ, TWK and NOP respectively.

It can be said that the best result in terms of the MSE performance metric is given by the TWK+NOP and ANN model. There is no statistically significant difference between these two models. These models are followed by the NOP, TWK+NOP, JIQ and WIQ models in terms of performance, respectively.

Table 6. Average performance values of models for SPT

Model	MAD	MSE	MT	NT	ME	NE
TWK	377	438156	131	1297	246	8704
NOP	387	567960	192	2190	195	7810
TWK+NOP	308	**305371**	152	4014	156	5986
JIQ	349	394630	142	3049	208	6951
WIQ	669	622800	**75**	**593**	593	9407
ANN	**211**	312044	92	4076	**119**	**5924**

Table 7. Average performance values of models for EDD

Model	MAD	MSE	MT	NT	ME	NE
TWK	396	**439933**	**125**	**1210**	271	8790
NOP	492	596478	162	1564	331	8436
TWK+NOP	446	503898	138	1324	307	8676
JIQ	380	497366	160	1723	220	8277
WIQ	**275**	481953	183	2436	**92**	**7564**
ANN	324	473826	166	1960	158	8040

When evaluated in terms of MT performance metric, the best result is given by the WIQ model. Then, the ANN model gives the second best result. The others, in order of success in terms of this performance metric, are the TWK, JIQ, TWK+NOP and NOP models.

Considering the ME performance metric, based on statistical analyses, it can be said that each model produces different results from each other. The most successful model among these is the ANN model. This model is followed by the models of TWK+NOP, NOP, JIQ and WIQ respectively.

Table 7 shows the performance values of the models using the EDD priority rule. Considering the MAD performance metric, it can be said that each model produces different values from each other with statistical analyses. The best result was produced by the WIQ model. This model is followed by the ANN, JIQ, TWK, TWK+NOP and NOP models, respectively.

It can be said that there are three groups of models that are different from each other considering the MSE performance metric. Of these, the first group of models, namely, the TWK, ANN ve WIQ models, yielded the best results.

When the MT performance metric is analysed, the best result is given by the TWK model. Then, the TWK+NOP model, followed by the third group consisting of JIQ, NOP and ANN, and finally followed by the WIQ model.

Similarly, in terms of the ME performance metric, the statistical analyses show that each model produces different values from each other and the best result is given by the WIQ model. This model is followed by ANN model.

4. Conclusions

In this study, an artificial neural network approach, which is thought to produce better results as an alternative to due date determination methods in dynamic job shop scheduling, is presented and its feasibility is demonstrated. The feasibility of the artificial neural network model in due date determination is demonstrated. In terms of both the shortest processing time first and earliest due date first priority rules, the neural network gave good results in terms of several performance metrics.

It was observed that the artificial neural network generally gave better results in the shortest processing time performance metric. It has been shown that the shortest processing time priority rule gives the best results in terms of mean absolute deviation, mean square error and mean earliness performance metrics. Again, it was seen that the artificial neural network was among the models that gave the best results in terms of the mean square error performance metric together with the earliest due date first priority rule.

In this study, the artificial neural network model used to determine the due date is back-propagation artificial neural network model. Single layer is used as hidden layer. It may be possible to obtain better results by using two or more hidden layers. In addition, a fully connected artificial neural network was used in this study. That is, each process element in each layer is connected with each process element of the next layer. Instead, a semi-connected network model can be used, that is, a network model in which a processing element is connected to only one or a few process elements in the next layer. In this way, better results can be obtained. In the due date prediction of the artificial neural network, various data of the job and job shop were used as input. It may be possible to get better results by using more information about the work and job shop as input or by not using some of the inputs used. The artificial neural network used in this study is a back-propagation artificial neural network. Using other artificial neural network models or deep learning models instead of this network may provide better prediction results.

References

Baker, K. R., “Introduction to Sequencing and Scheduling”, John Wiley & Sons, New York, 1974.

Baykasoglu, A., Gokcen, M., "Gene expression programming based due date assignment in a simulated job shop", Expert Systems with Applications, Cilt 26, Sayı 10, Sayfa 12143-12150, Aralık 2009.

Birman, M., Mosheiov, G., “A note on a Due Date Assignment on a Two Machine Flow Shop”, Computers and Operations Research, No. 31, s. 473-480, 2004.

Biskup, D., Jahnke, H., “Common Due Date Assignment for Scheduling on a Single Machine with Jointly Reducible Processing Times”, International Journal of Economics”, No. 69, s. 317-322, 2001.

Chen, T., “An Intelligent Hybrid System for Wafer Lot Output Time Prediction”, Advanced Engineering Informatics, No. 21, s. 55-65, 2007.

Cheng, T. C. E., Gupta, M. C.,” Survey of Scheduling Research Involving Ddue Date Determination Decisions”, European Journal of Operational Research, No. 38, s. 156-166, 1989.

Cheng, T. C. E., Chen, Z. L., Shakhlevich, N. V., “Common Due Date Assignment and Scheduling with Ready Times”, Computers and Operations Research, No. 29, s. 1957-1967, 2002.

French, S, “Sequencing and Scheduling: An Introduction to the Mathematics of the Job Shop”, John Wiley & Sons, New York, 1982.

Kohonen, T., “State of the Art in Neural Computing”, IEEE First International Conference on Neural Networks, No. 1, s. 79-90, 1987.

Gupta, J. N. D., Kruger, K., Lauff, V., Werner, F., Sotskov, Y. N., “Heuristics for Flow Shops with Controllable Processing Times and Assignable Due Dates”, Computers and Operations Research, No. 29, s. 1417-1439, 2002.

Inal, A., F., Sel, C., Aktepe, A., Turker, A., K., Ersoz, S., “A Multi-Agent Reinforcement Learning Approach to the Dynamic Job Shop Scheduling Problem”, Sustainability, Vol. 15 (10), 8262, 2023.

Kianpour, P., Gupta, D., Krishnan, K., K., Gopalakrishnan, B., “Automated job shop scheduling with dynamic processing times and due dates using project management and industry 4.0”, Journal of Industrial and Production Engineering, Cilt 38, Sayı 7, 2021.

Min, L., Cheng, W., “Genetic Algorithms fort he Optimal Common Due Date Assignment and the Optimal Scheduling Policy in Paralel Machine Earliness/Tardiness Scheduling Problems”, Robotics and Computer Integrated Manufacturing, No. 22, s. 279-287, 2006.

Mosheiov, G., “A Common Due Date Assignment Problem on Paralel Identical Machines”, Computers and Operations Research, N0. 28, s. 719-732, 2001.

Mosheiov, G., Oron, D., “Due Date Assignment and Maintenance Activitty Scheduling Problem”, Mathematical and Computer Modelling, No. 44, s. 1053-1057, 2006.

Mosheiov, G., Sarig, A., “The minmax due-date assignment problem with acceptable lead-times”, Annals of Operations Research, Vol. 343, pp. 401–410, 2024.

Negnevitsky, M., “Artificial Intelligence: A Guide to Intelligent Systems”, Addison-Wesley, 2002.

Neuralware Inc., “NeuralWorks Professional II/Plus: Neural Computing, Pittsburg, 1990.

Orhunbilge, N., “Uygulamalı Regresyon ve Korelasyon Analizi”, Avcıol Basım Yayın, İstanbul, 2000.

Oztemel, E., “Yapay Sinir Ağları”, Papatya Yayıncılık, İstanbul, 2003.

Philipoom, P. R., Rees, L. P., Wıegmann, L., “Using Neural Networks to Determine Internally Set Due Date Assignments for Shop Scheduling”, Decision Sciences, No. 25-5/6, s. 825-851, 1994.

Philipoom, P. R., “The Choice of Dispatching Rules in a Shop Using Internally Set Due Dates with Quoted Leadtime and Tardiness Costs”, International Journal of Production Research, No. 7, s. 1641-1655, 2000.

Sabuncuoglu, I, Comlekci, A., “Operation Based Flow Time Estimation in a Dynamic Job Shop”, Omega, No. 30, s. 423-442, 2002.

Sha, D. Y., Hsu, S. Y., “Due Date Assignment in Wafer Fabrication Using Artificial Neural Networks”, International Journal of Advanced Manufacturing Technology, No. 23, s. 768-775, 2004.

Sha, D. Y., Liu, C. H., “Using Data Mining for Due Date Assignment in a Dynamic Job Shop Environment”, International Journal of Advanced Manufacturing Technology, No. 25, s. 1164-1174, 2005.

Shabtay, D., Steiner, G., “Two Due Date Assignment Problems in Scheduling a Single Machine”, Operations Research Letters, No. 43, s. 683-691, 2006.

Song, D. P., Hicks, C., Earl, C. F., “Product Due Date Assignment for Complex Assemblies”, International Journal of Economics, No. 76, s. 243-256, 2002.

Tarı, R., “Ekonometri”, Alfa Basım Yayım, İstanbul, 1999.

Tasgetiren, M. F., “Atölye Tipi Çizelgeleme Problemi için Bir Uzman-Yapay Sinir Ağı Modeli”, Doktora Tezi, İstanbul Üniversitesi, 1996.

Teymourıfar, A, Ozturk, G., “New dispatching rules and due date assignment models for dynamic job shop scheduling problems”, International Journal of Manufacturing Research, Cilt 13, Sayı 4, 2018.

Veral, E. A., “Computer Simulation of Due Date Setting in Multimachine Job Shops”, Computers and Industrial Engineering, No. 41, s. 77-94, 2001.

Vinod, K., T., Prabagaran, S., Joseph, O., A., “Dynamic due date assignment method: A simulation study in a job shop with sequence-dependent setups”, Journal of Manufacturing Technology Management, Cilt 30, Sayı 6, 2019.

Wang, C. S., Uzsoy, R., “A Genetic Algoritm to Minimize Maximum Lateness on a Batch Processing Machine”, Computers and Operations Research, No. 29, s. 1621-1640, 2002.

Wang, X., Liu, W., Lu, L., Zhao, P., Zhang, R., “Due date assignment scheduling with positional-dependent weights and proportional setup times”, Mathematical Biosciences and Engineering, Vol. 19, Issue 5, pp. 5104-5119, 2022.

Xiao, W. Q., Lı, C. L., “Approximation Algorithms for Common Due Date Assignment and Job Schedulin on Parallel Machines”, IIE Transactions, No. 34, s. 467-477, 2002.

Yang, S. Wang, D., “A New Adaptive Neural Network and Heuristics Approach for Job Shop Scheduling”, Computers and Operations Research, No. 28, s. 955-971, 2001.

Yang, H., Sun, Q., Saygin, C., Sun, S., "Job shop scheduling based on earliness and tardiness penalties with due dates and deadlines: an enhanced genetic algorithm", The International Journal of Advanced Manufacturing Technology, Vol. 61, pp. 657–666, 2012.

Zhao, C., “Common duedate assignment and single-machine scheduling with release times to minimize the weighted

number of tardy jobs", Japan J. Indust. Appl. Math., Vol. 33, pp. 239–249, 2016.

Ek (Appendix)

Regression analysis of the TWK regression model with the SPT priority rule:

Regression

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	P	.	Enter

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,654	,428	,428	681,26581

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	3474552446,846	1	3474552446,846	7486,273	,000
	Residual	4640766870,155	9999	464123,099		
	Total	8115319317,001	10000			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	**P**	6,963	,080	,654	86,523	,000

Statistical test outputs of the models for the MAD performance measure with the SPT first priority rule:

Oneway

Test of Homogeneity of Variances
OMS

Levene Statistic	df1	df2	Sig.
346,785	5	59994	,000

ANOVA (OMS)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1182740505,285	5	236548101,057	8792,803	,000
Within Groups	1613986659,963	59994	26902,468		
Total	2796727165,248	59999			

Post Hoc Tests

Multiple Comparisons
Dependent Variable: OMS
Tukey HSD

(I) Model	(J) Model	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	**2**	-9,75590(*)	2,31959	,000	-16,3663	-3,1455
	3	69,38010(*)	2,31959	,000	62,7697	75,9905
	4	27,87910(*)	2,31959	,000	21,2687	34,4895
	5	-291,60910(*)	2,31959	,000	-298,2195	-284,9987
	6	166,68980(*)	2,31959	,000	160,0794	173,3002
2	**1**	9,75590(*)	2,31959	,000	3,1455	16,3663
	3	79,13600(*)	2,31959	,000	72,5256	85,7464
	4	37,63500(*)	2,31959	,000	31,0246	44,2454
	5	-281,85320(*)	2,31959	,000	-288,4636	-275,2428
	6	176,44570(*)	2,31959	,000	169,8353	183,0561
3	**1**	-69,38010(*)	2,31959	,000	-75,9905	-62,7697
	2	-79,13600(*)	2,31959	,000	-85,7464	-72,5256
	4	-41,50100(*)	2,31959	,000	-48,1114	-34,8906
	5	-360,98920(*)	2,31959	,000	-367,5996	-354,3788
	6	97,30970(*)	2,31959	,000	90,6993	103,9201
4	**1**	-27,87910(*)	2,31959	,000	-34,4895	-21,2687
	2	-37,63500(*)	2,31959	,000	-44,2454	-31,0246
	3	41,50100(*)	2,31959	,000	34,8906	48,1114
	5	-319,48820(*)	2,31959	,000	-326,0986	-312,8778
	6	138,81070(*)	2,31959	,000	132,2003	145,4211
5	**1**	291,60910(*)	2,31959	,000	284,9987	298,2195
	2	281,85320(*)	2,31959	,000	275,2428	288,4636
	3	360,98920(*)	2,31959	,000	354,3788	367,5996
	4	319,48820(*)	2,31959	,000	312,8778	326,0986
	6	458,29890(*)	2,31959	,000	451,6885	464,9093
6	**1**	-166,68980(*)	2,31959	,000	-173,3002	-160,0794
	2	-176,44570(*)	2,31959	,000	-183,0561	-169,8353
	3	-97,30970(*)	2,31959	,000	-103,9201	-90,6993
	4	-138,81070(*)	2,31959	,000	-145,4211	-132,2003
	5	-458,29890(*)	2,31959	,000	-464,9093	-451,6885

* The mean difference is significant at the .05 level.