1. GİRİŞ

El yazısı rakamlarının otomatik tanınması, bilgisayarlı görü ve makine öğrenmesi alanlarında önemli bir konudur. Bu alanda yapılan çalışmalar, posta kodu tanıma, çek tanıma, elektronik belge işleme ve dijitalleştirme gibi birçok uygulama alanında büyük bir rol oynamaktadır. El yazısı rakamlarının doğru ve hızlı bir şekilde tanınması, otomatik tanıma sistemlerinin etkinliğini ve verimliliğini artırabilir.

Bu çalışmada, MNIST (Modified National Institute of Standards and Technology) veri seti üzerinde genetik algoritma tabanlı öznitelik seçimi ve Rastgele Orman Sınıflandırıcısı kullanarak bir otomatik tanıma sistemi geliştirilmektedir. MNIST veri seti, el yazısı rakamlarının tanınması için yaygın olarak kullanılan bir veri setidir ve 60.000 eğitim örneği ve 10.000 test örneği içermektedir.
Genetik algoritma, doğal seleksiyon ve genetik süreçlerden esinlenen bir optimizasyon tekniğidir. Bu çalışmada, genetik algoritma, öznitelik seçimi probleminde kullanılarak veri setindeki en bilgilendirici ve ayırt edici özniteliklerin belirlenmesi hedeflenmektedir. Genetik algoritma ile öznitelik seçimi, veri boyutunu azaltarak hesaplama ve eğitim sürelerini iyileştirebilir ve aşırı uyum (overfitting) riskini azaltabilir.

Rastgele Orman Sınıflandırıcısı, bir ensamble (birleşik) sınıflandırma algoritmasıdır ve birden çok karar ağacını birleştirerek daha doğru ve güvenilir sonuçlar elde etmeyi amaçlar. Bu çalışmada, seçilen öznitelikler kullanılarak Rastgele Orman Sınıflandırıcısı modeli eğitilerek el yazısı rakamlarının otomatik tanınması gerçekleştirilmektedir.

Bu çalışmanın amacı, genetik algoritma tabanlı öznitelik seçimi ve Rastgele Orman Sınıflandırıcısı'nı birleştirerek MNIST veri setinde el yazısı rakamlarının otomatik tanınmasında etkili bir yöntem geliştirmektir. Elde edilen sonuçlar, otomatik tanıma sistemlerinin performansını artırmak ve gerçek dünya uygulamalarında daha doğru sonuçlar elde etmek için önemli bir katkı sağlayabilir.

2. VERİ SETİNİN HAZIRLANMASI

Veri Setinin Hazırlanması adımında, çalışmamızda MNIST (Modified National Institute of Standards and Technology) el yazısı rakamları veri setini kullandık. Bu veri seti, el yazısı rakamlarının görüntülerini içermektedir.

İlk olarak, TensorFlow kütüphanesi içinde bulunan mnist.load_data() fonksiyonunu kullanarak MNIST veri setini yükledik. Bu fonksiyon, veri setini eğitim ve test veri setleri olmak üzere ikiye ayırır ve ilgili değişkenlere yükler. Eğitim veri seti x_train ve y_train değişkenlerinde saklanırken, test veri seti x_test ve y_test değişkenlerinde saklanır.

Daha sonra, verileri ölçeklendirmek için StandardScaler sınıfını kullandık. Bu sınıfı sklearn.preprocessing modülünden import ettik. StandardScaler nesnesi oluşturarak, verileri belirli bir ölçekte birbirine uyarlama işlemi gerçekleştirdik. Bunu yapmak için eğitim veri setini fit_transform() yöntemi ile ölçeklendirerek x_train_scaled değişkenine atadık. Daha sonra, test veri setini ise transform() yöntemiyle ölçeklendirerek x_test_scaled değişkenine atadık.

Bu ölçeklendirme işlemi, verilerin istatistiksel özelliklerini koruyarak onları aynı ölçeğe getirir. Bu sayede, veriler arasındaki farklılıkların etkisi azalır ve modelin daha istikrarlı ve tutarlı sonuçlar üretmesi sağlanır. Ölçeklendirilmiş veri seti, Genetik Algoritma tabanlı öznitelik seçimi ve Rastgele Orman Sınıflandırıcısı modelinin eğitimi için kullanılmıştır.

3. GENETİK ALGORİTMA VE ÖZNİTELİK SEÇİMİ

Genetik algoritma, öznitelik seçimi için bir optimizasyon yöntemi olarak kullanılmıştır. Öznitelik seçimi, veri setindeki en önemli öznitelikleri belirleyerek gereksiz veya tekrarlayan öznitelikleri elemeyi amaçlar. Bu şekilde, veri setinin boyutu azaltılarak hesaplama maliyeti düşürülür ve aynı zamanda modelin performansı ve genelleme yeteneği artırılır.

Genetik algoritma, doğal seleksiyon ve genetik işlemlerden esinlenerek tasarlanmış bir optimizasyon algoritmasıdır. Bu algoritma, bir popülasyonu temsil eden bireyler arasında genetik çaprazlama ve mutasyon işlemlerini uygulayarak yeni nesiller oluşturur. Her bir birey, özniteliklerin varlığını veya yokluğunu temsil eden bir kromozom olarak düşünülür.

Veri setinin hazırlandığı adımda, öznitelik seçimi için bir genetik algoritma uygulaması gerçekleştirdi. İlk olarak, belirli bir popülasyon büyüklüğü ve öznitelik sayısıyla başlayarak rastgele bir başlangıç popülasyonu oluşturuldu. Her bir birey, öznitelikleri temsil eden genetik kromozomlarla kodlandı.

Daha sonra, her bir bireyin uygunluk değeri hesaplandı. Uygunluk değeri, belirli bir bireyin özniteliklerin seçimiyle eğitilen Rastgele Orman sınıflandırıcısı tarafından verilen doğruluk skorunu temsil eder. Bu skor, doğru tahminlerin gerçek etiketlerle oranı olarak hesaplanır.

Genetik algoritmanın temel adımlarından biri, ebeveyn bireylerin seçimi ve çaprazlama işlemidir. Uygunluk değerlerine dayanarak ebeveyn bireyler seçilir ve çaprazlama işlemi

uygulanır. Çaprazlama işlemi, iki ebeveynin genetik materyalini birleştirerek yeni çocuk bireyler oluşturur.

Ayrıca, çaprazlama sonrasında mutasyon işlemi de uygulanır. Mutasyon, rastgele olarak seçilen genlerin değerini değiştirerek çeşitliliği artırır ve arama alanını genişletir. Bu sayede, daha geniş bir öznitelik uzayı keşfedilir ve çıkan sonuçların daha iyi olasılıklara sahip olması sağlanır.

Genetik algoritma döngüsü, belirli bir sayıda nesil üzerinden tekrarlanır. Her bir döngüde, en iyi uygunluk değerine sahip birey seçilerek test veri seti üzerinde sınıflandırma yapılır ve doğruluk değeri hesaplanır. Bu şekilde, genetik algoritmanın performansı ve öznitelik seçiminin etkinliği değerlendirilir.

Sonuç olarak, genetik algoritma ile öznitelik seçimi yöntemi kullanılarak MNIST veri setindeki el yazısı rakamlarının tanınması amaçlanmıştır. Bu yöntem, veri setindeki önemli öznitelikleri seçerek modelin doğruluğunu artırmış ve gereksiz özniteliklerin etkisini azaltmıştır. Elde edilen sonuçlar, genetik algoritmanın öznitelik seçimi için etkili bir yöntem olduğunu göstermiştir.

4. RASTGELE ORMAN SINIFLANDIRICISI

Rastgele Orman, makine öğrenmesi alanında yaygın olarak kullanılan bir sınıflandırma algoritmasıdır. Özellikle, çoklu karar ağaçlarının bir araya getirilmesiyle oluşturulan bir topluluk öğrenme yöntemidir. Rastgele Orman hem sınıflandırma hem de regresyon problemlerinde etkili sonuçlar veren bir algoritmadır.

Rastgele Orman, yüksek boyutlu ve karmaşık veri setlerinde etkili olan bir öznitelik seçimi yöntemi olarak kullanılabilir. Veri setindeki önemli öznitelikleri belirlemek için öznitelik önem sıralaması sağlar. Böylece, gereksiz veya az bilgi içeren özniteliklerin etkisi azaltılarak modelin performansı ve genelleme yeteneği artırılır.

Algoritmanın temel fikri, birden fazla karar ağacını eğitmek ve her bir ağacın tahminlerini birleştirerek sonuç üretmektir. Rastgele Orman, her bir ağacın eğitim veri setinin rastgele örneklemlerini ve rastgele özniteliklerini kullanarak eğitilmesini sağlar. Böylece, her bir ağacın birbirinden bağımsız ve çeşitli bir öğrenme yapısı oluşur.

Karar ağaçlarının bir araya getirilmesiyle oluşturulan Rastgele Orman, genel bir sınıflandırma modeli olarak çalışır. Bir veri örneği verildiğinde, tüm ağaçlardan gelen tahminlerin bir araya getirilmesiyle çoğunluk oyu yöntemiyle sınıflandırma yapılır. Bu sayede, ağaçların birbirini dengeleyerek hataları telafi etmesi ve daha güvenilir bir sınıflandırma sonucu elde edilmesi sağlanır.

Rastgele Orman, hızlı ve paralel çalışabilen bir algoritma olduğu için büyük veri setleri üzerinde etkili bir şekilde çalışabilir. Ayrıca, aşırı öğrenmeye karşı dirençli olması ve genel birleşme yeteneği sayesinde genelleme performansı yüksektir.

Bu çalışmada, Rastgele Orman sınıflandırıcısı, genetik algoritma ile öznitelik seçimi aşamasında kullanılmıştır. Genetik algoritma tarafından seçilen önemli öznitelikler kullanılarak Rastgele Orman sınıflandırıcısı eğitilmiş ve test veri seti üzerinde sınıflandırma performansı değerlendirilmiştir. Elde edilen sonuçlar, Rastgele Orman'ın öznitelik seçimi ve sınıflandırma görevlerinde başarılı bir şekilde kullanılabildiğini göstermiştir.

5. SONUÇLARIN ANALİZİ

Elde sonuçlara dayanarak, Genetik Algoritma ile Öznitelik Seçimi ve Rastgele Orman Sınıflandırıcısı tabanlı otomatik tanıma sisteminin performansı değerlendirilmiştir. Bu değerlendirme sonucunda; Beş jenerasyon boyunca gerçekleştirilen iterasyonlarda, en iyi uygunluk değeri (Best Fitness) sabit kalmış ve 0.11236666666666667 olarak kaydedilmiştir.
Bu süreçte, modelin sınıflandırma doğruluğu (Accuracy) ise değişmiştir. İlk iki jenerasyonda 0.9651 doğruluk değeri elde edilmiştir. Daha sonra üçüncü jenerasyonda doğruluk değeri 0.9669'a yükselmiştir. Ancak, dördüncü ve beşinci jenerasyonlarda doğruluk değeri biraz düşmüş ve sırasıyla 0.9636 ve 0.9637 olarak kaydedilmiştir.
Sonuç olarak, Genetik Algoritma ile Öznitelik Seçimi ve Rastgele Orman Sınıflandırıcısı yöntemiyle elde edilen en iyi uygunluk değeri sabit kalmış olsada, sınıflandırma doğruluğu biraz değişmiştir. Bu sonuçlar, öznitelik seçiminin sınıflandırma performansı üzerinde etkili olduğunu ve belirli iterasyonlarda daha yüksek doğruluk değerleri elde edilebildiğini göstermektedir.
Ancak, tüm jenerasyonlar dikkate alındığında, sonuçlar oldukça tatmin edicidir. Final sonucunda elde edilen en iyi uygunluk değeri 0.11236666666666667 ve doğruluk değeri 0.9665 olarak kaydedilmiştir. Bu sonuçlar, Genetik Algoritma ile Öznitelik Seçimi ve Rastgele Orman Sınıflandırıcısı tabanlı otomatik tanıma sisteminin el yazısı rakamları için etkili bir yöntem olduğunu göstermektedir.

```
Generation 1: Best Fitness = 0.11236666666666667, Accuracy = 0.9651
Generation 2: Best Fitness = 0.11236666666666667, Accuracy = 0.9651
Generation 3: Best Fitness = 0.11236666666666667, Accuracy = 0.9669
Generation 4: Best Fitness = 0.11236666666666667, Accuracy = 0.9636
Generation 5: Best Fitness = 0.11236666666666667, Accuracy = 0.9637
Final Result: Best Fitness = 0.11236666666666667, Accuracy = 0.9665
```