# BURSA TEKNİK ÜNİVERSİTESİ

# BLM0478 Derin Öğrenmeye Giriş Dersi
# Proje Raporu

## 2023-2024 Bahar


**Mehmet Emir ERDEM**
**20360859091**

**Mehmet Mert FİDAN**
**19360859018**

## 1. Giriş

BLM0478 Derin Öğrenmeye Giriş Dersi Projesi olarak bir makale seçilmesi ve bu makalede problem için uygulanan çözümün taklit edilerek tekrardan yapılması istenmiştir. Seçilen makale "LSTM Tabanlı Derin Ağlar Kullanılarak Diyabet Hastalığı Tahmini" olarak belirlenmiştir. Bu makalede UCI Machine Learning Repository web sitesinde bulunan Pima Indians Diabetes isimli veri kümesi kullanılmıştır. Yöntem olarak Evrişimli Sinir Ağları ve LSTM algoritmaları birlikte kullanılmıştır. Bu yöntem sonucunda %86,45 oranında doğruluk değeri elde edilmiştir. Bu yapılan çalışma diyabet teşhisinde literatürdeki diğer çalışmalara kıyasla yüksek performans sergilediği belirtilmektedir.

## 2. Veri Kümesi

Çalışmada kullanılan veri kümesi UCI Machine Learning Repository web sitesinde Pima Indians Diabetes veri kümesi olarak bulunmaktadır. Veri seti, 268 tanesi diyabet pozitif ve geri kalanı diyabet negatif olan 768 kadın hastanın kayıtlarını içermektedir. Veri setinde, bir hastanın diyabetli olup olmadığını tanısal olarak tahmin etmek için Gebelik sayısı, Glikoz, Kan Basıncı, Beden kitle indeksi, Deri Kalınlığı, İnsülin, Diyabet Soy ağacı ve Yaş gibi sekiz değişken ve sonuç olarak adlandırılan bir hedef değişken vardır. Veri setinin açıklaması Tablo 1'de verilmiştir. (Veri Setinin UCI sayfasına https://archive.ics.uci.edu/dataset/34/diabetes, Kaggle sayfasına https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database web sayfalarından ulaşılabilir.)

| No | Özellikler | Açıklama | Aralık |
|---|---|---|---|
| 1 | Gebelik sayısı | Bir katılımcının hamile kalma sayısı | 0–17 |
| 2 | Glikoz | Oral glikoz tolerans testinde 2. saat plazma glikoz konsantrasyonu | 0-199 |
| 3 | Kan Basıncı | Diyastolik kan basıncı | 0-122 |
| 4 | Beden kitle indeksi | Vücut kitle indeksi (kg cinsinden ağırlık / (m cinsinden yükseklik) | 0-67,1 |
| 5 | Deri Kalınlığı | Deri kıvrım kalınlığı | 0-99 |
| 6 | İnsülin | 2. Saatteki insülin değeri | 0-846 |
| 7 | Diyabet Soy ağacı | Diyabet prognozunda kullanılan bir özellik | 0,078–2,42 |
| 8 | Yaş | Katılımcıların yaşı | 21-81 |
| 9 | Sınıf | İkili değişken (0, 268 örneğin diyabetik olmadığını gösterir. 1 ise kalan 500 örneğin diyabetik olduğunu gösterir). | 1/0 |

*Tablo 1: Pima Indian Diabetes Veri Kümesi*

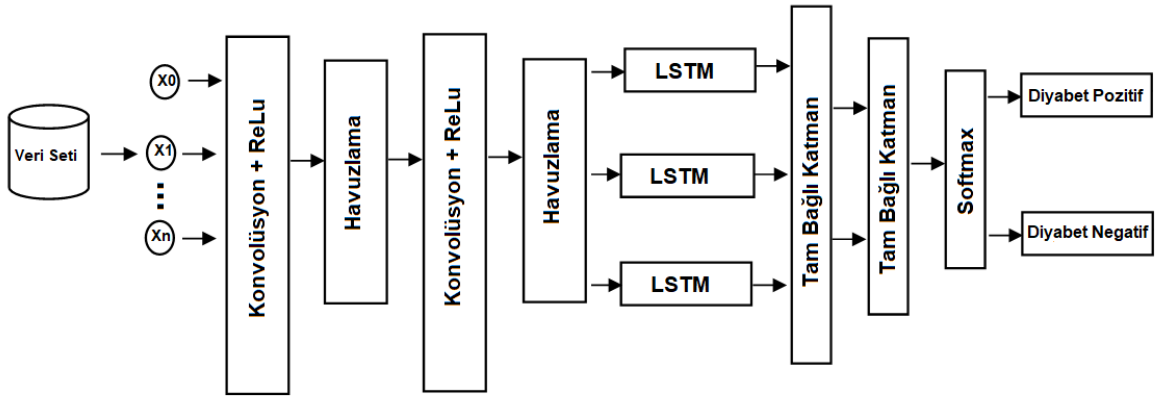
## 3. Kullanılan Yöntem

Önerilen yöntem 3 adımdan oluşmaktadır. Öncelikle veriler CNN katmanlarından geçirilmiştir. Daha sonra buradan elde edilen özellikler LSTM'e girdi olarak verilmiştir. Son

olarak Softmax katmanı kullanılarak sınıflandırma yapılmıştır. Önerilen Yöntem Şekil 3'te verilmiştir. Tasarlanan CNN mimarisi, 1 giriş katmanı, 2 konvolüsyon katmanı, 2 havuzlama katmanı, 2 tam bağlantılı katmanı ve 1 çıkış katmanından oluşur. Tasarlanan mimaride toplam 6 katman bulunmaktadır. Birinci konvolüsyon katmanında 128 adet konvolüsyon filtresi, ikinci konvolüsyon katmanında 64 adet konvolüsyon filtresi bulunmaktadır. Aktivasyon fonksiyonu olarak ReLu kullanılmıştır. Ayrıca her konvolüsyon ve ReLu işleminden sonra 2x1 boyutunda maksimum havuzlama yapılmıştır. Konvolüsyon katmanlarından sonra veriler LSTM katmanından geçirilmiştir. Son olarak tamamen bağlantı katmanlar ve Softmax bağlanarak model tamamlanmıştır. Tasarlanan CNN mimarisi Tablo 2'de verilmiştir.

| No | Katman İsmi | Açıklama | Özellikler |
|---|---|---|---|
| 1 | Giriş | Giriş vektörü | - |
| 2 | 'conv1' | Konvolüsyon | 1 boyutunda 128 adet konvolüsyon filtresi |
| 3 | 'relu1' | ReLu | - |
| 4 | 'pool1' | Maksimum Havuzlama | 2x1 havuzlama |
| 5 | 'conv2' | Konvolüsyon | 1 boyutunda 64 adet konvolüsyon filtresi |
| 6 | 'relu2' | ReLu | - |
| 7 | 'pool2' | Maksimum Havuzlama | 2x1 havuzlama |
| 8 | 'fc1' | Tam Bağlı Katman | 1024 nöron |
| 9 | 'relu7' | ReLu | - |
| 10 | 'drop7' | Dropout | 50% dropout |
| 11 | 'fc2' | Tam Bağlı Katman | 512 nöron |
| 12 | Çıkış | Softmax | 2 sınıf |

*Tablo 2: Tasarlanan CNN mimarisi*



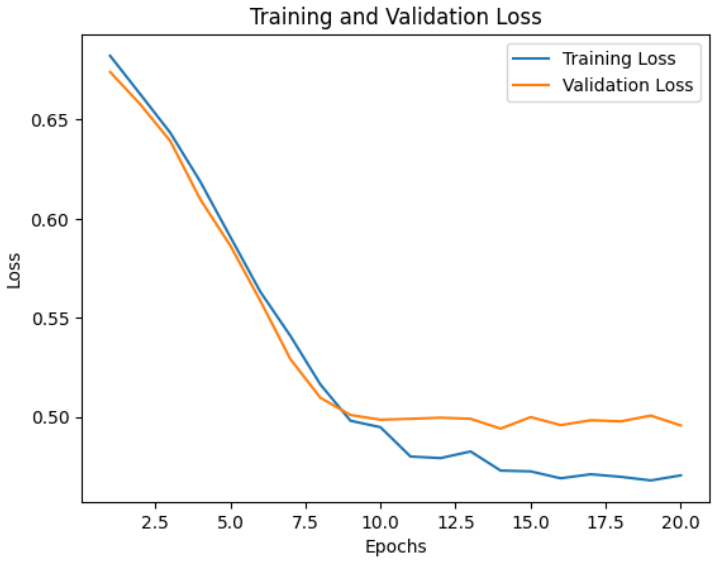

*Şekil 1: Önerilen Yöntem*

## 4. Uygulama

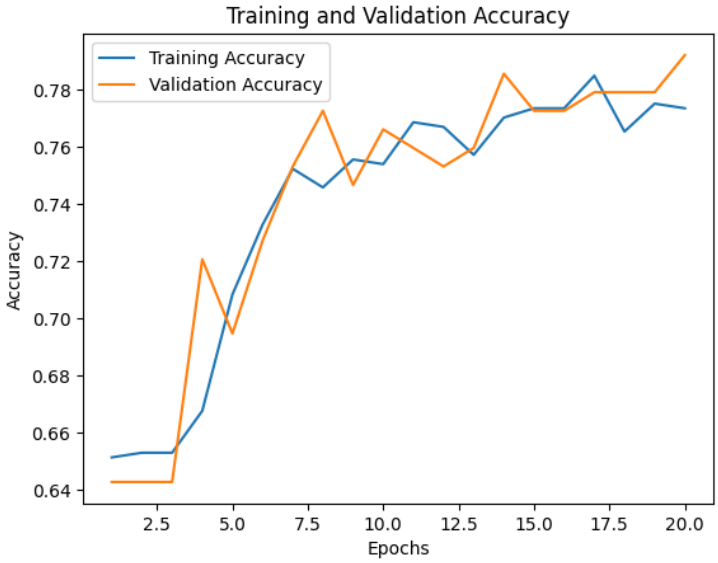
Projede kullanılan yöntemler taklit edilerek Python üzerinden denemeler gerçekleştirilmiştir. Projede CNN, LSTM ve CNN+LSTM şeklinde 3 adet yöntem kullanılmıştır. Ekstra olarak veri seti dağılımında %70 train %30 test olarak ve %80 train %20 test olarak 2 şekilde olmak üzere eğitimler denenmiştir. Bu durumda 6 farklı deneme yapıldığı söylenebilir.

CNN ve CNN+LSTM neredeyse aynı şekilde uygulandığı söylenebilir. Sadece Yöntemde de belirtildiği üzere 3 adet paralel şekilde 1 LSTM katmanı ikinci havuzlama katmanı ve birinci tam bağlı katman arasına eklenmiştir. Öncelikle csv dosyasında bulunan veriler koda aktarılır. Veri kümesindeki etiketler ve öznitelikler ayrılır. Veri kümesindeki eğitim ve test seti oranı %80 train %20 test olarak ayarlanır. Veriyi standartlaştırma işlemi yapılır. Bu projede StandardScaler() tercih edilmiştir. StandardScaler() işleminde her bir özelliğin ortalaması 0, standart sapması 1 olacak şekilde verileri standartlaştırır. Bu durumda modelin öğrenmesi verilerin standartlaştırılmış olması sebebiyle iyileşir. Daha sonra modelin giriş katmanı oluşturulur. Verilerin öznitelikleri sayısı boyutunda girdi alınır. Sırasıyla 128 konvolüsyon filtresinden oluşan, Kernel boyutu 3 olarak belirlenmiş ve aktivasyon fonksiyonu olarak relu kullanılmıştır. Maxpooling işlemi havuzlama boyutu 2 olacak şekilde uygulanmıştır. Tekrardan aynı özelliklere sahip konvolüsyon katmanı 64 konvolüsyon filtresi olacak şekilde uygulanmıştır. Aynı özellikteki başka bir maxpooling katmanı tekrardan uygulanmıştır. Buradaki çıktılar 3 adet paralel LSTM katmanına verilmiştir. Her bir LSTM hücre grubu aynı girişleri almıştır. Burada oluşan çıktı düzleştirme (flatten) katmanına verilmiştir. Düzleştirme katmanı tam bağlı katmanlara geçiş için gereklidir. 1024 nörona sahip relu aktivasyon fonksiyonlu tam bağlı katman eklenmiştir. %50 oranında Dropout eklenmiştir. Bu fonksiyon gelen çıktıların %50 oranında rastgele bir şekilde devre dışı bırakır. 512 nörona ve relu aktivasyon fonksiyonuna sahip tam bağlı katman eklenmiştir. Çıkış katmanı olarak softmax aktivasyon fonksiyonuyla beraber 2 nöronlu bir çıkış uygulanmıştır. Bunun sebebi diyabet var veya diyabet yok olarak sınıflandırma olması sebebiyledir. Model oluşturulduktan sonra Adam optimizasyon algoritmasıyla beraber 0.0001 öğrenme oranı kullanılmıştır. Loss fonksiyonu olarak sparse_categorical_crossentropy tercih edilmiştir. Çıkışın binary olmasına rağmen softmax kullanılması sebebiyle categorical_crossentropy kullanılması gerekmektedir. One-hot Encoder veya herhangi bir encoder kullanılmaması sebebiyle sparse_categorical_crossentropy kullanılmaktadır. Veriler başta standartlaştırılma işleminden geçirilmesi sebebiyle ve Conv1D fonksiyonu 3 boyutlu tensör girdi olarak beklemesi sebebiyle verilere reshape işlemi uygulanır. Bu sayede veriler 3 boyutlu bir tensöre dönüştürülür. Model.fit fonksiyonunda epoch sayısı 20 ve batch boyutu 64 olarak ayarlandıktan sonra eğitim başlatılır. Gerekli grafikler değerler ve matrisler çizdirilir.

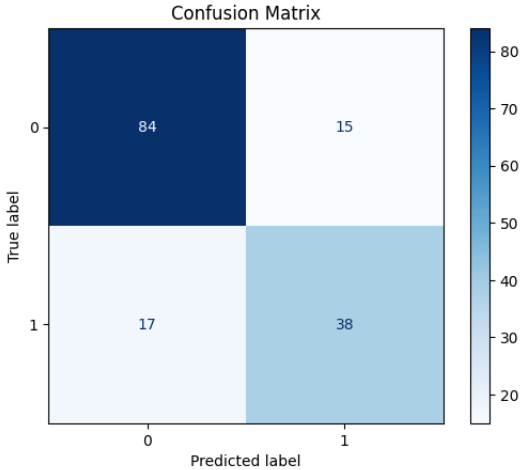
Yapılan CNN+LSTM uygulamasındaki çıktı grafikleri aşağıdaki gibidir.




*Şekil 2: CNN+LSTM Loss Grafiği*

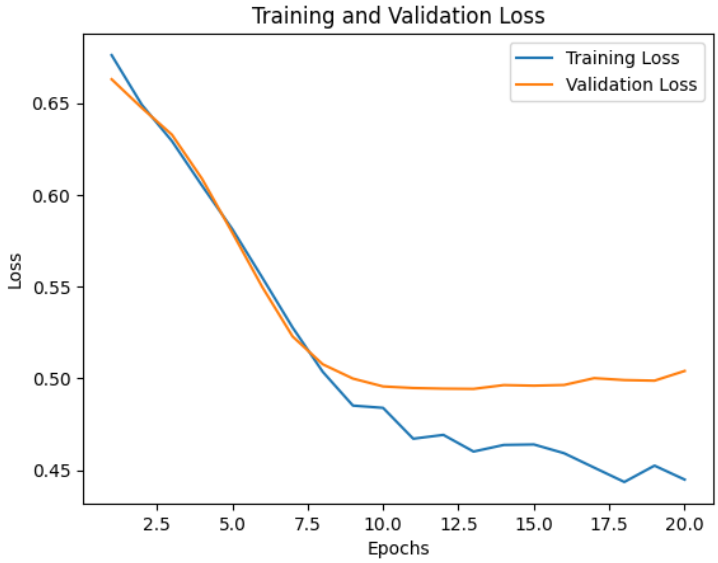



*Şekil 3: CNN+LSTM Accuracy Grafiği*

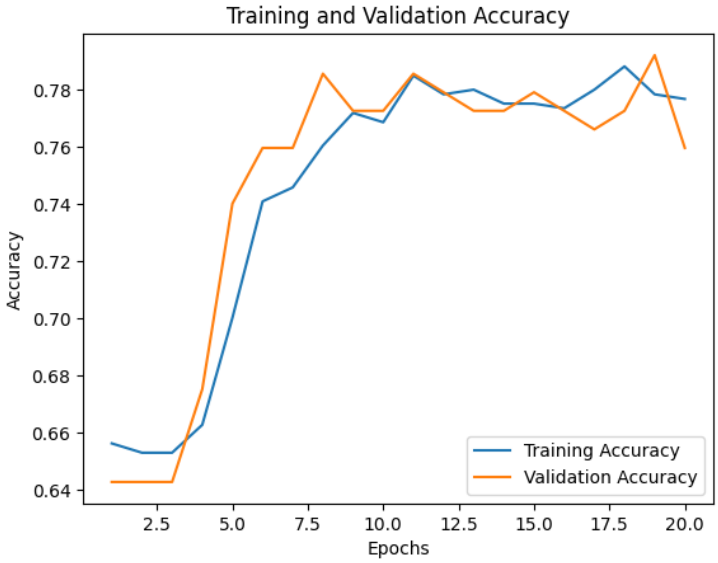



*Şekil 4: CNN+LSTM Karmaşıklık Matrisi*

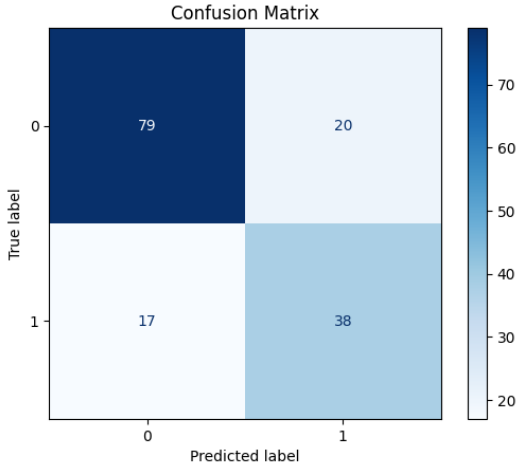
CNN+LSTM uygulamasının sadece CNN yapılmış versiyonu LSTM katmanının çıkarıldığı versiyondur. CNN ile yapılmış uygulamanın çıktıları aşağıdaki gibidir.




*Şekil 5: CNN Loss Grafiği*



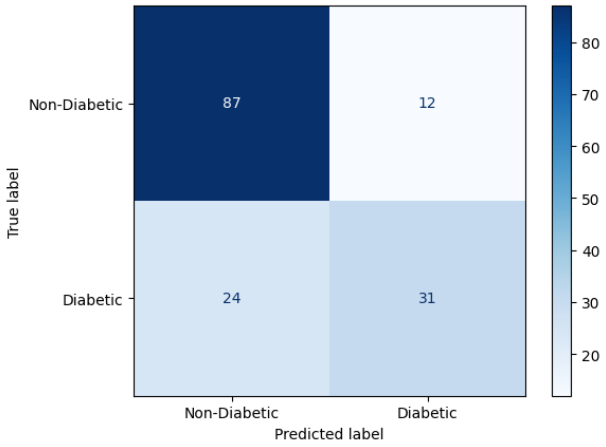

*Şekil 6: CNN Accuracy Grafiği*




*Şekil 7: CNN Karmaşıklık Matrisi*

Sadece LSTM kullanılan uygulamada işleyiş biraz daha farklıdır. Öncelikle makalede LSTM'in uygulamalarında nasıl kullanıldığı ile ilgili detaylı bir bilgi verilmemiştir. Bu sebeple tahmini şekilde bir uygulama yazılmıştır. Veriler yüklenmiştir. Etiketler ve öznitelikler ayrılmıştır. Veri StandardScale() ile standartlaştırılmıştır. Veri seti oranı %80 train %20 test olarak ayarlanmıştır. Veriler reshape fonksiyonuyla uygun input için uygun formata ayarlanmıştır. 2 adet 64 hücrelik LSTM katmanı kullanılmıştır. %50 oranda dropout eklenmiştir. Çıkış fonksiyonu olarak sigmoid kullanılmıştır. Bu sebeple loss fonksiyonu binary_crossentropy seçilmiştir. Optimizasyon algoritması olarak Adam seçilmiştir. 20 epoch sayısı ve 64 batch boyutu ile eğitilmiştir. Daha sonra accuracy loss grafikleri çizdirilip, classification report alınmıştır.


*Şekil 8: LSTM Loss ve Accuracy Grafikleri*




*Şekil 9: LSTM Karmaşıklık Matrisi*

## 5. Sonuçlar

Genel olarak skorların değerlendirilmesine bakıldığında %20 test veri olarak ayrılmış modellerde sonuçlar aşağıdaki gibidir.

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Non-Diabetic | 0.83 | 0.85 | 0.84 | 99 |
| Diabetic | 0.72 | 0.69 | 0.70 | 55 |
| accuracy | | | 0.79 | 154 |
| macro avg | 0.77 | 0.77 | 0.77 | 154 |
| weighted avg | 0.79 | 0.79 | 0.79 | 154 |

*Şekil 10: CNN+LSTM Sınıflandırma Raporu %20*

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Non-Diabetic | 0.82 | 0.80 | 0.81 | 99 |
| Diabetic | 0.66 | 0.69 | 0.67 | 55 |
| accuracy | | | 0.76 | 154 |
| macro avg | 0.74 | 0.74 | 0.74 | 154 |
| weighted avg | 0.76 | 0.76 | 0.76 | 154 |

*Şekil 11: CNN Sınıflandırma Raporu %20*

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Non-Diabetic | 0.78 | 0.88 | 0.83 | 99 |
| Diabetic | 0.72 | 0.56 | 0.63 | 55 |
| accuracy | | | 0.77 | 154 |
| macro avg | 0.75 | 0.72 | 0.73 | 154 |
| weighted avg | 0.76 | 0.77 | 0.76 | 154 |

*Şekil 12: LSTM Sınıflandırma Raporu %20*

Veri seti oranının %30 test olarak ayrılma durumunda elde edilen sonuçlar aşağıdaki gibidir.

| Classification Report: | precision | recall | f1-score | support |
|---|---|---|---|---|
| Non-Diabetic | 0.82 | 0.79 | 0.81 | 151 |
| Diabetic | 0.64 | 0.68 | 0.65 | 80 |
| accuracy | | | 0.75 | 231 |
| macro avg | 0.73 | 0.73 | 0.73 | 231 |
| weighted avg | 0.76 | 0.75 | 0.75 | 231 |

*Şekil 13: CNN+LSTM Sınıflandırma Raporu %30*

| Classification Report: | precision | recall | f1-score | support |
|---|---|---|---|---|
| Non-Diabetic | 0.82 | 0.74 | 0.78 | 151 |
| Diabetic | 0.59 | 0.69 | 0.63 | 80 |
| accuracy | | | 0.72 | 231 |
| macro avg | 0.70 | 0.71 | 0.70 | 231 |
| weighted avg | 0.74 | 0.72 | 0.73 | 231 |

*Şekil 14: CNN Sınıflandırma Raporu %30*

| Classification Report: | precision | recall | f1-score | support |
|---|---|---|---|---|
| Non-Diabetic | 0.78 | 0.84 | 0.81 | 151 |
| Diabetic | 0.65 | 0.55 | 0.59 | 80 |
| accuracy | | | 0.74 | 231 |
| macro avg | 0.71 | 0.70 | 0.70 | 231 |
| weighted avg | 0.73 | 0.74 | 0.73 | 231 |

*Şekil 15: LSTM Sınıflandırma Raporu %30*

Genel olarak sonuçlara bakıldığında test verisinin %20 olarak ayarlandığı uygulamalarda sonuçların daha iyi olduğu gözlemlenmektedir. Yöntem olarak bakılırsa en iyi yöntem CNN+LSTM olarak görülmektedir. Makalede elde edilen sonuçlar aşağıdaki gibidir.

| Verinin Eğitim ve Test için Farklı oranlarda Bölünmesi | Model | Doğruluk % | Kesinlik % | F-skoru % |
|---|---|---|---|---|
| %70- %30 | ESA | 82,47 | 83,14 | 83,56 |
| | LSTM | 83,77 | 84,46 | 84,23 |
| | ESA+LSTM | 85,21 | 84,94 | 58,14 |
| %80- %20 | ESA | 83,25 | 83,56 | 83,69 |
| | LSTM | 85,21 | 85,33 | 85,41 |
| | ESA+LSTM | 86,45 | 87,00 | 88,23 |

*Şekil 16: Makalede Elde Edilen Değerler*

Uygulamada elde edilen sonuçlar ve makaledeki sonuçlar farklıdır. Bunun sebebi makalede bütün uygulamanın en ince ayrıntısına kadar verilmemesidir. Dışarıdan biri uygulama sonucunda tahmin ve deneme yanılma yöntemiyle ilerlemesi sebebiyle elde edilen değerler değişebilir. Ayrıca genel olarak accuracy değeri makaleden daha yüksek de yapılabilir fakat veri kümesinde 768 verinin 268 diyabet ve 500 diyabet olmayan veri olması sebebiyle problem vardır. Çünkü veri kümesindeki etiket sayıları eşit derecede paylaşılmamıştır.

## 6. Kaynaklar

- Kullanılan Diabetes Veri Kümesi Bağlantıları: https://archive.ics.uci.edu/dataset/34/diabetes ve https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database bağlantılarından ulaşılabilir.