T.C.
DÜZCE ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GELİŞTİRİLEN YENİ FİLTRELERİN VE TEMEL FREKANS TESPİT YÖNTEMİNİN DERİN ÖĞRENME İLE KONUŞMA DUYGU ANALİZİNDE UYGULANMASI

CEVAHİR PARLAK

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
DOÇ. DR. YUSUF ALTUN

DÜZCE, 2022

T.C.
DÜZCE ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GELİŞTİRİLEN YENİ FİLTRELERİN VE TEMEL FREKANS TESPİT YÖNTEMİNİN DERİN ÖĞRENME İLE KONUŞMA DUYGU ANALİZİNDE UYGULANMASI

Cevahir PARLAK tarafından hazırlanan tez çalışması aşağıdaki jüri tarafından Düzce Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Yusuf ALTUN

Düzce Üniversitesi

Jüri Üyeleri

Doç. Dr. Yusuf ALTUN
Düzce Üniversitesi _____

Doç. Dr. Selman KULAÇ
Düzce Üniversitesi _____

Doç. Dr. Abdullah Talha KABAKUŞ
Düzce Üniversitesi _____

Dr. Öğretim Üyesi Murat İSKEFİYELİ
Sakarya Üniversitesi _____

Dr. Öğr. Üyesi Veli BAYSAL
Bartın Üniversitesi _____

Tez Savunma Tarihi: 21/06/2022

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün aşamalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı, yine bu tezin çalışılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

21 Haziran 2022

Cevahir Parlak

TEŞEKKÜR

Doktora öğrenimimde ve bu tezin hazırlanmasında gösterdiği her türlü destek ve yardımdan dolayı çok değerli hocam Doç. Dr. Yusuf ALTUN'a en içten dileklerimle teşekkür ederim.

Tez çalışmam boyunca değerli katkılarını esirgemeyen Tez İzleme Komitesi üyeleri Doç. Dr. Selman KULAÇ ve Doç. Dr. Abdullah Talha KABAKUŞ'a da şükranlarımı sunarım.

Bu çalışma boyunca yardımlarını ve desteklerini esirgemeyen sevgili aileme, çalışma arkadaşlarıma, Lisansüstü Eğitim Enstitüsü çalışanlarına sonsuz teşekkürlerimi sunarım.

21 Haziran 2022 **Cevahir Parlak**

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ vii
ÇİZELGE LİSTESİ ix
KISALTMALAR xii
SİMGELER xv
ÖZET xvi
ABSTRACT xvii
EXTENDED ABSTRACT xviii
1. GİRİŞ 1
2. İLGİLİ ÇALIŞMALAR 5
3. METOTLAR 15
3.1. NVIDIA CNN MODELİ 16
3.2. 1B KONVOLUSYON KATMANLI CNN MODELİ 16
3.3. LSTM 16
3.4. SVM 18
3.5. ÖNERİLEN EFB FİLTRE BANKASI 19
3.6. ÖNERİLEN HDM METODU VE KULLANILAN METOTLAR 20
3.6.1. HDM 22
3.6.2. Otokorelasyon 22
3.6.3. Kepstrum 23
3.6.4. YIN 23
3.6.5. YAAPT 24
3.6.6. CREPE 24
3.6.7. FCN 24
4. VERİSETLERİ VE DENEYSEL KURULUM 25
4.1. EMOSTAR 25
4.2. EMODB 26
4.3. IEMOCAP 26
4.4. MELD 27
4.5. TEMEL FREKANS VERİSETLERİ 28
4.6. ÖZNİTELİK SEÇME 30
4.6.1. Information Gain Öznitelik Seçici 30
4.6.2. CFS Subset Öznitelik Seçici 30
4.7. DENGESİZ DAĞILIMLI VERİ İŞLEME YÖNTEMLERİ 31
5. SONUÇLAR 32
5.1. TÜM ÖZNİTELİKLERLE YAPILAN DENEYLER 32
5.2. INFORMATION GAIN ÖZNİTELİK SEÇİCİ DENEYLERİ 39
5.3. CFS SUBSET EVALUATOR ÖZNİTELİK SEÇİCİ DENEYLERİ 46

5.4. VERİ TÜRETME DENEYLERİ 52
5.5. DENGESİZ DAĞILIMLI VERİSETİ DENEYLERİ 57
5.5.1. SMOTE, Veri Türetme ve Veri Azaltma Deneyleri 57
5.5.2. AdaboostM1, Bagging ve SVM Karşılaştırması 60
5.6. TEMEL FREKANS DENEYLERİ 60
5.6.1. Genişbant ve Darbant Temel Frekans Deneyleri 62
5.6.2. Temel Frekans ile Cinsiyet Tespit 64
6. SONUÇLAR VE GELECEK ÇALIŞMALAR 66
7. KAYNAKLAR 68
ÖZGEÇMİŞ 78

ŞEKİL LİSTESİ

Sayfa No

Şekil 1.1. MFCC ve Mel filtreleri hesaplama adımları. 2
Şekil 1.2. EFB hesaplama adımları. 3
Şekil 2.1. Farklı LSTM modellerin karşılaştırılması [67]. 10
Şekil 3.1 Önerilen EFB filtrelerinin görselleştirilmesi. 20
Şekil 5.1. 7-duygulu EmoDB üzerindeki deney sonuçlarının çubuk grafikleri. 33
Şekil 5.2. 7-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri. 34
Şekil 5.3. 4-duygulu EmoDB üzerindeki deney sonuçlarının çubuk grafikleri. 34
Şekil 5.4. 4-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri. 35
Şekil 5.5. 4-duygulu EmoSTAR üzerindeki deney sonuçlarının çubuk grafikleri. 35
Şekil 5.6. 4-duygulu EmoSTAR üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri. 36
Şekil 5.7. 4-duygulu IEMOCAP üzerindeki deney sonuçlarının çubuk grafikleri. 36
Şekil 5.8. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri. 37
Şekil 5.9. 4-duygulu IEMOCAP üzerindeki deney sonuçlarının çubuk grafikleri. 37
Şekil 5.10. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri. 38
Şekil 5.11. 4-duygulu MELD üzerindeki deney sonuçlarının çubuk grafikleri. 39
Şekil 5.12. 4-duygulu MELD üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri. 39
Şekil 5.13. 7-duygulu EmoDB üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 40
Şekil 5.14. 7-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 41
Şekil 5.15. 4-duygulu EmoDB üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 41
Şekil 5.16. 4-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 42
Şekil 5.17. 4-duygulu EmoSTAR üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 43
Şekil 5.18. 4-duygulu EmoSTAR üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 43
Şekil 5.19. 4-duygulu IEMOCAP üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 44
Şekil 5.20. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 44
Şekil 5.21. 4-duygulu MELD üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 45
Şekil 5.22. 4-duygulu MELD üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri. 46
Şekil 5.23. 7-duygulu EmoDB üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 47
Şekil 5.24. 7-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

.....47
Şekil 5.25. 4-duygulu EmoDB üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 48
Şekil 5.26 4-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 49
Şekil 5.27. 4-duygulu EmoSTAR üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 49
Şekil 5.28. 4-duygulu EmoSTAR üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 50
Şekil 5.29. 4-duygulu IEMOCAP üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 50
Şekil 5.30. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 51
Şekil 5.31. 4-duygulu MELD üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 52
Şekil 5.32. 4-duygulu MELD üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri. 52
Şekil 5.33. 4-duygulu IEMOCAP verisetinde NVIDIA CNN ve SVM ile veri türetme sonuçları. 53
Şekil 5.34. CNN2 modeli ve EFB filtreleri ile EmoSTAR veri kümesi için başarı oranı ve kayıp grafikleri. 55
Şekil 5.35. CNN2 modeli ve EFB filtreleri ile EmoDB veri kümesi için başarı oranı ve kayıp grafikleri. 55
Şekil 5.36. CNN2 modeli ve EFB filtreleri ile IEMOCAP veri kümesi için başarı oranı ve kayıp grafikleri.....56

ÇİZELGE LİSTESİ

Sayfa No

Çizelge 2.1. En İyi Ksa Modelinin Diğer Ysa Modelleri İle Karşılaştırılması [36]. 5
Çizelge 2.2. Çeşitli Verisetlerinde Lvcsr Sonuçları [39].....6
Çizelge 2.3. 4-Duygulu Iemocap Verisetinde Sınıflandırma Sonuçları [55].....7
Çizelge 2.4. Doğaçlama 4-Duygulu Iemocap Veri Kümesindeki Mevcut Yöntemlerin Doğruluk Karşılaştırması [60]. 8
Çizelge 2.5. IEMOCAP Verisetinde Modellerin Başarı Oranları [64]. 9
Çizelge 2.6. IEMOCAP Verisetinde Farklı Modellerin Başarı Oranları [66]. 9
Çizelge 3.1. NVIDIA CNN Modeli. 16
Çizelge 3.2. 1B CNN Modelinin Ayrıntılı Şeması. 17
Çizelge 3.3. Deneylerimizde Kullanılan LSTM Modeli. 18
Çizelge 3.4. Deneylerimizde Kullanılan Bidirectional LSTM Modeli. 18
Çizelge 3.5. Önerilen EFB Filtre Bankalarının Frekans Bölgeleri. 19
Çizelge 4.1. Emostar Verisetinde Duygu Sınıflarının Dağılımı (T=Türkçe, İ=İngilizce) [107]. 25
Çizelge 4.2. EmoDB Verisetinde Duygu Dağılımı [107].....26
Çizelge 4.3. IEMOCAP Veri Kümesinde Duygulara Göre Örnek Sayıları. 27
Çizelge 4.4. MELD Verisetindeki Duygu Sınıflarının Dağılımı. 28
Çizelge 4.5. Temel Frekans Deneylerinde Kullanılan Verisetleri. 29
Çizelge 5.1. 7-Duygulu EmoDB Üzerindeki Deney Sonuçları (%ACC).....33
Çizelge 5.2. 7-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC). 34
Çizelge 5.3. 4-Duygulu EmoDB Üzerindeki Deney Sonuçları (%ACC).....34
Çizelge 5.4. 4-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC). 35
Çizelge 5.5. 4-Duygulu EmoSTAR Üzerindeki Deney Sonuçları (%ACC). 35
Çizelge 5.6. 4-Duygulu EmoSTAR Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC). 36
Çizelge 5.7. 4-Duygulu IEMOCAP Üzerindeki Deney Sonuçları (%ACC). 36
Çizelge 5.8. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC). 37
Çizelge 5.9. 4-Duygulu IEMOCAP Üzerindeki Deney Sonuçları (%ACC). 37
Çizelge 5.10. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC). 38
Çizelge 5.11. 4-Duygulu MELD Üzerindeki Deney Sonuçları (%ACC). 38
Çizelge 5.12. 4-Duygulu MELD Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC). 39
Çizelge 5.13. 7-Duygulu EmoDB Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 40
Çizelge 5.14. 7-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 41
Çizelge 5.15. 4-Duygulu EmoDB Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 41
Çizelge 5.16. 4-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 42
Çizelge 5.17. 4-Duygulu EmoSTAR Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 42
Çizelge 5.18. 4-Duygulu EmoSTAR Üzerindeki Prozodik Öznitelikler Eklenerek

Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 43
Çizelge 5.19. 4-Duygulu IEMOCAP Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 44
Çizelge 5.20. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 44
Çizelge 5.21. 4-Duygulu MELD Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 45
Çizelge 5.22. 4-Duygulu MELD Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC). 45
Çizelge 5.23. 7-Duygulu EmoDB Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 46
Çizelge 5.24. 7-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 47
Çizelge 5.25. 4-Duygulu EmoDB Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 48
Çizelge 5.26. 4-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 48
Çizelge 5.27. 4-Duygulu EmoSTAR Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 49
Çizelge 5.28. 4-Duygulu EmoSTAR Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 50
Çizelge 5.29. 4-Duygulu IEMOCAP Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 50
Çizelge 5.30. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 51
Çizelge 5.31. 4-Duygulu MELD Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 51
Çizelge 5.32. 4-Duygulu MELD Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC). 52
Çizelge 5.33. 4-Duygulu Doğaçlama IEMOCAP Veri Kümesinde 7 Kata Kadar Veri Türetme Başarı Oranı (% ACC) Sonuçları. 53
Çizelge 5.34. 4-Duygulu 4490 Örnekli IEMOCAP Veri Kümesinde Modellerin Saniye Olarak CPU Sınıflandırma Hızları (Saniye). 54
Çizelge 5.35. 4-Duygulu 4490 Örnekli IEMOCAP Veri Kümesinde Modellerin Saniye Olarak GPU Sınıflandırma Hızları (Saniye). 54
Çizelge 5.36. 4-Duygulu EmoDB Veri Kümesinde Milisaniye Olarak Ortalama Öznitelik Çıkarma Hızları. 55
Çizelge 5.37. CNN2 Sınıflandırıcısı Ve EFB Filtreleri İle EmoSTAR, EmoDB, IEMOCAP Ve MELD Verisetlerindeki Karmaşıklık Matrisleri. 57
Çizelge 5.38. SMOTE, Veri Türetme, Veri Azaltma Uygulanmış Ve Orijinal 4-Duygulu MELD Veri Kümesindeki Örnek Sayıları. 57
Çizelge 5.39. SMOTE, Veri Türetme, Veri Azaltma Ve Orijinal MELD Verisetlerinin NVIDIA CNN Ve EFB Öznitelik Kümesi İle Elde Edilen Karmaşıklık Matrisleri. 58
Çizelge 5.40. EFB, MFCC Ve Mel Filtre Öznitelik Setlerinde 4-Duygulu MELD Veri Kümesinde NVIDIA CNN (CNN1) İle SMOTE, Veri Türetme (Tür), Veri

Azaltma (Az) Ve Orijinal (Orj) Veri Sonuçları. 59
Çizelge 5.41. EFB, MFCC Ve Mel Filtre Öznitelik Kümelerini Kullanan 4-Duygulu MELD Veri Kümesinde Adaboostm1, Bagging Ve SVM Sınıflandırıcılarının Sonuçları. 60
Çizelge 5.42 Hillenbrand, Teksas Ve TIMIT Veri Kümeleri İçin Cinsiyete Göre Ortalama F0 Değerleri. 61
Çizelge 5.43. Hillenbrand Verisetinde Uygulanan Temel Frekans Tespit Yöntemlerinin Saniye Olarak Ortalama Hızları.....62
Çizelge 5.44. Genişbant Hillenbrand, Teksas Ve TIMIT Verisetlerindeki Sonuçlar. 63
Çizelge 5.45. Darbant Hillenbrand, Teksas Ve TIMIT Veri Kümelerinde Deneysel Sonuçlar. 63
Çizelge 5.46. Genişbant TIMIT Veri Kümesinde Cinsiyet Algılama Sonuçları. 64
Çizelge 5.47. Darbant TIMIT Veri Kümesinde Cinsiyet Algılama Sonuçları. 65

KISALTMALAR

ABO	Ağırlıksız Başarı Oranı
AC	Autocorrelation
ACC	Accuracy
Adaboost	Adaptive Boosting
ADT	Audio Degradation Toolbox
AMDF	Average Magnitude Difference Function
AUC	Area Under Curve
BERT	Bidirectional Encoder Representations From Transformers
BiRNN	Bidirectional Recurrent Neural Network
BO	Başarı Oranı
BREF	Large Vocabulary Spoken Corpus for French
CEMO	Emotional Corpora for Emergency Call Centers
CFS	Correlation Based Feature Selection
CNN	Convolutional Neural Network
Conv1d	1-Dimensional Convolution
Conv2d	2-Dimensional Convolution
CPU	Central Processing Unit
CREPE	Convolutional Representation for Pitch Estimation
CTC	Connectionist Temporal Unit
DCT	Discrete Cosine Transform
DHBB	Deep Hierarchical Dual BiLSTM
DHBL	Deep Hierarchical BiLSTM and LSTM
DHLB	Deep Hierarchical LSTM and BiLSTM
DHLL	Deep Hierarchical LSTM and LSTM
DNN	Deep Neural Network
EFB	Emotional Filter Banks
EGG	Electroglottograph
EIH	Ensemble Interval Histograms
EmoDB	Berlin Emotional Database
f0	Temel Frekans
FC	Fully Connected
FCN	Fully Convolutional Network
FFN	Feed Forward Network
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GloVe	Global Vectors for Word Representations
GMM	Gaussian Mixture Models
GPE	Gross Pitch Error
GPU	Graphics Unit Interface
GRU	Gated Recurrent Unit
HDM	Harmonic Differences Method
HMM	Hidden Markov Model
Hz	Hertz
IEMOCAP	Interactive Emotional Dyadic Motion Capture Database

IFFT	Inverse Fast Fourier Transform
JPEG	Joint Photographic Experts Group
KHO	Kelime Hata Oranı
kHz	Kilo Hertz
KSA	Konvolusyonal Sinir Ağı
LDA	Linear Discriminant Analysis
LDVT	Lincoln Digital Voice Terminal
Librosa	Laboratory for the Recognition and Organization of Speech and Audio
LogMel	Logarithmic Mel
LPC	Linear Predictive Coding
LSTM	Long Short-Term Memory
LVCSR	Large Scale Continuous Speech Recognition
MAT	Modulated Attention Transformer
mcRBM	Mean Covariance Restricted Boltzmann Machine
MELD	Multimodal EmotionLines Dataset
MFCC	Mel Filter Cepstral Coefficients
MIT	Massachusetts Institute of Technology
MNT	Modulated Normalization Transformer
NCCF	Normalized Cross Correlation Function
NLPAUG	Natural Language Processing Augmentation
NT	Naïve Transformer
P	Projection
PCM	Pulse Code Modulation
PLP	Perceptual Linear Prediction
PR AUC	Precision Recall Area Under Curve
PTDB-TUG	Pitch Tracking Database-Technical University of Graz
pYIN	Probabilistic YIN
RAPT	Robust Algorithm for Pitch Detection
RBM	Restricted Boltzmann Machine
RELU	Rectified Linear Unit
RMSprop	Root Mean Squared Propagation
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
RWC	Real World Computing
SAIL	Stanford Artificial Intelligence Laboratory
SAIL	Speech Analysis and Interpretation Laboratory University of Southern California
SGD	Stochastic Gradient Descent
SHO	Ses Hata Oranı
SHRP	Subharmonic to Harmonic Ratio Pitch Detector
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Oversampling Technique
SNR	Sound-to-Noise Ratio
SPIRE	Surface Prior Information Reflectance Estimation
STRAIGHT	Speech Transformation and Representation based on Adaptive Interpolation of Weighted spectrogram

SVM	Support Vector Machines
Tanh	Tangent Hyperbolic
TDNN	Time Delay Neural Network
TEMPO	Time Domain Excitation Extraction Based on a Minimum Perturbation Operator
TIMIT	Texas Instrument Massachusetts Institute of Technology
UA	Unweighted Average
VDE	Voice Detection Error
VTLP	Vocal Tract Length Perturbation
WA	Weighted Average
WER	Word Error Rate
WSOLA	Waveform Similarity Overlap Add
XGBoost	Extended Gradient Boost
YAAPT	Yet Another Algorithm for Pitch Tracking
YSA	Yapay Sinir Ağları

SİMGELER

α	Önvurgu katsayısı
e_{10}	Gerçek değerden yüzde 10'dan fazla sapan örnek sayısı
$\hat{f}_j$	Tahmin edilen frekans değeri
f_j	Gerçek frekans değeri
f0	Temel Frekans

ÖZET

GELİŞTİRİLEN YENİ FİLTRELERİN VE TEMEL FREKANS TESPİT YÖNTEMİNİN DERİN ÖĞRENME İLE KONUŞMA DUYGU ANALİZİNDE UYGULANMASI

Cevahir PARLAK
Düzce Üniversitesi
Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı
Doktora Tezi
Danışman: Doç. Dr. Yusuf ALTUN
Haziran 2022, 77 sayfa

Bu tez çalışmasında konuşma duygu tanıma uygulamaları için yeni filtre bankaları ve insan sesi temel frekans tespiti için yeni bir metot önerilmektedir. Yeni filtre bankalarının konuşma duygu tanıma uygulamalarında büyük gelişmelerin önünü açması beklenmektedir. Günümüze kadar pek çok farklı filtre bankası konuşma tanıma uygulamaları için önerilmiştir. Ancak bu modeller genellikle çok fazla parametre içermekte veya karmaşık bazı matematiksel işlemlere gereksinim duymaktadırlar. MFCC (Mel Frequency Cepstral Coefficients) katsayıları Mel filtre bankalarından türetilirken DCT (Discrete Cosine Transform) uygulanmaktadır. Ayrıca MFCC katsayılarını akustik olarak yorumlamak hemen hemen imkansızdır. Mel filtre bankaları daha kolay yorumlanabilmesine rağmen çok fazla sayıda parametre içermektedir. Önerilen EFB (Emotional Filter Banks) filtre bankaları daha kolay yorumlanabildiği gibi hesaplama yönünden de daha hızlıdırlar. Bu çalışmada bu filtre bankalarını SVM-SMO (Support Vector Machine-Sequential Minimal Optimization) ve Derin Yapay Sinir Ağı modelleri ile uygulayıp MFCC ve Mel filtre bankaları ile EmoSTAR, EmoDB (Berlin Emotional Database), IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) ve MELD (Multimodal EmotionLines Dataset) verisetleri üzerinde uygulayıp karşılaştıracağız. Özellik seçme ve veri türetme uygulamaları da ayrıca incelenecektir. Temel frekans tespiti için HDM (Harmonic Differences Method) metodu önerilecek olup genişbant ve darbant (telefon) konuşma için araştırılacaktır. HDM harmonikler arasındaki farkı temel alarak çalışmaktadır. Temel frekans için Hillenbrand ve Texas Sesli verisetleri ile TIMIT (Texas Instruments Massachusetts Institute of Technology) verisetinin sesli kısmının tamamı kullanılacaktır. HDM algoritması otokorelasyon, kepstrum, YIN, YAAPT (Yet Another Algorithm for Pitch Tracking), CREPE (Convolutional Representation for Pitch Estimation) ve FCN (Fully Convolutional Network) metotları ile karşılaştırılacaktır. Sonuçlar harmonikler arasındaki farkların temel frekans için iyi bir seçim olduğunu ve HDM metodunun diğerlerine göre çoğunlukla daha başarılı sonuçlar üretebildiğini göstermektedir.

Anahtar sözcükler: Filtre bankası, Konuşma duygu tanıma, Konvolusyonal sinir ağları, LSTM, SVM, Temel frekans

ABSTRACT

APPLICATION OF NOVEL FILTER BANKS AND FUNDAMENTAL FREQUENCY DETECTION METHOD IN SPEECH EMOTION RECOGNITION WITH DEEP LEARNING

Cevahir PARLAK
Duzce University
Institute of Science, Department of Computer Engineering
Doctoral Thesis
Supervisor: Assoc. Prof. Yusuf ALTUN
June 2022, 77 pages

In this manuscript, a novel filter bank design, named EFB and a pitch determination algorithm, HDM, are proposed. The proposed filter banks are aimed to replace current state-of-the-art MFCC and Mel filter banks. We hope that EFB filters will have great impact over the speech emotion recognition applications. Today, most of the speech processing applications use Mel filters or its transformed and reduced version MFCC. There are various other filter banks proposed to imitate the human ear structure. However, these models have too many redundant frequency regions. MFCC contains fewer coefficients but computation of DCT is a setback of speed. Another disadvantage of these filters is the difficulty to interpret the MFCC values. It is very hard to gain an insight by inspecting the Mel filters or MFCC. The EFB filter banks are not only fast and easy to compute compared to the Mel and MFCC, but it also provides more insights about the underlying structure of the speech waveform. In this study, EFB filter bank is implemented on emotional speech datasets (EmoSTAR, EmoDB, IEMOCAP, MELD) with various Deep Learning Architectures and SVM-SMO classifier to compare them with MFCC and Mel filter banks. We also investigate feature selection and data augmentation methods. Prosodic features are used very extensively in speech emotion applications. For this part, we developed a novel fundamental frequency calculation method called HDM which exploits the intervals between the harmonics of vowel speech sounds. We test the HDM against some of the prominent algorithms such as autocorrelation, CREPE, YIN, YAAPT, cepstrum, and FCN on Hillenbrand Vowel dataset, Texas Vowel dataset, and vowel part of TIMIT dataset for narrowband telephony speech as well as wideband speech.

Keywords: Convolutional neural networks, Filter banks, Fundamental frequency, LSTM, Speech emotion recognition, SVM

EXTENDED ABSTRACT

APPLICATION OF NOVEL FILTER BANKS AND FUNDAMENTAL FREQUENCY DETECTION METHOD IN SPEECH EMOTION RECOGNITION WITH DEEP LEARNING

Cevahir PARLAK
Duzce University
Institute of Science, Department of Computer Engineering
Doctoral Thesis
Supervisor: Assoc. Prof. Yusuf ALTUN
June 2022, 77 pages

1. INTRODUCTION

In this thesis, a novel filter bank called EFB for speech emotion recognition and a novel pitch determination method called HDM for wideband and narrowband speech signals are proposed. The proposed novel filter banks are destined to open great improvements over the speech emotion recognition studies. Feature extraction is an important part in artificial intelligence tasks. Properly designed feature sets help the classifier models to spur the accuracies on the datasets. Mel filter banks and MFCC are still state-of-the-art feature sets in speech processing applications even though nearly a half century spanned over their debut. MFCC features are calculated out of Mel filters using DCT. Most of the speech processing implementations use 13 of these MFCCs instead of 40 Mel filters. Different alternatives have been suggested for Mel filter banks and MFCCs including PLP (Perceptual Linear Prediction), Seneff filters, Lyon Auditory Model, EIH (Ensemble Interval Histograms) auditory model; however, none of them were able to overtake the lead from MFCC. In speech production, vocal folds generate the periodic sound source, and the rest of the vocal tract (tongue, teeth, lips, nasal cavity, and jaw) constitute the filter part of the speech. Filter banks may use triangular, rectangular, or other shapes over the frequency region. In Mel scale, usually 40 triangular filters are considered for computation. The center frequency of Mel filter banks may depend upon the type of the application (speech, music, emotion, etc.). Frequency regions of Mel filters are not strictly defined and can be set variously according to the experiments. MFCC can be thought of an attribute reduction and transformation from 40 Mel filters and is an analogy of jpeg (Joint Photographic Experts Group) conversion from bitmap image files. The proposed

EFB filters are faster and easier to compute and can be interpreted better than MFCC and Mel filters. We run these filter banks with NVIDIA CNN (Convolutional Neural Network), 1D (1-Dimensional) CNN, LSTM (Long-Short Term Memory), Bidirectional LSTM and SVM-SMO classifier to prove the effectiveness of them against MFCC and Mel filter banks. We also handle feature selection, data augmentation, and data imbalance problem to prove the advantage and superiority of the proposed novel EFB filter banks.

Pitch is an extremely complex attribute of speech and has an important role in the human-computer applications and human conversational relations. Detection of fundamental frequency has a very debated background. Many different methods have been developed. However, it persists to be a daunting task specifically in noisy, narrowband (telephone), multi-talker, and multi-pitch tasks with sustainable resolution and quick implementation due to extraordinarily complex nature of frequency band. The vibrations of vocal cords are called fundamental frequency, and it can display quite different values among girls, boys, women, and men. Therefore, precise determination is a vital point in many applications including early detection of pathological symptoms and human-computer interaction. Missing fundamental concept is important in our daily phone talks. In phone signals, the spectrum of the frequency domain is squeezed between 300 Hz (Hertz) and 4000 Hz. However, speaker's gender, emotion and identity can still be identified very discriminatively. If we take the autocorrelation of a signal comprising harmonics at 200 Hz, 300 Hz, 400 Hz, 500 Hz and 600 Hz, we still perceive the pitch of the signal as 100 Hz instead of 200 Hz despite the lack of 100 Hz component in the signal.

2. MATERIALS AND METHODS

In this study, we propose a new filter bank strategy called EFB for speech emotion recognition to compete against the MFCC and Mel filters. The advantages of novel features can be delineated as follows:

1: EFB is nearly 40% faster than MFCC and Mel filters in the feature extraction stage.

2: EFB is easier to compute and classify. EFB has only 13 filter banks compared to 40 Mel filters.

3: EFB has a better interpretation particularly against the MFCC. The frequency regions of EFB hold the traces of vowel and consonant formant regions; therefore, it provides more information over the background of the speech samples. Formants are the distinctive frequencies of vowels and consonants.

4: The frequency regions of EFB are given beforehand. Mel filters are widely adopted but no concrete agreement is set on the borders of these triangular filters. Therefore, every study can implement different settings and get different outcomes. This is a setback on the comparison of separate models.

5: EFB filters are usually ahead of MFCC in terms of classification accuracy and very competitive versus Mel filters.

Speech emotion recognition is the arena we will test the proposed filter banks against the Mel Filter banks and MFCC. We have created a MATLAB library to extract the statistical features from a speech sample. We included some statistical functions and their delta acceleration coefficients with EFB filters, MFCC, and Mel filters.

Machine learning algorithms like SVM, HMM (Hidden Markov Model) have been used to classify datasets for a very long time. However, Deep Neural Networks specifically Convolutional Neural Networks gained significant superiority against the traditional methods during the last 15 years. CNN was invented earlier, however, only after the developments in GPU (Graphical Processing Unit) hardware, Deep Neural Networks started to dominate the field. Using a GPU, CNN models can be run much faster compared to the CPU (Central Processing Unit) implementations. CNNs also enable end-to-end (without hand-engineered features) studies. CNNs are designed for image processing applications mainly but in the last years, they are getting more popular in time series applications and are able to produce better results. In time series implementations of CNN models, time series can be rearranged like 2-D image data. CNN models extract the edges in the first layer, edges and corners in the second layer and nonlinear complicated features in the third layer. CNN uses convolutional layers and maximum pooling layers. Max pooling and convolutional layers can be considered as an attribute extraction stage and these attributes can be used for classification. A softmax layer usually follows the convolutional and max pooling layers. Convolutional layers are not completely connected to the input as opposed to the traditional neural networks. Specific parts of the input data are connected to some parts of the hidden layers. CNN models are expensive in terms of memory and computational load. Convolution is invariant against translation except rotation. For rotation invariance, data augmentation can be used by rotating the image left and right.

LSTM networks currently are the state-of-the-art technology in time-series classification

applications, namely, speech processing applications. LSTM networks are developed to overcome the short memory problems of HMM and GMM (Gaussian Mixture Model) based models to learn the longer series and to better calibrate the features they obtained by tackling the exploding and vanishing gradients problems.

Support Vector Machines are powerful classifiers used in machine learning applications. Soft margin SVM allows a slack (a small error) in the classification. It works well with a small margin of separation, optimality is guaranteed due to the nature of convex optimization, effective in high dimensional spaces, and memory efficient. However, SVM's performance on the large datasets is not satisfactory due to the high training time and SVM is also vulnerable to noisy samples and outliers.

In this study, EmoSTAR, EmoDB, IEMOCAP, and MELD datasets are used for emotional speech implementations. Human emotion is conveyed through many ways such as speech, hand gestures, body language, ambiance, face gestures. Therefore, some emotional datasets include other data together with speech utterances. This additional information comprises text and video. The IEMOCAP and MELD datasets are large datasets compared to the EmoSTAR and EmoDB but, in this study, we eliminated some emotional classes due to the large imbalance between the classes and selected angry, happy, neutral, and sad samples in most of the experiments. IEMOCAP and MELD are acted audiovisual datasets and provide textual information as well.

In this thesis, a new fundamental frequency detection algorithm called HDM is developed. We evaluated autocorrelation (AC), YIN, cepstrum, CREPE, YAAPT, and FCN methods on narrowband (telephone) and wideband vowel data. In the pitch determination experiments, we employed 3 vowel datasets: Hillenbrand Vowel dataset, vowel part of TIMIT dataset, and Texas Vowel dataset.

3. RESULTS AND DISCUSSIONS

In the novel filter bank implementation of this study, we compare EFB, MFCC, and Mel features using SVM-SMO classifier of Weka, NVIDIA CNN model, a 1-D CNN model, a LSTM model, and a BiLSTM model. CNN, LSTM, and BiLSTM models are run in Python 3.6.5 with Keras 2.2.4 and Tensorflow 1.7.1. Number of features in EFB and MFCC feature set is 573, and 1761 in Mel filter bank feature set. EFB filters perform better than MFCC and Mel filters in most cases.

We also run feature selection methods to show the effectiveness of proposed filter banks.

CFS (Correlation Based Feature Selection) Subset Feature Selection method can reduce the number of features very significantly while preserving or even sometimes increasing the performance. EFB filters achieved outstanding performances with feature selection. EFB filters are faster than MFCC and Mel filters. For Information Gain Attribute selection in IEMOCAP dataset, EFB filters perform better than MFCC and Mel filters and has the highest overall accuracy with NVIDIA CNN model. Number of selected features are very close for EFB and MFCC. Imbalanced datasets constitute a great challenge for the machine learning tasks. In this thesis, we also deal with imbalanced dataset problem leveraging undersampling, data augmentation, SMOTE techniques, and other evaluation metrics such as F-measure and Kappa. We show that SMOTE and data augmentation are quite successful for tackling the imbalanced data problem.

Voicing Detection Error (VDE) and Gross Pitch Error (GPE) are the most used error measures in pitch determination. In this study, we also evaluate e_{10} (stands for the number of speech samples oscillating more than 10% from the predetermined ground truth fundamental frequency value of the signal) and gender detection error metrics.

Ground truth fundamental frequency values of Hillenbrand and Texas Vowel datasets are supplied beforehand. Thus, a solid comparison can be made with our estimated values. On the other hand, TIMIT dataset has no ready-to-use ground truth pitch values. TIMIT is a very large dataset compared to Hillenbrand and Texas Vowel datasets. In wideband datasets, HDM and FCN compete, in narrowband datasets, convolutional neural networks produce very low results whereas HDM is capable of producing high accuracy results in wideband and also in narrowband (phone) speech samples. Our experiments are extended to the narrowband phone speech signals. We apply bandpass filter to all of our speech corpora twice to completely cancel the frequencies above 3400 Hz and below 400 Hz.

The proposed HDM algorithm is better than all other algorithms in narrowband TIMIT dataset in GPE and e_{10} error evaluations. YAAPT and autocorrelation perform well in e_{10} error. YAAPT is specifically designed for narrowband telephone signals with missing fundamental frequency. FCN and CREPE are almost blind in band passed speech data. We must bear in mind that training of FCN and CREPE is performed solely with the wideband data. FCN and CREPE must also be trained on the narrowband telephone speech data. CNNs are very smart in wideband human voice, but they are too slow versus HDM, autocorrelation and cepstrum. The fastest method is HDM followed by cepstrum and autocorrelation in our fundamental frequency experiments.

4. CONCLUSIONS AND OUTLOOK

In this study, we conducted experiments with the proposed EFB filters against Mel filters and MFCC features. EFB filters achieved better or at least comparable results in these experiments using SVM-SMO, NVIDIA CNN model, 1-D CNN model, LSTM, and BiLSTM model. We can safely claim that we are not obliged to the Mel filters or MFCC features in speech emotion processing applications. The novel filter banks are easy-to-compute, much faster, and better interpretable compared to the MFCC and Mel filter banks. Mel filters have 40 different triangular frequency regions to be calculated and MFCC usually uses 13 out of 40 Mel filters. The novel filter banks comprise only 13 different frequency regions based on the formant regions of the phones. These experiments concretely prove that EFB filter banks can be used as a replacement for MFCC and Mel filter banks. In the future studies, we will test EFB filters on speech recognition applications.

Our pitch detection experiments show that the spacings between the harmonics of a speech sound can be trusted to extract the pitch value of narrowband and wideband data. The novel algorithm shines particularly in the vowel part of the TIMIT corpus. HDM algorithm has the speed advantage over the others. CREPE and FCN are doing a good job in wideband data, but they run very slowly in comparison to the HDM, autocorrelation, and cepstrum methods. The results of CREPE and FCN are very low in narrowband experiments. As a future study, temporal smoothing can be incorporated inside HDM. Testing the HDM in musical and noisy data is another tough challenge to be tackled. Pitch refinement techniques such as frequency reassignment or synchro-squeezing can also be included in the HDM implementation.

1. GİRİŞ

Bu tezde, insan konuşması için yeni bir temel frekans belirleme algoritması ve Konvolusyonal Sinir Ağları [1]-[3], LSTM [4]-[6] ağları gibi Derin Öğrenme mimarilerini ve SVM [7], [8] sınıflandırıcılarını kullanarak mevcut MFCC [9] ve Mel filtre [10] bankalarının yerini alabilecek konuşma duygu tanımaya yönelik yeni bir filtre bankası stratejisi geliştirmeye çalışacağız. Kullanılacak veritabanları, konuşma duygu tanıma için EmoSTAR [11], EmoDB [12], IEMOCAP [13], MELD [14], temel frekans belirleme için ise Hillenbrand Sesli Harf veritabanı [15], Texas Sesli Harf veritabanı [16] ve TIMIT [17] veritabanının sesli harf bölümüdür.

Konuşma sinyalinin işlenmesi önvurgu (preemphasis) filtresi ile başlar. Önvurgu işlemi, gelen konuşma sinyalinin yüksek frekanslı bileşenlerini güçlendirerek spektrumu düz hale getirmek için kullanılır. Bu nedenle, sinyalin genel enerji dağılımı bir FIR (Finite Impulse Response) filtresi kullanılarak dengelenir. Önvurgu filtresinin sinyale etkisi, $x(k)$ sinyal, $y(k)$ sinyalin cevabı, α önvurgu katsayısı olmak üzere Denklem 1.1'de verilmektedir.

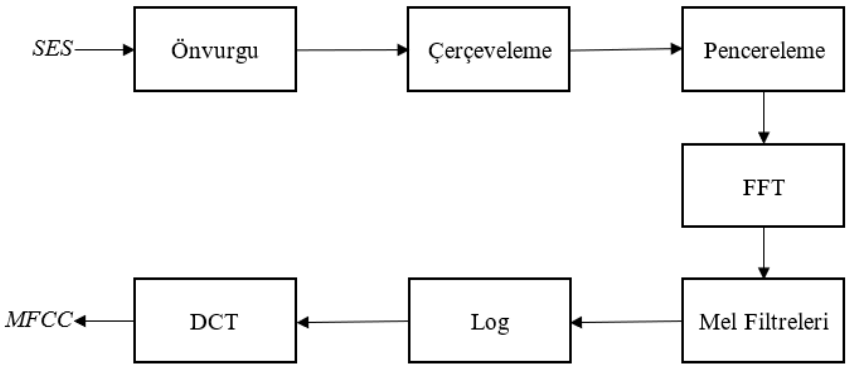
$$y(k) = x(k) - \alpha * x(k-1) \qquad 0{,}90 < \alpha < 1 \tag{1.1}$$

Bir sonraki adım sinyalin çerçevelenmesidir. Çerçeveleme adımında, önvurgulanmış konuşma sinyali N çerçeveye ayrılır ve her çerçevede belli sayıda nokta vardır (512, 1024, 2048, ...). Her çerçeve bir M (128, 256, ...) boyutuyla önceki çerçeveyle çakışır. Çakışmanın amacı çerçeveden çerçeveye geçişi yumuşatmaktır. Çerçevenin sınırlarındaki sinyal süreksizliklerini en aza indirmek için Hamming, Hanning veya Blackman gibi pencereler uygulanabilir. Pencere adımından sonra spektral bilgileri ve her frekans bandındaki enerji miktarlarını çıkarmak gerekir.

Konuşma hakkında anlaşılması gereken temel nokta, bir insan tarafından üretilen seslerin dil, dişler, dudaklar, çene ve hatta burun dahil olmak üzere ses sisteminin şekline göre filtrelenmiş olmasıdır. Ses yolunun bu şekli üretilecek sesin belirlenmesinde temel unsurdur. Ses yolu şekli doğru bir şekilde belirlenirse, üretilen sesin doğru bir temsilini elde edebiliriz. Ses kanalının şekli kendini kısa süreli güç spektrumunun zarfında gösterir

ve MFCC'nin rolü bu zarfı doğru bir şekilde temsil etmektir.

MFCC 1980'lerde Davis ve Mermelstein tarafından tanıtıldı ve o zamandan beri ses işleme uygulamalarında yaygın olarak kullanılmaktadır. MFCC'den önce, LPC (Linear Predictive Coding) [18], [19] otomatik konuşma tanıma için ana özellik türüydü. MFCC ve Mel filtre hesaplaması Şekil 1.1'de tasvir edilmektedir.

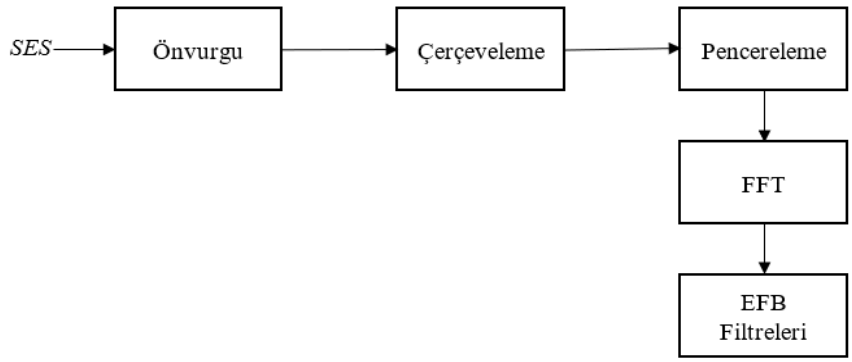


Şekil 1.1. MFCC ve Mel filtreleri hesaplama adımları.

Konuşma sinyali FFT (Fast Fourier Transform) [20] ile frekans boyutuna dönüştürüldükten sonra FFT bileşenlerinin şiddetleri hesaplanır ve daha sonra bu şiddet spektrumu Mel ölçeğinde çakışan üçgen filtrelerle çarpılır. Bu çalışmada frekans bantlarından enerji toplamak için 40 üçgen filtre oluşturulmuştur. Bu filtreler 1000 Hz'in altında doğrusal aralıklı, geri kalanı ise 1000 Hz'in üzerinde logaritmik aralıklıdır. Sinyalin Mel spektrumu elde edildikten sonra, bir sonraki adım Mel kepstral katsayılarını belirlemektir. Bunu yapmak için, filtre bankası şiddetlerinin logaritmaları alınır ve daha sonra DCT [21] kullanılarak 13 katsayı (ilk katsayı sinyalin enerjisi ile ilgilidir) hesaplanır (13 delta ve 13 çift delta katsayısı ile 39 boyutlu vektör). Genellikle, DCT'den alacağımız son 12 katsayı konuşma tanıma için yeterlidir ancak duygu veya cinsiyet tanıma gibi uygulamalarda ilk terim sinyalin enerjisini gösterdiği için ihmal edilemez. Duygu tanımada sinyalin enerjisi duygu sınıflarıyla güçlü bir ilişki içindedir.

Bu çalışmada, konuşma duygu tanıma için MFCC ve Mel filtre bankalarının yerini alabilecek EFB adlı yeni bir filtre bankası önerilmektedir. Yeni filtre bankasının konuşma duygu tanıma uygulamaları üzerinde büyük gelişmelerin önünü açacağı düşünülmektedir. Konuşma tanıma uygulamalarını modellemek için birçok filtre bankası önerilmiştir, ancak bu modeller çok fazla filtre içermekte veya hesaplamak için hantal matematiksel işlemlere ihtiyaç duymaktadırlar. MFCC, DCT'nin hesaplanmasını gerektirir ve MFCC katsayılarını yorumlamak da çok zordur. Mel filtrelerini yorumlamak daha kolay olsa da

çok fazla filtre içermektedir. Yeni filtre bankaları daha hızlı ve daha kolay hesaplanabildiği gibi, MFCC ve Mel filtrelerine kıyasla daha iyi yorumlanabilirler. Önerilen filtre bankalarını MFCC ve Mel filtre bankalarıyla karşılaştırmak için NVIDIA CNN modeli [22], 1-B CNN modeli [23], LSTM, BiLSTM [24] ve SVM-SMO sınıflandırıcıları ile uyguluyoruz. Ayrıca, önerilen filtre bankalarının etkinliğini göstermek için özellik seçimi ve veri türetme deneyleri uyguluyoruz. EFB filtre bankalarının hesaplanması Şekil 1.2'de gösterilmiştir.



Şekil 1.2. EFB hesaplama adımları.

Veri türetme, makine öğrenimi uygulamalarında kullanılan çok güçlü bir tekniktir. Bu çalışmada konuşma duygu işlemede veri türetmeyi uygulamaya çalışacağız. Derin ağlar, iyi performans elde etmek için büyük miktarda eğitim verisine ihtiyaç duyar. Veri türetme rastgele döndürme, kaydırma, kesme, çevirme, uzatma, daraltma gibi yöntemlerle uygulanır. Konuşma sinyali işlemede zaman çarpıtma, frekans maskeleme ve zaman maskeleme olmak üzere veri türetmenin 3 temel yolu vardır. Bu çalışmada MATLAB 2021 audioDataAugmenter [25] sınıfını 4-duygulu (Kızgın, Üzgün, Mutlu, Nötr) doğaçlama IEMOCAP veri kümesinde kullandık.

Sınıflandırma görevlerinde en uygun öznitelik kümesini bulmak araştırmacıların en temel çabalarından biridir. Aşırı fazla veya gereğinden az sayıda özellik kullanmak performansı düşürebilir. Öznitelik seçiciler veri kümesinin boyutunu azaltır ve ayrıca aşırı öğrenmeyi önler. Bu çalışmada Information Gain [26] ve CFS Subset Öznitelik Seçme [27] yöntemlerini Weka [28] aracı ile kullanacağız.

İnsan konuşmasının ana karakteristiklerinden biri, konuşmacı hakkında cinsiyet ve duygusal durum veya bir müziğin tonları gibi bazı önemli ipuçlarını tanımlamamıza yardımcı olan temel frekanstır. Temel frekans algılama algoritmaları bir kişinin konuşmasının temel periyodunu veya müzikal bir tonu çıkarmak için yaygın olarak

kullanılır. Bir sinyalin temel frekansını çıkarmak için geliştirilmiş sayısız algoritma bulunmaktadır. Bu çalışmada FCN [29], CREPE [30], otokorelasyon [31], kepstrum [32], YIN [33], YAAPT [34] ve önerilen HDM [35] yöntemlerini kullandık.

Bu metinde genişbant ve darbant (telefon) konuşması için HDM adlı yeni bir temel frekans belirleme algoritması önerilmiştir. Telefon konuşmasında frekans bandı, daha uzun mesafeli iletim sağlamak için 3000 Hertz bant genişliğinde kısıtlanmaktadır. Konuşma 300 Hertz ve 4000 Hertz arasında filtrelenmektedir. Bu, mevcut temel frekans algılama algoritmalarından bazılarının verimli çalışmasını önleyecektir. Önerilen f0 algılama algoritması, bu tür kısıtlamaların üstesinden gelmek için tasarlanmıştır. Konuşma sinyalinin harmonik bileşenleri arasındaki farkları ses temel frekans değerini bulmak için kullanıyoruz. Deneylerimizde Hillenbrand ve Texas Sesli Harf veritabanları ile TIMIT veritabanının sesli harf bölümünün tamamını kullanacağız. Algoritmamızı otokorelasyon, kepstrum, YIN, YAAPT, CREPE ve FCN algoritmalarıyla karşılaştırıyoruz. Sonuçlar, harmonik aralıkların temel frekans tespiti için sağlam bir seçim olduğunu ve çoğu durumda otokorelasyon, kepstrum, YIN, YAAPT, CREPE ve FCN'ye göre çok daha üstün olduğunu göstermektedir.

2. İLGİLİ ÇALIŞMALAR

Konuşma duygu analizinde çalışmalar son 20 yılda oldukça büyük mesafe kaydetmiştir. İlk çalışmalarda duygu sınıfları oldukça kısıtlı (olumlu, olumsuz) verisetleri kullanılırken günümüzde çok çeşitli konuşma duygu türleri (kızgın, mutlu, nötr, üzgün, korku, tiksinti, can sıkıntısı vb.) içeren verisetleri kullanılmaktadır. MFCC ve Mel filtreleri gibi insan sesine ait öznitelikler konuşma duygu tanıma için de kullanılmakta ancak bunlara istatistiksel fonksiyonlardan (ortalama değer, standart sapma, eğrilik, diklik, vb..) elde edilen özellikler eklenmektedir. Prozodik özellikler olarak adlandırılan temel frekansa dayalı özniteliklerde ses duygu analizinde faydalı olmaktadır. Konuşma duygu veritabanları konusunda oldukça fazla ilerleme kaydedilmesine rağmen normal konuşma veritabanlarına göre örnek sayıları oldukça azdır. Konuşma duygu verisi elde etmek hem etiketleme hem de istenilen duyguyu üretebilme açısından oldukça zahmetlidir. Önceleri sınıflandırıcı olarak SVM gibi metotlar kullanırken günümüzde CNN ve LSTM gibi derin öğrenme metotları daha ağırlıklı olarak kullanılmaktadırlar.

[36] çalışmasında, TIMIT veri kümesinde karma Sinir Ağı-Hidden Markov Model ağına sahip KSA (Konvolusyonal Sinir Ağı) kullanılmıştır. Bu, konuşma verileriyle ilgili KSA kullanan ilk makaledir. Mel filtreleri, birinci ve ikinci delta özellikleri çıkarılmıştır. TIMIT çekirdek test setinde Ses Hata Oranları (SHO) açısından 3 gizli katmanlı en iyi KSA'nın diğer YSA (Yapay Sinir Ağları) ile karşılaştırılması Çizelge 2.1'de verilmektedir. Konvolusyonal Sinir Ağı metodu %20,07'lik Ses Hata Oranı ile DBN (Deep Belief Net) [37] öneğitimli ve mcRBM (Mean Covariance Restricted Boltzmann Machine) [38] özellik çıkarma kullanan Yapay Sinir Ağı modelinden daha iyi sonuç elde etmiştir.

Çizelge 2.1. En İyi Ksa Modelinin Diğer Ysa Modelleri İle Karşılaştırılması [36].

Metot	**SHO (%)**
YSA (3 gizli katman ve 1000 nöronlu)	22,95
KSA [36]	**20,07**
YSA (DBN ön eğitimi ile)	20,70
YSA (DBN ön eğitimi ve mcRBM özellik çıkarma ile)	20,50

[39]'da Ko vd., 0,9'luk, 1,0'lik ve 1,1'lik hız faktörlerini kullanarak farklı sinyaller üreten bir veri türetme öneriyorlar. 100 ila 5500 saat arasında değişen İngilizce ve Mandarin verilerini kullanan 5 farklı LVCSR (Large Vocabulary Continuous Speech Recognition) deneyinin sonuçlarını, Switchboard [40], GALE [41], TED-LIUM [42], Aspire [43] ve Librispeech [44] veri kümeleri ile sunmaktadırlar. Dört gizli katmanlı zaman gecikmeli sinir ağı kullanılmıştır. Ağa giriş olarak 40 MFCC ve 100 i-vektör [45] özellikleri verilmektedir. Eğitim için 18 GPU kullanılmaktadır. Veri türetme için farklı VTLP (Vocal Tract Length Perturbation) [46] faktörleri, WSOLA (Waveform Similarity Overlap-Add) [47] bazlı hız değişimi ve hız değişimi kullanılmıştır. Çizelge 2.2'de gösterildiği gibi, hız değişimi KHO (Kelime Hata Oranı) olarak VTLP ve WSOLA'dan daha iyi performans göstermektedir.

Çizelge 2.2. Çeşitli Verisetlerinde Lvcsr Sonuçları [39].

Veriseti	Veriseti Hacmi	Referans Değer (%)	KHO (Hız-değişimi) (%)	Bağıl İyileştirme (%)
Switchboard	300 saat	20,7	**19,3**	6,7
GALE	100 saat	17,51	**17,16**	2,0
Librispeech	960 saat	12,93	**12,51**	3,2
TED-LIUM	118 saat	17,9	**17,2**	3,9
Aspire	5500 saat	30,8	**30,7**	0,32

[48]'de, Mel filtre bankaları TIMIT veri kümesinde CNN ve RNN-CTC (Recurrent Neural Network-Connectionist Temporal Classification) [49] kombinasyonu ile kullanılıyor. Önceki bir çalışmada CNN-HMM %26,3 hata oranı ile denendi. HMM [50] parçası RNN-CTC modeliyle değiştirilmiştir. CNN modeli Caffe [51] ile, CTC ise SAIL (Speech Analysis and Interpretation Laboratory) [52] Python kütüphanesi ile uygulanmıştır. 25 pencereli, 128-256-384-384 Konvolusyon katmanlı ve 1024-512 Yoğun katmanlı CNN modeli %22,1 çerçeve hatası elde etti. 2048 girişli 4 katmana sahip CNN modeli CTC kullanarak %29,4 hece hatasına ulaşmıştır.

[53]'te Toth, mel ölçekli zaman frekansı gösterimine sahip TIMIT veri kümesinde maksimum etkinleştirme (maxout) işlevlerine, düşürme (dropout) ve hiyerarşik modellere sahip CNN kullanmaktadır. Daha önce bir LSTM modeli %17,7 ses hata oranına ulaşmış, başka bir CNN modeli %17,4 ses hata oranına sahiptir. Bu çalışma hiyerarşik CNN modelini sundu ve önceki tüm çalışmalardan daha iyi performans göstererek %16,5 ses hata oranına ulaştı.

Michalek [54] tarafından yapılan çalışmada, TIMIT veri kümesinde birkaç DNN, Zaman Gecikmeli Ağ ve LSTM mimarisi değerlendirilmektedir. Zaman gecikmeli sinir ağı, sürekli verileri sınıflandırmak için tasarlanmış bir ağdır. İlk olarak fonemlerin konuşma tanıma sistemlerinde sınıflandırılması önerildi. MFCC, delta, ikinci delta LDA ile işlenmiş ve son özellik vektör boyutu 40 olmuştur. 8 gizli katman ve 2048 nöron ile Feed Forward Network, ortalama ses hata oranı %16,91'e ulaşmıştır. Filtre sayısı 512 olan Zaman Gecikme Ağı ortalama ses hata oranı %18,20'e ulaşmıştır. En başarılı ağ, 4 katmanlı ve 1024 LSTM birimine sahip ve ortalama ses hata oranı %15,58 olan LSTM'dir.

[55] çalışması içinde iki aşamalı bir derin öğrenme mimarisi önerilmiştir. LSTM tabanlı bağımsız bir aşama ve dönüştürücüye dayalı ikinci bir hiyerarşik aşama kullanılmıştır. Akustik ve dilsel özellikleri birleştirmek için Dikkat (Attention) ve Doğrusal modülasyon uygulanmıştır. IEMOCAP, MELD, MOSI [56] ve MOSEI [57] veri kümeleri Dönüştürücüler ve modülasyona dayalı iki mimari ile uygulanmıştır. Dilsel ve akustik veriler duygu analizi için birleştirilmiş ve akustik ve metinsel özelliklerin projeksiyonu referans değer olarak alınmıştır. NT (Naive Transformer), MAT (Modulated Attention Transformer) ve MNT (Modulated Normalization Transformer) yeni modeller olarak çalıştırılmıştır. 300 boyutlu GloVe (Global Vectors for Word Representations) [58] özellikleri çıkarılmış ve tipik bir metinden konuşma sistemiyle aynı prosedüre sahip mel-spektrogramlarla birleştirilmiştir. IEMOCAP veriseti için 4-duygulu (Kızgın, Mutlu (Mutlu+Heyecanlı), Nötr, Üzgün) sınıflandırma başarı oranı sonuçları Çizelge 2.3'te gösterilmiştir. MAT sınıflandırıcı %74 BO (Başarı oranı) ile en iyi sonuçları elde etmiştir.

Çizelge 2.3. 4-Duygulu Iemocap Verisetinde Sınıflandırma Sonuçları [55].

Model	**Hassasiyet (%)**	**BO (%)**	**F1 (%)**
MAT	**74**	**74**	**74**
MNT	72	72	72
NT	71	70	70
P	69	67	67

NT modeli bile, yalnızca bir LSTM ve projeksiyon katmanından oluşan P (Projection) modeline kıyasla önemli performans artışı sağlamaktadır. MAT, yaklaşım analizi için daha uygun bulunmuştur. MNT, MELD için daha uygun görünüyor. MAT için olası bir sorun, 48 katmana kadar kullanan BERT [59] gibi son NLP çözümlerine kıyasla sığ

mimarilerle çalışılmasıdır. Sunulan veri kümesi kapsamında, bu tür mimarileri eğitmek için yeterli örnek yoktur.

[60]'da uç füzyonunu temel alan derin bir sinir ağında çok ölçekli alan dikkati (Multi scale area attention) duygusal özelliklere katılmak için uygulanmaktadır. Ses Yolu Uzunluğu Değişimi (VTLP), veri seyrekliği ile başa çıkmak için veri türetme olarak IEMOCAP veri kümesi üzerinde kullanılmaktadır. Nlpaug (Natural Language Processing Augmentation) [61] veri türetme kitaplığı ile özgün veriler 7 kat çoğaltılmıştır. Yazarlar doğaçlama IEMOCAP veri kümesinde dört tür duygu (nötr, heyecan, üzüntü ve öfke) verilerini kullandılar. Veriseti %80 eğitim seti ve %20 test seti olarak sınıflandırıcılara sunulmuştur. Öznitelik olarak, 80 Logmel (Logarithmic Mel) özelliği Librosa (Laboratory for the Recognition and Organization of Speech and Audio) [62] ses işleme kitaplığı kullanılarak çıkarılmıştır. Alan dikkat özellikleri de eklenmiştir. Tüm veri örnekleri 2 saniyelik bölümlere ayrılmış, örnekler eğitimde 1 saniye ve testte 1,6 saniye uzunluğunda segmentlere ayrılmıştır. Çalışma, Çizelge 2.4'te gösterildiği gibi %79,34 BO (Başarı Oranı) ve %77,54 ABO (Ağırlıksız Başarı Oranı) ile IEMOCAP veri kümesinde son teknoloji doğruluk oranlarını elde etmiştir.

Çizelge 2.4. Doğaçlama 4-Duygulu Iemocap Veri Kümesindeki Mevcut Yöntemlerin Doğruluk Karşılaştırması [60].

Metot	**BO (%)**	**ABO (%)**
Attention	71,75	68,06
Multitask learning+Attention	76,40	70,10
Self-Attention	70,17	70,85
Head Fusion	76,18	76,36
Area Attention	**79,34**	**77,54**

[63] çalışmasında, Satt vd., kızgın, mutlu, nötr ve üzgün örneklerle IEMOCAP'in doğaçlama bölümünü kullanmaktadır. En iyi CNN modelinin 5 katman içerdiği bulunmuş (2-8 katman denenmiş) ve en iyi karma topolojinin 3 konvolusyon katmanı ve bir LSTM katmanı içerdiği tespit edilmiştir (1-6 konvolusyon katmanı ve 1-2 LSTM katmanı denendi). Gürültüsüz 300 zaman adımlı güç log spektrogramları ile 3 konvolusyon katmanlı CNN+LSTM kullanarak %68,8 BO ve %59,4 ABO elde ettiler. Gürültülü deneyler de uygulanmıştır. Deneylerde 0-4000 Hz frekans bölgesi kullanılmıştır.

[64] uygulamasında, IEMOCAP veri kümesi kullanılarak uçtan uca konuşma duygu tanıma ile ilgili derin öğrenme uygulamaları anlatılmaktadır. Kızgın, mutlu, nötr ve üzgün duygusal sınıflar değerlendirilmektedir. Özellik vektörü olarak 40 Mel katsayısı kullanılmaktadır. Çizelge 2.5'te gösterilen 5 gizli katmanlı FC (Fully Connected) DNN (Deep Neural Network), ReLU (Rectified Linear Unit) [65] ve terk katmanlarına sahip 1024 nöron, bazı Conv (Convolutional) mimariler ve LSTM modelleri kullanılmıştır.

Çizelge 2.5. IEMOCAP Verisetinde Modellerin Başarı Oranları [64].

Model	**BO (%)**	**ABO (%)**
LSTM-RNN(256)×2-Softmax (256)	61,71	58,05
FC(1024)×5-Softmax (1024)	62,55	58,78
Conv(16×10×10)-Conv(32×10×10)-FC(716)×2-Softmax (716)	**64,78**	**60,89**

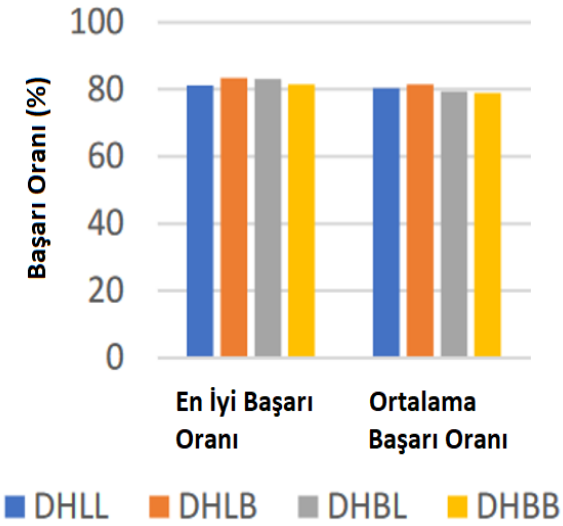
[66] içinde Tripathi vd., duygu algılama için CNN ile konuşma özellikleri olarak spektrogramları ve MFCC'leri kullandılar. Spektrogram/MFCC tabanlı (Model 2A, Model 3, Model 4A) dahil olmak üzere birçok farklı CNN modeli uygularlar. Deneylerinde, IEMOCAP veri kümesini 4 duygusal sınıfla (kızgın, heyecan (mutluluk), nötr ve üzüntü) kullandılar. Çizelge 2.6'da başarı oranları gösterilmektedir.

Çizelge 2.6. IEMOCAP Verisetinde Farklı Modellerin Başarı Oranları [66].

Model	**Giriş**	**BO (%)**	**ABO (%)**
Satt	Spektrogram	68,8	59,4
Model 2A	Spektrogram	71,2	61,9
Model 3	MFCC	71,6	59,9
Model 4A	Spektrogram+MFCC	**73,6**	**62,9**

[67]'de yazarlar Tamil dili veri kümesindeki Derin Hiyerarşik LSTM ve BiLSTM modellerini karşılaştırdılar. Veri kümesi, mobil uygulamalar aracılığıyla 10 kişiden toplanan 1400 duygusal konuşma örneği içerir. Duygusal sınıflar kızgın, mutlu, nötr, üzgün, korku, tiksinti ve can sıkıntısıdır. Sınıflandırma için MFCC, MFCC deltası, Bark spektrumu, spektral diklik, spektral eğrilik özellikleri çıkarılır. DHLL (Deep Hierarchical LSTM and LSTM), DHLB (Deep Hierarchical LSTM and BiLSTM), DHBL (Deep Hierarchical BiLSTM and LSTM), DHBB (Deep Hierarchical Dual BiLSTM) olarak 4 farklı modeli doğruluk, kayıp, eğitim süresi ve değerlendirme süresi ile karşılaştırdılar.

Sonuçlar, DHLB'nin Şekil 2.1'de gösterildiği gibi en iyi model olduğunu göstermektedir.



Şekil 2.1. Farklı LSTM modellerin karşılaştırılması [67].

[68]'de yazarlar uç-uca konuşma duygu sistemini denediler. IEMOCAP ve CEMO (Emotional Corpora for Emergency Call Centers) [69] veri kümelerini kullandılar. Acil çağrı merkezlerinde üretilen CEMO korku, kızgın, nötr ve rahatlama sınıfları ile 485 konuşmacıdan Fransızca (2 saat 16 dakika) 440 diyalog içermektedir. Bu veri kümelerinde yeni bir CNN-BiLSTM modeli denendi ve 4 duyguda 5 kat çapraz doğrulama ile IEMOCAP'te %63 ABO elde ettiler. CEMO veri kümesinde, 4 sınıf için ABO %45,6'dır.

[70]'da Pakyurek vd. MFCC histogramlarını EmoDB ve Youtube videoları ve popüler televizyon programlarından yeni oluşturdukları İngilizce PAU verisetinde SVM sınıflandırıcı ile denediler. 13 MFCC, delta ve ikinci delta katsayıları da hesaplanarak öznitelik setleri oluşturulmuş ve Üzgün, Mutlu ve Nötr duygu sınıfları üstünde sınıflandırmalar yapılmıştır. Sonuçlar MFCC histogramlarının performansı arttırdığını göstermiştir.

[71]'de Zhang vd., 64 Mel spektrogramı ve delta katsayısı kullanarak EmoDB ve IEMOCAP veri kümelerinde Dikkat mekanizmasına sahip bir Konvolusyonal Sinir Ağı ve çift yönlü LSTM modeli uyguladı. Doğaçlama IEMOCAP veri kümesinin 4 duygulu (kızgın, mutlu, nötr, üzgün) kısmında ağırlıksız başarı oranı %68,50 ve EmoDB'de %87,86'ya ulaştılar. Ayrıca, aşırı öğrenme sorunuyla mücadele etmek için veri türetme uyguladılar ve dengesiz veri dağılımının üstesinden gelmek için veri kümesinin sınıfları

için farklı ağırlıklar atadılar. 64 mel spektrogram görüntüsü, 227x227x3 boyutunda giriş olarak CNN'ye verilmiştir. Önerdikleri modellerini diğer CNN, LSTM modelleriyle karşılaştırdılar ve önerilen Deep CNN+LSTM+Attention modeli en iyi sonucu %68,50 ile elde etmiştir.

[72] içinde, önerilen temel frekans yöntemi, aynı sinyalin bir parçasına sahip bir sinyalin otomatik olarak ilişkili olmasına dayanır. Çalışma sırasında frekans tahminleri hesaplanır ve sinyalin bir parçası algılanarak segment güncelleştirilir. Temel frekansların düşük hata oranı ve basit uygulama ile hızlı tahmini, gerçek zamanlı konuşma sinyali işleme için uygundur. Algoritmanın amacı, örneklenmiş bir konuşma sinyalinin temel frekansını tahmin etmektir. Bu, sinyal ve kendi segmenti arasındaki otomatik düzeltmelerin hesaplanmasıyla elde edilir. Edinburg Evaluation veritabanı [73], Keele Konuşma veritabanı [74] ve PTDB-TUG (Pitch Tracking Database-Technical University of Graz) [75] veritabanı kullanılmıştır. Otokorelasyon, YIN, SHRP (Subharmonic to Harmonic Ratio Pitch Detector) [76] algoritmaları karşılaştırılmıştır. Gürültü de çalışmalara dahil edilmiştir.

[77]'da yazarlar, harmonik oranlar ve kepstrum analizi yaklaşımlarını birleştiren BaNa adlı hibrit gürültüye dayanıklı bir temel frekans algılama algoritması sunarlar. Birkaç temel frekans adayı arasındaki temel frekans değerini tanımlamak için maliyet işlevine sahip bir Viterbi [78] algoritması kullanmaktadırlar. BaNa algoritmasının doğruluğunu ve son teknoloji temel frekans algılama algoritmasını değerlendirmek için bir gürültü veritabanı ile çevrimiçi bir konuşma veritabanı kullandılar. Sonuçlar, araştırılan her türlü gürültü ve SNR değeri için BaNa'nın en iyi temel frekans algılama doğruluğunu elde ettiğini göstermektedir. Ayrıca, BaNa algoritmasının 0 dB sinyal-gürültü oranında dahi yaklaşık %80 temel frekans algılama doğruluk oranı elde ettiği gösterilmiştir.

[79] içinde yazarlar, hataları azaltmak için bir dizi değişik algoritmanın uygulandığı otokorelasyon yöntemini temel alan YIN algoritmasını önerdiler. Algoritmanın birkaç arzu edilen özelliği vardır. Hata oranları, bir laringograf sinyaliyle birlikte kaydedilen bir konuşma veritabanı üzerinden değerlendirildiği gibi, en iyi rakip yöntemlerden yaklaşık üç kat daha düşüktür. Frekans aralığında bir üst sınır yoktur, bu nedenle algoritma tiz sesler ve müzik için uygundur. Algoritma nispeten basittir ve verimli olarak düşük gecikme süresiyle uygulanabildiği gibi ayar yapılması gereken birkaç parametre içerir. Kullanılan 5 veritabanı toplam 1,9 saat konuşma içeriyor ve bunların %48'i sesli olarak etiketlenmiştir. Veritabanları Japonca (30), İngilizce (14) ve Fransızca (4) olmak üzere

toplam 48 konuşmacı (24 erkek, 24 kadın) tarafından üretildi. Her biri konuşmayla birlikte kaydedilmiş bir laringograf dalga formu içeriyordu. Otokorelasyon, çapraz korelasyon, altharmonik toplama [80], kepstrum, normalleştirilmiş otokorelasyon ve TEMPO (Time Domain Excitation Extraction Based on a Minimum Perturbation Operator) [81] yöntemleri YIN ile kıyaslanmış ve YIN yöntemi, tüm veritabanlarında diğer yöntemlerden daha iyi performans göstermiştir.

Geleneksel YIN, monofonik f0 tahmini için basit ama etkili bir yöntemdir ve bu etki alanındaki en popüler yöntemlerden biri olmaya devam etmektedir. Kısa süreli hataları ortadan kaldırmak için, frekans tahmincilerinin çıkışları genellikle daha yumuşak bir temel frekans eğrisi elde etmek için tekrar işlenir. YIN metodunun bir eksikliği, bu tür işlem sonrasında sinyalin alternatif yorumlarına geri dönememesidir, çünkü yöntem çerçeve başına sadece bir temel frekans tahmini sunar. Bu sorunu gidermek için yazarlar YIN metodunu ilişkili olasılıklara sahip birden çok temel frekans adayının çıktısını verecek şekilde değiştirerek pYIN (probabilistic YIN) metodunu önerdiler (pYIN Aşama 1) [82]. Bu olasılıklar doğal olarak YIN eşik parametresindeki bir öndağılımdan kaynaklanır. Bu olasılıkları, geliştirilmiş bir temel frekans izi (pYIN Aşama 2) üretmek için Viterbi ile deşifre edilen gizli bir Markov modelinde gözlem olarak kullanırlar. pYIN metodunun YIN metoduna göre ek hesaplama karmaşıklığı düşüktür. RWC (Real World Computing) Müzik Veritabanı [83] ile mevcut olan f0 temel frekans parçalarından şarkı sözleri sentezlediler ve bunları 44,1 kHz örnekleme hızında doğrusal ses dosyaları olarak kaydettiler. Parçalar, veritabanının popüler müzik alt bölümünün 100 adet tam uzunlukta şarkısını kapsıyor. Bu tür gürültüsüz veriler olmadan daha gerçekçi bir durumu simüle etmek için, orijinal ses dosyalarına ek olarak ADT [84] (Audio Degradation Toolbox) ile beş ön ayar kullanarak sesin kalitesini düşürdüler. Verilerin tamamı 30 saatten fazla ses içerir. Verilen tüm sonuçlar bu tam veri kümesindedir. Beta parametre dağılımları ile önerilen pYIN yönteminin üç farklı sürümünü çalıştırdılar.

[85] içinde, spektral temel frekans izleme için, potansiyel olarak kayıp olan temel frekansı kısmen geri yüklemek için doğrusal olmayan işleme kullanılarak YAAPT yönteminin yeni bir sürümü sunuluyor. Bu algoritmanın ilk sürümü 2002'de sunulmuştur. Spektrumdaki harmonik tepeler arasındaki aralığı belirlemek için frekans etki alanı değiştirilmiş otokorelasyon kullandılar. Frekans etki alanı spektral parça daha sonra Talkin tarafından önerilen NCCF (Normalized CrossCorrelation Function) [86] kullanılarak elde edilen zaman alanı aralığı adaylarını iyileştirmek için kullanılır.

Dinamik programlama hem yerel hem de geçiş maliyetlerini kullanarak tüm adaylar arasında en iyi temel frekans parçasını bulmak için kullanılır. Algoritma, Keele veritabanı kullanılarak değerlendirildi. YAAPT, gürültüsüz stüdyoda %1,59, telefon konuşmasında %2,69 hata oranı elde etti ve YAAPT'nin önceki 2002 sürümünü, doğrusal olmayan spektral yöntemi ve NCCF yöntemini geride bıraktı.

[30] içinde yazarlar, doğrudan zaman verisi dalga formu üzerinde çalışan derin konvolusyonal sinir ağına dayanan bir temel frekans izleme algoritması olan CREPE'i önermektedir. Önerilen modelin pYIN'e yakın veya daha iyi performans göstererek son teknoloji sonuçlar ürettiğini gösteriyorlar. Ayrıca, modelin genelleştirilebilirliğini gürültü sağlamlığı açısından değerlendiriyorlar. RWC-synth ve MedleyDB [87] verisetlerini kullanmışlardır. RWC-synth, RWC Müzik Veritabanından sentezlenen 6,16 saatlik ses içerir. Bu veri kümesindeki sinyallerin az sayıda sinüzoidin sabit bir toplamı kullanılarak sentezlendiğini, veri kümesinin son derece homojen olduğunu ve aşırı basitleştirilmiş bir senaryoyu temsil ettiğini belirtmek önemlidir. Algoritmaları daha gerçekçi (ancak yine de kontrol edilen) koşullar altında değerlendirmek için kullandıkları ikinci veri kümesi, MedleyDB'den alınarak yeniden sentezlenen 230 monofonik gövdeden oluşan bir koleksiyondur ve orijinal parçanın tınısını ve dinamiklerini koruyan mükemmel bir f0 etiketlemesi ile sentezlenmiş bir parça oluşturmak için bir analiz/sentez yaklaşımı kullanılmıştır. Bu veri kümesi 25 enstrümanlı 230 parçadan oluşur ve toplam 15,56 saat ses içerir. CREPE'i, monofonik temel frekans izlemede kullanılan mevcut pYIN ve SWIPE (Sawtooth Waveform Inspired Pitch Estimator) [88] algoritmalarıyla karşılaştırmışlardır. ADT tarafından sağlanan pub, beyaz, pembe ve kahverengi gürültü uygulanmıştır.

[29]'da yazarlar Konvolusyonal Sinir Ağlarını temel alan FCN adlı ham ses verisini kullanan bir temel frekans izleme metodu sunmaktadırlar. Metot önceki CREPE metoduna benzer bir yol izlemekte ancak hesaplama açısından daha fazla avantaj sağlamaktadır. Yazarlar veritabanı olarak mükemmel bir f0 etiketleme olan veritabanı kullanmak amacıyla laringograf ile etiketleme yapan PTDB-TUG gibi veritabanlarını kullanmak istemiş ancak yaptıkları deneyler laringograf çıktılarının da bazı durumlarda güvenilir bir temel frekans tahmini yapılamayacak kadar bozuk olabileceğini göstermiştir. İkinci bir yöntem olarak mevcut veritabanındaki ses kayıtlarını hesaplanan temel frekans değeriyle tekrar sentezleyen bir vokoder kullanmışlardır. Veritabanı olarak TIMIT ve BREF (Large Vocabulary Spoken Corpus for French) [89] kullanılmış ve temel

frekans önce CREPE ile tespit edildikten sonra vokoder ile bu temel frekans üzerinden tekrar sentezlenmiştir. Veriseti %60 eğitim, %20 doğrulama, %20 test olmak üzere 3 kısma ayrılmıştır. Sonuçlar CREPE ve SWIPE ile karşılaştırılmış ve FCN metodu %97,36 ile daha başarılı sonuçlar üretmiştir.

Günümüzde temel frekans tespit metotları derin öğrenme algoritmaları ile geliştirilmeye devam edilmektedir. Normal konuşma verilerinde oldukça başarılı sonuçlar elde edilmesine rağmen çok konuşmacılı, gürültülü ve darbant telefon verilerinde yeterli başarı oranlarına erişmek hala önemli bir sorun olarak durmaktadır. Konvolusyonal Sinir Ağları doğrudan zaman verisinden temel frekans tespitinde oldukça başarılı olmaktadırlar. Ancak zaman ve hesaplama maliyetleri standart metotlara göre oldukça pahalı olmaktadır.

3. METOTLAR

SVM, HMM gibi makine öğrenimi yöntemleri, veri kümelerini sınıflandırmak için çok uzun zamandır kullanılmaktadır. Bununla birlikte, derin sinir ağları özellikle konvolusyonal sinir ağları, son 15 yılda geleneksel yöntemlere karşı önemli bir üstünlük kazanmıştır. Daha önce icat edilmiş olsalar da ancak GPU donanımındaki gelişmelerden sonra, Derin Sinir Ağları alana hâkim olmaya başladı. CNN modelleri paralel çalışmaya uygun yapıları nedeniyle GPU ile CPU uygulamalarına kıyasla çok daha hızlı çalıştırılabilir. Konvolusyonal Sinir Ağları uçtan-uça (el yapımı özellikler olmadan) çalışmalara da olanak tanır. Uçtan-uça çalışmalar doğrudan ham konuşma sinyallerinde çalıştırılır. Konuşma verileri sınıflandırıcıya 1 boyutlu, 2 boyutlu hatta 3 boyutlu bir veri olarak gönderilebilir. Konvolusyonal Sinir Ağları öncelikle görüntü işleme uygulamaları için tasarlanmıştır, ancak son yıllarda, zaman serisi uygulamalarda da giderek popüler hale gelmektedir ve daha iyi sonuçlar üretebilmektedir. CNN modellerinin zaman serisi uygulamaları, zaman serisinin 2-B görüntü verisi olarak düzenlenmesiyle uygulanabilir. CNN modelleri, birinci katmanda kenarları, ikinci katmanda kenarları ve köşeleri ve üçüncü katmanda ise doğrusal olmayan karmaşık özellikleri belirlerler. Konvolusyon ve maksimum havuzlama (max pooling) katmanları [90] bir öznitelik çıkarma aşaması olarak düşünülebilir ve bu özellikler sınıflandırma için kullanılabilir. Konvolusyon ve maksimum havuzlama katmanlarını genellikle bir softmax sınıflandırıcı izler. Tam olarak bağlanmış katmanların aksine, konvolusyonal sinir ağlarında giriş verilerinin bazı bölümleri gizli katmanların bazı bölümlerine bağlanır. Geri besleme aşamasında, bir konvolusyonun geri çevrimi başka bir konvolusyondur. Maksimum havuzlamanın geri beslemesi için, en yüksek değeri veren girişin gradyanı geri döndürülür. Bu nedenle, ileri besleme işlemi sırasında ilgili girişin dizinini kaydetmek gerekebilir. Konvolusyonal Sinir Ağı modelleri çok fazla sayıda katmanlarla tasarlanabilir olduklarından bellek ve işlemci hesaplama yükü açısından oldukça pahalıdırlar ve ciddi kaynak sorununa neden olabilirler. Konvolusyon işlemi, döndürme dışındaki ölçeklendirmelere karşı değişmezdir. Döndürme değişmezliği için, görüntüyü sola ve sağa döndürerek veri türetme kullanılabilir. Bu modellerde birçok farklı aktivasyon fonksiyonları (sigmoid, hiperbolik tanjant, ReLU vb.) kullanılmaktadır [3].

3.1. NVIDIA CNN MODELİ

NVIDIA tarafından önerilen CNN modeli görüntü işlemede ve özellikle sürücüsüz otomobillerde oldukça popülerdir [91]-[94] ve Çizelge 3.1'de tasvir edilmektedir. Bu tezde, konuşma duygu tanıma için kategorik çapraz entropi kaybı, grup boyutu 32, Adam optimize edici, başarı oranı (ACC: Accuracy) ölçümü ve varsayılan öğrenme oranı 0,001 ile kullanılmıştır. Çevrim sayısı (epoch) 1000'dir.

Çizelge 3.1. NVIDIA CNN Modeli.

Katman	**Çıkış**	**Parametre Sayısı**
Giriş (10,58)	(374, 10, 58, 1)	0
Conv2D (32, (3, 3), aktivasyon='relu')	(Boş, 8, 56, 32)	320
Conv2D (64, (3, 3), aktivasyon='relu')	(Boş, 6, 54, 64)	18496
MaxPooling2D (pool_size=(2, 2))	(Boş, 3, 27, 64)	0
Dropout (0,25)	(Boş, 3, 27, 64)	0
Flatten	(Boş, 5184)	0
Dense (128, activation='relu')	(Boş, 128)	663680
Dropout (0,5)	(Boş, 128)	0
Dense (sınıf sayısı, aktivasyon='softmax')	(Boş, 7)	903

3.2. 1B KONVOLUSYON KATMANLI CNN MODELİ

Bu bölümde, çalışmamıza yeni bir CNN modeli ekliyoruz. Bu model, NVIDIA'nın 2B Konvolusyonal katmanları yerine 1 Boyutlu Konvolusyon katmanlarını kullanan daha derin bir modeldir. Model mimarisi Çizelge 3.2'deki gibidir:

Kategorik çapraz entropi kaybı, 16 grup normalizasyonu, başarı oranı (ACC) metriği ile kullanılmıştır. Bu model için çevrim sayısı 500'dür. Model, öğrenme oranı düzlükte kayıp gözleme ile azalan minimum 0,000001; 0,9 faktörlü; 20 bekleme parametreleri ile kullanılmıştır.

3.3. LSTM

LSTM, Hochreichter tarafından 1997 yılında geliştirilmiştir ve şu anda zaman serisi sınıflandırma uygulamaları için en son teknolojidir. LSTM'nin HMM ve GMM tabanlı modellere karşı en büyük avantajı, daha uzun serileri hatırlama ve elde ettiği özellikleri daha iyi kalibre edebilmesidir. Vanilya RNN olarak bilinen ilk modellerde, ağın çıktısı,

geçmişi daha iyi hatırlaması için tekrar girişe geri gönderilir. Bu sağlam bir düşünce gibi görünse de pratikte, bu ağlar birkaç adımdan daha gerisini hatırlayamamıştır. LSTM ve GRU (Gated Recurrent Unit) [6] gibi varyantları bu sorunu çözmek için tasarlanmıştır.

Çizelge 3.2. 1B CNN Modelinin Ayrıntılı Şeması.

Katman	**Çıkış**	**Parametre Sayısı**
Conv1D (256, 8, padding='same', 'relu')	(Boş, 573, 256)	2304
Conv1D (256, 8, padding='same', BatchNormalization(),'relu')	(Boş, 573, 256)	524544
Dropout (0,25)	(Boş, 573, 256)	0
MaxPooling1D (pool_size=(8))	(Boş, 71, 256)	0
Conv1D (128, 8, padding='same', 'relu')	(Boş, 71, 128)	262272
Conv1D (128, 8, padding='same', 'relu')	(Boş, 71, 128)	131200
Conv1D (128, 8, padding='same', 'relu')	(Boş, 71, 128)	131200
Conv1D (128, 8, padding='same', BatchNormalization(), 'relu')	(Boş, 71, 128)	131200
Dropout (0,25)	(Boş, 71, 128)	0
MaxPooling1D (pool_size=(8))	(Boş, 8, 128)	0
Conv1D (64, 8, padding='same', 'relu')	(Boş, 8, 64)	65600
Conv1D (64, 8, padding='same', 'relu')	(Boş, 8, 64)	32832
Flatten()	(Boş, 512)	0
Dense (sınıf sayısı, 'softmax', SGD (decay=0, lr=0,0001, nesterov=False, momentum=0)	(Boş, sınıf_ sayısı)	2052

Bugün konuşma işleme uygulamalarının çoğu LSTM tabanlı modellerle diğer geleneksel HMM, GMM modellerine kıyasla yüksek performanslara ulaşmaktadırlar. Son birkaç yılda CNN modelleri LSTM modelleriyle rekabet etmeye başladı. HMM modelleriyle LSTM ağlarını içeren bazı yeni çalışmalar vardır. LSTM yapıları belleğe sahip ağlar olarak düşünülebilir. LSTM modeli, Derin Sinir Ağlarının patlayan ve yokolan gradyan sorunlarına da iyi bir çözümdür. LSTM, patlayan gradyanları gradyan kırpma ve yokolan gradyanları ise unut kapısı ile çözmektedir. Unut kapısı çıkışı sıfır olduğunda, bellekteki bilgi temizlenir. Unut kapısı çıktısı 1 olduğunda, belleğin tüm içeriği korunur. LSTM modelleri, unut kapısı 0 değerine gitme eğiliminde olduğunda unut kapısının 1 değerine nasıl çıkarılabileceğini öğrenebilen mekanizmalardır. Diğer kapılar benzer işlevlere sahiptir ve belleğe ne kadar veri ekleneceğini belirlerler. GRU, LSTM'nin basitleştirilmiş bir çeşididir. LSTM ve GRU'nun performansı çok benzerdir, ancak GRU basitleştirilmiş yapısı nedeniyle LSTM'den daha hızlıdır. Önceki gizli durum ve geçerli giriş durumu bilgileri sigmoid işlevinden geçirilir. Gizli katman ve geçerli giriş, ağı düzenlemek ve değerleri -1 ile +1 arasında sıkıştırmak için bir tanh (Hiperbolik Tanjant) işlevinden

geçirilir. Daha sonra tanh ve sigmoid çıkışları çarpılır. Tanh çıkışının önemli kısmına sigmoid çıkış ile karar verilir ve tutulur. Hücre durumu ağın belleğidir ve unut kapısıyla noktasal olarak çarpılır. Hücre durumunun sıfıra yakın değerlerle çarpılan bölümleri silinir ve yeni hücre durumumuzu elde etmek için giriş kapısının çıktısını hücre durumuna ekleriz. Çıkış kapısı bir sonraki gizli durumu belirler. Gizli durum önceki girişler hakkında bilgiye sahiptir ve tahminler için kullanılabilir [5].

RNN iyi bir fikir olmasına rağmen, yokolan ve patlayan gradyanlar gibi çeşitli sorunları vardır. Geri kullanım aşamasında, gradyanlar zincir kuralı ile hesaplanır ve bu nedenle gradyanlar küçükse, bu küçük sayıları defalarca çarpar. Bu işlem gradyanların sıfıra doğru gitmesine neden olur. Gradyan değerleri sıfıra yakınsa, bunun ağların ağırlıklarının iyileştirilmesi üzerinde hiçbir etkisi olmaz ve ağ hiçbir yere yakınsamaz. Yokolan gradyan sorununun aksine, ağırlık matrisinin çarpımı, ağırlıklar 1'den büyükse zincir kuralı nedeniyle patlayan gradyan sorununa neden olur [4]-[6].

LSTM ağları genellikle tek yönde çalıştırılır; bazı uygulamalarda, ağı ileri ve geri çalıştırmak yararlı olabilir. Schuster ve Paliwal 1997'de Bidirectional RNN'yi [95] sundular. BiRNN'de ileri ve geri bağlantılar vardır.

Deneylerimizde, 100 gizli birim, 1000 çevrim, RMSprop (Root Mean Squared Propagation) optimize edici, grup boyutu 32, karesel ortalama hata kaybı, başarı oranı (ACC) ölçümüne sahip Çizelge 3.3 ve Çizelge 3.4'teki LSTM ve Bidirectional LSTM modellerini kullandık.

Çizelge 3.3. Deneylerimizde Kullanılan LSTM Modeli.

Katman	**Çıktı**	**Parametre Sayısı**
LSTM	(Boş, 200)	544800
Dense (softmax)	(Boş, 7)	1407

Çizelge 3.4. Deneylerimizde Kullanılan Bidirectional LSTM Modeli.

Katman	**Çıktı**	**Parametre Sayısı**
BiLSTM	(Boş, 200)	544800
Dense (softmax)	(Boş, 7)	1407

3.4. SVM

Destek Vektör Makineleri (SVM), makine öğrenimi uygulamasında kullanılan güçlü

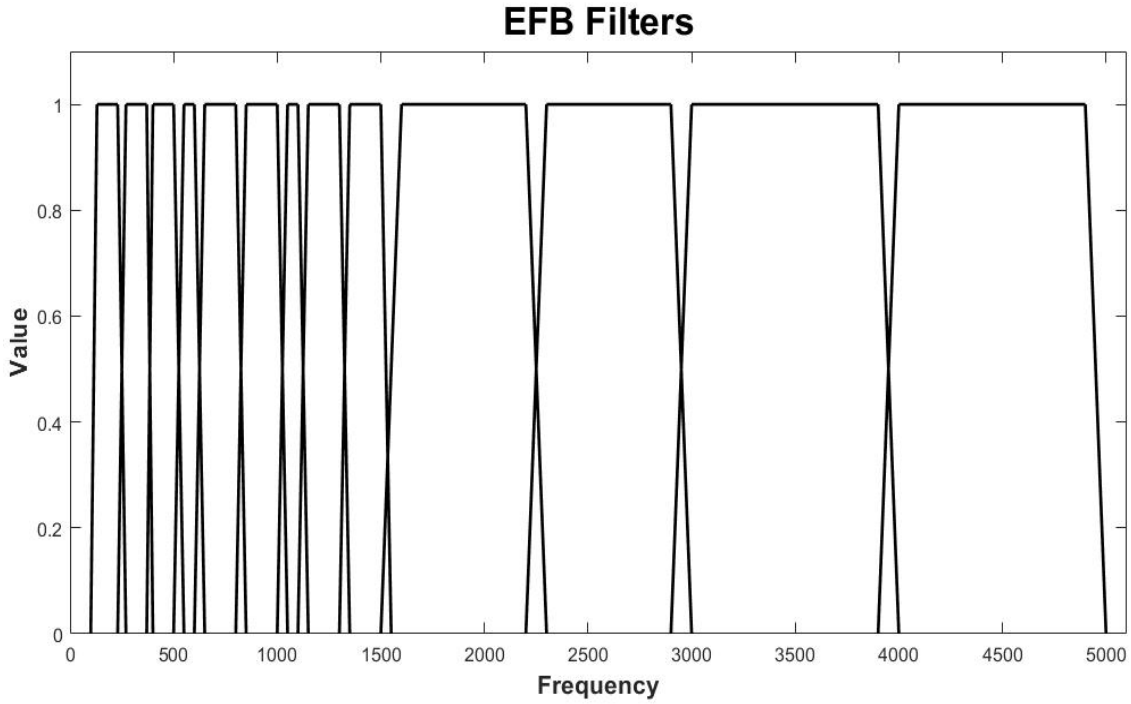
sınıflandırıcılardır. SVM'nin temellerini yalnızca iki sınıflı doğrusal olarak ayrılabilir verilerle başlayarak anlayabiliriz. İki ayrı veri sınıfının n boyutlu uzayda bir hiper düzlemle ayrılabileceğini varsayabiliriz. Amaç, ikili sınıflandırıcı modeli öğretmektir. Genellikle, doğrusal ayırma her zaman mümkün değildir veya mümkün olsa bile, verilerdeki aykırılıklar veya gürültüler için birkaç istisna yaparak daha iyi bir çözüm elde edilebilir. Soft Margin SVM, eğitim verilerinde belli bir miktar hata oranına izin verir. SVM, doğrusal olarak ayrılabilir ve doğrusal olarak ayrılamaz veriler için kullanılabilir. Verilerin etiketlendiği ve etiketsiz olduğu alanlarda kullanılabilir. SVM, Çekirdek Hilesi (Kernel Trick) yardımıyla doğrusal verileri doğrusal olmayan verilere dönüştürebilir. Bu çalışmada Weka aracının polinom çekirdekli SVM-SMO sınıflandırıcısını kullanacağız [7], [8].

3.5. ÖNERİLEN EFB FİLTRE BANKASI

Konuşma tanıma için Mel, Plp, Seneff, Lyon ve EIH gibi birçok filtre bankası önerilmiştir. Bu çalışmada, MFCC ve Mel filtrelerine alternatif konuşma duygu tanıma için EFB adlı yeni bir filtre bankası stratejisi öneriyoruz. Yeni özelliklerin avantajı, doğruluğu artırırken çok daha hızlı, hesap ve yorumlanmalarının daha kolay olmasıdır. Bu filtre bankalarının frekans bölgeleri Hertz olarak Çizelge 3.5 ve Şekil 3.1'de gösterilmektedir.

Çizelge 3.5. Önerilen EFB Filtre Bankalarının Frekans Bölgeleri.

Filtre No	f1	f2	f3	f4
1	100	130	230	270
2	230	270	370	400
3	370	400	500	550
4	500	550	600	650
5	600	650	800	850
6	800	850	1000	1050
7	1000	1050	1100	1150
8	1100	1150	1300	1350
9	1300	1350	1500	1550
10	1500	1600	2200	2300
11	2200	2300	2900	3000
12	2900	3000	3900	4000
13	3900	4000	4900	5000



Şekil 3.1 Önerilen EFB filtrelerinin görselleştirilmesi.

3.6. ÖNERİLEN HDM METODU VE KULLANILAN METOTLAR

Temel frekans tespiti, kapsamlı bir araştırma alanıdır. Günümüzde temel frekans algılama yöntemleri genişbantlı, gürültüsüz insan konuşmasında oldukça başarılıdır. Ancak özellikle darbant (telefon konuşması) [85], çoklu ses perdeli [96] ve gürültülü konuşma sinyallerinde [97] çözülmesi gereken çok fazla zorlu engel vardır. Otokorelasyon, kepstrum, dalgacık tabanlı ve son olarak CNN yöntemleri de dahil olmak üzere çeşitli teknikler uygulanmıştır. Bununla birlikte, temel frekans belirleme için ilginç bir ipucu son derece ihmal edilir ve çok nadiren kullanılır. Harmonik aralıklar, insan konuşması olağanüstü karmaşık periyodik olsa da temel frekans belirleme için çok güçlü delil parçalarıdır. Bu yöntem ilk olarak Seneff [98] tarafından LDVT (Lincoln Digital Voice Terminal) ile harmonik aralıklardan elde edilen temel frekans adaylarına ağırlık faktörleri atayarak 210 Hz-1050 Hz bandı arasındaki gerçek zamanlı verilerde darbantlı konuşma için kullanılmıştır. Seneff'in çalışmaları çok eski olduğu gibi veri kümesi ve sonuçlar da net olarak sunulmamıştır. 21. yüzyıla kadar başka çalışmalar da bulunmaktadır. Wu [99] bir çizelge ile gitar sesinde harmonik aralıkları en büyük ortak bölen kullanarak denemiştir. Dziubiński ve Kostek çeşitli müzik aletlerinde harmonik setler matrisini yapay sinir ağı kullanarak uygulamıştır [100]. Bu yöntemi hem genişbant hem de darbant konuşma (400 Hz-3400 Hz) sinyalleri için büyük konuşma veri kümelerinde

uygulayacağız ve histogram kullanarak harmonikler arasındaki en fazla tekrarlanan farkı belirleyerek doğruluğu ve hızı arttırırken bazı sınırlamaları ortadan kaldıracağız.

Temel frekans algılamanın otokorelasyon modeli, Licklider'ın "dubleks" [101] modeline kadar uzanır. Periyodiklik temel frekans teorisi, Helmholtz ve von Bekesy gibi bilim adamları tarafından desteklenen "Place Pitch" teorisini arka planda bırakmış görünüyordu. Bu teori Ritsma [102] tarafından ayrıntılı incelenmiştir ve bazı sınırlamaları gösterilmiştir.

Temel frekans, insan konuşmasının olağanüstü karmaşık ve belirgin bir özelliğidir ve insan konuşmalarında olduğu gibi insan-bilgisayar etkileşiminde de önemli rol oynar. Temel frekans algılama, yüzyıldan fazla bir süredir çok güçlü ve tartışmalı bir geçmişe sahiptir. Sayısız yöntem önerilmiştir, ancak özellikle frekans spektrumunun son derece karmaşık yapısı nedeniyle makul çözünürlük ve hızlı uygulama ile darbantlı (telefon), gürültülü, çok sesli ve çok konuşmacılı uygulamalarda hala zorlu bir görevdir. Temel frekans konuşmacı hakkında kimlik, cinsiyet, duygusal durum veya bir müzik aletinin tonları gibi bazı önemli ipuçlarını tanımlamamıza yardımcı olur. Duygu ve cinsiyet tanıma, konuşma sentezi, insan-bilgisayar etkileşimi, patolojik bozukluk semptomlarının erken evrelerde tespiti gibi çok geniş bir uygulama yelpazesine sahiptir. Temel frekans, ses teli titreşiminin sıklığıdır. Erkekler, kadınlar ve çocuklar arasında yüksek oranda değişebilir. Belli bir dizi harmoniklere sahip bir sinyalde ilk harmonik silinse bile temel frekans aynı olarak algılanır ve otokorelasyonları da aynıdır. Bu olay kayıp temel frekans (missing fundamental frequency) olarak bilinir.

Temel frekans tespit metotları ile ilgili ilk çalışmalar 1950'ler civarında AMDF (Average Magnitude Difference Function) [103] ile başladı. AMDF otokorelasyondaki çarpmanın yerine çıkarmanın kullanıldığı bir fonksiyondur ve bilgisayarların henüz yeterli hızda olmadığı o zamanlarda hesaplama maliyeti açısından iyi bir çözümdü. Kayıp temel frekans tartışması, temel frekansın hesaplanmasında yoğun çalışmalara yol açtı ve bir sinyalin temel frekansını çıkarmak ve izlemek için birçok algoritma geliştirildi. En dikkat çekenler arasında otokorelasyon, kepstrum, harmonik çarpanlar [104], TEMPO, RAPT (Robust Algorithm for Pitch Tracking), dalgacık tabanlı [105], PRAAT [106], YIN, YAAPT, pYIN, CNN tabanlı (FCN, CREPE) yöntemler sayılabilir. Diğer deneylerde olduğu gibi, temel periyot izleme deneylerinde de uygun veri kümelerinin oluşturulması önemli bir husustur. Telefon konuşmasında, bant genişliğinden tasarruf etmek için sinyal genellikle 300 Hz ile 4000 Hz arasında bant geçişli olarak filtrelenir. Bu çalışmada 400

Hz ile 3400 Hz arasındaki tüm frekansları silmek için iki kez bant geçirgen filtre uygulanmıştır.

Temel periyot veri kümeleri için bir diğer önemli konu da sinyalin gerçek temel periyodunun tam ve doğru olarak belirlenmesidir. Elle düzenleme, EGG (ElectroGlottoGraph), Laryngograph yaygın olarak kullanılmaktadır. Her birinin bazı avantajları ve dezavantajları vardır ancak genellikle bir uzman tarafından elle düzenlemeye ihtiyacımız olmaktadır. Bazı durumlarda, görsel uzman incelemesi ile bile temel frekansı tespit etmek oldukça zor olabilir. EGG, laringogram, özellikle diferansiyel laringogram mevcut yöntemleri kullanarak otomatik temel periyot hesaplamayı kolaylaştıran bir sinyal sağlar.

3.6.1. HDM

Harmonik farklar yöntemi (HDM), en çok tekrarlayan farkı bulmak amacıyla sinyalin güç spektrumunun harmonikleri arasındaki farkları değerlendirmeye çalışır. Harmonik aralıkları Seneff LDVT, Wu gitar sesi, Dziubiński ve Kostek ise çeşitli müzik aleti sesleri için denemiştir. Seneff, tepenin altındaki alanı frekans spektrumunda kullanarak 8 tahminli bir liste oluşturup ortanca filtre ile zamansal yumuşatma uygulamıştır. Biz uygulamamızda zamansal yumuşatma olmadan en çok tekrarlayan farkı bulmaya çalışıyoruz. Uygulamadaki birçok adımı ortadan kaldırarak ve yöntemi genişbant ve darbant ses dalgalarına uygulamak suretiyle genelleştirerek doğruluğu arttırıyoruz. Bu çalışmada MATLAB 2019a versiyonu HDM, otokorelasyon, kepstrum, YIN ve YAAPT yöntemlerini uygulamak için kullanılmaktadır. Python 3.6.5, CREPE ve FCN yöntemlerinin uygulamaları için kullanılmıştır. HDM algoritmasının adımları aşağıdaki gibidir [35]:

1. FFT ile frekans spektrumu çıkarma
2. Frekans spektrumunun içinde tepe noktalarının tespiti
3. Tepe noktalarından gereksiz olanların listeden kaldırılması
4. Listedeki frekans değerleri arasındaki farkların tespiti
5. En çok tekrar eden farkın bir histogram ile tespiti

3.6.2. Otokorelasyon

Otokorelasyon metodu temel frekans tespiti için kullanılan en eski tekniktir ve zamanla yeni sürümleri geliştirilmiştir. Otokorelasyon işlevindeki ana tepe noktası sıfır gecikme

konumundadır (k = 0). Bir sonraki tepenin konumu periyodun ve temel frekansın bir tahminini verir ve yükseklik sinyalin periyodikliğinin bir göstergesini verir. Otokorelasyon yöntemi genellikle güvenilir bir temel frekans tahmini oluşturmak için en az iki veya daha fazla periyot verisi gerektirir. Son yıllarda sadece tek bir periyotu kullanan otokorelasyon yöntemleri de geliştirilmiştir. Bir sinyalin otokorelasyonu simetri özelliğiyle beraber Denklem 3.1'deki gibi formüle edilir.

$$A_{ac}(k) = \sum_{j=k}^{M-1} x_j x_{j-k}\,, k\ duraksama\ katsayısı \tag{3.1}$$

3.6.3. Kepstrum

Bir konuşma sinyalinin kepstrumu, sesli konuşma için temel periyoda karşılık gelen bir tepe noktaya sahiptir, ancak sessiz harflerde tepe oluşmaz. Kepstrum, spektrum kelimesinden terslenerek türetilmiştir ve frekansın tersine işaret etmektedir. Kepstrum analizörü hem bir temel frekans hem de sesli-sessiz (Voiced-Unvoiced) harf ayıtedici olarak işlev görebilir. Kepstral temel frekans algılama, faz bozulmasına karşı duyarsız olması ve konuşma sinyalinin gürültüsüne ve genlik bozulmasına karşı da dirençli olması gibi önemli avantajlara sahiptir. Yöntem, konuşma sinyalinde temel frekansın varlığını gerektirmez ve birkaç farklı temel frekans periyodu varsa birkaç ayrı kepstral tepe noktası verecektir. Kepstrum $O(NlogN)$ karmaşıklığa sahiptir. Bu nedenle otokorelasyondan daha hızlıdır. Kepstrum işleminde zaman sinyaline iki kere Fourier dönüşümü uygulanmaktadır. Otokorelasyondan farkı arada logaritma alma işleminin olmasıdır. Aradaki logaritmanın çıkarılması otokorelasyonu verecektir. Güç kepstrumu, $\mathcal{F}$ Fourier dönüşümü, τ zaman olmak üzere Denklem 3.2'deki gibi formüle edilir.

$$K_G = |\mathcal{F}^{-1}\{log\ (|\mathcal{F}\{f(\tau)\}|^2)\}|^2 \tag{3.2}$$

3.6.4. YIN

Otokorelasyon metodunu temel alır ve hataları azaltmak için bazı düzenlemeler önerir. Otokorelasyonun aksine tepeler değil çukur bölgelere bakar. Algoritmanın birkaç arzu edilen özelliği vardır. Frekans arama aralığında bir üst sınır yoktur, bu nedenle algoritma yüksek frekanslı tiz sesler ve müzik aleti seslerinde temel frekans tespiti için uygundur. Algoritma nispeten basit olmakla beraber verimli ve düşük gecikme süresiyle

uygulanabilir, ayrıca ayar yapılması gereken sadece birkaç parametre içerir.

3.6.5. YAAPT

YAAPT metodunda kayıp temel frekansı kısmen geri yüklemek için doğrusal olmayan işleme kullanılarak yeni bir yöntem sunulmaktadır. Spektrumdaki harmonik tepeler arasındaki aralığı belirlemek için frekans düzlemi değiştirilmiş otomatik düzeltme kullanılır. Frekans etki alanı spektral penceresi daha sonra Talkin tarafından önerilen NCCF kullanılarak elde edilen zaman düzlemi aralığı adaylarını iyileştirmek için kullanılır. Dinamik programlama tüm adaylar arasında en iyi temel periyot parçasını bulmak için kullanılır. YAAPT özellikle darbant telefon konuşma ses örnekleri için önerilmiş bir metottur.

3.6.6. CREPE

CREPE, doğrudan zamansal dalga formu üzerinde çalışan derin konvolusyonal sinir ağına dayanır. pYIN'e yakın veya daha iyi performans göstererek son teknoloji sonuçlar ürettiği gösterilmiştir. RWC Synth ve MedleyDB veri kümesi, CREPE'in performansını pYIN algoritmasına göre değerlendirmek için kullanılmıştır.

3.6.7. FCN

FCN, CREPE'in geliştirilmiş bir modelidir ve konvolusyonal sinir ağlarına dayanmaktadır. CREPE gibi, giriş olarak ham dalga formu kullanmaktadır.

4. VERİSETLERİ VE DENEYSEL KURULUM

Bu bölümde deneylerimizde kullanılan verisetlerini sunuyoruz. EmoSTAR, EmoDB, IEMOCAP ve MELD verisetlerini konuşma duygu tanıma için önerilen yeni filtre bankaları için kullandık. EmoSTAR ve EmoDB küçük boyutlu veri kümeleridir. IEMOCAP ve MELD verisetleri, EmoDB ve EmoSTAR verisetleri ile karşılaştırıldığında büyük veri kümeleridir. Teksas Sesli Harf veri kümesi, TIMIT veri kümesinin sesli harf kısmı ve Hillenbrand Sesli Harf veri kümesi temel frekans tespit deneyleri için kullanılmıştır.

4.1. EMOSTAR

EmoSTAR, Türkçe ve İngilizce dilinde 393 kızgın, nötr, mutlu ve üzgün konuşma örneğinden oluşmaktadır. Nötr konuşmalar çoğunlukla televizyon haber kanallarından geliyor. Kızgın örnekler, TV dizilerinden ve filmlerden elde edilen konuşma örnekleridir. Üzgün örnekler, insanların gerçekten ağlayıp konuştuğunu gösteren internet videolarındandır, ancak bazıları rol yapılan üzgün konuşma örnekleridir. EmoSTAR'ın mutlu söylemleri Golden Globe, Amerikan Music Awards ve Oscar gibi çeşitli törenlerin ödül sahiplerinin konuşmalarından toplandı ve çoğunlukla doğal, spontane mutluluk örneklerini temsil ediyor. EmoSTAR, 106 farklı konuşmacı tarafından konuşulan 2,2 ila 14,5 saniye arasında değişen uzunluklara sahip farklı cümleler içerir. Bu nedenle, konuşmacı özgü özellikler açısından çok çeşitli bir veri kümesidir. Konuşmacı başına cümle sayısı 1 ile 29 arasında değişir. EmoSTAR veri kümesinde, örneklerin çoğu videonun sahipleri tarafından etiketlendirilmiştir; ancak, analizimiz için kendi değerlendirmelerimizi de yaptık. EmoSTAR veri kümesinin duygu türlerinin ayrıntılı sayısı Çizelge 4.1'de gösterilmiştir [107].

Çizelge 4.1. Emostar Verisetinde Duygu Sınıflarının Dağılımı (T=Türkçe, İ=İngilizce) [107].

	Kızgın	**Mutlu**	**Nötr**	**Üzgün**
Kadın	40 İ	37 İ	37 İ - 20 T	51 İ - 19 T
Erkek	33 İ - 30 T	45 İ	35 İ - 34 T	12 İ
Toplam=393	103	82	126	82

4.2. EMODB

EmoDB, Çizelge 4.2'de gösterildiği gibi 7 duygu türünde 535 konuşma örneğinden oluşmaktadır. EmoDB veriseti, 5 uzun ve 5 kısa tümceden oluşmaktadır ve Almanca dilindedir. 7 duygunun sınıflandırılmasında EmoDB'deki en yüksek başarı oranı %89,71'dir ve Parlak tarafından 997 öznitelikle SVM-SMO sınıflandırıcı kullanılarak elde edilmiştir [108].

Çizelge 4.2. EmoDB Verisetinde Duygu Dağılımı [107].

	Kızgın	**Mutlu**	**Üzgün**	**Korku**	**Nötr**	**Sıkkın**	**Tiksinme**
Kadın	67	44	37	33	40	46	35
Erkek	60	27	25	36	39	35	11
Toplam=535	127	71	62	69	79	81	46

EmoDB'de algı testleri 20 denek tarafından yapıldı. Konuşma örnekleri bilgisayar tarafından hakemlere sunuldu ve konuşmacının duygusu yanı sıra bu duyguyu ifade ederken ne kadar ikna edici olduğuna karar vermeleri istendi. Her konuşma örneğini bir kez dinleyebilirler. Analizde kullanılmak üzere hakemler tarafından belirli doğruluk ve doğallık oranına sahip olan örnekler kabul edilip veritabanına dahil edildi.

4.3. IEMOCAP

İnsan duyguları konuşma, el hareketleri, beden dili, ambiyans, yüz hareketleri gibi birçok farklı yolla aktarılır. Bu nedenle, bazı duygusal veri kümeleri konuşma örnekleriyle birlikte metin, resim ve video verileri de içerir. IEMOCAP veri kümesi, Güney Kaliforniya Üniversitesi'ndeki Konuşma Analizi ve Yorumlama Laboratuvarı (SAIL) tarafından 10 aktörden görsel-işitsel olarak toplanmıştır. Aktörlerin ellerinde, yüzlerinde ve kafalarında, ortaya çıkan spontane konuşmalar sırasında el ve yüz hareketlerinden ilgili bilgileri yakalamak için sensörler vardır. IEMOCAP yaklaşık 12 saatlik ses ve video verisi içerir. İki kişilik iletişim konuşmaları hedeflenmiştir. Ayrıca metinsel bilgiler de sağlar. IEMOCAP rol yapılarak oluşturulmuş bir veri kümesidir ve 10039 örnek içerir, ancak bu çalışmada sınıflar arasındaki büyük dengesizlik nedeniyle bazı duygusal sınıfları kullanmadık. Çizelge 4.3'te gösterildiği gibi sadece 2 tiksinti örneği, sadece 3 sınıfı belirsiz örnek, 40 korku örneği, 107 sürpriz örnek, 1421 kızgın örnek ve 1708 nötr örnek vardır. IEMOCAP verisetinde, Xxx kısaltması, hakemlerin söylenişin duygusal

sınıfı üzerinde anlaşamadığı örnekleri gösterir ve çok fazla Xxx etiketli örnek vardır. Etiketleme 3 hakem ile yapılır. Bir örneği tam olarak etiketlemek için en az 2 hakem etiket üzerinde anlaşmalıdır. Etiketleme tutarlılığını analiz etmek için Fleiss Kappa [109] istatistikleri kullanılmıştır. Numunelerin maksimum uzunluğu 34,1388 saniye minimum uzunluğu 0,5849 saniye, ortalama uzunluğu 4,4601 saniye ve standart sapma 3,0647 saniyedir. Toplam uzunluk 44774,72406 saniyedir.

Çizelge 4.3. IEMOCAP Veri Kümesinde Duygulara Göre Örnek Sayıları.

	Doğaçlama	**Senaryo**	**Toplam**
Kızgın	289	814	1103
Mutlu	284	311	595
Nötr	1099	609	1708
Üzgün	608	476	1084
Heyecanlı	663	378	1041
Şaşkın	971	878	1849
Sürpriz	60	47	107
Korku	8	32	40
Tiksinti	1	1	2
Diğer	1	2	3
Xxx	800	1707	2507
Toplam	**4784**	**5255**	**10039**

Duygu ortaya çıkarırken senaryolu oturumlar ve doğaçlama senaryolar kullanılmıştır. Senaryolu oturumlarda (oyunların kullanımı) denekler ezberlenmiş senaryoları prova etmiştir. Doğaçlama senaryolarda, denekler belirli duyguları ortaya çıkarmaya çalışmaktadırlar. Deneylere yedi profesyonel aktör ve üç üniversite son sınıf öğrencisi katıldı. Cinsiyet dengesi 5 kadın ve 5 erkek oyuncu seçilerek sağlanır. Mutluluk için evlenmek, üzgün duygu için kanser olan arkadaşını kaybetmek gibi farklı senaryolar kullanıldı. Kategorik duygular kızgın, mutlu, nötr, üzgün, tiksinti, korku, şaşkın, heyecanlı, sürpriz ve diğer duygulardır. Güç, aktivasyon ve baskınlık duygusal sınıflandırması da yapılmıştır. IEMOCAP veri kümesinin Fleiss Kappa değeri 0,40'tır.

4.4. MELD

MELD, rol yapılarak elde edilmiş bir görsel-işitsel veri kümesidir ve metinsel bilgiler de sağlar. Diyalog başına ikiden fazla konuşmacı ile çok gruplu ve çok kişili duygusal

veriseti eksikliğine çözüm olmak için tasarlanmış olup EmotionLines [110] veri kümesinin gelişmiş bir sürümüdür. EmotionLines veri kümesi yalnızca metinsel veriler içerir. Bir televizyon dizisinden seçilen 1.433 diyalog ve 13.708 duygusal konuşma örneğinden oluşur. Örnekler duygu (kızgın, nötr, üzgün...) ve kanaat (olumlu, olumsuz, nötr) etiketleri ile etiketlenmiştir. Örneklerin maksimum uzunluğu 304,98 saniye, minimum uzunluğu 0,0835 saniye, ortalama uzunluğu 3,20 saniyedir. Toplam uzunluk 43993,66 saniyedir. Etiketleme Mutlu, Üzgün, Korku, Kızgın, Sürpriz, Tiksinti, Nötr duygularını kapsar. Ayrıca olumlu, olumsuz ve nötr kanaatlara göre de sınıflandırma yapılır. Duygu etiketlemelerinde kızgın, tiksinti, korku, üzüntü olumsuz olarak, mutlu olumlu olarak etiketlenir. Sürpriz bazen olumlu bazen olumsuz olarak kabul edilir. Kanaat etiketlemesinde Fleiss Kappa 0,91'e ulaşmıştır. Örnekler, konuşma ve yazmada yüksek İngilizce yeterliliğine sahip yüksek lisans öğrencisi olan üç etiketleyici ile etiketlenmiştir. Son duygu etiketine karar vermek için çoğunluk oylama planı kullanıldı. Genel Fleiss kappa puanı 0,43'tür ve orijinal EmotionLines kappa puanı olan 0,34'ten daha yüksektir. Çok kısa uzunlukta videoların kullanımının ve dağıtımının adil kullanım kategorisine girdiğini doğrulamak için bir hukuk bürosuna danışılmıştır. MELD verisetindeki duygu sınıflarının dağılımı Çizelge 4.4'te ayrıntılı olarak verilmektedir.

Çizelge 4.4. MELD Verisetindeki Duygu Sınıflarının Dağılımı.

Duygu	**Örnek Sayısı**
Kızgın	1607
Mutlu	2308
Nötr	6436
Üzgün	1002
Korku	358
Sürpriz	1636
Tiksinti	361
Toplam	**13708**

4.5. TEMEL FREKANS VERİSETLERİ

Temel frekans çalışmalarında Hillenbrand Sesli Harf Veriseti, Texas Sesli Harf Veriseti ve TIMIT veri kümesinin sesli harf kısmını kullanıyoruz. Hillenbrand Sesli Harf veri kümesi, erkek, kadın, erkek çocuk ve kız çocuk konuşma örnekleri dahil olmak üzere 1668 sesli harften oluşmaktadır. Örneklerin toplam uzunluğu 223,878 saniyedir ve

dosyaların uzunluğu 0,1280 ile 0,4295 saniye arasında değişmektedir. Ortalama uzunluk 0,134 ve standart sapma 0,0305 saniyedir. Texas Sesli Harf veri kümesi, çocuk, kadın ve erkek sesli harfleri olan 3314 örnekten oluşan bir kümedir. Toplam uzunluk 698,4711 saniyedir ve dosyaların uzunluğu 0,069 ile 1,516 saniye arasında değişmektedir. Ortalama uzunluk 0,210 saniyedir. TIMIT veriseti eğitim, geliştirme ve test setlerine ayrılır. Bu çalışmada, TIMIT kümesinin SA örnekleri de dahil olmak üzere tüm sesli harflerini kullanıyoruz. TIMIT SA örnekleri, aşırı öğrenme sorunlarına yol açması nedeniyle sınıflandırma uygulamalarında kullanılmaz, ancak bu bizim durumumuzda geçerli değildir. TIMIT sesli harf seti, 24017 kadın ve 54357 erkek sesi dahil olmak üzere toplam 78374 örnek içerir. Toplam uzunluk 7521,265 saniyedir, dosyaların uzunluğu 0,0046 ile 0,4834 saniye arasındadır, ortalama uzunluk 0,0959 saniyedir. Cinsiyet dağılımı Çizelge 4.5'te gösterilmiştir. Tüm verisetlerinde örnekler mono formattadır ve örnekleme frekansı 16 kHz'tir.

Çizelge 4.5. Temel Frekans Deneylerinde Kullanılan Verisetleri.

	Erkek	**Kadın**	**Kız Çocuk**	**Erkek Çocuk**
Hillenbrand	540	576	228	324
Texas	1232	1110	972	
TIMIT	54357	24017		

Hillenbrand ve Texas Sesli Harf verisetleri temel frekans gerçek değerlerini hazır olarak vererek bize tahmini değerlerin doğruluğunu test etme açısından kolaylık sağlamaktadırlar. Hillenbrand Sesli Harf verisetinde, temel frekans gerçek değeri otokorelasyon ve ardından el düzenlemesi kullanılarak, Texas Sesli Harf verisetinde ise yarı otomatik LPC analizi ve görsel uzman incelemesi ile hesaplanmıştır. TIMIT verisetinde transkripsiyonlar MIT'nin SPIRE (Surface Prior Information Reflectance Estimation) programı kullanılarak elde edilir ve daha sonra deneyimli akustik fonetikçiler tarafından elle doğrulanır. Ancak TIMIT sesli harfleri için temel frekans gerçek değerleri hesaplanmamıştır. Bu nedenle, HDM, otokorelasyon, kepstrum, YAAPT, CREPE ve FCN yöntemleri tarafından bulunan f0 değerlerinin ortalamasını TIMIT veri kümesi tarafından sağlanan cinsiyet etiketleriyle birlikte tutarlı bir şekilde kullanarak temel frekans gerçek değerlerini oluşturduk. TIMIT veri kümesinin bazı örneklerinde, frekans spektrumunda görsel uzman incelemesi ile dahi temel frekansı bulmak çok zordur. Harmonikler arası aralıklar neredeyse rastgele bir şekilde yayılmaktadır.

4.6. ÖZNİTELİK SEÇME

Çalışmanın bu bölümünde Information Gain ve Correlation Based Feature Selection Subset Evaluator öznitelik seçicileri tanıtıyoruz. Öznitelik seçme belirli sayıdaki özniteliklerden bazı kriterleri kullanarak gereksiz öznitelikleri ayıklar ve öznitelik sayısını azaltmayı hedefler. Sınıflandırıcılar bu sayede daha hızlı çalışır ve daha iyi sonuçlarda elde edebilirler.

4.6.1. Information Gain Öznitelik Seçici

Information Gain öznitelik seçici, özniteliklerin sınıflara göre bilgi kazançlarını ölçerek değerlendirir. İyi öznitelikler maksimum bilgi kazancı verir ve ilgisiz öznitelikler hiç veya az bilgi kazancı verir. Bilgi kazancı entropideki azalmayı ölçen bir metriktir. Entropi, veri kümesindeki safsızlıktır. Bir veri kümesini böldüğümüzde, iyi bir kriter kullanmalıyız. Bölünme ne kadar iyi ise, sınıfların görünürlüğü de o kadar iyidir. Bu çalışmada Information Gain öznitelik seçme metodunu Weka aracındaki Ranker arama yöntemi ile kullandık. Ranker, entropi gibi yöntemlerle öznitelikleri ayrı ayrı değerlendirmek için kullanılır. Ranker, özniteliklerin alt kümelerini bulmak için bir arama algoritması değildir, bunun yerine öznitelikleri bilgi kazancına göre sıralar. Ayrıca, önceden tanımlanmış bir eşik değeri kullanarak öznitelikleri seçer. Düşük dereceli öznitelikleri kaldırmak için bir eşik değeri ayarlanabilir [26], [27].

4.6.2. CFS Subset Öznitelik Seçici

CFS Subset Evaluator öznitelik alt kümelerini çift yönlü olarak değerlendirmek için korelasyon katsayılarını kullanır ve birbirleriyle en az ilgisi olan öznitelikleri bulmaya çalışır. CFS algoritmasının ana kriteri, çoğunlukla sınıflarla yüksek ilişkili ve birbirleriyle düşük ilişkili öznitelikleri bulmaktır. Gereksiz öznitelikler, sınıflarla düşük korelasyonları ve diğer özniteliklerden bir veya daha fazlası ile yüksek korelasyonları nedeniyle bırakılacaktır. Özellik seçimi ölçütleri, bu özelliğin tahmin gücüne bağlıdır. Bu tezde, CFS, Weka aracındaki LFS (Linear Forward Selection) algoritması ile uygulanmaktadır. LFS, BestFirst [111] aramasının bir uzantısıdır ve özellik alt alanı arama değerlendirmelerinin sayısını azaltır. LFS, her adımda öznitelik genişletmelerinin sayısını azaltır ve öznitelikleri ayrı ayrı sıralamak için iki çalışma moduna sahiptir. İlk mod, ileriye dönük en iyi ilk arama kullanarak ilk-k özelliklerinde çalışır. İkinci mod, üst sıralardaki k özelliklerden yeni bir özellik ekleyerek en iyi alt kümeyi genişletir. İşlem

modu, geri izleme derecesi, arama önbelleğinin boyutu, özelliklerin başlangıç listesi ve k değeri herhangi bir karar algoritması tarafından belirlenebilir [27].

4.7. DENGESİZ DAĞILIMLI VERİ İŞLEME YÖNTEMLERİ

Dengesiz dağılımlı veriler makine öğrenimi uygulamaları için çok ciddi bir sorundur. Dengesiz veri kümelerinde, bazı sınıfların örneklerinin sayısı diğer sınıflardan çok düşüktür ve bu durum, sınıflandırıcıların örnek sayısı çok olan sınıfı aşırı öğrenmesine neden olur. Sınıflandırma sonuçları önyargılı ve yanlış olma eğilimindedir. Aşırı öğrenme ve dengesiz karmaşıklık matrisi bu tür veri kümelerinin yaygın sorunlarıdır. Dengesizliğin nedeni farklı olabilir. Önyargılı örnekleme, ölçüm hataları ve örnekleme alanının sorunlu doğası bu nedenler arasındadır. Örneğin, kanserli hasta veri kümelerinde, doğal olarak kanserli olmayan örneklerin sayısı, kanserli örneklerin sayısına kıyasla çok yüksektir. Makine öğrenimi görevlerinde dengesizlik sorunuyla mücadele etmek için birçok teknik kullanılır. Örnek azaltma, örnek türetme, SMOTE (Synthetic Minority Oversampling Technique) [112], Bagging [113], AdaBoost (Adaptive Boosting) [114], XGBoost (Extended Gradient Boost) [115] gibi topluluk sınıflandırıcılı metotlar dengesizlik problemi için çözüm olabilir. Doğrudan başarı oranı kullanmak yerine Karmaşıklık Matrisi (Confusion Matrix), Hassasiyet (Precision), F1 skoru, F-Beta ölçütü, Kappa, Duyarlılık (Sensitivity), Özgüllük (Specificity), Geometrik Ortalama, Çapraz Entropi, BrierScore, ROC (Receiver Operating Characteristics), ROC AUC (Area Under Curve), PR (Precision-Recall) AUC metrikleri de dengesizlik problemlerinin çözümü için kullanılabilir [116]-[118].

5. SONUÇLAR

Bu bölümde, SVM, NVIDIA CNN, 1B CNN, LSTM ve Bidirectional LSTM kullanarak EmoSTAR, EmoDB, IEMOCAP ve MELD veri kümelerinde Information Gain Öznitelik seçici, CFS Subset Öznitelik seçici, Veri Türetme ve Dengesiz Veri Metotları deneylerimizi ele alıyoruz. SVM için Weka 3.6.10, Derin Öğrenme modelleri için Python 3.6.5 [119] ile Tensorflow 1.7.1 tabanlı [120] Keras 2.2.4 [121] sürümünü, öznitelik çıkarma için MATLAB 2019a sürümünü kullandık. Diğer derin öğrenme platformlarının bir karşılaştırması Kabakuş tarafından [122]'de verilmiştir. Ses dosyaları 512 örnekli çerçeveleme ve 256 örnekli pencereleme ile işlenmektedir. Önvurgu α=0,97 ile uygulanmıştır. Ses dosyaları örnekleme frekansı 16000 Hz olmak üzere tek kanallıdır. Tüm sınıflandırmalar %70 eğitim ve %30 test verileri ile yapılmıştır.

5.1. TÜM ÖZNİTELİKLERLE YAPILAN DENEYLER

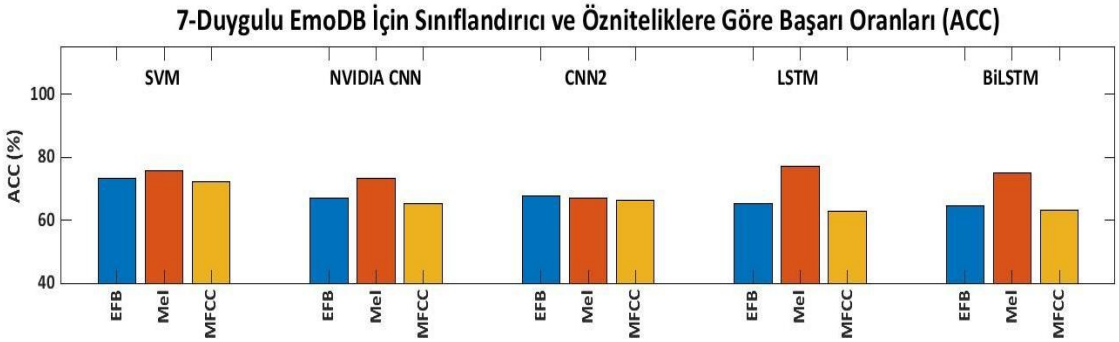
EmoSTAR, EmoDB, IEMOCAP ve MELD veri kümeleri üzerinde SVM, NVIDIA CNN modeli (CNN1), 1B konvolusyon katmanlarına sahip ikinci bir CNN modeli (CNN2), LSTM ve Bidirectional LSTM kullanarak konuşma duygu tanıma deneylerimizi gerçekleştirdik. Giriş olarak önerilen EFB filtrelerini, Mel filtrelerini ve MFCC özelliklerini kullandık. Mel ve MFCC özellikleri Auditory Toolbox [123] kullanılarak elde edilmiştir. EFB filtreleri de aynı koddan uyarlanmıştır. 4-duygulu (Kızgın, Mutlu, Nötr, Üzgün) ve 7 duygulu (Kızgın, Mutlu, Nötr, Üzgün, Korku, Tiksinti, Can Sıkıntısı) sınıflandırmalar çalıştırılmıştır. EFB ve MFCC öznitelik setleri 573 (13 filtre, 13 delta, 22 istatistik fonksiyon), Mel filtre öznitelik setleri ise 1761 (40 filtre, 40 delta, 22 istatistik fonksiyon) öznitelik içermektedir. Öznitelik çıkarma için kullanılan istatistik fonksiyonlar en büyük değer, en küçük değer, en büyük değer noktası, en küçük değer noktası, değer aralığı, ortalama, standart sapma, varyans, eğrilik, diklik, 1., 2., 3. çeyrekler, 1., 2., 3., çeyrek aralıkları, 90 ve 98 yüzdelikler, 1. ve 2. doğrusal yaklaşım katsayıları, doğrusal hata yaklaşıklığı, karesel hata yaklaşıklığıdır. Tam öznitelik setleri, HDM algoritması kullanılarak elde edilen temel frekansa dayalı prozodik özellikleri de (temel frekans, sıfır geçiş oranı, enerji) içermektedir. EFB ve MFCC Tam öznitelik setleri 705, Mel filtre Tam öznitelik setleri ise 1893 öznitelik içermektedir.

Sonuçlardan görülebileceği gibi önerilen EFB filtreleri Mel ve MFCC'ye göre daha iyi veya karşılaştırılabilir sonuçlar üretmekte, hesaplanmaları daha hızlı ve yorumlanmaları daha kolaydır. EFB filtreleri deneylerin çoğunluğunda MFCC özelliklerinden daha üstün olduğu gibi Mel filtre bankalarıyla da ciddi bir rekabet içindedir. SVM, Derin modellere göre oldukça başarılıdır. Algoritma sonuçları arasında büyük farklar vardır. Model karmaşıklığı mutlaka daha yüksek başarı oranı anlamına gelmemektedir. Başarı oranları, veri kümesi ve yönteme göre büyük ölçüde değişmektedir. Kırmızı değerler, belirtilen veri kümesinin en iyi performansını, kalın değerler belirtilen sütundaki en iyi değeri göstermektedir. Aşağıdaki çizelgelerde K, Kızgın; M, Mutlu; N, Nötr; Ü, Üzgün; S, Sıkkın; T, Tiksinti; Kr, Korku duygu sınıflarını temsil etmektedir. CNN1 NVIDIA CNN modelini, CNN2 ise 1B CNN modelini göstermektedir.

Çizelge 5.1 ve Şekil 5.1'den görüleceği gibi 7-duygulu EmoDB verisetinde en başarılı sonuç %77,02 ile Mel filtre bankaları ve LSTM ile elde edilmiştir. EFB bütün deneylerde MFCC'den daha iyi sonuçlar üretmiştir. SVM bu sınıflandırmalarda oldukça başarılı bir performans sergileyerek NVIDIA CNN (CNN1), CNN2 ve BiLSTM modellerinden daha iyi sonuçlar üretebilmiştir.

Çizelge 5.1. 7-Duygulu EmoDB Üzerindeki Deney Sonuçları (%ACC).

7-duygu (K,M,N,Ü,S,T,Kr)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB EFB	73,27	67,08	**67,70**	65,22	64,59
EmoDB Mel	**75,88**	**73,29**	67,08	77,02	**75,15**
EmoDB MFCC	72,33	65,22	66,46	62,73	63,35

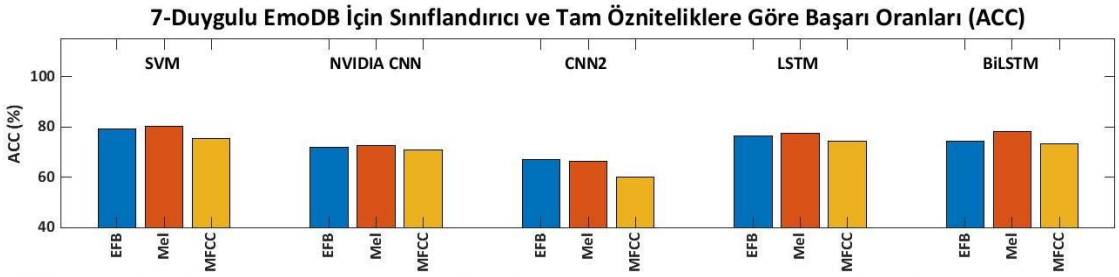


Şekil 5.1. 7-duygulu EmoDB üzerindeki deney sonuçlarının çubuk grafikleri.

Çizelge 5.2 ve Şekil 5.2'den görüleceği gibi 7-duygulu EmoDB verisetinde prozodik özellikler eklenerek en başarılı sonuç %80,37 ile Mel filtre bankaları ve SVM sınıflandırıcıyla elde edilmiştir. EFB bütün deneylerde MFCC'den daha iyi sonuçlar üretmiş ve Mel filtreleri ile de rekabet içindedir.

Çizelge 5.2. 7-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC).

7-duygu (K,M,N,Ü,S,T,Kr)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB Tam EFB	79,25	72,05	**67,08**	76,40	74,53
EmoDB Tam Mel	80,37	**72,67**	66,46	**77,64**	**78,26**
EmoDB Tam MFCC	75,32	70,81	60,25	74,53	73,29

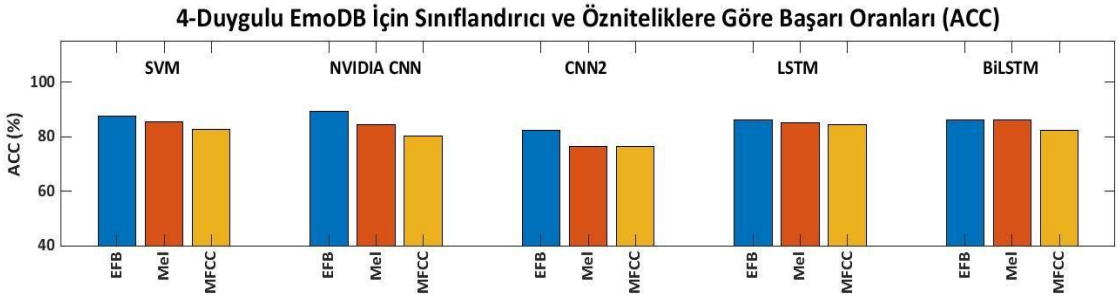


Şekil 5.2. 7-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri.

Çizelge 5.3 ve Şekil 5.3'ten görüleceği gibi 4-duygulu EmoDB verisetinde en başarılı sonuç EFB filtre bankaları ve NVIDIA CNN (CNN1) ile elde edilmiştir. EFB bütün deneylerde Mel filtreleri ve MFCC'den daha iyi sonuçlar üretmiştir.

Çizelge 5.3. 4-Duygulu EmoDB Üzerindeki Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB EFB	**87,61**	89,22	**82,35**	**86,27**	**86,27**
EmoDB Mel	85,54	84,31	76,47	85,29	**86,27**
EmoDB MFCC	82,89	80,39	76,47	84,31	82,35

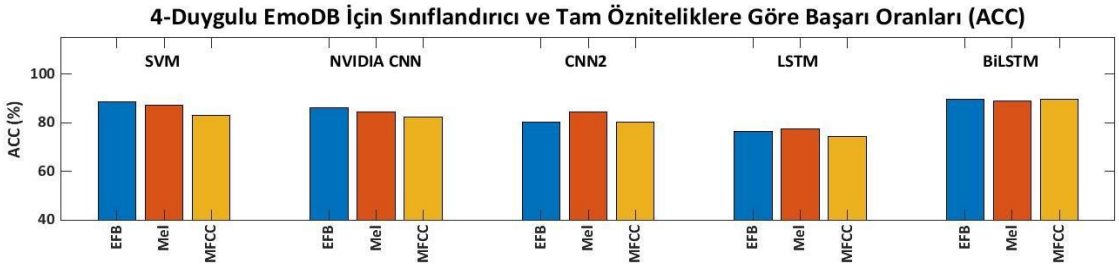


Şekil 5.3. 4-duygulu EmoDB üzerindeki deney sonuçlarının çubuk grafikleri.

Çizelge 5.4 ve Şekil 5.4'ten görüleceği gibi 4-duygulu EmoDB verisetinde prozodik öznitelikler eklenerek en başarılı sonuç EFB filtreleri ve MFCC ile BiLSTM kullanılarak elde edilmiştir. EFB, CNN2 ve BiLSTM dışındaki bütün deneylerde MFCC'den daha iyi sonuçlar üretmiştir.

Çizelge 5.4. 4-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB Tam EFB	**88,79**	**86,27**	80,39	76,40	89,83
EmoDB Tam Mel	87,31	84,31	**84,31**	**77,64**	88,98
EmoDB Tam MFCC	83,18	82,35	80,39	74,53	89,83

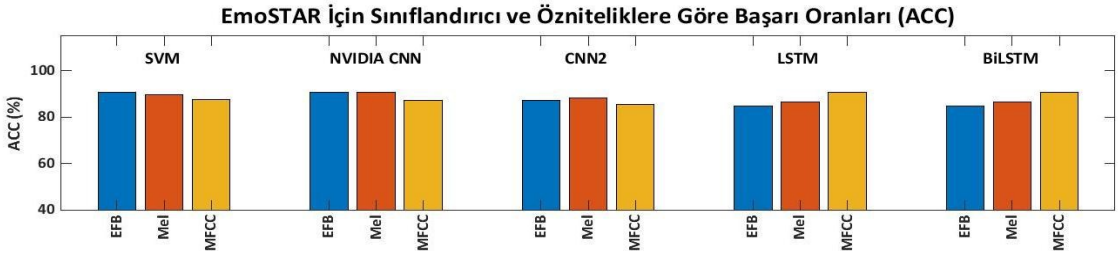


Şekil 5.4. 4-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri.

Çizelge 5.5 ve Şekil 5.5'ten görüleceği gibi 4-duygulu EmoSTAR verisetinde EFB, MFCC ve Mel filtreleri değişik sınıflandırıcılarla %90,68 ile en başarılı sonuçları elde edebilmişlerdir. Bu deneylerde EFB SVM, NVIDIA CNN (CNN1) ve CNN2 ile MFCC'den daha iyi sonuçlar elde edebilmiştir.

Çizelge 5.5. 4-Duygulu EmoSTAR Üzerindeki Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoSTAR EFB	**90,58**	90,68	87,29	84,75	84,75
EmoSTAR Mel	89,56	90,68	**88,14**	86,44	86,44
EmoSTAR MFCC	87,53	87,29	85,59	90,68	90,68

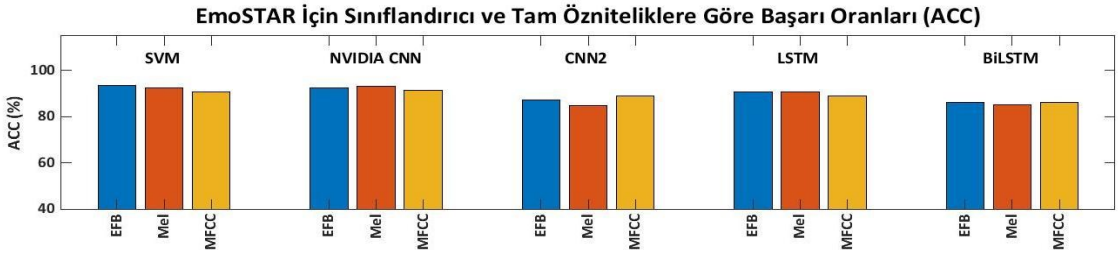


Şekil 5.5. 4-duygulu EmoSTAR üzerindeki deney sonuçlarının çubuk grafikleri.

Çizelge 5.6 ve Şekil 5.6'dan görüleceği gibi 4-duygulu EmoSTAR verisetinde prozodik öznitelikler eklenerek EFB, %93,63 ile SVM sınıflandırıcıda MFCC ve Mel filtrelerinden daha başarılı sonuç elde etmiştir. Bu deneylerde EFB SVM, NVIDIA CNN (CNN1) ve CNN2 ile MFCC'den daha iyi sonuçlar elde edebilmiştir.

Çizelge 5.6. 4-Duygulu EmoSTAR Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoSTAR Tam EFB	93,63	92,37	87,29	**90,68**	**86,27**
EmoSTAR Tam Mel	92,36	**93,22**	84,75	**90,68**	85,29
EmoSTAR Tam MFCC	90,58	91,53	**88,98**	88,98	**86,27**

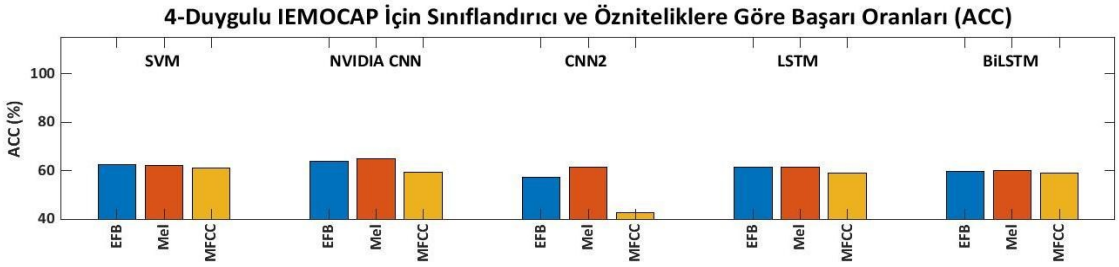


Şekil 5.6. 4-duygulu EmoSTAR üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri.

Çizelge 5.7 ve Şekil 5.7'den görüleceği gibi 4-duygulu IEMOCAP verisetinde Mel filtreleri, %65,03 ile NVIDIA CNN (CNN1) sınıflandırıcısında MFCC ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. Bu deneylerde EFB bütün deneylerde MFCC'den daha iyi sonuçlar elde edebilmiş ve Mel filtreleri ile rekabet halindedir.

Çizelge 5.7. 4-Duygulu IEMOCAP Üzerindeki Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP EFB	**62,37**	63,85	57,31	61,32	59,69
IEMOCAP Mel	62,08	65,03	**61,54**	**61,47**	**60,13**
IEMOCAP MFCC	60,99	59,32	42,84	59,02	59,02

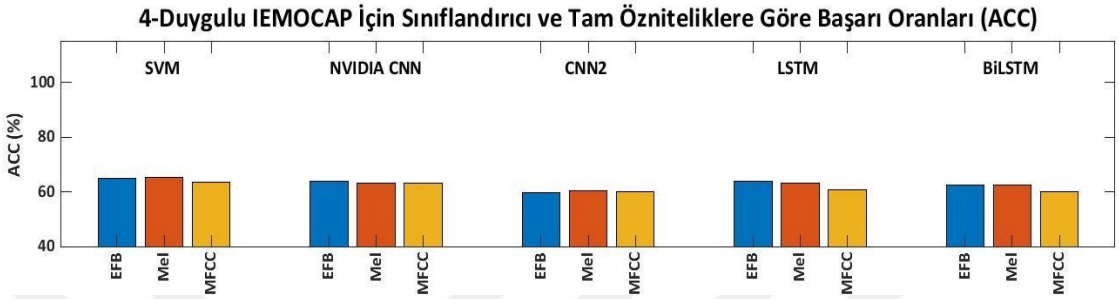


Şekil 5.7. 4-duygulu IEMOCAP üzerindeki deney sonuçlarının çubuk grafikleri.

Çizelge 5.8 ve Şekil 5.8'den görüleceği gibi 4-duygulu IEMOCAP verisetinde prozodik öznitelikler eklenerek Mel filtreleri %65,30 ile SVM sınıflandırıcısında MFCC ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. EFB, CNN2 dışında kalan deneylerde MFCC'den daha iyi sonuçlar elde edebilmiş ve Mel filtreleri ile rekabet halindedir.

Çizelge 5.8. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP Tam EFB	64,81	**64,07**	59,61	**64,07**	62,36
IEMOCAP Tam Mel	65,30	63,25	**60,58**	63,18	**62,66**
IEMOCAP Tam MFCC	63,54	63,25	59,99	60,95	60,13

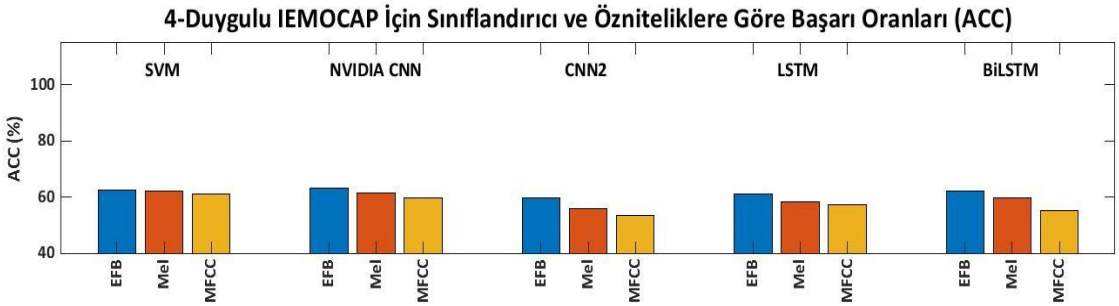


Şekil 5.8. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri.

Çizelge 5.9 ve Şekil 5.9'da IEMOCAP verisetinde eğitim seti %70, test seti %30 olarak ayrıldıktan sonra doğrulama (validation) setini de eğitim setinin %30'u olacak şekilde seçerek deneyimizi gerçekleştiriyoruz. Çizelge 5.9 ve Şekil 5.9'dan görüleceği gibi 4-duygulu IEMOCAP verisetinde EFB filtreleri, %63,32 ile NVIDIA CNN (CNN1) sınıflandırıcısında MFCC ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. Bu deneylerde EFB bütün deneylerde Mel filtreleri ve MFCC'den daha iyi sonuçlar elde edebilmiştir.

Çizelge 5.9. 4-Duygulu IEMOCAP Üzerindeki Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP EFB	**62,37**	63,32	**59,83**	**61,02**	**62,06**
IEMOCAP Mel	62,08	61,54	55,90	58,42	59,61
IEMOCAP MFCC	60,99	59,68	53,30	57,46	55,38

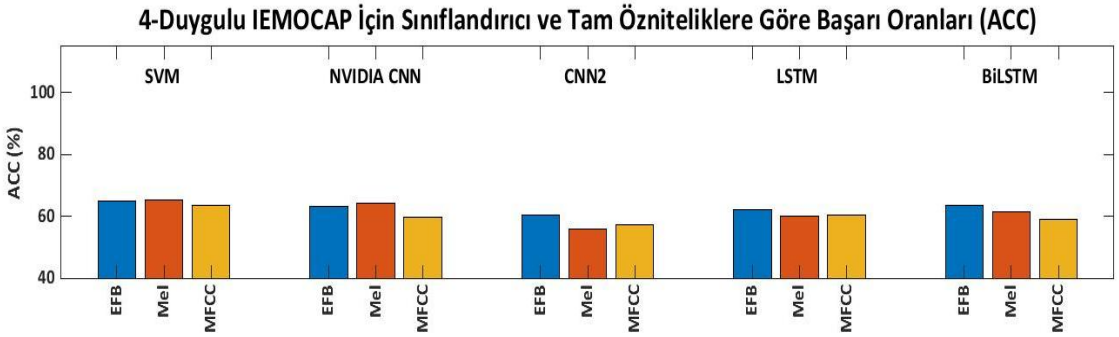


Şekil 5.9. 4-duygulu IEMOCAP üzerindeki deney sonuçlarının çubuk grafikleri.

Çizelge 5.10 ve Şekil 5.10'da IEMOCAP verisetinde eğitim seti %70, test seti %30 olarak ayrıldıktan sonra doğrulama (validation) setini de eğitim setinin %30'u olacak şekilde seçerek deneyimizi gerçekleştiriyoruz. Çizelge 5.10 ve Şekil 5.10'dan görüleceği gibi 4-duygulu IEMOCAP verisetinde prozodik öznitelikler eklenerek Mel filtreleri %65,30 ile SVM sınıflandırıcısında MFCC ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. EFB, bütün deneylerde MFCC'den daha iyi sonuçlar elde edebilmiştir. EFB filtreleri CNN2, LSTM ve BiLSTM sınıflandırıcılarda Mel filtrelerinden daha iyi sonuçlar üretebilmiştir. SVM bu deneyde bütün öznitelik setleri için en başarılı sınıflandırıcı olarak kendini göstermektedir.

Çizelge 5.10. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP Tam EFB	64,81	63,25	**60,28**	**62,13**	**63,47**
IEMOCAP Tam Mel	65,30	**64,29**	56,05	59,98	61,61
IEMOCAP Tam MFCC	63,54	59,83	57,23	60,50	59,09

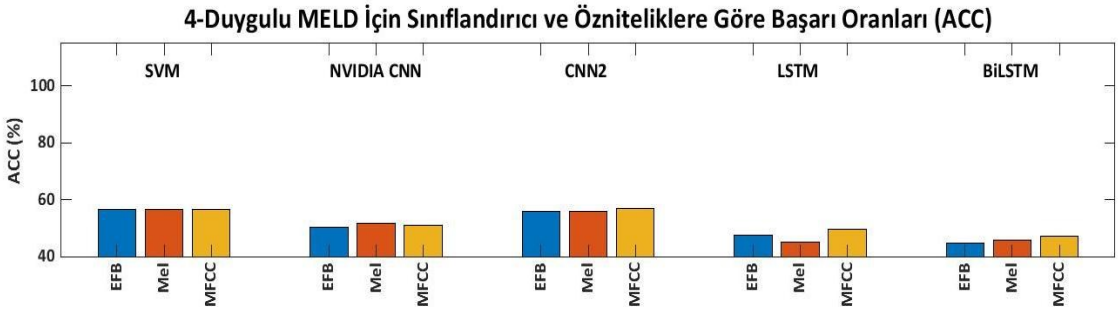


Şekil 5.10. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri.

Çizelge 5.11 ve Şekil 5.11'den görüleceği gibi 4-duygulu MELD verisetinde MFCC filtreleri, %57,08 ile CNN2 sınıflandırıcısında Mel filtreleri ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. SVM sınıflandırıcının performansı oldukça yüksektir ve MFCC öznitelik setinin CNN2 sınıflandırıcısı dışındaki tüm deneylerde en yüksektir.

Çizelge 5.11. 4-Duygulu MELD Üzerindeki Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
MELD EFB	56,68	50,47	55,81	47,65	44,80
MELD Mel	56,63	**51,70**	55,81	45,01	45,95
MELD MFCC	**56,69**	51,20	57,08	**49,50**	**47,36**

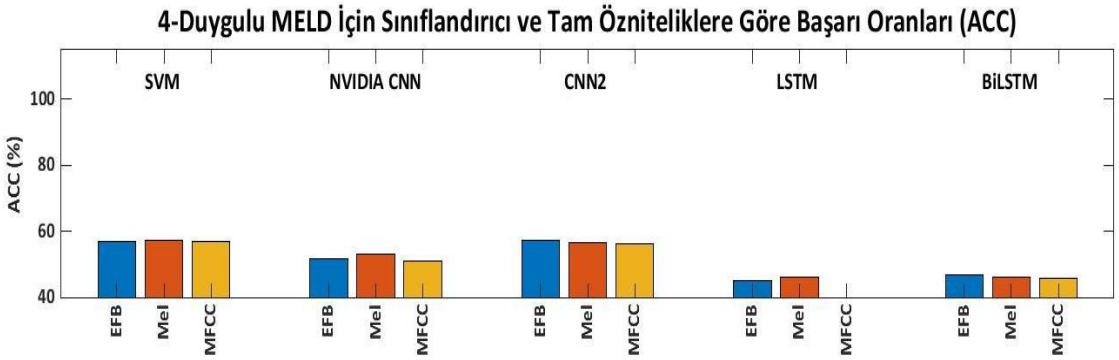


Şekil 5.11. 4-duygulu MELD üzerindeki deney sonuçlarının çubuk grafikleri.

Çizelge 5.12 ve Şekil 5.12'den görüleceği gibi 4-duygulu IEMOCAP verisetinde prozodik özellikler eklenerek Mel filtreleri, %57,25 ile SVM sınıflandırıcısında MFCC ve EFB filtrelerinden daha başarılı sonuç elde etmiştir.

Çizelge 5.12. 4-Duygulu MELD Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
MELD Tam EFB	56,79	51,64	**57,19**	45,10	**46,71**
MELD Tam Mel	57,25	**53,20**	56,46	**46,15**	46,24
MELD Tam MFCC	57,00	51,17	56,14	39,87	45,86



Şekil 5.12. 4-duygulu MELD üzerindeki prozodik öznitelikler eklenerek yapılan deney sonuçlarının çubuk grafikleri.

5.2. INFORMATION GAIN ÖZNİTELİK SEÇİCİ DENEYLERİ

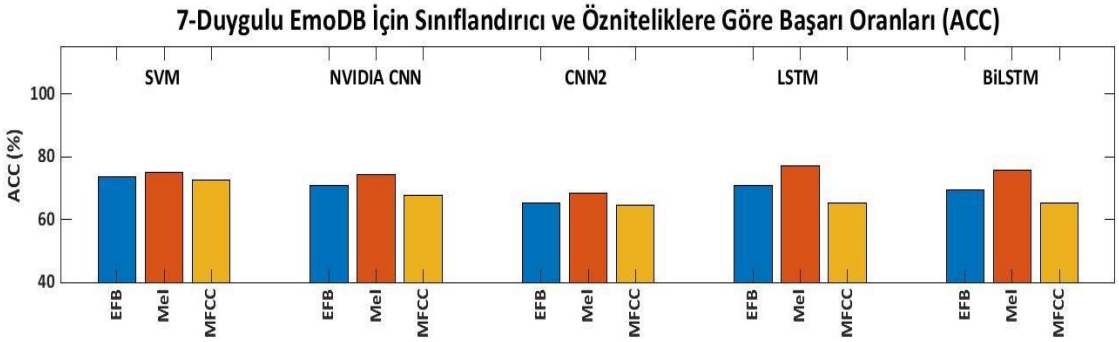
Bilgi Kazancı, özelliklerin etkinliği hakkında bir fikir edinmek için çok kullanışlı ve yaygın olarak kullanılan bir metriktir. Deneylerimizde, eşik olarak 0,01'den fazla bilgi kazancına sahip öznitelikleri Weka aracını kullanarak seçtik. Information Gain'de seçilen özniteliklerin sayısı CFS Subset Evaluator öznitelik seçicisine kıyasla çok yüksektir. Yeni filtre bankaları, CFS Subset Evaluator öznitelik seçicisine kıyasla yüksek sonuçlar

üretmektedir. Sınıflandırıcılar arasında belirgin bir üstünlük farkı olmamasına rağmen EFB filtreleri Mel filtreleri ve MFCC'ye göre daha başarılıdır. Bu bölümdeki çizelgelerde parantez içindeki sayılar Information Gain öznitelik seçici ile seçilen öznitelik sayısını göstermektedir.

Çizelge 5.13 ve Şekil 5.13'ten görüleceği gibi 7-duygulu EmoDB verisetinde Mel filtreleri, %77,02 ile LSTM sınıflandırıcısında MFCC ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. EFB filtrelerinde seçilen öznitelik sayısı 401, Mel filtrelerinde seçilen öznitelik sayısı 1134'e göre oldukça azdır. Bu deneylerde EFB, bütün deneylerde MFCC'den daha iyi sonuçlar elde edebilmiştir. EFB filtreleri en iyi sonucu %73,64 ile SVM sınıflandırıcı ile elde etmiştir. Mel filtreleri bu deneylerde bariz bir üstünlük göstermektedir. Derin öğrenme modellerinde MFCC ile Mel filtreleri arasındaki farklar oldukça dikkat çekicidir ve LSTM sınıflandırıcı da 11,80'e ulaşmıştır.

Çizelge 5.13. 7-Duygulu EmoDB Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

7-duygu (K,M,N,Ü,S,T,Kr)	**SVM**	**CNN1**	**CNN2**	**LSTM**	**BiLSTM**
EmoDB EFB (401)	73,64	70,81	65,22	70,81	69,57
EmoDB Mel (1134)	**75,14**	**74,53**	**68,32**	77,02	**75,78**
EmoDB MFCC (311)	72,52	67,70	64,60	65,22	65,22

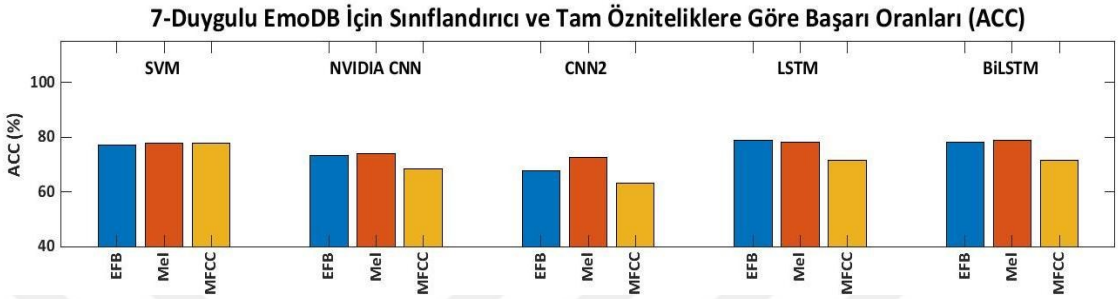


Şekil 5.13. 7-duygulu EmoDB üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.14 ve Şekil 5.14'ten görüleceği gibi 7-duygulu EmoDB verisetinde prozodik öznitelikler eklenerek EFB ve Mel filtreleri, %78,88 ile sırasıyla LSTM ve BiLSTM sınıflandırıcılarında en başarılı sonucu elde etmişlerdir. Bu deneylerde EFB, SVM dışındaki deneylerde MFCC'den daha iyi sonuçlar elde edebilmiştir. Ayrıca EFB filtrelerinde seçilen öznitelik sayısı 471, Mel filtrelerinde seçilen öznitelik sayısı 1204'e göre oldukça azdır.

Çizelge 5.14. 7-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

7-duygu (K,M,N,Ü,S,T,Kr)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB Tam EFB (471)	77,19	73,29	67,70	78,88	78,26
EmoDB Tam Mel (1204)	77,75	**73,91**	**72,67**	78,26	78,88
EmoDB Tam MFCC (381)	**77,94**	68,32	63,35	71,43	71,43

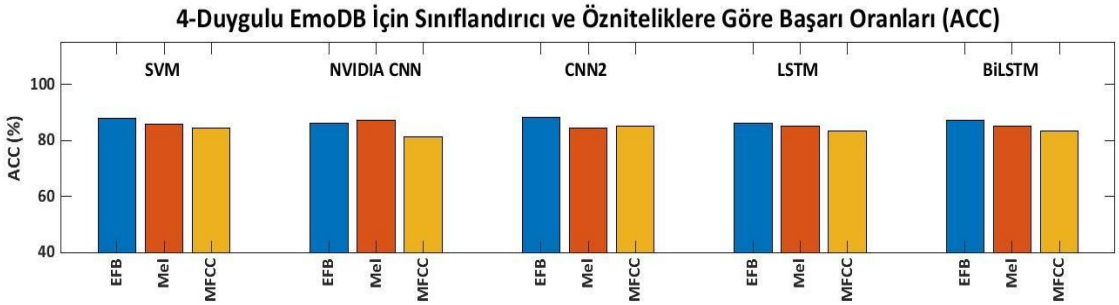


Şekil 5.14. 7-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.15 ve Şekil 5.15'ten görüleceği gibi 4-duygulu EmoDB verisetinde EFB filtreleri, %88,24 ile CNN2 sınıflandırıcısında MFCC ve Mel filtrelerinden daha başarılı sonuç elde etmiştir. Bu deneylerde EFB, bütün deneylerde MFCC'den daha iyi sonuçlar elde edebilmiştir. Ayrıca EFB filtrelerinde seçilen öznitelik sayısı 412, Mel filtrelerinde seçilen öznitelik sayısı 1166'ya göre oldukça azdır.

Çizelge 5.15. 4-Duygulu EmoDB Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB EFB (412)	**87,90**	86,27	88,24	**86,27**	**87,25**
EmoDB Mel (1166)	85,84	**87,25**	84,31	85,29	85,29
EmoDB MFCC (365)	84,36	81,37	85,29	83,33	83,33

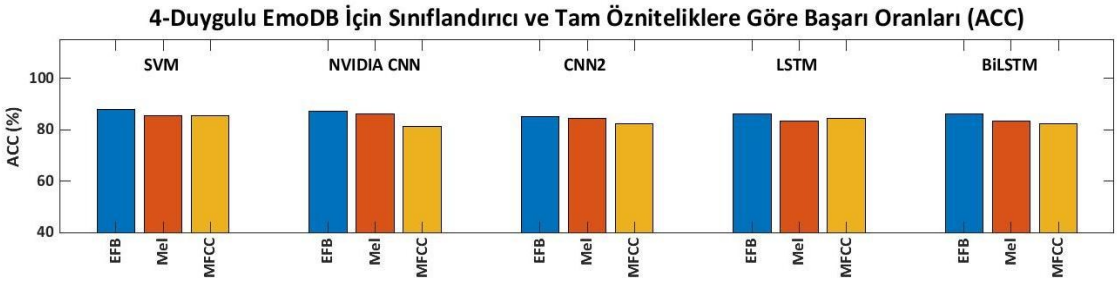


Şekil 5.15. 4-duygulu EmoDB üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.16 ve Şekil 5.16'dan görüleceği gibi 4-duygulu EmoDB verisetinde prozodik öznitelikler eklenerek EFB filtreleri %87,90 başarı oranı ile SVM sınıflandırıcısında Mel filtreleri ve MFCC öznitelik setlerinden daha başarılı sonuçlar elde etmiştir. Bu kısımda EFB bütün deneylerde MFCC ve Mel filtrelerinden daha iyi sonuçlar elde edebilmiştir. Ayrıca EFB filtrelerinde seçilen öznitelik sayısı 482, Mel filtrelerinde seçilen öznitelik sayısı 1236'ya göre oldukça azdır. SVM sınıflandırıcının başarısı ise oldukça dikkat çekicidir. EFB ve MFCC veri kümelerinden seçilen öznitelik sayısı birbirine oldukça yakındır.

Çizelge 5.16. 4-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB Tam EFB (482)	87,90	**87,25**	**85,29**	**86,27**	**86,27**
EmoDB Tam Mel (1236)	85,54	86,27	84,31	83,33	83,33
EmoDB Tam MFCC (435)	85,54	81,37	82,35	84,31	82,35

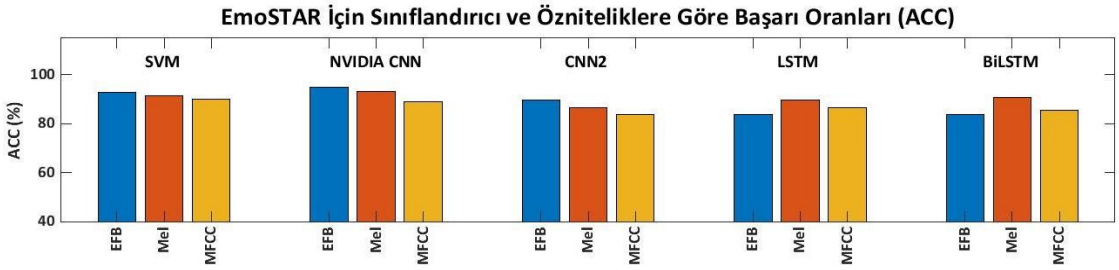


Şekil 5.16. 4-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.17 ve Şekil 5.17'den görülebileceği gibi 4-duygulu EmoSTAR veri kümesinde EFB filtreleri NVIDIA CNN (CNN1) sınıflandırıcısını kullanarak %94,92 başarı oranı ile Mel filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. Ayrıca EFB filtrelerinde seçilen öznitelik sayısı 427, Mel filtrelerinde seçilen öznitelik sayısı 1259'a göre oldukça azdır. EFB daha az sayıda öznitelikle daha iyi sonuçlar üretebilmektedir.

Çizelge 5.17. 4-Duygulu EmoSTAR Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoSTAR EFB (427)	**92,87**	94,92	**89,83**	83,90	83,90
EmoSTAR Mel (1259)	91,34	93,22	86,44	**89,83**	**90,68**
EmoSTAR MFCC (386)	90,07	88,98	83,90	86,44	85,59

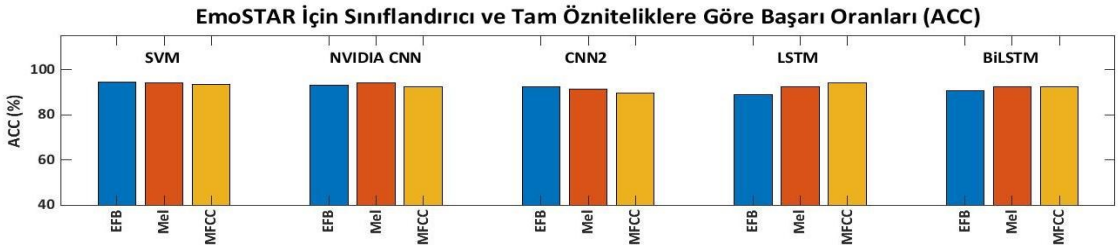


Şekil 5.17. 4-duygulu EmoSTAR üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.18 ve Şekil 5.18'den görüleceği gibi 4-duygulu EmoSTAR verisetinde prozodik öznitelikler eklenerek EFB filtreleri %94,65 ile SVM sınıflandırıcısında Mel filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. Bu deneylerde EFB, LSTM ve BiLSTM dışındaki deneylerde MFCC'den daha iyi sonuçlar elde edebilmiştir. Ayrıca EFB filtrelerinde seçilen öznitelik sayısı 516, Mel filtrelerinde seçilen öznitelik sayısı 1348'e göre oldukça azdır.

Çizelge 5.18. 4-Duygulu EmoSTAR Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	**SVM**	**CNN1**	**CNN2**	**LSTM**	**BiLSTM**
EmoSTAR Tam EFB (516)	94,65	93,22	**92,37**	88,98	90,68
EmoSTAR Tam Mel (1348)	94,14	**94,07**	91,53	92,37	**92,37**
EmoSTAR Tam MFCC (475)	93,63	92,37	89,83	**94,07**	**92,37**

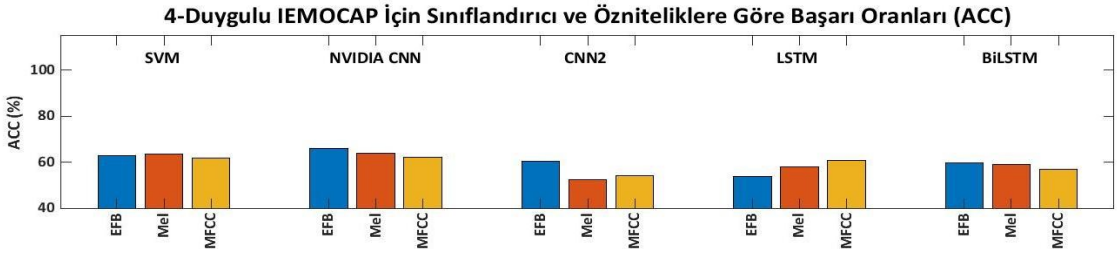


Şekil 5.18. 4-duygulu EmoSTAR üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.19 ve Şekil 5.19'dan görüleceği gibi 4-duygulu IEMOCAP verisetinde EFB filtreleri %66,14 ile Mel filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. Bu oran aynı zamanda IEMOCAP verisetinde bütün deney sınıflarındaki en yüksek başarı oranıdır. Bu deneylerde EFB, LSTM ve SVM dışındaki deneylerde MFCC ve Mel filtrelerinden daha iyi sonuçlar elde edebilmiştir. Ayrıca EFB filtrelerinde seçilen öznitelik sayısı 455, Mel filtrelerinde seçilen öznitelik sayısı 558'den azdır.

Çizelge 5.19. 4-Duygulu IEMOCAP Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP EFB (455)	62,93	66,14	**60,28**	53,97	**59,76**
IEMOCAP Mel (558)	**63,42**	63,84	52,56	**57,98**	58,87
IEMOCAP MFCC (459)	61,89	62,06	54,05	60,73	57,02

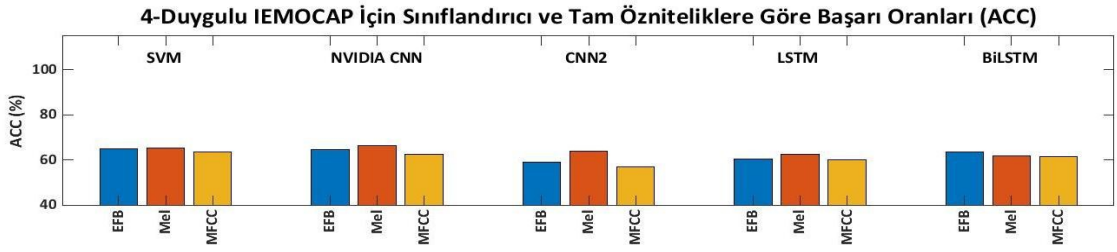


Şekil 5.19. 4-duygulu IEMOCAP üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.20 ve Şekil 5.20'den görüleceği gibi 4-duygulu IEMOCAP verisetinde prozodik öznitelikler eklenerek Mel filtreleri %66,44 ile NVIDIA CNN (CNN1) sınıflandırıcısında en başarılı sonucu elde etmiştir. EFB yaklaşık olarak aynı öznitelik sayısına sahip olmasına rağmen bütün deneylerde MFCC'den daha iyi sonuçlar elde edebilmiştir. SVM sınıflandırıcının diğer derin öğrenme modellerine kıyasla başarısı oldukça iyidir ve Mel filtrelerinin NVIDIA CNN (CNN1) modeli hariç tüm deneylerde daha yüksektir.

Çizelge 5.20. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP Tam EFB (554)	65,10	64,51	59,02	60,50	**63,70**
IEMOCAP Tam Mel (797)	**65,21**	66,44	**63,99**	**62,36**	61,99
IEMOCAP Tam MFCC (558)	63,42	62,58	57,09	60,13	61,32

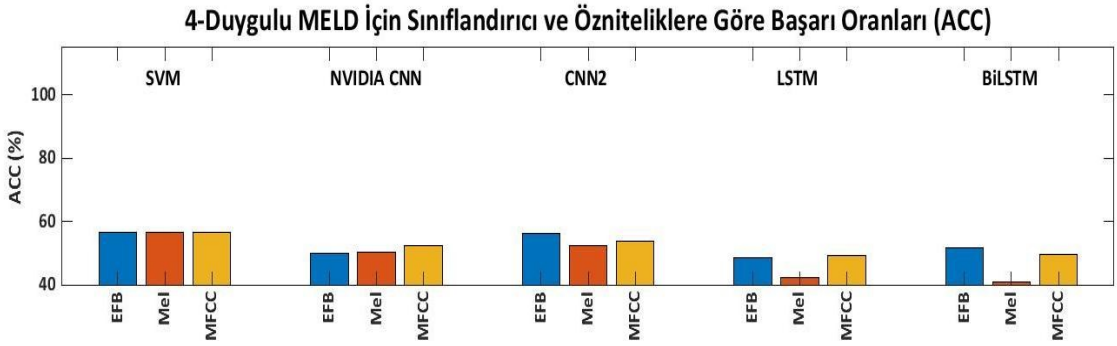


Şekil 5.20. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.21 ve Şekil 5.21'den görüleceği gibi 4-duygulu MELD verisetinde Mel filtreleri %56,69 ile SVM sınıflandırıcısında EFB filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. EFB ve MFCC verisetlerinin başarısı SVM sınıflandırıcıda Mel filtreleriyle hemen hemen aynıdır. Bu deneylerde seçilen öznitelik sayısı bütün verisetlerinde oldukça yakındır. SVM bu kısımda bütün deneylerde en yüksek performansı sergilemektedir. Mel filtreleri kullanan özellik seti bu kısımda derin öğrenme modellerinde MFCC özelliklerinden daha düşük bir performans ortaya koymuştur. Özellikle LSTM ve BiLSTM modellerde MFCC öznitelik setinin Mel filtrelerine göre oldukça bariz bir üstünlüğü gözükmektedir.

Çizelge 5.21. 4-Duygulu MELD Üzerindeki Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	**SVM**	**CNN1**	**CNN2**	**LSTM**	**BiLSTM**
MELD EFB (204)	56,68	49,91	**56,28**	48,47	**51,61**
MELD Mel (271)	56,69	50,46	52,58	42,34	41,02
MELD MFCC (221)	56,68	**52,49**	53,72	**49,35**	49,79

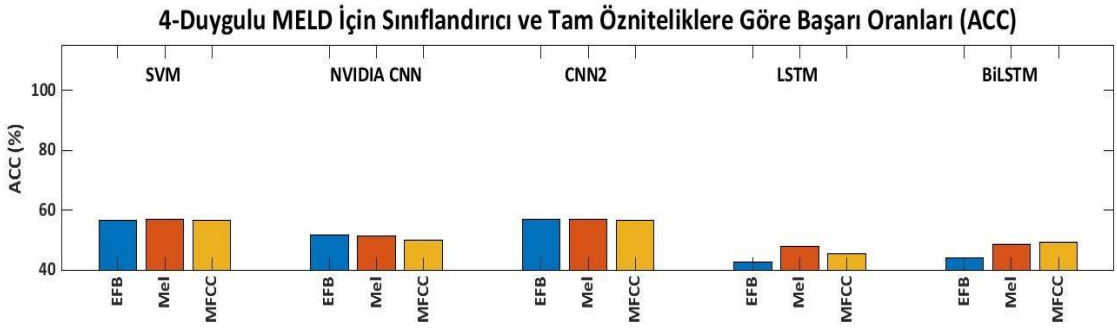


Şekil 5.21. 4-duygulu MELD üzerindeki Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.22 ve Şekil 5.22'den görüleceği gibi 4-duygulu MELD verisetinde prozodik öznitelikler eklenerek Mel filtreleri %57,05 ile CNN2 sınıflandırıcısında EFB filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir.

Çizelge 5.22. 4-Duygulu MELD Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan Information Gain Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	**SVM**	**CNN1**	**CNN2**	**LSTM**	**BiLSTM**
MELD Tam EFB (315)	56,71	**51,79**	56,99	42,78	44,07
MELD Tam Mel (365)	**56,79**	51,29	57,05	**47,92**	48,65
MELD Tam MFCC (271)	56,73	49,88	56,69	45,60	**49,35**



Şekil 5.22. 4-duygulu MELD üzerindeki prozodik öznitelikler eklenerek yapılan Information Gain öznitelik seçici deney sonuçlarının çubuk grafikleri.

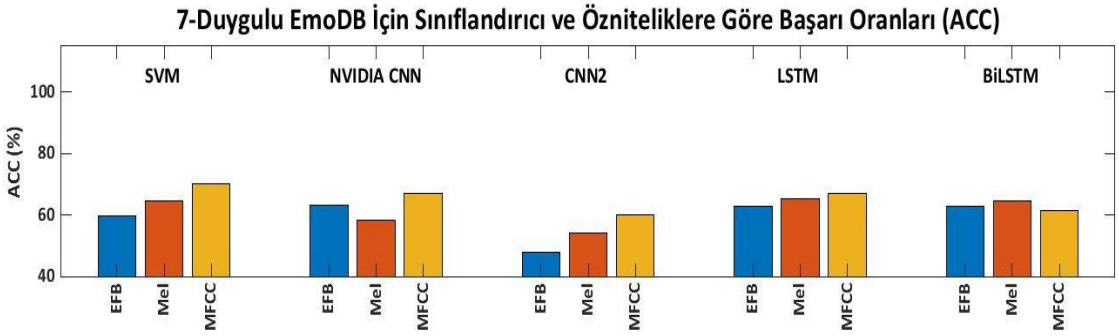
5.3. CFS SUBSET EVALUATOR ÖZNİTELİK SEÇİCİ DENEYLERİ

CFS Subset Evaluator inceleyeceğimiz ikinci öznitelik seçim algoritmasıdır. CFS algoritmasının önemli bir özelliği, Information Gain gibi diğer öznitelik seçicilerine kıyasla öznitelik sayısını çok yüksek oranda azaltabilmesidir. CFS Subset Evaluator, orijinal setten 50 kat daha küçük bir öznitelik seti üretebilir ve daha yüksek başarı oranları sağlayabilir [103]. Ancak bu çalışmada beklenen performanstan uzak kalmıştır. Bu tablolarda parantez içindeki öznitelik sayısı, bütün öznitelik kümesinden seçilen öznitelik sayısını gösterir. Örneğin, EFB özellik setinde 573 özniteliğimiz bulunmaktadır. Çizelge 5.23'ün ilk satırından görüldüğü gibi, CFS Subset Evaluator öznitelik seçici toplam 573 öznitelikten 37 öznitelik seçmektedir. CFS Subset Evaluator öznitelik seçme metodunu Weka veri madenciliği uygulamasında Linear Forward Selection yöntemiyle çalıştırıyoruz.

Çizelge 5.23 ve Şekil 5.23'ten görüleceği gibi 7-duygulu EmoDB verisetinde MFCC %70,09 ile SVM sınıflandırıcısında Mel filtreleri ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. Ancak EFB filtrelerinde seçilen öznitelik sayısı 37, MFCC özniteliklerinde seçilen öznitelik sayısı 156'ya göre oldukça azdır. Mel filtrelerindeki öznitelik sayısı da 1761'den 92'ye kadar düşmüştür.

Çizelge 5.23. 7-Duygulu EmoDB Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

7-duygu (K,M,N,Ü,S,T,Kr)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB EFB (37)	59,62	63,35	47,83	62,73	62,73
EmoDB Mel (92)	64,48	58,39	54,04	65,22	**64,60**
EmoDB MFCC (156)	70,09	**67,08**	**60,25**	**67,08**	61,49

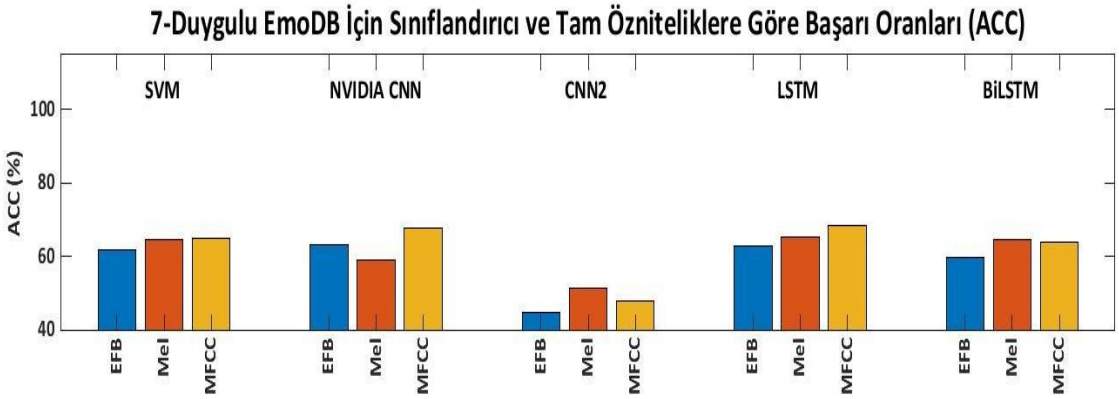


Şekil 5.23. 7-duygulu EmoDB üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.24 ve Şekil 5.24'ten görüleceği gibi 7-duygulu EmoDB verisetinde prozodik öznitelikler eklenerek MFCC %68,32 ile LSTM sınıflandırıcısında Mel filtreleri ve EFB filtrelerinden daha başarılı sonuç elde etmiştir. EFB filtrelerindeki seçilen öznitelik sayısı (38) MFCC öznitelik kümesinde seçilen öznitelik sayısından (71) daha azdır.

Çizelge 5.24. 7-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

7-duygu (K,M,N,Ü,S,T,Kr)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB Tam EFB (38)	61,68	63,35	44,72	62,73	59,63
EmoDB Tam Mel (92)	64,48	59,01	**51,55**	65,22	**64,60**
EmoDB Tam MFCC (71)	**65,04**	**67,70**	47,83	68,32	63,98

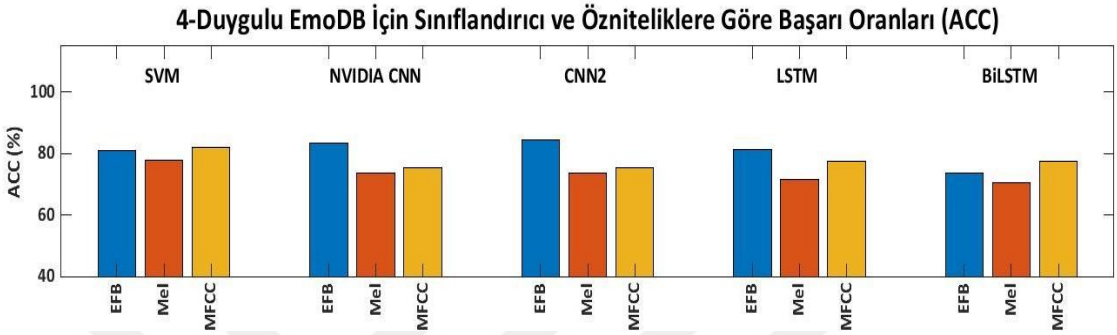


Şekil 5.24. 7-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.25 ve Şekil 5.25'ten görüleceği gibi 4-duygulu EmoDB verisetinde EFB %84,31 ile CNN2 sınıflandırıcısında Mel filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. Ayrıca EFB filtrelerinde seçilen öznitelik sayısı 33, MFCC özniteliklerinde seçilen öznitelik sayısı 132'ye göre oldukça azdır.

Çizelge 5.25. 4-Duygulu EmoDB Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB EFB (33)	81,12	**83,33**	84,31	**81,37**	73,53
EmoDB Mel (33)	77,87	73,53	73,53	71,57	70,59
EmoDB MFCC (132)	**82,00**	75,49	75,49	77,45	**77,45**

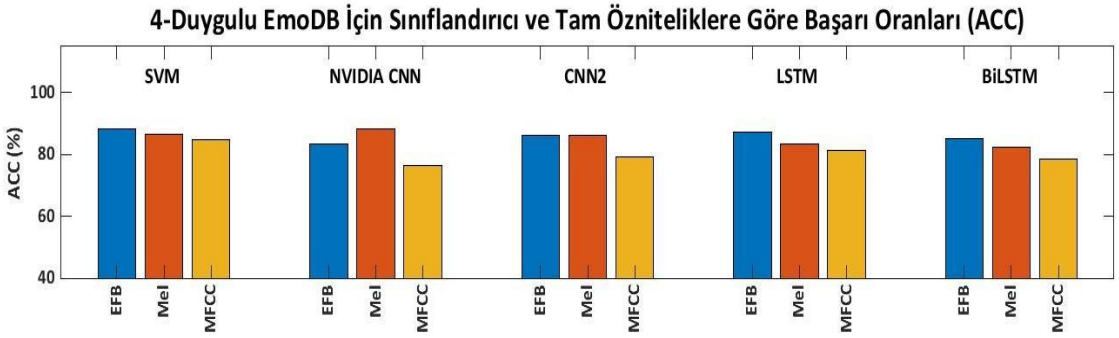


Şekil 5.25. 4-duygulu EmoDB üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.26 ve Şekil 5.26'dan görüleceği gibi 4-duygulu EmoDB verisetinde prozodik öznitelikler eklenerek Mel filtreleri %88,24 ile NVIDIA CNN (CNN1) sınıflandırıcısında EFB filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. EFB filtreleri bütün deneylerde MFCC'den daha başarılı olduğu gibi %88,20 ile Mel filtrelerinin sonucuna oldukça yakındır. Öznitelik sayısında büyük oranda bir düşüş olmasına rağmen başarı oranındaki azalma nispeten azdır. EFB filtreleri CNN2 sınıflandırıcıda Mel filtreleri ile aynı sonuca ulaşabilmiş, LSTM ve BiLSTM sınıflandırıcılarda ise oldukça belirgin bir şekilde daha yüksek sonuçlar üretmiştir. EFB filtrelerinde seçilen öznitelik sayısı (62) Mel filtrelerinde seçilen öznitelik sayısının (113) neredeyse yarısıdır. Bu sonuçlar öznitelik sayısının gereğinden fazla olmasının sınıflandırıcıların başarısını olumsuz etkileyebileceğini göstermektedir. EFB filtreleri bütün sınıflandırıcılarda MFCC özellik setinden daha başarılı olmuştur.

Çizelge 5.26. 4-Duygulu EmoDB Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoDB Tam EFB (62)	**88,20**	83,33	**86,27**	**87,25**	**85,29**
EmoDB Tam Mel (113)	86,43	88,24	**86,27**	83,33	82,35
EmoDB Tam MFCC (51)	84,95	76,47	79,41	81,37	78,43

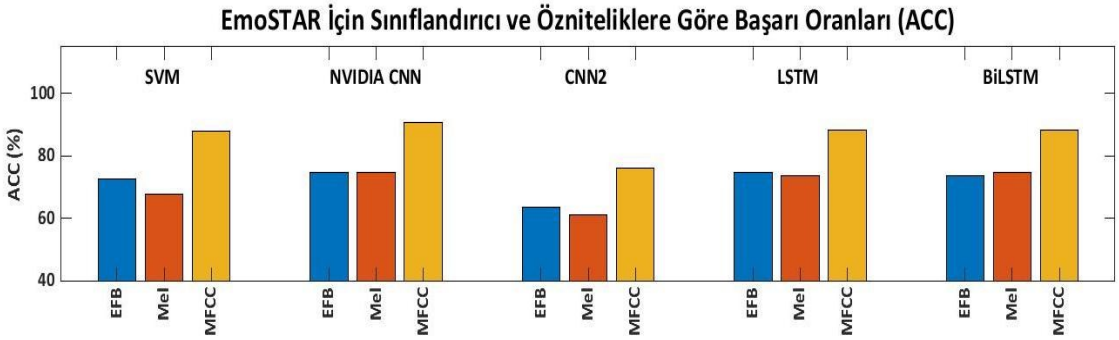


Şekil 5.26 4-duygulu EmoDB üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.27 ve Şekil 5.27'den görüleceği gibi 4-duygulu EmoSTAR verisetinde MFCC %90,68 ile NVIDIA CNN (CNN1) sınıflandırıcısında EFB ve Mel filtrelerinden daha başarılı sonuç elde etmiştir. Ancak EFB filtrelerinde seçilen öznitelik sayısı 23, MFCC özniteliklerinde seçilen öznitelik sayısı 108'e göre oldukça azdır. MFCC öznitelik seti EFB ve Mel filtrelerine göre oldukça yüksek başarı oranlarına ulaşabilmiştir. CNN2 sınıflandırıcısında bu fark %15,25'e çıkmıştır.

Çizelge 5.27. 4-Duygulu EmoSTAR Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	**SVM**	**CNN1**	**CNN2**	**LSTM**	**BiLSTM**
EmoSTAR EFB (23)	72,77	74,58	63,56	74,58	73,73
EmoSTAR Mel (29)	67,68	74,58	61,02	73,73	74,58
EmoSTAR MFCC (108)	**87,78**	90,68	**76,27**	**88,14**	**88,14**

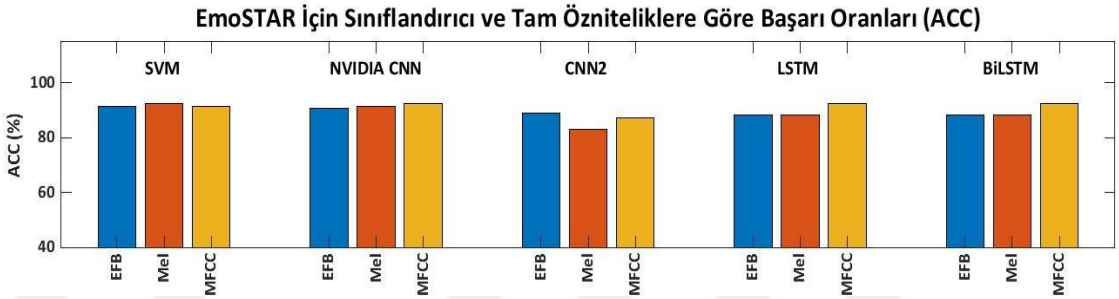


Şekil 5.27. 4-duygulu EmoSTAR üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.28 ve Şekil 5.28'den görüleceği gibi 4-duygulu EmoSTAR verisetinde prozodik öznitelikler eklenerek MFCC %92,37 ile NVIDIA CNN (CNN1), LSTM ve BiLSTM sınıflandırıcılarında EFB ve Mel filtrelerinden daha başarılı sonuç elde etmiştir.

Çizelge 5.28. 4-Duygulu EmoSTAR Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
EmoSTAR Tam EFB (78)	91,34	90,68	**88,98**	88,14	88,14
EmoSTAR Tam Mel (99)	**92,36**	91,53	83,05	88,14	88,14
EmoSTAR Tam MFCC (54)	91,60	92,37	87,29	92,37	92,37

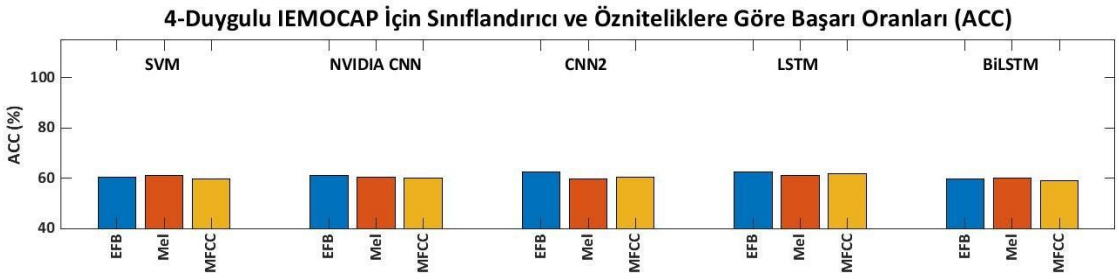


Şekil 5.28. 4-duygulu EmoSTAR üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.29 ve Şekil 5.29'dan görüleceği gibi 4-duygulu IEMOCAP verisetinde EFB filtreleri %62,51 ile LSTM sınıflandırıcısında Mel filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. EFB filtreleri bütün deneylerde MFCC'den daha başarılıdır.

Çizelge 5.29. 4-Duygulu IEMOCAP Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP EFB (68)	60,49	**61,17**	**62,50**	62,51	59,69
IEMOCAP Mel (118)	**61,06**	60,43	59,76	61,10	**60,13**
IEMOCAP MFCC (59)	59,71	59,99	60,28	61,69	59,02



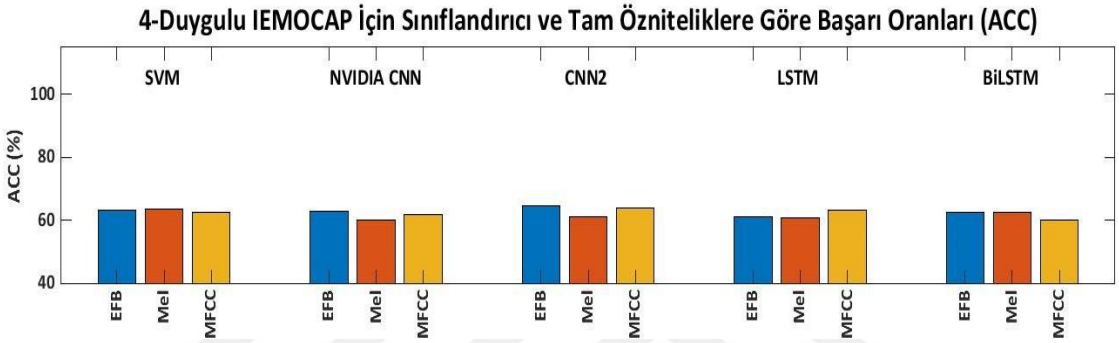
Şekil 5.29. 4-duygulu IEMOCAP üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.30 ve Şekil 5.30'dan görüleceği gibi 4-duygulu IEMOCAP verisetinde prozodik öznitelikler eklenerek EFB filtreleri %64,66 ile CNN2 sınıflandırıcısında Mel

filtreleri ve MFCC'den daha başarılı sonuç elde etmiştir. EFB filtreleri LSTM dışındaki bütün deneylerde MFCC'den daha başarılıdır. Öznitelik sayılarında dikkat çekici bir azalma olmasına rağmen başarı oranları nispeten korunabilmiştir.

Çizelge 5.30. 4-Duygulu IEMOCAP Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
IEMOCAP Tam EFB (89)	63,09	**62,88**	64,66	61,25	62,36
IEMOCAP Tam Mel (121)	**63,63**	59,99	61,24	60,95	**62,66**
IEMOCAP Tam MFCC (85)	62,44	61,69	64,06	**63,18**	60,13

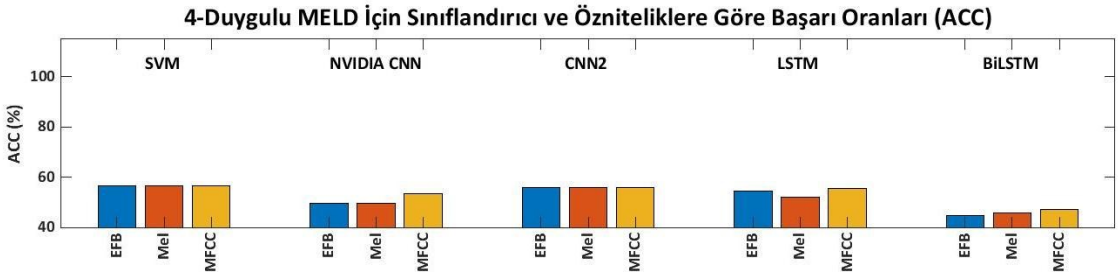


Şekil 5.30. 4-duygulu IEMOCAP üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.31 ve Şekil 5.31'den görüleceği gibi 4-duygulu MELD verisetinde bütün öznitelik setleri SVM sınıflandırıcıda %56,68 ile aynı en yüksek sonucu elde etmiştir. MFCC diğer bütün deneylerde Mel filtreleri ve EFB filtrelerinden daha başarılıdır. Öznitelik sayılarında önemli oranda düşüş elde edilmiş ancak başarı oranları neredeyse sabit kalmış hatta bazı deneylerde daha da artmıştır. SVM sınıflandırıcı derin öğrenme modellerine açık bir üstünlük sağlamıştır. Bu durumun oluşmasında MELD verisetindeki dengesiz veri dağılımının etkili olduğu söylenebilir. MELD verisetinde nötr örneklerin sayısı diğer örneklere göre çok fazladır. Bu açıdan MELD verisetindeki başarı oranları dengesiz bir niteliğe sahiptir.

Çizelge 5.31. 4-Duygulu MELD Üzerindeki CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	SVM	CNN1	CNN2	LSTM	BiLSTM
MELD EFB (69)	56,68	49,77	55,90	54,43	44,80
MELD Mel (100)	56,68	49,59	55,87	52,06	45,95
MELD MFCC (51)	56,68	**53,41**	**55,93**	**55,61**	**47,36**

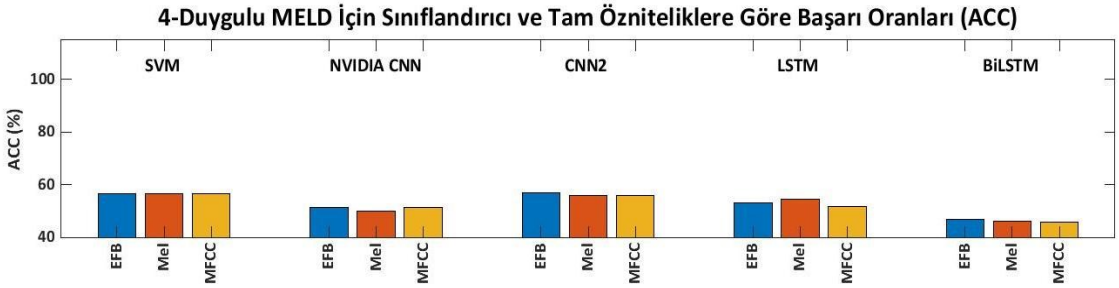


Şekil 5.31. 4-duygulu MELD üzerindeki CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

Çizelge 5.32 ve Şekil 5.32'den görüleceği gibi 4-duygulu MELD verisetinde prozodik öznitelikler eklenerek EFB öznitelik seti CNN2 sınıflandırıcıda %56,87 ile en yüksek sonucu elde etmiştir.

Çizelge 5.32. 4-Duygulu MELD Üzerindeki Prozodik Öznitelikler Eklenerek Yapılan CFS Subset Evaluator Öznitelik Seçici Deney Sonuçları (%ACC).

4-duygu (K,M,N,Ü)	**SVM**	**CNN1**	**CNN2**	**LSTM**	**BiLSTM**
Meld Tam EFB (63)	**56,68**	51,35	56,87	53,26	**46,71**
Meld Tam Mel (72)	**56,68**	50,03	55,75	**54,43**	46,24
Meld Tam MFCC (66)	**56,68**	**51,41**	56,07	51,67	45,86



Şekil 5.32. 4-duygulu MELD üzerindeki prozodik öznitelikler eklenerek yapılan CFS Subset Evaluator öznitelik seçici deney sonuçlarının çubuk grafikleri.

5.4. VERİ TÜRETME DENEYLERİ

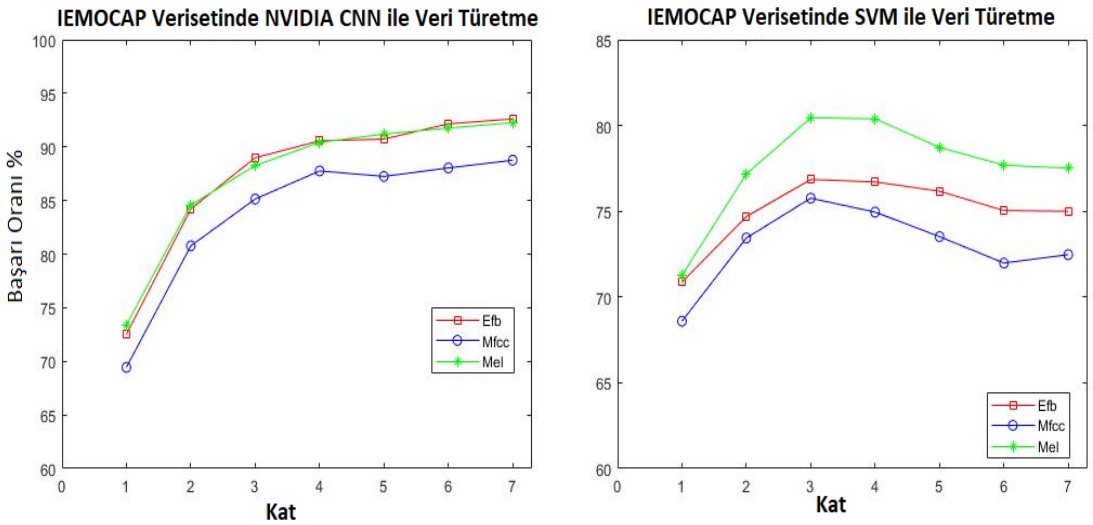
Çalışmanın bu bölümünde, 4-duygulu doğaçlama IEMOCAP verisetinde MATLAB audioDataAugmenter sınıfı ile veri türetmenin sonuçlarını sunuyoruz. Sonuçlar Çizelge 5.33 ve Şekil 5.33'te verilmiştir. En iyi sonuçlar NVIDIA CNN (CNN1) ile EFB ve Mel filtre bankası özellikleri kullanılarak elde edilmiştir. Örnek sayısını 7 kata kadar çıkardık. 4-duygu sınıflı (Kızgın, Mutlu, Nötr, Üzgün) doğaçlama IEMOCAP (2280 örnek) veri kümesinde MATLAB 2021 audioDataAugmenter sınıfı AugmentationMode=sequential,

NumAugmentations=9, TimeStretchProbability=0,8; SpeedupFactorRange=[1,3;1,4], VolumeControlProbability=0,8; VolumeGainRange = [-5,5], PitchShiftProbability=0,8; TimeShiftProbability=1, TimeShiftRange=[-0,3;0,3], AddNoiseProbability=0, SNRRange=[-20,40] parametreleri ile kullanıldı. IEMOCAP verisetinde en son örnek sayısı 15960'tır.

Çizelge 5.33. 4-Duygulu Doğaçlama IEMOCAP Veri Kümesinde 7 Kata Kadar Veri Türetme Başarı Oranı (% ACC) Sonuçları.

		1	**2**	**3**	**4**	**5**	**6**	**7**
EFB	**SVM**	69,88	75,80	76,16	76,42	75,17	74,87	74,16
	NVIDIA CNN	72,51	84,21	**88,99**	**90,57**	90,73	**92,15**	**92,61**
	LSTM	67,25	79,97	86,11	86,11	88,04	86,74	89,05
MFCC	**SVM**	68,12	73,97	74,75	75,00	75,17	72,24	71,72
	NVIDIA CNN	69,44	80,77	85,14	87,76	87,25	88,04	88,76
	LSTM	69,73	81,57	85,18	87,31	87,66	88,91	89,09
Mel	**SVM**	68,71	76,16	79,48	79,13	78,65	77,53	76,60
	NVIDIA CNN	**73,39**	**84,58**	88,26	90,42	**91,20**	91,76	92,27
	LSTM	69,15	79,53	82,06	84,61	85,05	85,28	84,92

SVM ve NVIDIA CNN (CNN1) sınıflandırıcıları ile 4-duygulu IEMOCAP veri kümesindeki veri türetme Şekil 5.33'te tasvir edilmektedir. Veri türetme belli bir aşamaya kadar sonuçları arttırıcı yönde etki etmekte daha sonra ise başarı oranı sabit kalmakta veya bir miktar azalma eğilimi göstermektedir.



Şekil 5.33. 4-duygulu IEMOCAP verisetinde NVIDIA CNN ve SVM ile veri türetme sonuçları.

Çizelge 5.34 ve Çizelge 5.35'te, belirtilen 4490 örnekli 4-duygulu IEMOCAP veri kümesinde kullanılan yöntemlerin CPU ve GPU sınıflandırma hızı karşılaştırmasını saniye olarak sunuyoruz. Kullanılan sistem Intel i7 (6700 HQ, 8 çekirdek, 2,6 GHz), 16 GB RAM (2133 MHz), NVIDIA GeForce GTX 950M (4 GB DDR3 RAM), 256 GB SSD, 1 TB HDD özelliklerine sahiptir.

Çizelge 5.34. 4-Duygulu 4490 Örnekli IEMOCAP Veri Kümesinde Modellerin Saniye Olarak CPU Sınıflandırma Hızları (Saniye).

	Özellik sayısı	**SVM**	**NVIDIA CNN**	**CNN2**	**LSTM**	**BiLSTM**
EFB Filters	573	5,47	3,00	75,02	1,00	1,00
Mel Filters	1761	33,94	10,00	227,72	2,00	3,00
MFCC	573	5,28	3,00	75,02	1,00	1,00

Çizelge 5.35. 4-Duygulu 4490 Örnekli IEMOCAP Veri Kümesinde Modellerin Saniye Olarak GPU Sınıflandırma Hızları (Saniye).

	Özellik sayısı	**SVM**	**NVIDIA CNN**	**CNN2**	**LSTM**	**BiLSTM**
EFB Filters	573	-	2,00	20,00	1,00	2,00
Mel Filters	1761	-	4,00	53,01	2,00	3,00
MFCC	573	-	2,00	20,00	1,00	2,00

Çizelge 5.34'te geçen zaman değerleri, SVM için Weka'daki model oluşturma (build) süresini gösterir. NVIDIA CNN (CNN1), CNN2, LSTM ve BiLSTM zamanları 1 çevrim için geçen süredir. NVIDIA CNN (CNN1), LSTM ve BiLSTM için 1000 çevrim, CNN2 için 500 çevrim kullanılmıştır. CNN2 açık arayla en çok zaman alan modeldir. EFB ve MFCC zamanlaması, eşit sayıda özellik nedeniyle hemen hemen aynıdır. GPU genellikle BiLSTM uygulamaları dışında CPU'dan daha hızlıdır. LSTM ağları hem CPU hem de GPU'yu çok yoğun kullanır. BiLSTM ile yaptığımız deneylerde CPU kullanımı %100'e çok yakın ve GPU kullanımı %30'un üzerindedir; ancak CNN deneylerinde GPU kullanımı (%60) CPU kullanımından (%30) daha yüksektir. CNN modelleri paralel GPU mimarisi için daha uygun gözükmektedir. CPU uygulamaları için Tensorflow 1.7.1 ve Keras 2.2.4, GPU için, Keras 2.2.4 ile uyumluluk sorunları nedeniyle Tensorflow 1.9.0 kullanıldı. Weka, SVM deneyleri için GPU kullanım imkânı sunmamaktadır.

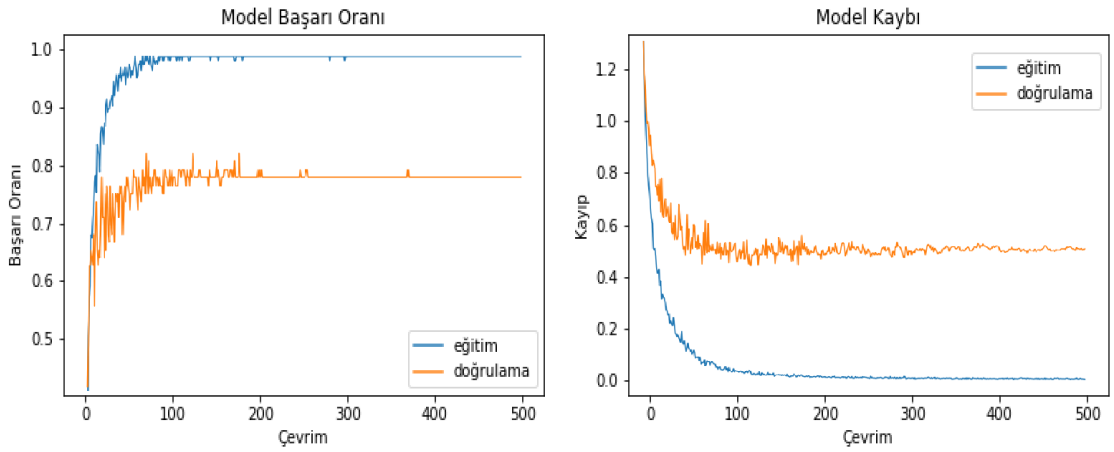
Çizelge 5.36'da 4-duygulu doğaçlama EmoDB veri kümesinde 5 ardışık çalıştırmanın ortalama öznitelik çıkarma hızı karşılaştırmasını milisaniye olarak sunuyoruz. EFB öznitelik seti yaklaşık %40 oran ile en hızlı yöntemdir. Bu oran MFCC ve Mel filtre

öznitelik setlerine göre çok önemli bir hız iyileştirmesidir. MFCC ve Mel filtreleri arasındaki zaman farkı Mel filtrelerine uygulanan log ve DCT işlemlerinden kaynaklanmaktadır.

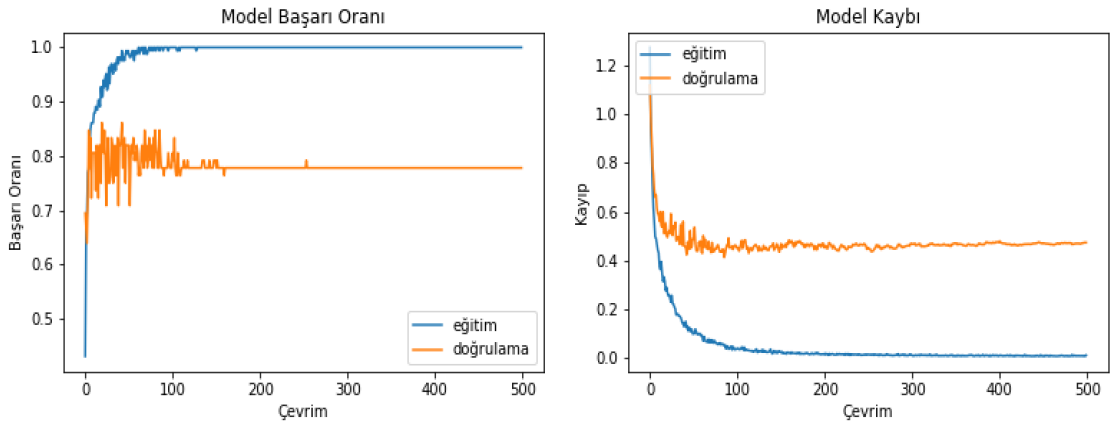
Çizelge 5.36. 4-Duygulu EmoDB Veri Kümesinde Milisaniye Olarak Ortalama Öznitelik Çıkarma Hızları.

	EFB	**MFCC**	**Mel**
Hız (ms)	**1042**	1457	1431

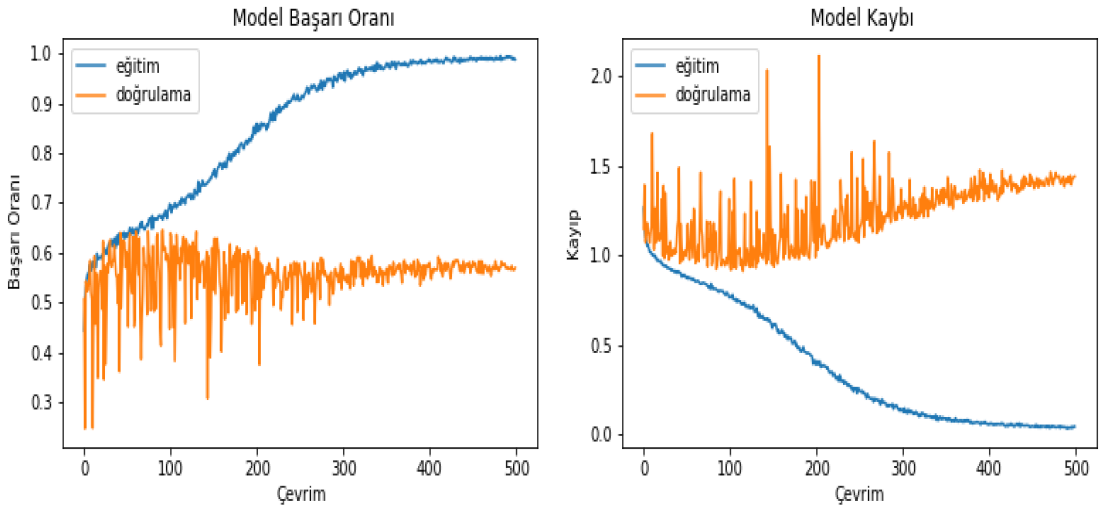
EmoSTAR, EmoDB, IEMOCAP veri kümeleri için CNN2 modeli ile elde edilen başarı oranı ve kayıp grafikleri sırasıyla Şekil 5.34, 5.35, 5.36'da gösterilmiştir.



Şekil 5.34. CNN2 modeli ve EFB filtreleri ile EmoSTAR veri kümesi için başarı oranı ve kayıp grafikleri.



Şekil 5.35. CNN2 modeli ve EFB filtreleri ile EmoDB veri kümesi için başarı oranı ve kayıp grafikleri.



Şekil 5.36. CNN2 modeli ve EFB filtreleri ile IEMOCAP veri kümesi için başarı oranı ve kayıp grafikleri.

Çizelge 5.37'de EFB filtreleri öznitelik kümesi ve CNN2 sınıflandırıcı kullanarak EmoSTAR, EmoDB, IEMOCAP ve MELD verisetlerinde elde edilen karmaşıklık matrisleri verilmektedir. Özellikle MELD verisetinde diğer ses örneklerine göre çok daha fazla sayıda bulunan nötr örneklerdeki aşırı öğrenme oldukça belirgindir. Çizelge 5.37'de K, Kızgın; M, Mutlu; N, Nötr; Ü, Üzgün sınıfı temsil etmektedir. MELD verisetindeki başarı oranlarının %50'nin üstünde olduğunu belirtmekte de fayda vardır. Başarı oranının yüksekliği tek başına modellerinin başarısını göstermek açısından yeterli olmamaktadır. EmoSTAR verisetinde en başarılı duygu 23 Gerçek Pozitif (True Positive) ve sadece 1 Hatalı Negatif (False Negative) ile Üzgün sınıfıdır. Kızgın sınıfında da Gerçek Pozitif (23 örnek) oldukça yüksek ve Hatalı Negatif (1 Mutlu, 0 Nötr, 3 Üzgün olmak üzere 4 örnek) oldukça düşüktür. Mutlu sınıfı kızgın ve nötr sınıflarla daha çok karışmaktadır. Benzer durum EmoDB içinde geçerlidir. IEMOCAP verisetinde de Üzgün sınıfı en kolay tanınan sınıftır ve 87 Gerçek Pozitife karşı sadece 16 Hatalı Negatif örnek bulunmaktadır. Kızgın sınıfı Üzgün sınıfından sonra en iyi tanınan sınıftır. Kızgın ve Üzgün sınıflarının her ikisi de Mutlu ve Nötr sınıflarla karışmaktadır. Nötr sınıfın Üzgün sınıf olarak tanınma oranı da oldukça yüksektir. MELD verisetindeki karmaşıklık matrisinin değerlendirilmesi oldukça tutarsız sonuçlar üretmektedir. Görüldüğü gibi sınıflar arasındaki dengesizlik nötr sınıfındaki aşırı örnek sayısı nedeniyle oldukça belirgindir. Diğer 3 verisetinde en başarılı sınıf olan Üzgün sınıfı burada oldukça başarısız olmuştur. Nötr sınıfındaki Gerçek Pozitif oranı diğer sınıflara göre oldukça yüksektir. Bu konu veri türetme ve dengesiz dağılımlı verisetleri çalışmalarında ayrıca incelenmektedir.

Çizelge 5.37. CNN2 Sınıflandırıcısı Ve EFB Filtreleri İle EmoSTAR, EmoDB, IEMOCAP Ve MELD Verisetlerindeki Karmaşıklık Matrisleri.

	EmoSTAR				**IEMOCAP**			
	K	**M**	**N**	**Ü**	**K**	**M**	**N**	**Ü**
K	**27**	1	0	3	**263**	27	49	3
M	3	**18**	3	1	15	**42**	44	37
N	0	0	**3**	3	61	95	**399**	209
Ü	0	1	0	**23**	4	6	6	**87**
	EmoDB				**MELD**			
K	**37**	7	0	0	**128**	93	150	38
M	3	**7**	0	0	8	**23**	28	3
N	0	2	**23**	3	356	589	**1719**	247
Ü	0	0	2	**18**	3	5	8	**8**

5.5. DENGESİZ DAĞILIMLI VERİSETİ DENEYLERİ

Tezimizin bu bölümünde, 4-duygulu (Kızgın, Mutlu, Nötr ve Üzgün) MELD veri kümesindeki dengesiz veri sorununa çözüm için Örnek Azaltma, SMOTE, Veri Türetme, Bagging ve AdaboostM1 metotlarını uyguladık. Ayrıca Hassasiyet, F-skor ve Kappa metriklerini de değerlendirmeye aldık.

5.5.1. SMOTE, Veri Türetme ve Veri Azaltma Deneyleri

Çizelge 5.38'de 4-duygulu MELD verisetinin, SMOTE, Veri Türetme, Veri Azaltma ve orijinal örnek sayısı sunulmaktadır. SMOTE ve veri türetme uygulanmış MELD veri kümesinde, veri kümesini dengeli hale getirmek için sınıf etiketleri yaklaşık eşit dağıtılmıştır. Veri azaltma işleminde, çoğunluk sınıfı örneklerinden bazılarını kaldırdık ve her sınıftan yalnızca 1000 örnek kullandık.

Çizelge 5.38. SMOTE, Veri Türetme, Veri Azaltma Uygulanmış Ve Orijinal 4-Duygulu MELD Veri Kümesindeki Örnek Sayıları.

	Kızgın	**Mutlu**	**Nötr**	**Üzgün**	**Toplam**
MELD Orijinal	1607	2308	6434	1002	11351
MELD SMOTE	6428	6416	6434	6432	25710
MELD Türetilmiş	6428	6416	6434	6432	25710
MELD Azaltılmış	1000	1000	1000	1000	4000

SMOTE'un öznitelik kümesine uygulandığı bilinmelidir. Öte yandan veri türetme zamansal ses dalgasına uygulanmaktadır. SMOTE için Weka aracını, veri türetme için ise MATLAB 2021 audioDataAugmenter sınıfını kullandık.

Çizelge 5.39'da NVIDIA CNN (CNN1) sınıflandırıcı ve EFB öznitelik setini kullanarak SMOTE, Veri Türetme, Veri Azaltma ve orijinal veri kümelerinin MELD verisetindeki karmaşıklık matrislerini sunuyoruz. SMOTE ve Veri Türetme uygulanmış verisetleri daha tutarlı karmaşıklık matrisleri üretmektedir. Çizelgede 5.39'da K, Kızgın; M, Mutlu; N, Nötr; Ü, Üzgün sınıfı temsil etmektedir. SMOTE ve Veri türetme sonucunda elde edilen yeni öznitelik verisetleri aynı örnek sayısına sahip olduğundan SMOTE ve Veri Türetme sonuçları doğrudan kıyaslanabilirler. Orijinal karmaşıklık matrisinde Nötr-Nötr eşleşmesindeki aşırı öğrenme dikkat çekicidir. SMOTE ve Veri türetme bu tür aşırı öğrenme sorunlarının üstesinden gelmek için sıklıkla kullanılan yöntemlerdir ve bir sonraki deneydeki sonuçlar bu başarıyı göstermektedir. SMOTE ve Veri türetme uygulanmış verisetlerinin karmaşıklık matrislerindeki iyileşmeler oldukça belirgindir ve Üzgün sınıfı Gerçek Pozitiflerin Hatalı Negatiflere oranına göre en iyi belirlenen sınıf olmuş ve Kızgın sınıfı ikinci sırada yer almaktadır. SMOTE uygulanmış verisetinde 1523 Gerçek pozitif örneğe karşı 555 Hatalı Negatif örnek bulunmaktadır. Üzgün sınıfı en çok Nötr sınıf ile karışmaktadır. Kızgın sınıf ise en fazla Mutlu ile karışmaktadır. Veri türetme uygulanmış verisetinde ise 1348 Gerçek Pozitif örneğe karşı 362 Hatalı Negatif örnek bulunmaktadır. Üzgün sınıfı en çok Kızgın sınıf ile karışmıştır. Kızgın sınıf en çok Üzgün sınıf ile karışmaktadır.

Çizelge 5.39. SMOTE, Veri Türetme, Veri Azaltma Ve Orijinal MELD Verisetlerinin NVIDIA CNN Ve EFB Öznitelik Kümesi İle Elde Edilen Karmaşıklık Matrisleri.

	SMOTE				**Veri Türetme**			
	K	**M**	**N**	**Ü**	**K**	**M**	**N**	**Ü**
K	**1313**	295	169	141	**1270**	184	138	192
M	182	**1016**	252	118	210	**1018**	213	164
N	216	378	**1385**	170	300	525	**1542**	247
Ü	149	269	137	**1523**	164	148	50	**1348**
	Veri Azaltma				**Orijinal**			
K	**126**	84	51	58	**105**	77	121	23
M	76	**87**	79	58	62	**113**	199	35
N	46	68	**79**	56	314	503	**1534**	214
Ü	59	70	94	**109**	14	17	51	**24**

Çizelge 5.40'ta, NVIDIA CNN (CNN1) sınıflandırıcı EFB, MFCC ve Mel filtre öznitelik kümeleri ile kullanılarak MELD-SMOTE (SMOTE), MELD Veri Türetilmiş (Tür), MELD Veri Azaltılmış (Az) ve Orijinal MELD (Orj) veri kümelerinde elde edilen sonuçlar gösterilmektedir. SMOTE ve veri türetme tekniğinin dengesiz veri kümesi sorununun üstesinden gelmek için çok başarılı olduğu söylenebilir. SMOTE ve veri türetme sadece karmaşıklık matrislerini dengeli hale getirmekle kalmamış başarı oranlarında da %17'ye varan oranlarda iyileştirme sağlamıştır. Ayrıca hassasiyet, F-skor ve kappa değerlerinde de ciddi ilerlemeler elde edilmektedir. Mel filtrelerinde SMOTE ile Kappa değeri %64,19'a, F-skor değeri ise %73,09'a, hassasiyet ise %73,77'ye kadar çıkmıştır.

Çizelge 5.40. EFB, MFCC Ve Mel Filtre Öznitelik Setlerinde 4-Duygulu MELD Veri Kümesinde NVIDIA CNN (CNN1) İle SMOTE, Veri Türetme (Tür), Veri Azaltma (Az) Ve Orijinal (Orj) Veri Sonuçları.

		Başarı(%)	**Hassasiyet(%)**	**F-skor(%)**	**Kappa(%)**
EFB	**SMOTE**	**67,89**	67,74	**67,57**	**57,20**
	Tür	67,13	**68,16**	67,07	56,15
	Az	33,41	33,27	33,21	11,27
	Orj	52,14	45,87	47,35	10,90
MFCC	**SMOTE**	65,68	66,07	65,04	54,24
	Tür	**66,05**	**66,84**	**65,48**	**54,70**
	Az	33,08	33,05	33,06	10,73
	Orj	52,37	46,18	47,88	11,97
Mel	**SMOTE**	**73,16**	**73,77**	**73,09**	**64,19**
	Tür	66,79	67,27	66,57	55,70
	Az	33,00	33,33	33,00	10,78
	Orj	53,17	45,29	46,83	10,05

EFB özellik kümesi SMOTE ile MFCC ise Veri Türetme ile daha iyi sonuçlar elde etmiştir. En iyi sonuçlar Mel filtrelerinde SMOTE kullanılarak elde edilmiştir. Veri türetmede ise EFB filtrelerinin başarısı Mel filtrelerinden daha yüksektir. Örnek azaltılmış veri kümeleriyle karmaşıklık matrislerinde bir miktar gelişme sağlanmış gözükse de orijinal veri kümelerine kıyasla başarı oranları oldukça düşüktür. Örnek azaltma sadece Kappa puanını biraz artırmıştır. SMOTE ve veri türetme Başarı Oranı, Hassasiyet, F-skor ve Kappa değerlerine iyileştirici yönde etki etmektedir ve orijinal verisetine göre oldukça yüksek artışlar sağlayabilmektedirler.

5.5.2. AdaboostM1, Bagging ve SVM Karşılaştırması

Çizelge 5.41'den görülebileceği gibi, AdaboostM1 ve Bagging sınıflandırıcıları sınıflandırma metriklerinde 4-duygulu (Kızgın, Nötr, Mutlu, Üzgün) MELD verisetinde dengesiz dağılım problemine herhangi bir iyileştirme sağlamamaktadır. Özellikle bütün sınıflandırıcılardaki Kappa değerlerinin düşüklüğü oldukça dikkat çekicidir. AdaboostM1 ve Bagging sınıflandırıcıları Weka aracı ile %70 eğitim, %30 test seti olacak şekilde uygulanmıştır.

Çizelge 5.41. EFB, MFCC Ve Mel Filtre Öznitelik Kümelerini Kullanan 4-Duygulu MELD Veri Kümesinde Adaboostm1, Bagging Ve SVM Sınıflandırıcılarının Sonuçları.

		EFB	**MFCC**	**Mel**
Başarı (%)	**Adaboost M1**	57,23	57,23	57,23
	Bagging	55,24	56,62	56,32
	SVM	57,18	**57,59**	57,50
Hassasiyet (%)	**Adaboost M1**	32,80	32,80	32,80
	Bagging	45,20	45,50	47,20
	SVM	36,20	**59,80**	42,90
F_skor (%)	**Adaboost M1**	41,70	41,70	41,70
	Bagging	45,70	**46,20**	46,00
	SVM	41,70	42,90	44,10
Kappa (%)	**Adaboost**	0	0	0
	Bagging	5,51	**7,03**	6,18
	SVM	-0,02	1,93	4,08

5.6. TEMEL FREKANS DENEYLERİ

Bu bölümde tezimizin temel frekans hesaplama ile ilgili deneylerini sunuyoruz. Temel frekans algılama algoritmalarını değerlendirmek için birçok hata ölçütü kullanılmaktadır. GPE (Gross Pitch Error) en çok kullanılanlar arasındadır. Bu çalışmada ek olarak iki farklı hata ölçüsü sunuyoruz. Bunlardan ilki, gerçek temel frekans değerinden %10'dan daha fazla sapmaya sahip örnek sayısını gösteren e_{10} hata ölçütüdür. Temel frekans

tahmininin bir diğer yararlı uygulaması cinsiyet tespitidir. Cinsiyet tespit hatası temel frekans hesaplama algoritmalarının değerlendirilmesi ve karşılaştırılması için çok yararlı olabilir. GPE ve e_{10} hataları sırasıyla Denklem (5.1) ve Denklem (5.2)'de tanımlanmaktadır.

$$GPE = \frac{1}{K}\sum_{j=1}^{K}|\hat{f}_j - f_j| \tag{5.1}$$

$$e_{10} = \#\left\{|\hat{f}_j - f_j| > \frac{f_j}{10}, j | j = 1..K\right\} \tag{5.2}$$

Hillenbrand ve Texas Sesli veri kümelerinde gerçek temel frekans değerleri hesaplanmış olarak verilir ve bize tahminlerimizle sağlam bir karşılaştırma yapma imkânı sağlar. Erkek sesleri kesinlikle çocuk ve kadınlara kıyasla daha düşük f0 değerlerine sahiptir. Ancak sadece f0 değerlerini kullanarak erkek çocuk, kız çocuk ve kadın seslerini ayırmak imkansızdır. Çizelge 5.42'den görüldüğü gibi çocuk ve kadın sesleri temel frekans yönünden oldukça benzerdir. Kadın, erkek çocuk ve kız çocuk ayırımı için başka verilerin ele alınması gereklidir. Hillenbrand veri kümesinde erkek çocuklar için ortalama f0 236 Hz, kız çocuklar için 238 Hz, yetişkin erkekler için 131 Hz ve kadınlar için de 220 Hz'tir. Teksas Sesli veri kümesinde çocuk, erkek ve kadın sınıfları vardır. Çocuk, erkek ve kadın örnekleri için maksimum f0 sırasıyla 392, 202 ve 271 Hz'tir. Çocuk, erkek ve kadın örnekleri için minimum f0 sırasıyla 105 (k7bree05.wav adlı örnek, bu bir çocuk için beklenmedik bir değer olduğu için elle kontrol ettik), 82 ve 141 Hz'tir. TIMIT veri kümesinde, yalnızca erkek ve kadın sınıfları vardır ve ortalama değerler Çizelge 5.42'de gösterildiği gibi diğer iki veri kümesiyle uyum içindedir. TIMIT veri kümesinde, temel frekans gerçek değerleri hazır olarak verilmemektedir. Bu çalışmada kullanılan yöntemlerin bulduğu tahmini temel frekans değerlerinin ortalaması TIMIT tarafından sağlanan cinsiyet bilgisi ile beraber kullanılarak gerçek değer olarak kullanılmaktadır.

Çizelge 5.42 Hillenbrand, Teksas Ve TIMIT Veri Kümeleri İçin Cinsiyete Göre Ortalama F0 Değerleri.

	Erkek	**Kadın**	**Erkek Çocuk**	**Kız Çocuk**
TIMIT	119	207	-	
Texas	110	217	245	
Hillenbrand	131	220	236	238

Çizelge 5.43'ten görülebileceği gibi HDM, en hızlı metottur ve kepstrum tarafından takip edilmektedir. Zamanlamalar, Hillenbrand veri kümesinde ardışık 10 çalıştırmanın ortalama değeridir. Otokorelasyon ve kepstrum için Naotoshi SEO'nun [124] uygulamalarını kullandık, ancak daha iyi sonuçlar elde etmek için maksimum temel frekans değerini 500 Hz olarak değiştirdik. YIN, YAAPT, FCN ve özellikle CREPE hız açısından HDM, kepstrum ve otokorelasyondan çok uzak kalmaktadırlar. HDM, Hamming penceresi kullanmadan dahi neredeyse aynı sonuçları üretir, ancak diğer metotlar Hamming penceresi olmadığı durumda daha kötü sonuçlara sahiptir, bilhassa kepstrum, Hamming penceresi uygulanmaması durumunda TIMIT veri kümesindeki hata marjını iki katına çıkarmaktadır.

Çizelge 5.43. Hillenbrand Verisetinde Uygulanan Temel Frekans Tespit Yöntemlerinin Saniye Olarak Ortalama Hızları.

	AC	**YIN**	**KEPS**	**YAAPT**	**CREPE**	**FCN**	**HDM**
Hız	0,32	24,26	0,22	13,89	440,29	137,65	**0,18**

5.6.1. Genişbant ve Darbant Temel Frekans Deneyleri

Çizelge 5.44'ten görüldüğü gibi, genişbant Hillenbrand veri kümesinde, önerilen HDM en küçük GPE'ye ve AC en küçük e_{10} hatasına sahiptir. Otokorelasyon çok eski bir teknik olmasına rağmen, bu veri kümesinde oldukça başarılıdır. Hillenbrand verisetinde kepstrum e_{10} hatasında ikinci, HDM ise üçüncü sıradadır. Teksas veri kümesinde, FCN en küçük GPE hatasına ve AC en küçük e_{10} hatasına sahiptir. TIMIT veri kümesinde FCN, GPE'de %5,82 ile en iyi yöntemdir ve %6,39 ile HDM tarafından çok yakından takip edilmektedir. Kepstrum %11,27 ile GPE hatasında üçüncü sıradadır ve CREPE'ten daha iyi bir performans gösterebilmiştir. Öte yandan TIMIT verisetinde önerilen HDM e_{10} hatasında %4,68 ile en iyi yöntemdir ve %5,29 ile FCN tarafından takip edilmektedir. CREPE %6,84 ile üçüncü sıradadır. YIN çok fazla hatalı sonuç ürettiğinden daha iyi sonuçlar bulabilmek için birçok farklı parametre denedik. AC ve kepstrum uygulamalarında 500 Hz üst sınır frekansı uyguladığımızı, aksi takdirde bu yöntemlerin daha kötü sonuçlar ürettiğini vurgulamamız gerekiyor. Diğer yöntemlerde üst sınır uygulanmamıştır. Konvolusyonal Sinir Ağı yöntemleri genişbant uygulamalarında

oldukça başarılıdır, ancak sonraki bölümde göreceğimiz gibi, darbant telefon konuşma verilerinde neredeyse kördürler.

Çizelge 5.44. Genişbant Hillenbrand, Teksas Ve TIMIT Verisetlerindeki Sonuçlar.

	GPE			**e_{10}**		
	TIMIT	**Texas**	**Hillenbrand**	**TIMIT**	**Texas**	**Hillenbrand**
FCN	**5,82**	**8,17**	9,10	5,29	7,42	8,75
YIN	43,88	16,67	16,71	12,58	8,90	8,63
KEPS	11,27	14,53	7,86	7,45	11,38	6,00
YAAPT	16,00	12,87	17,18	15,13	14,76	13,55
CREPE	12,33	8,45	8,97	6,84	8,12	9,53
AC	22,74	10,36	8,17	8,79	**7,09**	**4,68**
HDM	6,39	9,46	**7,65**	**4,68**	11,41	6,41

Deneylerimizi telefon konuşmalarına doğru genişleterek Çizelge 5.45'te sunuyoruz. Bu amaçla, 400 Hz'in altındaki ve 3400 Hz'in üzerindeki frekansları tamamen kaldırmak için veri kümelerimize iki kez bant geçiren filtre uygulayacağız. Bazı telefon konuşmalarında, bu bant genişliği 300 Hz ile 4000 Hz arasında uygulanabilir. Veri kümelerimizde en yüksek f0 değeri 392 Hz olduğu için eşik değer olarak 400 Hz'i seçerek veri kümelerindeki bütün örneklerden temel frekansı kaldırıyoruz. Bu algoritmanın amaçlarından biri de budur. Darbant Hillenbrand veri kümesinde kepstrum, Çizelge 5.45'te gösterildiği gibi her iki hata türünde en başarılı algoritmadır. HDM GPE'de ikinci, AC ise e_{10} hata ölçüsünde ikinci sırada yer alıyor. CREPE ve FCN darbant konuşmada oldukça yüksek hata oranları sergilemektedir.

Çizelge 5.45. Darbant Hillenbrand, Teksas Ve TIMIT Veri Kümelerinde Deneysel Sonuçlar.

	GPE			**e_{10}**		
	TIMIT	**Texas**	**Hillenbrand**	**TIMIT**	**Texas**	**Hillenbrand**
FCN	194,79	191,81	185,53	67,28	71,73	70,08
YIN	160,97	130,29	72,64	40,83	34,85	25,12
KEPS	75,17	**37,89**	**11,23**	25,36	**17,50**	**7,37**
YAAPT	27,10	45,83	51,93	28,14	42,18	50,18
CREPE	162,44	134,87	111,52	50,66	46,11	42,51
AC	69,47	55,71	37,33	23,30	22,24	15,53
HDM	**15,26**	39,42	34,72	**17,88**	29,06	19,30

Darbant Teksas veri kümesinde kepstrum tüm hata türlerinde en başarılı algoritmadır. HDM, GPE'de ikinci, AC ise e_{10} hata türünde ikinci sıradadır. Darbant TIMIT veri kümesinde, yeni HDM algoritmamız GPE ve e_{10} hata ölçülerindeki diğer tüm yöntemlerden daha üstündür. YAAPT, GPE'de ikinci, AC ise e_{10} hatasında ikincidir. YAAPT metodu öncelikle telefon konuşması için tasarlanmıştır. CREPE ve FCN darbant konuşmasında neredeyse işe yaramamaktadır. Bunun nedeni, eğitimlerinin sadece genişbant konuşma örneklerinde yapılmış olması olabilir. Darbantlı telefon konuşmaları için de eğitilmeleri gerekmektedir. Ayrıca FCN ve CREPE'in uçtan-uca algoritmalar olduğunu ve frekans bilgisi kullanmadan doğrudan ham ses dalgasını giriş olarak aldıklarını unutmamalıyız. Yararlı temel frekans bilgilerinin çoğu sinyalin düşük frekans kısmında (0-400 Hz) gizlidir ve bu veriler olmadan FCN ve CREPE temel frekans belirleme için gerekli özellikleri çıkaramamaktadır. Konvolusyonal katmanlarla frekans bilgisi elde etmek oldukça zor ve hesaplama maliyeti yüksek bir işlemdir.

5.6.2. Temel Frekans ile Cinsiyet Tespit

Cinsiyet algılama, temel frekans tespitinin önemli bir uygulamasıdır. Cinsiyet temel frekans ile sınırlı olmasa da değeri ile oldukça ilişkilidir. Temel frekansın erkekler, kadınlar, erkek çocuklar ve kız çocuklar için belirli aralıkları vardır. Bu nedenle, f0 algoritmalarının cinsiyet değerlendirmesi, algoritmanın sağlamlığı için iyi bir ölçüdür. Burada genişbant ve darbant TIMIT veri kümesi için cinsiyet algılama hatalarını sunuyoruz. TIMIT veri kümesinde, cinsiyet bilgileri konuşma örneğinin adının ilk harfi ile verilir. Çizelge 5.46 ve Çizelge 5.47'den görüldüğü gibi, genişbant ve darbant TIMIT veri kümesinde, HDM önemli bir farkla cinsiyet algılamada en iyi yöntemdir, genişbantta FCN ikinci ve CREPE üçüncüdür. TIMIT veri kümesinde 24017 kadın ve 54357 erkek örneği bulunmaktadır. Çizelge 5.46 ve Çizelge 5.47'den, HDM'nin özellikle erkek örneklerdeki başarısının daha iyi olduğu sonucuna varabiliriz.

Çizelge 5.46. Genişbant TIMIT Veri Kümesinde Cinsiyet Algılama Sonuçları.

	Kadın Hatalı	**Erkek Hatalı**	**Hata (%)**
FCN	1724	1302	3,86
AC	2403	5331	9,86
YIN	2674	8136	13,79
YAAPT	2483	5475	10,15
CREPE	2295	3435	7,31
CEPS	3219	2999	7,93
HDM	**1476**	**832**	**2,94**

Çizelge 5.47. Darbant TIMIT Veri Kümesinde Cinsiyet Algılama Sonuçları.

	Kadın Hatalı	**Erkek Hatalı**	**Hata (%)**
FCN	17717	34398	66,49
AC	**3919**	15091	24,25
CEPS	4248	16258	26,16
CREPE	10331	29155	50,38
YAAPT	5124	6129	14,35
YIN	7359	24393	40,51
HDM	6561	**363**	**8,83**

6. SONUÇLAR VE GELECEK ÇALIŞMALAR

Bu çalışmada, konuşma duygu tanıma için önerilen EFB filtreleri ile Mel ve MFCC özniteliklerini karşılaştırıcı deneyler yapılmış, insan sesi temel frekans tespitinde kullanılmak üzere yeni bir metot (HDM) geliştirilmiştir. HDM metodu otokorelasyon, kepstrum, YIN, YAAPT, FCN ve CREPE yöntemleriyle kıyaslanmıştır. EFB filtreleri için EmoSTAR, EmoDB, IEMOCAP ve MELD verisetleri, temel frekans deneyleri için ise TIMIT, Hillenbrand ve Texas sesli harf verisetleri kullanılmıştır. İkinci bölümde ilgili çalışmalar ele alınmıştır. CNN, LSTM ve BiLSTM modellerini konuşma tanıma ve konuşma duygu tanıma alanlarında Mel filtreleri, MFCC, spektrogram özellikleri ile IEMOCAP, MELD verisetlerinde uygulayan çalışmalar özetle sunulmuştur. Derin öğrenme metotlarının başarılarının gün geçtikçe arttığı gözlemlenmiştir. Temel frekansla ilgili çalışmalarda YIN, YAAPT, CREPE, FCN, otokorelasyon, kepstrum metotlarından bahsedilmiştir. CREPE ve FCN derin öğrenmeye dayalı temel frekans tespit yöntemleridir ve doğrudan ses verisi üzerinde çalışmaktadırlar. Üçüncü bölüm ise tezde uygulanan metotlara ayrılmıştır. Derin öğrenme metotlarından CNN, LSTM ve BiLSTM metotları konuşma duygu tanıma deneylerinde denenerek, SVM metodu ile karşılaştırılmıştır. CNN metodu olarak 2 boyutlu konvolusyon katmanları kullanan NVIDIA modeli ve 1 boyutlu konvolusyon katmanları kullanan bir model denenmiştir. Genel olarak NVIDIA modeli ve SVM daha başarılı olarak ön plana çıkmaktadır. Temel frekans deneylerinde de önerilen HDM metodu diğer metotlarla Hillenbrand, Texas ve TIMIT sesli harf verisetlerinde test edilmiştir. Dördüncü bölüm ise tezde kullanılan EmoSTAR, EmoDB, IEMOCAP ve MELD verisetleri, Information Gain, CFS Subset Evaluator özellik seçme metotları, dengesiz dağılımlı verisetlerinde uygulanan yöntemler ile ilgili bilgi verilmiştir. Beşinci bölüm deneylerin sonuçlarını sunmaktadır. Deney sonuçları tüm özniteliklerle yapılan, Information Gain öznitelik seçme metodu ile yapılan, CFS Subset Evaluator öznitelik seçme metodu ile yapılan deneylerle, veri türetme, dengesiz dağılımlı veriseti ve temel frekans deneylerini kapsamaktadır. SVM ve NVIDIA CNN modeli genel olarak daha başarılı gözükmektedir. 1B CNN modeli daha karmaşık bir model olmasına rağmen başarı oranlarında genelde daha basit bir model olan NVIDIA CNN modelinin gerisinde kalmaktadır. Model karmaşıklığı daha yüksek başarı

oranını garanti etmemektedir.

EFB filtreleri, SVM, NVIDIA CNN (CNN1), 1B CNN (CNN2), LSTM ve Bidirectional LSTM modelleri kullanılarak karşılaştırılmış ve daha iyi veya kıyaslanabilir sonuçlar elde etmiştir. Konuşma duygu işleme uygulamalarındaki çalışmalar Mel filtreleri veya MFCC öznitelikleri ile sınırlı değildir. Yeni filtre bankaları, MFCC ve Mel filtre bankalarına kıyasla çok daha hızlı hesaplanabilir ve daha kolay yorumlanabilir özelliklerdir. Önerilen yeni filtre bankaları sadece 13 farklı frekans bölgesinden oluşmaktadır. Bu deneyler, EFB filtre bankalarının MFCC ve Mel filtre bankalarının yerine kullanılabileceğini somut olarak kanıtlamaktadır. Dengesiz veri sorunu birçok farklı yönden ele alınmış, örnek azaltma, SMOTE, veri türetme, AdaboostM1 ve Bagging yaklaşımları uygulanmıştır. Veri türetme ve SMOTE başarı oranlarını oldukça yükseltmektedir. İleriki çalışmalarda yeni filtre bankalarını diğer konuşma işleme (konuşma tanıma, konuşmacı tanıma, kelime tanıma vb..) deneylerinde uygulamayı düşünmekteyiz.

Temel frekans deneysel sonuçları, önerilen harmonikler arası farkların genişbant ve darbant konuşmalarda temel frekansı tespit etmek için güvenle kullanılabileceğini göstermektedir. Yeni algoritma, özellikle büyük TIMIT veri kümesinde önemli başarı gösteriyor. Hızlı Fourier dönüşümünün doğal bir çözünürlük sorunu vardır, ancak bu tezde, uygulamanın düşük çözünürlüğüne rağmen, sonuçlar oldukça tatmin edicidir. HDM algoritması en hızlı yöntemdir. FCN ve CREPE genişbant verilerinde oldukça iyi performans gösteriyorlar, ancak diğer yöntemlere kıyasla çok yavaştırlar. Bu nedenle, gerçek zamanlı uygulamalar için kullanılamazlar, ancak gerçek temel frekans değerini belirlemede çok yardımcı olabilirler. İlginç bir bulgu, darbant konuşmasında FCN ve CREPE algoritmalarının çok hatalı sonuçlar üretmesidir. Kepstrum ve otokorelasyon algoritmaları YIN, YAAPT, CREPE ve FCN ile karşılaştırıldığında çok eskidir, ancak bazı durumlarda daha iyi tahminler sunabiliyor. Bundan sonraki çalışmalarda, HDM'de zamansal yumuşatmayı hayata geçirmeyi planlıyoruz. Gelecekteki bir diğer çalışma, HDM'nin gürültülü ortamlarda ve müzik seslerindeki yeteneğini test etmektir. Temel frekans hassaslığını arttırma, HDM'nin içine dahil edilebilecek başka bir tekniktir.

7. KAYNAKLAR

[1] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Networks*, c. 1, sayı 2, ss. 119-130, 1988, doi: 10.1016/0893-6080(88)90014-7.

[2] Y. LeCun, , B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard ve L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, c. 1*,* sayı 4, ss. 541-551, 1989, doi: 10.1162/neco.1989.1.4.541.

[3] Stanford University. (2022, 1 February). *Convolutional neural networks (CNNs / ConvNets)*. Stanford University. Erişim: https://cs231n.github.io/convolutional-networks/.

[4] S. Hochreiter ve J. Schmidhuber, "LSTM can solve hard long time lag problems," *Advances in Neural Information Processing Systems* 9, Denver, Colorado, USA, 1996, ss. 473-479.

[5] F. F. Li, J. Johnson ve S. Yeung, (2014, 1 May). *Lecture 10: Recurrent neural networks*. Stanford University. Erişim: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture10.pdf.

[6] J. Chung, C. Gulcehre, K. Cho ve Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555 [cs.NE]*, 2014.

[7] C. Cortes ve V. Vapnik, "Support-vector networks," *Machine Learning*, c. 20*,* sayı 3, ss. 273-297, 1995, doi: https://doi.org/10.1007/BF00994018.

[8] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, c. 2, sayı 2, ss.121-167, 1998, doi: https://doi.org/10.1023/A:1009715923555.

[9] S. Davis ve P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, c. 28, sayı 4, ss. 357-366, 1980.

[10] J. Volkmann, S. S. Stevens ve E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, c. 8, sayı 3, ss. 185-190, 1937, doi: 10.1121/1.1915893.

[11] C. Parlak, B. Diri ve F. Gürgen, "A cross-corpus experiment in speech emotion recognition," *15th Annual Conference of the International Speech Communication Association, SLAM@INTERSPEECH,* Penang, Malaysia, 2014, ss. 58-61.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier ve B. Weiss, "A database of German emotional speech," *Eurospeech, 9th European Conference on Speech Communication and Technology, INTERSPEECH,* Lisbon, Portugal, 2005, c. 5, ss. 1517-1520.

[13] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee ve S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, c. 42, sayı 4, ss. 335-359, 2008, doi: 10.1007/s10579-008-9076-6.

[14] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria ve R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508 [cs.CL],* 2018.

[15] J. Hillenbrand, L. A. Getty, M. J. Clark ve K. Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical Society of America*, c. 97, sayı 5, ss. 3099-3111, 1994, doi: 10.1121/1.409456.

[16] P. F. Assmann ve T. M. Nearey, "Perception of front vowels: The role of harmonics in the first formant region," *The Journal of the Acoustical Society of America*, c. 81, sayı 2, ss. 520-534, 1987, doi: 10.1121/1.394918.

[17] J. S. Garofolo, vd., "TIMIT acoustic phonetic continuous speech corpus," *Web Download. Philadelphia: Linguistic Data Consortium,* 1993.

[18] P. Elias, "Predictive coding-I," *IRE Transactions on Information Theory*, c. 1, sayı 1, ss. 16-24, 1955, doi: 10.1109/tit.1955.1055116.

[19] B. Atal ve M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, c. 27, sayı 3, ss. 247-254, 1979, doi: 10.1109/tassp.1979.1163237.

[20] J. W. Cooley ve J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, c. 19, sayı 90, ss. 297-301, 1965, doi: 10.2307/2003354.

[21] N. Ahmed, T. Natarajan ve K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, c. 100, sayı 1, ss. 90-93, 1974, doi: 10.1109/T-C.1974.223784.

[22] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, D. L. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao ve K. Zieba, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316 [cs.CV],* 2016.

[23] Puthran, M. (2021, 12 January). *Speech emotion analyzer*. Github. Erişim: https://github.com/Miteshputhranneu/Speech-Emotion-Analyzer.

[24] A. Graves, S. Fernández ve J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," *15th International Conference on Artificial Neural Networks*, Warsaw, Poland, 2005, ss. 799-804, doi: 10.1007/11550907_126.

[25] Matlab, audioDataAugmenter (2022, 1 Şubat). *Matworks*. Erişim: https://www.mathworks.com/help/audio/ref/audiodataaugmenter.html.

[26] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski ve M. W. Mahoney, "Feature selection methods for text classification," *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA, 2007, ss. 230-239.

[27] M. A. Hall, "Correlation-based feature selection for machine learning," Doktora tezi, Department of Computer Science, University of Waikato, Hamilton, New

Zealand, Apr. 1999.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann ve I. H. Witten, "The WEKA data mining software: an update," *ACM Special Interest Group on Knowledge Discovery in Data (SIGKDD) Explorations Newsletter,* c. 11, sayı 1, ss. 10-18, 2009, doi: 10.1145/1656274.1656278.

[29] L. Ardaillon ve A. Roebel, "Fully-convolutional network for pitch estimation of speech signals," *Proc. Twentieth Annual Conference of the International Speech Communication Association, INTERSPEECH,* Graz, Austria, 2019, ss. 2005-2009, doi: 10.21437/Interspeech.2019-2815.

[30] J. W. Kim, J. Salamon, P. Li ve J. P. Bello, "CREPE: A convolutional representation for pitch estimation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, Calgary, Alberta, Canada, 2018, ss. 161-165, doi: 10.1109/ICASSP.2018.8461329.

[31] J. C. R. Licklider, "Periodicity pitch and related auditory process models," *International Audiology*, c. 1, sayı 1, ss. 11-34, 1962, doi: 10.3109/05384916209074592.

[32] A. M. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, c. 41, sayı 2, ss. 293-309, 1967, doi: 10.1121/1.1910339.

[33] A. De Cheveigné ve H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, c. 111, sayı 4, ss. 1917-1930, 2002, doi: 10.1121/1.1458024.

[34] K. Kasi ve S. A. Zahorian, "Yet another algorithm for pitch tracking," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE, Renaissance Orlando Resort, Orlando, Florida, USA, 2002, c. 1, ss. I-361-I-364, doi: 10.1109/ICASSP.2002.5743729.

[35] C. Parlak ve Y. Altun, "Harmonic Differences Method for Robust Fundamental Frequency Detection in Wideband and Narrowband Speech Signals," *Mathematical Problems in Engineering*, c. 2021, ss. 1-17, 2021, doi: 10.1155/2021/6658951.

[36] O. Abdel-Hamid, A. R. Mohamed, H. Jiang ve G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, Kyoto, Japan, 2012, ss. 4277-4280, doi: 10.1109/ICASSP.2012.6288864.

[37] G. E. Hinton ve R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, c. 313, sayı 5786, ss. 504-507, 2006, doi: 10.1126/science.1127647.

[38] G. Dahl, M. A. Ranzato, A. R. Mohamed ve G. E. Hinton, "Phone recognition with the mean-covariance Restricted Boltzmann Machine," *Advances in Neural Information Processing Systems*, NIPS, Vancouver, Canada, 2010, ss. 469-477.

[39] T. Ko, V. Peddinti, D. Povey ve S. Khudanpur, "Audio augmentation for speech recognition," *Sixteenth Annual Conference of the International Speech Communication Association,* Dresden, Germany, 2015, ss. 3586-3589, doi: 10.21437/Interspeech.2015-711.

[40] J. J. Godfrey ve E. Holliman, Switchboard-1 Release 2 LDC97S62. Web

Download. *Philadelphia: Linguistic Data Consortium*, 1993.

[41] K. Walker, et al. (2013). GALE Phase 2 Chinese Broadcast News Speech LDC2013S08. Web Download. *Philadelphia: Linguistic Data Consortium, 2013*.

[42] A. Rousseau, P. Deléglise ve Y. Esteve, "TED-LIUM: an Automatic Speech Recognition dedicated corpus," *The International Conference on Language Resources and Evaluation, LREC,* İstanbul, Turkey, 2012, ss. 125-129.

[43] M. Gogate, K. Dashtipour, A. Adeel ve A. Hussain, "CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement," *Information Fusion*, c. 63, sayı November 2020, ss. 273-285, 2020, doi: 10.1016/j.inffus.2020.04.001.

[44] V. Panayotov, G. Chen, D. Povey ve S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, Brisbane, Queensland, Australia, 2015, ss. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[45] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel ve P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, c. 19, sayı 4, ss. 788-798, 2011, doi: 10.1109/tasl.2010.2064307.

[46] N. Jaitly ve G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," *Proc. of the 30th International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech and Language,* Atlanta, Georgia, USA, 2013, c. 117, ss. 21-25.

[47] W. Verhelst ve M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE, Minneapolis, Minnesota, USA, 1993, c. 2, ss. 554-557, doi: 10.1109/ICASSP.1993.319366.

[48] W. Song ve J. Cai, "End-to-end deep neural network for automatic speech recognition". California, USA, Stanford CS224D Reports, 2015.

[49] A. Graves, S. Fernández, F. Gomez ve J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, 2006, ss. 369-376, doi: 10.1145/1143844.1143891.

[50] L. E. Baum ve T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, c. 37, sayı 6, ss. 1554-1563, 1966, doi: 10.1214/aoms/1177699147.

[51] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, ve T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, Florida, USA, 2014, ss. 675-678, doi: 10.1145/2647868.2654889.

[52] A. Maas, Z. Xie, D. Jurafsky ve A. Y. Ng, "Lexicon-free conversational speech recognition with neural networks," *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Denver, Colorado, USA, 2015, ss. 345-354, doi:

10.3115/v1/N15-1038.

[53] L. Tóth, “Phone recognition with hierarchical convolutional deep maxout networks,” *EURASIP Journal on Audio, Speech, and Music Processing*, c. 2015, sayı 1, ss. 1-13, 2015, doi: 10.1186/s13636-015-0068-3.

[54] J. Michalek ve J. Vaněk, “A survey of recent DNN architectures on the TIMIT phone recognition task,” *International Conference on Text, Speech, and Dialogue,* Brno, Czech Republic, 2018, ss. 436-444.

[55] J. B. Delbrouck, N. Tits ve S. Dupont, “Modulated fusion using transformer for linguistic-acoustic emotion recognition,” *arXiv preprint arXiv:2010.02057 [cs.CL],* 2020.

[56] A. Zadeh, R. Zellers, E. Pincus ve L. P. Morency, “MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,” *arXiv preprint arXiv:1606.06259 [cs.CL],* 2016.

[57] A. Zadeh ve P. Pu, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers),* Melbourne, Australia, 2018, ss. 2236-2246, doi: https://aclanthology.org/P18-1208/.

[58] J. Pennington, R. Socher ve C. D. Manning, “Glove: Global vectors for word representation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Doha, Qatar, 2014, ss. 1532-1543, 10.3115/v1/D14-1162.

[59] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805,* 2018.

[60] M. Xu, F. Zhang, X. Cui ve W. Zhang, “Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation,” *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, Toronto, Ontario, Canada, 2021, ss. 6319-6323, doi: 10.1109/ICASSP39728.2021.9414635.

[61] E. Ma, (2022, 23 December). *Nlpaug: Data augmentation for NLP*. Erişim: https://github.com/makcedward/nlpaug.

[62] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenber ve O. Nieto, “Librosa: Audio and music signal analysis in python.” *Proceedings of the 14th Python in Science Conference,* Austin, Texas, USA, 2015, ss. 18-24, doi: 10.25080/Majora-7b98e3ed-003.

[63] A. Satt, S. Rozenberg ve R. Hoory, “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms,” *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, INTERSPEECH,* Stockholm, Sweden, 2017, ss. 1089-1093, doi: 10.21437/Interspeech.2017-200.

[64] H. M. Fayek, M. Lech ve L. Cavedon, ”Evaluating deep learning architectures for Speech Emotion Recognition,” *Neural Networks*, c. 92, 2017, ss. 60-68, 2017, doi: 10.1016/j.neunet.2017.02.013.

[65] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato ve Y. LeCun, “What is the best multi-

stage architecture for object recognition?," *2009 IEEE 12th International Conference on Computer Vision,* IEEE, Kyoto, Japan, 2009, ss. 2146-2153.

[66] S. Tripathi, A. Kumar, A. Ramesh, C. Singh ve P. Yenigalla, "Deep learning-based emotion recognition system using speech features and transcriptions," *arXiv preprint arXiv:1906.05681 [eess.AS],* 2019.

[67] B. Fernandes ve K. Mannepalli, "Speech Emotion Recognition Using Deep Learning LSTM for Tamil Language," *Pertanika Journal of Science & Technology*, c. 29, sayı 3, ss. 1915-1936, 2021, doi: 10.47836/pjst.29.3.33.

[68] T. Deschamps-Berger, L. Lamel ve L. Devillers, "End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings," *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII),* IEEE, Virtual Event, 2021, ss. 1-8, doi: 10.1109/ACII52823.2021.9597419.

[69] L. Vidrascu ve L. Devillers, "Detection of real-life emotions in call centers," *Ninth European Conference on Speech Communication and Technology, ISCA,* Lisbon, Portugal, 2005, ss. 1841-1844, doi: 10.21437/Interspeech.2005-582.

[70] M. Pakyurek, M. Atmis, S. Kulac ve U. Uludag, "Extraction of novel features based on histograms of MFCCs used in emotion classification from generated original speech dataset," *Elektronika ir Elektrotechnika*, c. 26, sayı 1, ss. 46-51, 2020, doi: 10.5755/j01.eie.26.1.25309.

[71] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu ve G. Dai, "Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition," *Frontiers in Physiology*, c. 12, 2021, doi: 10.3389/fphys.2021.643202.

[72] M. Staudacher, V. Steixner, A. Griessner ve C. Zierhofer, "Fast fundamental frequency determination via adaptive autocorrelation," *EURASIP Journal on Audio, Speech, and Music Processing*, c. 2016, sayı 1, ss. 1-8, 2016, doi: 10.1186/s13636-016-0095-8.

[73] P. Bagshaw, "Automatic prosodic analysis for computer aided pronunciation teaching," Doktora tezi, University of Edinburgh, Edinburgh, United Kingdom, 1994.

[74] F. Plante, G. Meyer ve W. Ainsworth, "A pitch extraction reference database," *Children*, c. 8, sayı 12, ss. 30-50, 1995.

[75] G. Pirker, M. Wohlmayr, S. Petrik ve F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," *Twelfth Annual Conference of the International Speech Communication Association,* ISCA, Florence, Italy, 2011, c. 3, ss. 1509-1512, doi: 10.21437/Interspeech.2011-317.

[76] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," *Sixth International Conference on Spoken Language Processing,* Beijing, China, 2000, c. 4, ss. 676-679.

[77] H. Ba, N. Yang, I. Demirkol ve W. Heinzelman, "BaNa: A hybrid approach for noise resilient pitch detection," *2012 IEEE Statistical Signal Processing Workshop (SSP),* IEEE, Ann Arbor, Michigan, USA, 2012, ss. 369-372, doi: 10.1109/SSP.2012.6319706.

[78] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal

decoding algorithm," *IEEE Transactions on Information Theory*, c. 13, sayı 2, ss. 260-269, 1967. doi: 10.1109/TIT.1967.1054010.

[79] A. De Cheveigné ve H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, c. 111, sayı 4, ss. 1917-1930, 2002, doi: 10.1121/1.1458024.

[80] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustical Society of America*, c. 83, sayı 1, ss. 257-264, 1988, doi: 10.1121/1.396427.

[81] H. Kawahara, "STRAIGHT-TEMPO: A universal tool to manipulate linguistic and para-linguistic speech information," *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation,* IEEE, 1997, c. 2, ss. 1620-1625, doi: 10.1109/ICSMC.1997.638234.

[82] M. Mauch ve S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, 2014, ss. 659-663, doi: 10.1109/ICASSP.2014.6853678.

[83] M. Goto, H. Hashiguchi, T. Nishimura ve R. Oka, "RWC music database: popular, classical and jazz music databases," *3rd International Conference on Music Information Retrieval, ISMIR,* Paris, France, 2002 c. 2, ss. 287-288.

[84] M. Mauch ve S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," *14th International Conference on Music Information Retrieval, ISMIR*, Curitiba, Brazil, 2013*,* ss. 83-88.

[85] S. A. Zahorian, P. Dikshit ve H. Hu, "A spectral-temporal method for pitch tracking," *Ninth International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, 2006*,* ss. 1710-1713, doi: 10.21437/Interspeech.2006-475.

[86] D. Talkin, W. B. Kleijn ve K. K. Palatal, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, ss. 495-518, 1995.

[87] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam ve J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research," *The 15th International Conference on Music Information Retrieval, ISMIR*, Taipei, Taiwan, 2014, c. 14, ss. 155-160.

[88] A. Camacho ve J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, c. 124, sayı 3, ss. 1638-1652, 2008, doi: 10.1121/1.2951592.

[89] J. L. Gauvain, L. Lamel ve M. Eskénazi, "Design considerations and text selection for BREF, a large French read-speech corpus," *Proc. First International Conference on Spoken Language Processing (ICSLP 1990)*, Kobe, Japan, 1990, c. 90, ss. 1097-1100.

[90] K. Yamaguchi, K. Sakamoto, T. Akabane ve Y. Fujimoto, "A neural network for speaker-independent isolated word recognition," *Proc. First International Conference on Spoken Language Processing (ICSLP 1990),* Kobe, Japan, 1990, ss. 1077-1080.

[91] M. Jafarzadeh ve Y. Tadesse, "Convolutional neural networks for speech controlled prosthetic hands," *2019 First International Conference on*

Transdisciplinary AI (TransAI), IEEE, Laguna Hill, California, USA, 2019, ss. 35-42, doi: 10.1109/TransAI46475.2019.00014.

[92] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li ve Y. Zhang, "Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, Barcelona, Spain, 2020, ss. 6124-6128, doi: 10.1109/ICASSP40776.2020.9053889.

[93] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen ve R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," arXiv preprint arXiv:1904.03288 [eess.AS], 2019.

[94] S. Rella ve T. Yarshney, (2021, 9 November). *Speech recognition: generating accurate domain-specific audio transcriptions using NVIDIA Riva*. Erişim: https://developer.nvidia.com/blog/speech-recognition-generating-accurate-transcriptions-using-riva/.

[95] M. Schuster ve K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, c. 45, sayı 11, ss. 2673-2681, 1997, doi: 10.1109/78.650093.

[96] Z. Jin ve D. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, c. 19, sayı 5, ss. 1091-1102, 2011, doi: 10.1109/tasl.2010.2077280.

[97] M. Wu, D. Wang ve G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, c. 11, sayı 3, ss. 229-241, 2003, doi: 10.1109/tsa.2003.811539.

[98] S. Seneff, "Real-time harmonic pitch detector," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, c. 26, sayı 4, ss. 358-365, 1978, doi: 10.1109/tassp.1978.1163118.

[99] L. Wu, "Guitar sound analysis and pitch detection," Stanford University, Stanford, California, USA, 2017.

[100] M. Dziubiński ve B. Kostek, "High accuracy and octave error immune pitch detection algorithms," *Archives of Acoustics*, c. 29, sayı 1, ss. 1-21, 2004.

[101] J. C. R. Licklider, "A duplex theory of pitch perception," *The Journal of the Acoustical Society of America*, c. 23, sayı 1, ss. 147-147, 1951, doi: 10.1121/1.1917296.

[102] R. J. Ritsma, "Existence region of the tonal residue. I," *The Journal of the Acoustical Society of America*, c. 34, sayı 9A, ss. 1224-1229, 1962, doi: 10.1121/1.1918307.

[103] M. Ross, H. Shaffer, A. Cohen, R. Freudberg ve H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, c. 22, sayı 5, ss. 353-362, 1974, doi: 10.1109/tassp.1974.1162598.

[104] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," *Symposium on Computer Processing in Communication,* ed. by the Microwave Institute, University of Brooklyn Press, New York, USA, 1970, c. 19, ss. 779-797.

[105] C. Wendt ve A. P. Petropulu, "Pitch determination and speech segmentation using the discrete wavelet transform," *1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World. ISCAS 96,* IEEE, Atlanta, Georgia, USA, 1996, c. 2, ss. 45-48, doi: 10.1109/ISCAS.1996.540348.

[106] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, c. 5, sayı 9, ss. 341-345, 2002.

[107] C. Parlak ve B. Diri, "Farklı verisetleri arasında duygu tanıma çalışması," *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, c. 16, sayı 48, ss. 21-29, 2014.

[108] C. Parlak ve B. Diri, "İnsan sesinden duygu tanıma," Yüksek lisans tezi, Bilgisayar Mühendisliği, Fen Bilimleri Enstitüsü, Yıldız Teknik Universitesi, Istanbul, Turkey, 2015.

[109] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, c. 76, sayı 5, ss. 378-382, 1971, doi: 10.1037/h0031619.

[110] S. Y. Chen, C. C. Hsu, C. C. Kuo ve L. W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," *arXiv preprint arXiv:1802.08379, 2018*.

[111] J. Pearl, *Heuristics: intelligent search strategies for computer problem solving,* Reading, Massachusetts, USA: Addison-Wesley Longman Publishing Co., Inc., 1984.

[112] N. V. Chawla, K. W. Bowyer, L. O. Hall ve W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, c. 16, ss. 321-357, 2002, doi: 10.1613/jair.953.

[113] L. Breiman, "Bagging predictors," *Machine Learning*, c. 24, sayı 2, ss. 123-140, 1996, doi: 10.1007/bf00058655.

[114] Y. Freund ve R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, c. 55, sayı 1, ss. 119-139, 1997, doi: 10.1006/jcss.1997.1504.

[115] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho ve K. Chen, "Xgboost: extreme gradient boosting," *R Package Version 0.4-2*, c. 1, sayı 4, ss. 1-4, 2015.

[116] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk ve F. Herrera, *Learning From Imbalanced Data Sets.* Berlin, Germany, Springer, 2018.

[117] J. Brownlee, *Imbalanced Classification With Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning,* San Fransisco, USA, Machine Learning Mastery, 2020.

[118] C. Ferri, J. Hernández-Orallo ve R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, c. 30, sayı 1, ss. 27-38, 2009, doi: 10.1016/j.patrec.2008.08.010.

[119] G. V. Rossum ve F. L. Drake, *Python 3 Reference Manual*. CreateSpace, Scotts Valley, California, USA, 2009.

[120] M. Abadi, P. Barham, J. Chen, J., Z. Chen, A. Davis, J. Dean, M. Devin, X. Zheng vd., "{TensorFlow}: A System for {Large-Scale} Machine Learning," *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16),* Savannah, Georgia, USA, 2016, ss. 265-283.

[121] F. Chollet vd. (2015). *Keras*. GitHub. Erişim: https://github.com/fchollet/keras.

[122] A. T. Kabakuş, "A Comparison of the State-of-the-Art Deep Learning Platforms: An Experimental Study," *Sakarya University Journal of Computer and Information Sciences*, c. 3, sayı 3, ss. 169-182, 2020, doi: 10.35377/saucis.03.03.776573.

[123] M. Slaney, "Auditory toolbox," Interval Research Corporation, Palo Alto, USA, Technical Report #1998-010, 1998.

[124] Seo, N. (2008, 1 April). *Enee632 project4 part i.* Erişim: http://note.sonots.com/SciSoftware/Pitch.html#v44c5761.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Cevahir PARLAK

Yabancı Dili : İngilizce

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Doktora	Bilgisayar Müh.	Düzce Üniversitesi	2022
Y. Lisans	Bilgisayar Müh.	Yıldız Teknik Üniversitesi	2015
Lisans	Bilgisayar Müh.	Sakarya Üniversitesi	2012
Lise	Fen	Kazım İşmen Lisesi	1986

YAYINLAR

Article Name: Harmonic Differences Method for Robust Fundamental Frequency Detection in Wideband and Narrowband Speech Signals

DOI: https://doi.org/10.1155/2021/6658951