T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SES OLAY TESPİT PROBLEMİNE DERİN ÖĞRENME TABANLI ÇÖZÜMLER

DOKTORA TEZİ

Furkan Yusuf YAVUZ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Bilim Dalı

HAZİRAN 2025

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SES OLAY TESPİT PROBLEMİNE DERİN ÖĞRENME TABANLI ÇÖZÜMLER

DOKTORA TEZİ

Furkan Yusuf YAVUZ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Bilim Dalı

Tez Danışmanı: Prof. Dr. Nejat YUMUŞAK

HAZİRAN 2025

Furkan Yusuf Yavuz tarafından hazırlanan "Ses Olay Tespit Problemine Derin Öğrenme Tabanlı Çözümler" adlı tez çalışması 16.06.2022 tarihinde aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı **Bilgisayar Mühendisliği** Bilim Dalı'nda Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı :	**Prof. Dr. Pakize ERDOĞMUŞ** Düzce Üniversitesi	.....
Jüri Üyesi :	**Prof. Dr. Nejat YUMUŞAK** (Danışman) Sakarya Üniversitesi	.....
Jüri Üyesi :	**Prof. Dr. Resul KARA** Düzce Üniversitesi	.....
Jüri Üyesi :	**Prof. Dr. Celal ÇEKEN** Sakarya Üniversitesi	.....
Jüri Üyesi :	**Prof. Dr. Ünal ÇAVUŞOĞLU** Sakarya Üniversitesi	.....

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum "SES OLAY TESPİT PROBLEMİNE DERİN ÖĞRENME TABANLI ÇÖZÜÜMLER" başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

(……/……/20…..).

(imza)

Furkan Yusuf Yavuz

everyone tells their own story

TEŞEKKÜR

Tez çalışmam süresince kıymetli yönlendirmeleriyle bana yol gösteren danışmanım Prof. Dr. Nejat YUMUŞAK'a en içten teşekkürlerimi sunarım.

Akademik yolculuğum boyunca yanımda olan, fedakârlıkları ve sürekli destekleriyle bu süreci mümkün kılan aileme derin bir şükran borçluyum.

Furkan Yusuf YAVUZ

İÇİNDEKİLER

Sayfa

KISALTMALAR

CNN : Convolutional Neural Network
DCT : Discrete Cousine Transform
DFT : Discrete Fourier Transform
DL : Deep Learning
DNN : Deep Neural Network
FFT : Fast Fourier Transform
FLAC : Free Lossless Audio Codec
GPU : Graphics Processing Unit
HMM : Hidden Markov Model
IoT : Internet of Things
IR : Infrared - Kızılötesi
LPCC : Linear Predictive Coding Cepstrum
LSTM : Long Short-Term Memory
MFCCs : Mel-Frequency Cepstral Coefficients
PSDS : Polyphonic Sound Detection Score
ReaLISED : Real-Life Indoor Sound Event Dataset
ReLU : Rectified Linear Unit
SED : Sound Event Detection
SNR : Signal-to-Noise Ratio
STFT : Short-Time Fourier Transform
SVM : Support Vector Machine
TN : True Negative
TP : True Positive
t-SNE : t-distributed Stochastic Neighbor Embedding
YZ : Yapay Zeka

SİMGELER

$\mathcal{L}_{BCE}$	: İkili Çapraz Entropi Kayıp fonksiyonu
2π/N	: Frekans çözünürlüğü ölçekleme faktörü
H	: Atlama boyutu (ardışık çerçeveler arasındaki adım)
i	: Örneğin indeksi
j	: İmajiner birim ($j^2 = -1$)
k	: Frekans indeksi
log()	: Doğal logaritma fonksiyonu
m	: Çerçeve indeksi
N	: Analiz penceresinin uzunluğu (her karedeki örnek sayısı)
n	: Örnek sırasa sayusu
w[n]	: n. kareye uygulanan pencere fonksiyonu (örn. Hamming, Hanning)
$\mathbf{y'_i}$	: i sırasındaki örnek için öngörülen olasılık (0 ile 1 arasında değer)
$\mathbf{y_i}$	: i sırasındaki örneğin gerçek etiketi (0 veya 1)
Σ (Sigma)	: Toplama operatörü, belirtilen aralık üzerinde toplama yapar

TABLO LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

SES OLAY TESPİT PROBLEMİNE DERİN ÖĞRENME TABANLI ÇÖZÜMLER

ÖZET

Mülk güvenliği için üretilen gelişmiş kamera sistemleri, fiziksel olarak kolayca devre dışı bırakılabilmeleri nedeniyle çoğu zaman yetersiz kalmaktadır. Bu sınırlamayı aşmak amacıyla, olası tehditlere ait ses verilerinin güvenlik sistemlerine dâhil edilmesi, geleneksel yöntemlere önemli bir katkı sunmaktadır. Ses tabanlı algılama sistemleri, düşük maliyetli donanımlar (örneğin mikrofonlar) ile kurulabilen kompakt yapıları sayesinde görsel tabanlı güvenlik çözümlerine etkili ve ölçeklenebilir bir alternatif oluşturmaktadır.

Bu çalışmada, adım sesi olaylarının sınıflandırılması amacıyla Evrişimli Sinir Ağı (CNN) tabanlı bir derin öğrenme yöntemi önerilmektedir. Geliştirilen model, çevresel sesler arasından adım seslerini tespit etmeye odaklanarak, mülk güvenliğine karşı tehditlerin erken fark edilmesini kolaylaştırmaktadır. Çalışmanın ilk aşamsında, gerçek yaşamdan çeşitli ses olaylarını içeren ReaLISED veri kümesi kullanılmıştır. Ancak bu veri kümesine dayalı ilk değerlendirme sonuçları görece düşük performans göstermiştir. Yapılan analizler, sınıf dengesizliği ve düşük kaliteli ya da tutarsız ses örneklerinin, modelin genelleştirilebilir örüntüler öğrenme kapasitesini sınırladığını ortaya koymuştur. Bu sorunları gidermek adına ikinci aşamada, düşük kaliteli veriler dikkatlice elenmiş ve Epidemic Sound platformundan temin edilen yüksek kaliteli adım sesi örnekleri ile veri kümesi zenginleştirilmiştir. Bu iyileştirme, sadece sınıf dağılımlarını dengelemekle kalmamış, aynı zamanda genel veri kalitesini artırarak modelin gerçek dünya koşullarında daha tutarlı ve güvenilir sınıflandırma performansı göstermesini sağlamıştır.

Her bir ses olayı, Mel-Frekans Kepstrum Katsayıları (MFCC) temsiline dönüştürülmüştür. Ayrıca, MFCC ile ilgili ayarlar üzerinde kapsamlı parametre optimizasyonu gerçekleştirilmiş; özellikle n_fft ve n_mfcc değerleri, öznitelik çözünürlüğü ve sınıflandırma doğruluğu açısından optimize edilmiştir. Elde edilen MFCC temsilleri, CNN modelinin eğitimi için görüntü benzeri girişler olarak kullanılmıştır.

Önerilen model, 17 farklı ses kategorisinden, adım sesi olaylarını %98 doğruluk oranı tespit etmeyi başarmıştır. Modelin sağlamlığı, 5 katlı ve 10 tekrar içeren Tekrarlı Katmanlı K-Fold Doğrulama yöntemi ile test edilmiş; F1-skorları 0.905 ile 0.992 arasında değişmiş ve ortalama 0.960 olarak hesaplanmıştır. Ek olarak, ResNet50, ResNet101, EfficientNetB0, EfficientNetB4 ve CRNN gibi güncel mimarilerle yapılan karşılaştırmalı değerlendirmeler, önerilen CNN modelinin etkinliğini pekiştirmiştir. Açık kaynaklı verilerin stratejik biçimde sisteme entegre edilmesi, şeffaflık ve tekrar edilebilirlik ilkelerini destekleyerek sistemin gerçek hayattaki ses tabanlı gözetim uygulamaları alanında taşıdığı pratik değeri güçlendirmiştir.

UTILIZING FOOTSTEP SOUND EVENT DETECTION BY USING CNNTECHNIQUES FOR ASSURING PROPERTY SECURITY

SUMMARY

Although camera-based security systems developed for property protection have reached an advanced level in recent years, they continue to exhibit a number of significant vulnerabilities that limit their overall effectiveness. One of the most pressing concerns is their susceptibility to deliberate tampering or sabotage. In particular, these systems can be easily disabled or rendered inoperative through direct physical intervention by intruders, such as covering, destroying, or disconnecting the cameras. This vulnerability severely undermines their reliability, especially in high-risk or unsupervised environments. As a result, the need for supplementary or alternative security solutions that do not solely rely on visual data has become increasingly evident. In response to this need, researchers and system developers have begun exploring audio-based security systems, which detect and classify environmental sounds such as footsteps, door movements, and glass breakage. Audio-based event detection systems offer a range of practical advantages over their visual counterparts. These include simpler and more flexible installation processes, significantly lower hardware costs due to the affordability and availability of microphones, and greater resistance to tampering or sabotage. Unlike cameras, microphones can be easily concealed, making them less likely to be noticed or targeted by intruders. Furthermore, audio sensors are capable of capturing events occurring outside the direct line of sight, thus overcoming some of the spatial limitations associated with camera systems. Taken together, these advantages highlight the potential of audio-based systems as a cost-effective, resilient, and complementary approach to enhancing property security.

This study proposes a deep learning-based approach for the classification of footstep sound events, addressing a critical component of intelligent surveillance systems. Footstep sounds are widely regarded as a primary acoustic indicator of human presence and can serve as early warnings of unauthorised entry, especially in secured or restricted areas. Unlike general-purpose sound event detection models, the model developed in this research is specifically optimised to distinguish footstep sounds from a variety of other environmental noises commonly found in indoor settings. The classification task is treated as a binary problem: detecting whether a given audio signal contains a footstep sound or not.

To develop and refine the model, a two-stage experimental framework was employed. In the initial stage, the model was trained and evaluated using the ReaLISED dataset, which contains a broad range of labeled environmental audio events recorded under realistic indoor conditions. While this dataset provided a useful starting point due to its variety and availability, the initial performance of the model was found to be suboptimal, especially in terms of generalisation and robustness. A deeper examination of the dataset revealed some major issues affecting model performance: significant class imbalance, the inclusion of low-quality and mislabeled audio samples.

The class imbalance was particularly problematic, as the dataset included a disproportionately small number of footstep events compared to other sound classes, causing the model to develop a bias toward the majority (non-footstep) class. Furthermore, some audio clips labeled as "walking" were actually recordings of running, jumping, or stair climbing—each of which exhibits distinct temporal and spectral characteristics that differ from normal footsteps. The dataset also contained segments with poor acoustic clarity, background interference, and faint or ambiguous signals, making it difficult for the model to extract meaningful features. These issues collectively hindered the model's ability to learn generalised patterns necessary for reliable footstep detection.

To address the limitations identified in the initial phase, a second experimental stage was designed with a strong emphasis on dataset enhancement and refinement. This stage not only sought to resolve the issue of class imbalance by ensuring an equal distribution of footstep and non-footstep audio samples but also significantly improved the quality and diversity of the dataset. Additional high-quality footstep recordings were sourced from Epidemic Sound, a professional sound library, and integrated with carefully selected samples from the original ReaLISED dataset. The resulting collection—termed the "Footsteps Sound Dataset"—was structured to ensure each class contained 333 audio samples, yielding a total of 666 clips. This deliberate balance between classes was essential to prevent the model from being biased toward dominant categories and to promote fair learning.

Prior to feeding the data into the classification model, all audio clips were processed into Mel-Frequency Cepstral Coefficient (MFCC) representations. These features, widely recognized for their ability to mimic the human auditory system's perception of sound, proved especially suitable for distinguishing footstep events from other types of audio. The MFCCs were structured as two-dimensional arrays, analogous to grayscale images, making them highly compatible with Convolutional Neural Networks (CNNs). A thorough hyperparameter optimisation process was conducted, focusing on key parameters such as n_fft (set to 2048) and n_mfcc (set to 20), which directly influenced the resolution and detail of the extracted features. These optimised parameters were chosen after a series of empirical trials and literature-guided benchmarking to ensure optimal performance in both training stability and generalisation.

The CNN model trained on these MFCC inputs demonstrated strong classification capabilities, accurately distinguishing footstep sounds from a background of 17 other environmental sound categories. The model achieved a classification accuracy of approximately 98% on the validation set. To further assess the model's generalisability and robustness, an extensive Repeated Stratified K-Fold Cross-Validation procedure was employed. This validation method, consisting of 5 folds repeated 10 times, ensured that the model's performance was not contingent on a specific data split and mitigated the risk of overfitting. The model achieved F1-scores ranging from 0.905 to 0.992, with an average of 0.960 and a low standard deviation, indicating not only high overall performance but also consistency across multiple data partitions and experimental runs. These results underscore the effectiveness of the proposed CNN-based approach for reliable footstep detection in diverse acoustic environments.

To comprehensively assess the proposed model's effectiveness, it was systematically benchmarked against several widely recognized deep learning architectures, each known for their performance in audio classification and computer vision tasks. These

models included ResNet50, ResNet101, EfficientNetB0, EfficientNetB4, and Convolutional Recurrent Neural Networks (CRNNs), which combine convolutional layers with recurrent units to capture both spatial and temporal dependencies in audio signals. The comparative evaluation was carried out using identical datasets and experimental conditions to ensure fairness. Performance metrics such as precision, recall, and F1-score were used as the primary indicators of model capability. The results from these experiments consistently revealed that the custom-designed CNN model exceeded the performance of all benchmarked architectures in footstep detection tasks. While some models, particularly CRNN and ResNet50, demonstrated results close to the proposed CNN, none were able to match its consistent and high-level classification accuracy across all trials. These findings emphasise the benefits of a specialised and task-focused architecture, tailored specifically for detecting footstep events within noisy, multi-class acoustic environments.

Another significant contribution of this study lies in the strategic integration of open-access audio data to construct a robust and diverse dataset. By incorporating professionally recorded samples from platforms such as Epidemic Sound and selectively combining them with the cleaned and curated portions of the ReaLISED dataset, this work ensured not only higher quality and balance in training data but also enhanced reproducibility and openness in research methodology. The use of publicly accessible sources contributes to the transparency and verifiability of the proposed approach, thereby supporting its adoption in both academic and applied contexts. Moreover, this strategy reinforces the practical relevance of the study by demonstrating that high-performing audio detection systems can be trained and validated using readily available resources, paving the way for real-world deployment in areas such as property security, smart buildings, and health monitoring systems. Ultimately, this research highlights a scalable, efficient, and highly accurate framework for footstep sound detection, offering both theoretical advancement and practical utility.

To further demonstrate the practical applicability of the proposed system, a real-world deployment scenario was developed and analysed. In this envisioned use case, multiple microphones are strategically positioned across various locations within a property—such as entryways, hallways, or near windows—to continuously capture environmental audio signals. These microphones transmit the collected data in real time to a central processing hub, which is responsible for preprocessing the audio, extracting relevant features, and running the trained CNN-based footstep detection model. Upon identifying acoustic patterns indicative of human footsteps, particularly during user-specified active monitoring periods, the system forwards the results to a control interface. This control system includes a user-friendly dashboard where property owners can configure system settings, define monitoring schedules, and manage notification preferences. When unexpected footstep activity is detected—especially during periods when no movement is anticipated—the system immediately triggers an alert mechanism. These alerts can be delivered through push notifications, emails, or other messaging services, ensuring that users are promptly informed of potential intrusions. The ability to selectively activate or deactivate the system adds flexibility and user control, enhancing both usability and reliability. This end-to-end framework highlights the model's efficiency and readiness for real-world application. Compared to traditional surveillance systems, which often rely solely on visual input and require complex hardware and infrastructure, the audio-based approach provides a more discreet, cost-effective, and tamper-resistant alternative. By leveraging the

strengths of deep learning and acoustic analysis, the proposed system offers a robust solution for intelligent, real-time security monitoring.

In conclusion, this study introduces a reliable, low-cost, and high-accuracy audio-based approach for enhancing property security, particularly through the detection of footstep sound events. The proposed CNN-based model demonstrates strong performance and robustness, effectively addressing common limitations found in traditional camera-based surveillance systems, such as high cost, limited coverage, and susceptibility to tampering. By functioning as a complementary layer to existing security infrastructure—or even as a standalone solution in specific contexts—this method significantly broadens the scope of intelligent surveillance technologies. Additionally, the carefully designed model architecture, the rigorous two-phase training process, and the extensive data preprocessing techniques employed throughout the study offer a solid foundation for future research. These contributions not only improve the feasibility of real-world audio-based detection systems but also serve as a valuable guideline for researchers and practitioners aiming to advance the field of sound-based monitoring and intelligent audio analysis in both security and broader acoustic event detection applications.

1. GİRİŞ

Bu bölümde öncelikle, çalışmanın bağlamı açıklanarak temel çerçeve ortaya konulacaktır. Ardından, çalışmanın ele aldığı problem tanımlanacak; bu problemin kapsamı ve beraberinde getirdiği zorluklar açık bir şekilde ifade edilecektir. Daha sonra, araştırmaya yön veren temel hedefler açıklanacak ve bu hedeflere ulaşmak için uygulanan yaklaşım ayrıntılarıyla sunulacaktır. Çalışma kapsamında elde edilen özgün katkılar özetlendikten sonra tezin kurgusu ve akışı sunulacaktır.

1.1. Mülk Güvenliği

Mülkiyet kavramı, en temel anlamıyla, insanlığın Neolitik Dönem'de göçebe yaşam tarzından yerleşik yaşama geçmesiyle ortaya çıkmıştır. Toplumlar tarım yapmaya ve hayvanları evcilleştirmeye başladıkça; toprak ve barınak gibi sabit kaynaklara sahip olma, bunları yönetme ve koruma ihtiyacı doğmuştur. Bu ihtiyaç, belirli bir mekânın sahipliğini ifade eden mülkiyet kavramının doğmasına neden olmuştur. Zamanla mülkiyet, ekonomik istikrarın ve toplumsal gelişimin merkezinde yer alan bir unsur haline gelmiştir. Hukuki ve kültürel sistemler de mülkiyetin düzenlenmesi etrafında şekillenmiş; böylece mülkiyet yalnızca maddi bir mesele olmaktan çıkıp medeniyetin temel taşlarından biri hâline gelmiştir. Günümüzde mülkiyet, bireysel servetin, güvenliğin ve toplumsal yapının temel birimi olarak kabul edilmektedir (Bowles ve Choi, 2019).

Günümüzde mülkiyet çeşitli kategorilere ayrılmakta ve her biri toplumda özgün bir rol üstlenmektedir. Konut mülkiyeti; evler ve apartmanlar gibi yapılar aracılığıyla temel barınma ihtiyacını karşılar ve çoğu aile için finansla ilgili en önemli varlığı temsil eder. Ofisler ve mağazalar gibi ticari mülkler, ekonomik faaliyetlerin ve alışverişin sürdürülebilirliğini sağlar. Sanayi mülkleri (örneğin fabrikalar, depolar, lojistik merkezleri) üretim ve dağıtımı mümkün kılarken; okullar ve hastaneler gibi kurumsal yapılar, temel kamu hizmetlerinin sunulmasında rol oynar. Tarım arazileri ise gıda üretimi ve doğal kaynak yönetimi açısından önem taşımaktadır. Bu mülkiyet türleri birlikte değerlendirildiğinde; yaşanılan meskenlerin, şehirlerin, iktisadi sistemlerin,

kamu altyapısının, kısaca tüm içtimai hayatın temelini oluşturmaktadır (S. J. Ball, 2022).

Günümüz uluslararası gayrimenkul piyasası, mülkiyetin ne denli büyük bir iktisadi değere sahip olduğunu açıkça göstermektedir. Gayrimenkul, günümüzde dünyanın en büyük varlık sınıfı konumundadır; toplam değeri 2015 yılında yaklaşık 217 trilyon dolardan, 2022 yılında 380 trilyon dolara yükselmiş ve bu miktar, uluslararası GSYİH'nin neredeyse dört katına ulaşmıştır (Savills World Research, 2016; Savills Research, 2022). Bu değerin yaklaşık %75'ini konut mülkleri oluştururken, geri kalanı ticari ve zirai varlıklardan meydana gelmektedir. Böylesine büyük bir servet birikimi, mülkiyetin hem ferdi refah hem de içtimai refah açısından merkezi rolünü vurgulamaktadır. Dolayısıyla, mülkiyete yönelik fiziki veya dijital her türlü tehdit, ciddi iktisadi ve içtimai sonuçlara yol açabileceğinden, mülkiyetin korunması ve etkin yönetimi büyük önem taşımaktadır (Fennelly, 2017).

Mülki varlıklarının korunması, sahip oldukları yüksek içtmati ve iktisadi değer nedeniyle kritik bir öneme sahiptir. Mülkiyeti tehdit eden unsurlar genel olarak iki ana grupta toplanabilir: hasar ve izinsiz giriş. Hasar; doğal afetler, kazalar veya kasıtlı eylemler sonucunda oluşabilir ve bir yapının bütünlüğünü ve iktisadi değerini ciddi şekilde zedeleyebilir. Deprem, sel ve yangın gibi olaylar, bu tür zararın başlıca örnekleridir. Kasten yapılan tahribat (örneğin vandalizm) da gayrimenkulün değerini ve işlevini olumsuz etkileyerek dayanıklı tasarım ve koruyucu önlemlerin gerekliliğini pekiştirmektedir (S. J. Ball, 2022).

İzinsiz giriş ise mülk sınırlarına bilinçli ve yetkisiz şekilde girilmesi anlamına gelir ve büyük ölçüde önlenebilir bir tehdit türüdür. Bu tür girişimler; hırsızlık, kasıtlı zarar verme veya mülk sakinlerini tehlikeye atma gibi suçlara zemin hazırlayabilir. Hatta boş alanlar bile işgal ya da malzeme hırsızlığı gibi risklere açıktır. Genellikle daha ciddi suçların başlangıç noktası olan izinsiz girişleri önlemek amacıyla, mülk sahipleri birden fazla katmandan oluşan güvenlik sistemleri (örneğin kilitler, alarmlar, çitler, güvenlik devriyeleri ve kamera sistemleri) kurarak olası zararları engellemeye çalışmaktadır (Fennelly, 2017).

Son yıllardaki teknolojik gelişmeler mülk yönetimini kökten değiştirmiş; ancak bu değişim beraberinde yeni güvenlik açıklarını da getirmiştir. Akıllı kilitler, kameralar, termostatlar ve sensörler gibi Nesnelerin İnterneti (IoT) olarak adlandırılan cihazlar,

kullanıcıya anlık kontrol kabiliyeti ve kolaylığı sağlasa da, çoğu zaman varsayılan zayıf şifrelerle, güncellenmemiş yazılımlarla ya da yetersiz şifreleme teknikleriyle piyasaya sunulmaktadır. Avrupa Siber Güvenlik Ajansı (ENISA) tarafından yayımlanan rapora göre, bilinen güvenlik açığına sahip IoT cihazlarının oranı 2023 yılında %14 iken, bu oran 2024'te %33'e yükselmiştir. Bu durum, akıllı kilit ve alarm sistemlerinin yetkisiz kişlerce siber yollarla ele geçirilmesine ve mülke fiziki olarak izinsiz girişe kapı aralamasına neden olabilmektedir.

İzinsiz girişin önlenmesi, mülk güvenliğinin ilk savunma hattını oluşturur; zira yetkisiz erişim, hırsızlık, vandalizm ve hatta şiddet gibi daha ağır tehlikelere zemin hazırlayabilir. Doğal afetler ya da kazalarla karşılaştırıldığında, izinsiz girişler kasıtlı olarak zayıf noktaları hedef almakta ve bu yöndeki teknikler mevcut güvenlik önlemlerine karşı sürekli güçlendirilmektedir. Bu nedenle güvenlik sistemlerinin de sürekli olarak gelişmesi, caydırma, tespit ve geciktirme işlevlerini yerine getirecek şekilde tasarlanması gerekmektedir (Fennelly, 2017). Bu kapsamda; kilitler, çitler, alarmlar, güvenlik görevlileri ve özellikle kamera sistemleri, modern mülk güvenliğinde standart koruma uygulamaları arasında yer almaktadır.

1.2. Problem Tanımı

Geleneksel güvenlik kameralarının istismar edebilebilecek ciddi zayıflıkları bulunmaktadır. İlk olarak, güvenlik kameraları sabit konumlandırıldıkları ve çevreden kolayca fark edilebildikleri için, tehdit unsuru tarafından açık hedef halindedirler. Ortamı önceden dikkatlice inceleyen biri, her bir kameranın görüş açısını ayrıntılı şekilde haritalandırarak kör noktaları belirleyebilir ve bu alanlara yönelerek izlenmeden hareket edebilir. Bu tür kör noktalar, kameralarla izlenemeyeceği için izinsiz girişlere sebebiyet verir. Ayrıca, kameraların sabit yapısı farkedilmesine ve tehdit unsurlarının kimliklerini maske takmak ya da görünüşlerini değiştirmek gibi basit yöntemlerle gizlemelerine imkân tanır; böylece kaydedilen görüntülerin delil niteliği azalır (Shams ve ark, 2025). Bu öngörülebilirlik ve görünürlük, sabit kamera sistemlerinin caydırıcılığını ve delil sağlama potansiyelini ciddi şekilde sınırlar.

İkinci olarak, güvenlik kameraları optik körleştirme (blinding) saldırılarına karşı oldukça savunmasızdır ve bu saldırılar genellikle basit ama etkili yöntemlerle gerçekleştirilebilir. Örneğin, düşük maliyetli ve kolayca temin edilebilen kızılötesi infrared (IR - kızılötesi) light emmiting diode (LED)'ler, insan gözüyle görülmeyen

ancak kameralar tarafından algılanabilen ışık yayarak kameranın sensörünü aşırı derecede ışığa maruz bırakabilir. Aynı şekilde, kameranın lensine doğrudan yöneltilen yüksek güçlü bir el feneri, sensörü geçici olarak kör edebilir ve kaydedilern görüntülerin kullanılmaz hâle gelmesine yol açabilir. Daha karmaşık saldırılar ise lazer işaretçilerin hassas şekilde sensöre yöneltilmesiyle gerçekleştirilebilir; bu da sensörün doygunluğa ulaşmasına ya da kalıcı hasar görmesine neden olabilir. Bu tür optik körleştirme teknikleri çoğu zaman gizlice uygulanabilir ve izleme yapan personel veya mülk sahipleri, olay gerçekleşene kadar kameranın etkisiz hâle geldiğinden habersiz olabilir (Costin, 2016). Bu tür zafiyetler, yalnızca kamera tabanlı gözetim sistemlerine güvenmenin ne derece riskli olduğunu açıkça göstermektedir. Ayrıca bu sorun, otonom araç teknolojilerinde de karşılaşılan en büyük güvenlik risklerinden biri olarak değerlendirilmektedir (Dh ve Ansari, 2020).

Üçüncü olarak, IR kameralar, düşük ışıklı ya da tamamen karanlık ortamlarda da görüntü sağlayabildikleri için bu soruna bir çözüm sunsa da, bu tür kameraların yüksek maliyetleri, yaygın kullanımını ciddi şekilde kısıtlamaktadır (Dhve Ansari, 2020). Bu iktisadi engel, kullanıcıları ya standart kameraların kısıtlarıyla yetinmeye ya da IR teknolojisinin yüksek maliyetine katlanmaya zorlamaktadır.

Son olarak, kamera sistemleri içerdiği teknolojik bileşenler nedeniyle ev güvenliği açısından görece yüksek maliyetli unsurlardır. Tüm bu zayıflıklar göz önüne alındığında, akıllı ev çağında sadece kameralara dayanarak izinsiz girişlerin önlenmesini sağlamak mümkün değildir.

1.3. Tezin Amacı

Çoğu izinsiz giriş algılama sistemi kamera ve görüntü işleme teknolojilerine dayanırken, insan adım sesinin kendine özgü akustik imzası da aynı derecede güçlü ancak çoğu zaman göz ardı edilen bir güvenlik göstergesidir. Mikrofonlar çevredeki sesleri her yönden algılayabildikleri için, kapalı devre kamera sistemlerinin özgü dar (ve bazen sabit) görüş açılarına kıyasla 360 derecelik kapsama alanı sunarlar.

Karmaşık bir işitsel ortam içinde ve belirli bir zaman aralığında adım sesi gibi belirli ses olaylarını izole edebilme ve tanıyabilme yeteneği, potansiyel izinsiz girişleri tespit etmek için güçlü bir araç sağlar. Örneğin, bir binada genellikle kimsenin bulunmadığı saatlerde ya da girişin yasak olduğu bir alanda bir adım sesi tespit edilirse, bu durum yetkisiz bir girişin güçlü bir göstergesi olabilir. Tespit edilen adım seslerinin

zamanlaması ve özellikleri analiz edilerek, hareket desenleri çıkarılabilir ve güvenlik personelini harekete geçirecek uyarılar oluşturulabilir. Bu şekilde, belirli akustik imzalara odaklanan akıllı bir güvenlik katmanı sağlanmış olur. Ayrıca, gizli mikrofonların veya yer kaplamayan ses sensörlerinin kullanımı bu bağlamda önemli avantajlar sunar. Bu ses sensörleri ortamda göze çarpmayacak şekilde yerleştirilebildiğinden, duvarlara monte edilen ve açıkça görülebilen kameraların aksine daha az dikkat çeker. Bu görünmezlik, onları daha önce belirtilen karşı önlemlere, örneğin LED veya lazer ile optik körleştirme, kör noktaların kullanılması, karşı büyük ölçüde dirençli hâle getirir. Kameraların aksine, bu ses sensörleri varlıklarını belli etmeden sürekli çevresel sesleri izleyebilir ve potansiyel davetsiz misafirlere tespit alanlarını belli etmez. Bu gizlilik, zorla giriş, hareket ve hatta fısıltı hâlindeki konuşmalar gibi kritik işitsel delillerin yakalanmasını kolaylaştırarak daha kapsamlı ve dayanıklı bir güvenlik çözümü sunar. Ayrıca, adım sesi tespiti için gereken donanım ve işlem gücünün görece düşük olması, bu sistemlerin çok kameralı ağlara kıyasla çok daha düşük satın alma, kurulum ve bakım maliyetleriyle uygulanabilmesini sağlar. Bu da özellikle bütçe dostu ev güvenlik sistemleri açısından cazip bir avantajdır.

Bu tezde, ses olayı tespiti bağlamında, mülk güvenliği amacıyla özgün bir Evrişimli Sinir Ağı (CNN) modelinin kullanımına odaklanılmıştır. Çalışma, adım sesi olaylarının tespitine odaklanarak, CNN tabanlı yaklaşımların gürültülü ve zorlu ortamlarda dahi ölçeklenebilirlik ve dayanıklılık sunduğunu göstermektedir. Ayrıca, MFCC kullanılarak sesin yapay sinir ağlarına uygun biçimde temsil edilmesi ve bu temsilin iyileştirilmesi, veri kümelerinin sınıf dengesi ve kalite bakımından geliştirilmesi gibi çalışmalar yapılmıştır. Modelin değerlendirilmesi için katmanlı çapraz doğrulama gibi yöntemler de uygulanarak gerçekçi tespit başarımı hedeflenmiştir.

Bu tez kapsamında cevap aranan temel araştırma soruları şu şekildedir: CNN mimarileri, gerçekçi iç mekân ortamlarında adım sesi olaylarını ne ölçüde doğru ve güvenilir bir şekilde tespit edebilir? İkili sınıflandırma amaçlı ses olay tespiti görevlerinde hangi mimari yapılandırmalar ve ses öznitelik çıkarım stratejileri en yüksek başarıyı sağlamaktadır? Son olarak, veri kümesi üzerinde yapılan iyileştirme ve çeşitlendirme işlemleri, sınırlı kaynaklara sahip ve gürültülü ses ortamlarında modelin genelleme yeteneğini arttırabilir mi?

Bu çalışma, ses tabanlı güvenlik çözümlerinin, geleneksel kamera sistemlerine kıyasla düşük maliyetli ve ölçeklenebilir bir alternatif olarak uygulanabilirliğini deneysel verilerle desteklemektedir. Geliştirilen CNN tabanlı derin öğrenme modeli, 17 farklı iç mekân ses kategorisi arasında adım sesi olaylarını %98 doğruluk ve %1 hata oranıyla başarıyla sınıflandırmıştır. Modelin dayanıklılığı, 5 katlı ve 10 tekrarlı Katmanlı K-Fold Doğrulama yöntemiyle test edilmiş ve ortalama F1 puanı 0.960 olarak elde edilmiştir. Farklı ve açık erişimli veri kümelerinin entegre edilmesi, yöntemin hem şeffaflık ve tekrarlanabilirlik yönünden güçlenmesini hem de gerçek dünyadaki akustik değişkenliğe uyum sağlayabilmesini sağlamıştır. Bu bulgular, ses olay tespit sistemlerinin geleneksel görsel gözetim teknolojilerini tamamlayıcı veya onların yerini alabilecek, güvenilir ve dikkat çekmeyen bir güvenlik katmanı olarak kullanılabileceğini ortaya koymaktadır.

1.4. Tezin Organizasyonu

Bu tezin geri kalan bölümleri şu şekilde yapılandırılmıştır: Bölüm 2, ses olay tespiti ve bu alanın güvenlik bağlamındaki uygulamaları üzerine kapsamlı bir literatür taraması sunmaktadır. Bölüm 3, tanımlar, karşılaşılan zorluklar, ön işleme teknikleri ve derin öğrenmenin ses olayları tespiti üzerindeki rolü gibi gerekli arka plan bilgilerini içermektedir. Bölüm 4'te ise, adım sesi olaylarının tespiti için önerilen yaklaşım ayrıntılı olarak ele alınmakta; veri kümesi hazırlığı, model geliştirme süreci ve iki aşamalı değerlendirme sonuçları açıklanmaktadır. Tezin son bölümü olan Bölüm 5'te ise bulgular özetlenmekte, sonuçların olası etkileri tartışılmakta ve gelecekteki araştırmalar için öneriler sunulmaktadır.

2. LİTERATÜR TARAMASI

Yapay zekâ (YZ) ve IoT teknolojileri ev güvenlik sistemlerine giderek daha fazla entegre edilse de, güncel araştırmalar kamera merkezli güvenlik sistemlerinin bazı önemli kısıtlarına dikkat çekmektedir. İlk olarak, bu sistemlerde kapsama alanı doğası gereği sınırlıdır: sabit kameralar sınırlı bir görüş açısına sahiptir ve bu durum, mülkün bazı bölümlerinin gözetim dışında kalmasına yol açan kör noktalara neden olur. En gelişmiş pan-tilt-zoom (döner-eğilir-zoom yapabilen) kameralar bile tüm yönleri aynı anda izleyemez; bu da davetsiz kişilerin kullanabileceği kaçınılmaz kör bölgeler yaratır (Mottakin ve ark, 2024). Özetle tek bir sabit kamerada genellikle kör noktalar bulunur ve kötü yerleştirme veya uygun olmayan açılar, kritik olayların tespit edilememesine yol açabilir. İkinci olarak, kameranın görüş açısı ve konumlandırılması etkinliğini doğrudan etkiler: Çok yüksekten ya da elverişsiz açılardan yapılan çekimlerde yüzler ya da önemli ayrıntılar görülemeyebilir ve bu da saldırganların tanınmasını zorlaştırır. Bu sınırlama, yalnızca belirli bakış açılarından bireylerin güvenli şekilde tespit edilebildiği akıllı izleme sistemlerinde belgelenmiştir. Üçüncü olarak, kameralar kasıtlı sabotaj ya da atlatma taktiklerine karşı savunmasızdır. Bir davetsiz misafir kör bir noktadan yaklaşabilir ya da cihaza fiziksel olarak müdahale ederek (örneğin, üzerini örtmek veya yönünü değiştirmek) kamera tespiti gerçekleşmeden sistemi devre dışı bırakabilir. Daha gelişmiş saldırılarda ise doğrudan temasa geçilmeden kameraları etkisiz hâle getirebilir; örneğin araştırmalar, bir kameranın sensörüne lazer ışığı tutulmasıyla kameranın görüntüleme işlevinin bozulabileceğini veya görüntülerin zarar görebileceğini göstermiştir (Costin, 2016). Bu bulgular, IoT destekli akıllı ev güvenlik sistemleri gelişmiş hâle gelse dahi, kamera bileşenlerinin kısıtlarının olduğunu ve temel zafiyetler taşıdığını ortaya koymaktadır. Bu zayıflıkların giderilmesi, daha sağlam bir akıllı ev güvenliği için kamera yerleşiminin optimize edilmesi, çoklu sensör birleşimi ve sensör saldırılarına karşı önlemler gibi yöntemlerle mümkün olabilir.

Ses olayları tespiti teknikleri, ses sinyalinin analizi sınucunda elde edilen bulgular kullanarak anormal veya izinsiz aktivitelerin belirlenmesi amacıyla güvenlik sistemlerinde kullanılmaktadır. Örneğin, Zieger ve ark. (2009), iç mekânda herhangi

bir ses olayını izleyerek sesin türüne bağlı olmaksızın davetsiz bir kişinin varlığını tespit eden ses tabanlı bir gözetim sistemi geliştirmiştir. Bu yaklaşım, iç mekânda yankı ve yansıma gibi zorluklara rağmen, çok sayıda dağınık yerleştirilmiş mikrofon kullanarak sesin konumunu ve şiddetini tahmin etmiştir. Ancak bu çalışma, belirli ses sınıflarını (örneğin adım sesleri) ayırt etmeye odaklanmamıştır. Daha sonraki araştırmalar, izinsiz giriş tespiti için basit akustik eşik yöntemlerini de incelemiştir. Örneğin Al-Khalli ve ark. (2023), ortam gürültüsü belirli bir eşiği aştığında alarm tetikleyen IoT tabanlı bir sistem önermiştir. Ancak yalnızca eşik değerine dayalı karar mekanizmaları, zararsız seslerin de bu eşiği geçmesi durumunda yanlış alarmlara yol açabilmektedir. Al-Khalli ve ekibi, bu tür ani dalgalanmaları yumuşatmak için kısa süreli ortalama / uzun süreli ortalama filtreleme tekniği uygulayarak bu sorunu hafifletmiş olsa da, yaklaşım yine de basit karar kriterlerinin sınırlı etkisini göstermektedir. Bu çalışmalar, güvenlik bağlamında daha sağlam ses olayları tespiti yöntemlerine olan ihtiyacı ortaya koymakta ve özellikle adım sesleri gibi belirli ses olaylarını sınıflandırmaya yönelik çalışmaları teşvik etmektedir.

İnsan konuşması, en çok çalışılan ses olaylarından biri olup, konuşma tanıma alanındaki gelişmeler genel olarak ses olayları tespiti araştırmalarına da önemli katkılar sağlamıştır. Özellikle derin öğrenme, son yıllarda otomatik konuşma ve konuşmacı tanıma performansını önemli ölçüde artırmıştır. Hourri ve Kharroubi (2020), metinden bağımsız konuşmacı tanıma amacıyla geleneksel spektral özellikleri (MFCC) daha ayrımcı derin özniteliklere dönüştüren yeni bir derin sinir ağı yaklaşımı geliştirmiştir. Bu derin sinir ağları (Deep Neural Network – DNN) tabanlı sistem, o dönemdeki ileri düzey i-vektör tabanlı yöntemlerle rekabet edebilir doğruluk elde etmiş ve öğrenilmiş özniteliklerin ses sınıflandırma görevlerindeki potansiyelini göstermiştir. Daha sonraki çalışmalarda, konuşmanın zamansal yapısını yakalayabilmek için evrişimsel ve tekrarlayan mimariler birleştirilmiştir. Mehrish ve ark. (2023), derin öğrenmenin konuşma işleme alanında nasıl bir devrim yarattığını kapsamlı biçimde incelemiştir. Bu inceleme, MFCC ve Gizli Markov Modelleri (HMM) gibi geleneksel yöntemlerden başlayarak Transformer, RNN, CNN ve difüzyon modelleri gibi modern yaklaşımlara uzanan gelişim sürecini ortaya koymaktadır. Graves ve ark. (2013) ise son dönem derin öğrenme tekniklerini ele alarak, akustik ve dil modellemesini birlikte optimize eden uçtan uca konuşma tanıma modellerine dikkat çekmiştir. Bu tür modeller, doğruluk oranlarında artış ve gürültüye

karşı dayanıklılık sağlayarak alanda önemli ilerlemelere aracı olmuştur. Bu çalışmalarla elde edilen bilgi birikimi (örneğin öznitelik öğrenimi, ağ mimarisi) adım sesi tespiti gibi diğer ses olaylarının tespiti uygulamalarına da aktarılabilecek niteliktedir.

Adım sesi olay tespiti, güvenlik ve ortam algılama uygulamaları kapsamında giderek artan bir ilgi görmektedir. İlk yaklaşımlar, adım sesi tespitini el ile çıkarılmış öznitelikler ve klasik makine öğrenmesi algoritmaları kullanarak adım sesi / arka plan gürültüsü şeklinde ikili sınıflandırma problemi olarak ele almıştır. Itai ve Yasukawa (2008), iç mekân adım seslerini algılamak amacıyla konuşma işleme tekniklerinden yararlanarak, LPCC (Linear Predictive Coding Cepstrum) özniteliklerini ve dinamik programlama yöntemini kullanmıştır. Yaklaşık 10 kişilik küçük ölçekli bir deneyde, %98 adım sesi tespit oranına ulaşılmış; ancak bu çalışma, sınırlı veri ve F-skoru gibi standart değerlendirme metriklerinin kullanılmaması nedeniyle değerlendirme için yetersiz kabul edilmektedir. Cai ve ark. (2010) da benzer şekilde, adım sesi tanımını bir makine öğrenmesi problemi olarak ele almıştır. Kendi veri kümelerini oluşturarak adım sesi (pozitif) ve adım dışı (negatif) örnekler etiketlenmiş; ardından MFCC öznitelikleri kullanılarak Weka aracı ile sınıflandırıcı eğitilmiştir. Verideki sınıf dengesizliğine rağmen (%18 adım sesi), model %90'ın üzerinde doğruluk ve hatırlama değerlerine ulaşmıştır. Nakadai ve ark. (2013) ise, adım sesi benzeri sesleri sınıflandırmak için geometrik ses özniteliklerini kullanmıştır. Dört sınıftan oluşan (yürüme, koşma, el çırpma, konuşma) kontrollü bir veri kümesiyle SVM sınıflandırıcı eğitilmiş; temiz veride %99'un üzerinde başarı sağlanmış, ancak gürültülü ortamlarda performans %56'nın altına düşmüştür. Bu durum, gürültülü ortamlarda akustik tespit için daha dayanıklı modellere ihtiyaç olduğunu göstermektedir.

Son yıllarda adım sesi tespitine yönelik çalışmalar, derin öğrenme ve dizisel modellemeye yönelerek daha dayanıklı sonuçlar elde etmeye odaklanmıştır. Summoogum ve Palaniappan (2021), yaşlı bireylerin sağlık takibi bağlamında ev ortamında adım sesi tespiti mekanizmasına sahip Long Short-Term Memory (LSTM) temelli bir DNN modeli önermiştir. Modeli eğitmek için 2.494 adım sesi ve 2.510 adım dışı ses içeren sentetik bir veri kümesi oluşturulmuş; bu küme gerçek hayatı yansıtacak şekilde çakışan iç mekân seslerini de içermektedir. Değerlendirme için zamanla değişen hassasiyet ve duyarlılığı ölçen Polyphonic Sound Detection Score (PSDS) metriği kullanılmış ve model, önceki çalışmaların 0.38–0.41 aralığındaki skorlarını

geride bırakarak 0.65 PSDS elde etmiştir. Liu ve Kong (2021) ise MFCC öznitelikleriyle HMM tabanlı bir yaklaşım önermiş, küçük ölçekli bir veri kümesiyle %100 doğruluk bildirmiştir. Ancak yalnızca 12 test örneğiyle elde edilen bu sonuç metodolojik detay eksikliği nedeniyle güvenilir bulunmamaktadır. Genel olarak, adım ses olaylarının tespiti, klasik sinyal işleme ve makine öğrenmesi yöntemlerinden, zamansal örüntüleri daha iyi yakalayabilen ve gürültülü ortamlarda genellenebilir derin öğrenme modellerine doğru evrilmiştir.

Adım sesi tespiti yalnızca olayın varlığını belirlemekle sınırlı değildir; bazı araştırmalar adım sesinin bireysel kimlik ya da cinsiyet gibi demografik özelliklerin tahmininde kullanılabileceğini göstermektedir. DeLoney (2008), biyometrik tanıma amacıyla adım sesi sinyallerini analiz etmiş ve yürüyüşe özgü akustik örüntülerin bireye dair ayırt edici bilgiler içerdiğini göstermiştir. Daha sonra, Algermissen ve Hörnlein (2021), CNN tabanlı bir model geliştirerek beş farklı bireyin adım seslerinden kimliklerini %94 F1 skoru ile ayırt edebilmiştir. Bu başarı, adım seslerindeki zamanlama ve frekans özelliklerinin ayırt edici gücünü ortaya koymuştur. Ancak sistem bireye özgü eğitildiği ve küçük bir veri kümesiyle test edildiği için ölçeklenebilirlik sınırlıdır. Bu tür sistemlerin geniş popülasyonlarda uygulanabilmesi için daha büyük ve çeşitli veri kümelerine ihtiyaç duyulmaktadır. Ayrıca, cinsiyet gibi demografik özellikler doğrudan test edilmese de, birey tanıma sistemlerinde dolaylı olarak bu bilgilerin de öğrenilebildiği değerlendirilmektedir. Sonuç olarak, güvenlik amacıyla adım sesi tespiti yalnızca bir davetsiz misafirin varlığını değil, aynı zamanda kim olabileceğine dair ipuçlarını da ortaya çıkarabilir. Bu yetenek, ileri düzey örüntü tanıma teknikleri ile desteklendiğinde güvenlik sistemlerine değerli ek bilgiler sağlayabilir; ancak yüksek doğruluk için geniş çaplı eğitilmiş modeller gereklidir.

3. MATERYAL VE YÖNTEM

3.1. Ses Olaylarının Tespiti (Sound Event Detection - SED)

SED, bir ses akışı içerisinde hangi ses olaylarının ne zaman gerçekleştiğini otomatik olarak tanımlamayı amaçlayan bir süreçtir. Pratikte bir SED sistemi, sürekli bir ses sinyalini işler ve her algılanan ses olayı için sembolik bir çıktı üretir; bu çıktı, olayın etiketini (örneğin "adım sesi", "cam kırılması") ve gerçekleşme zamanını içerir. Daha teknik bir ifadeyle, SED'nin amacı ses olaylarını zaman çizgisinde başlangıç ve bitiş saatlerini tespit etmek ve bunları önceden tanımlanmış sınıflara ayırmaktır. Örneğin insanlar, bir bebeğin ağlamasını ya da adım seslerini işiterek bu seslerin hangi olaya ait olduğunu kolayca çıkarabilirler; SED algoritmaları da bu yetiyi taklit etmeye çalışarak bu tür sesleri doğru zamanda tespit edip uygun etiketlerle sınıflandırmayı hedefler (örneğin "bebek ağlaması").

SED, konuşma veya müzik tanımadan daha kapsamlı olarak, tüm ses olaylarını hedef alır. Bunlar kuş cıvıltısından araba geçişine, insan adımından cam kırılmasına kadar çok geniş bir yelpazeyi kapsar. Çünkü teknik olarak ses çıkaran her nesne veya eylem bir ses olayı olarak değerlendirilebilir. Bu nedenle, SED sistemleri genellikle belirli bir uygulama alanına göre geliştirilir; örneğin şehir sesleri ya da ev içi sesleri gibi, önceden tanımlanmış bir ses olayları ontolojisine dayanır (Heittola, 2013). SED, sinyal işleme ve makine öğrenmesi alanlarında giderek öne çıkan bir araştırma konusudur. İlk yaklaşımlar klasik sinyal işleme yöntemlerine dayanırken, günümüzde modern derin öğrenme teknikleri ön plana çıkmıştır. Örneğin Çakır ve ark. (2015), gerçekçi ses ortamlarında zaman açısından çakışan ses olaylarını tespit etmek üzere çok etiketli (multi-label) sinir ağı tabanlı derin öğrenme modellerinden birini geliştirmiştir. Son yıllarda geliştirilen ileri düzey SED sistemlerinin büyük çoğunluğu, yüksek doğruluk elde etmek amacıyla evrişimsel (CNN) veya tekrarlayan (RNN) sinir ağı mimarilerini kullanmaktadır (Mesaros ve ark, 2021).

SED teknolojisi artık çeşitli gerçek dünya senaryolarında kullanılabilecek olgunluğa erişmiştir. En yaygın uygulama alanlarından biri, akıllı evler ve destekleyici yaşam ortamlarıdır. Bu sistemlerde ses tabanlı izleme, güvenlik ve güvenliği artırmak

amacıyla kullanılır. Modern akıllı ev sistemleri, kullanıcıların ya da yetkili birimlerin haberdar edilmesini sağlayacak şekilde belirli olayları otomatik olarak tespit eder. Örneğin mikrofonlarla donatılmış bir ev, bir camın kırılması ya da kapının zorlanarak açılması gibi bir hırsızlık girişimini veya bir yangın alarmını algılayabilir ve kameralar sabote edilmiş olsa bile ev sahibine anında bildirim gönderebilir (Bonet-Solà ve Alsina-Pagès, 2021). Ayrıca akıllı evler, konfor amaçlı (örneğin kapı zili ya da çamaşır makinesi bitiş sesi) ya da yaşlı bireylerin bakımı için (örneğin düşme ya da yardım çağrısı tespiti) SED'den yararlanmaktadır.

Bir diğer önemli kullanım alanı ise kamuya açık ya da ticari alanlarda ses tabanlı gözetimdir. Bu ortamlarda SED, durum farkındalığı sağlamak amacıyla kullanılır. Gözetim sistemleri; silah sesi, patlama, çığlık ya da saldırgan sesleri algılayarak, sokaklarda veya kampüslerde acil müdahale gerektiren olayların hızlıca fark edilmesini sağlar (Gencoglu ve ark, 2014). Şehir yönetimleri ve kolluk kuvvetleri, silah sesleri veya cam kırılmaları gibi olayları gerçekleştiği anda otomatik olarak tespit edebilen ses izleme ağlarına ilgi göstermektedir. Bu sistemler, adeta otomatik ses "tuzakları" gibi çalışmaktadır (Mesaros ve ark, 2021). Nitekim SED, son yıllarda güvenlik ve izleme sistemlerinde geniş ölçekte benimsenmiştir.

Buna ek olarak SED; akıllı şehir projelerinde (örneğin gürültü kirliliği izlemesi, araç kazalarının sesle tespiti) ve endüstriyel ortamlarda (örneğin makine arızaları ya da alarmların tespiti) da kullanılmaktadır. Özetle, SED'in "olayları duymaya" yönelik yetisi, yalnızca görsel izlemenin yetersiz kaldığı durumlarda değil, aynı zamanda görsel sistemleri tamamlayıcı bir güvenlik ve izleme katmanı gerektiğinde de değerli bir teknoloji olarak öne çıkmaktadır.

3.2. Zoruluklar

SED alanında son yıllarda hızlı ilerlemeler kaydedilmiş olsa da, gerçek yaşam seslerine yönelik dayanıklı sistemler geliştirmek hâlâ birçok zorluk barındırmaktadır. Bu zorlukların büyük bölümü, hem ses verisinin doğasında hem de öğrenme sürecinin karmaşıklığında kaynaklanmaktadır. Başlıca zorluklar aşağıda özetlenmiştir:

- **Sınıf İçi Çeşitlilik (Intra-Class Variability):** Aynı sınıfa ait ses örnekleri, akustik açıdan birbirinden oldukça farklı özellikler taşıyabilir. Örneğin bir "adım sesi", kişinin kilosuna ve yürüme biçimine, ayakkabı türüne veya zemin

malzemesine (çim, parke, çakıl vb.) bağlı olarak değişiklik gösterir. Benzer şekilde, bir "cam kırılma" sesi de camın kalınlığına veya kırılma şekline göre farklılık arz edebilir. Bu çeşitlilik, sınıfı temsil edecek tek bir prototipin öğrenilmesini zorlaştırır. Genel olarak ortam ses olayları durağan değildir; spektral özellikleri zamanla değişir ve belirli bir yapıya sahip değildir. Bu nedenle, modellerin her bir olay türü için çok çeşitli örnekleri genelleyebilecek şekilde eğitilmesi gerekir (Mesaros ve ark., 2021).

- **Çakışan Olaylar (Overlapping Events):** İzole ses kliplerinin aksine, gerçek hayat genellikle aynı anda gerçekleşen birden fazla ses olayını içerir. Bu polifonik yapı, bir ses parçasında örneğin bir yandan adım sesi varken, arka planda müzik ve konuşma seslerinin de bulunabileceği anlamına gelir. SED sistemlerinin, zaman açısından üst üste binen bu olayları ayırt edebilmesi gerekir. Seslerin aynı anda gerçekleşmesi, birbirini maskelemesi nedeniyle tespiti oldukça zorlaştırır. Bu durum akustik çakışma olarak değerlendirilebilir. Bu tür çakışmaları yönetebilmek için, tüm etkin sesleri aynı anda tanımlayabilen çok etiketli sınıflandırma modelleri ve karışık ses kaynaklarını ayrıştırabilen yapılar gereklidir. Erken dönem derin öğrenme çalışmaları, çakışan sesleri çok etiketli sinir ağlarıyla ele almış ve polifonik SED hâlen aktif bir araştırma alanı olmaya devam etmektedir (Mesaros ve ark., 2021).
- **Arka Plan Gürültüsü ve Akustik Koşullar:** Gerçek dünyada kullanılan SED sistemleri, karmaşık arka plan gürültüleriyle baş etmek zorundadır. Trafik uğultusu, rüzgâr, insan kalabalığı ya da elektrikli cihaz sesleri gibi ortam gürültüleri, hedef olayların sinyal-gürültü oranını (SNR) düşürerek tespiti zorlaştırır. Kapalı alanlarda yankı ve yansıma, olay seslerini daha da yayarak bozar. Hedef ses olayı (örneğin bir adım sesi veya kapı gıcırtısı), diğer seslere kıyasla oldukça zayıf olabilir. Bu durum, yanlış tespitlere ve yanlış alarmlara neden olabilir. Gürültüye dayanıklı öznitelik çıkarımı ve ön işleme teknikleri (filtreleme, spektral çıkarım gibi) genellikle bu aşamada gereklidir. Ancak, sessiz bir ofisten yoğun bir caddeye kadar farklı akustik koşullarda güvenilir tespit sağlamak hâlâ önemli bir zorluktur (Bonet-Solà ve Alsina-Pagès, 2021).
- **Etiketleme ve Sınıflandırma Belirsizlikleri:** Ses olaylarının tanımlanması ve etiketlenmesi çoğu zaman belirsizlik içerir. Konuşma gibi iyi tanımlanmış birimlere (örneğin kelime) sahip olmayan ortam sesler, evrensel bir ontolojiye sahip değildir. Aynı ses kaydı farklı uzmanlarca farklı şekilde etiketlenebilir

(örneğin duyulan bir çarpma sesi miydi, silah sesi miydi yoksa patlama mıydı?). Zaman içinde olayların başlangıç ve bitiş noktaları da kolayca belirlenemez; örneğin bir dizi adım sesi tek bir olay mı, yoksa ardışık birden fazla olay mı sayılmalıdır? Bu belirsizlikler eğitim verisini etkiler: Etiketler tutarsız veya örtüşen yapıdaysa, modellerin öğrenme süreci sekteye uğrayabilir. Standardize bir ses olayları taksonomisinin eksikliği, her veri kümesinin kendi sınıflarını tanımlamasına neden olur ve genellemeyi zorlaştırır. Etiketleme tutarsızlıkları ve insan kaynaklı hatalar, etiket gürültüsü (label noise) yaratır. Güçlü SED sistemleri, bu belirsizliklere dayanıklı olmalı ve hatta yalnızca sınıf varlığını belirten ama zaman bilgisi içermeyen zayıf etiketli verilerden öğrenebilecek şekilde tasarlanmalıdır (Mohino-Herranz ve ark., 2022).

- **Bağlam Değişimi ve Genelleme:** SED modelleri genellikle belirli bir ortama veya veri kümesine özel olarak eğitilir ve farklı bir bağlamda kullanıldıklarında performans düşüşü gözlemlenir. Bu durum, yukarıda belirtilen çeşitliliklere ve ontolojik farklılıklara ek olarak, kayıt koşullarındaki farklılıklardan da kaynaklanmaktadır. Örneğin, iç mekân adım sesleriyle eğitilmiş bir model (örneğin ReaLISED verisi), dış mekândaki çakıl taşları üzerindeki adım seslerini başarıyla tespit edemeyebilir. Ya da temiz silah sesleriyle eğitilen bir model, şehir ortamındaki silah seslerini algılamakta zorlanabilir. Bu nedenle her uygulama için özel veri toplama ve model ayarı gerekebilir. Hiçbir öznitelik seti ya da model, tüm akustik bağlamlarda evrensel başarı göstermemiştir. Bu nedenle bağlam uyumu ve transfer öğrenme (transfer learning) gibi yöntemlerle, farklı ortamlarda da doğru çalışabilen uyarlanabilir modeller geliştirmek önemli bir araştırma hedefidir (Mesaros ve ark., 2021).
- **Veri Azlığı ve Etiketleme Maliyeti:** Etkili bir SED sistemi geliştirmek, genellikle zaman damgalı etiketlere sahip geniş hacimli ses verisi gerektirir. Güçlü etiketler (her olayın başlangıç ve bitiş zamanını içeren) detaylı dinleme gerektirdiğinden, eğitimli uzmanlarca yapılmalı ve bu da süreci hem zaman alıcı hem de maliyetli hâle getirir. Bu nedenle pek çok SED veri kümesi ya küçük ölçeklidir ya da yalnızca olayların varlığını belirten zayıf etiketlere dayanır. Etiketli verinin azlığı, yarı denetimli öğrenme (semi-supervised learning) ve çoklu örnek öğrenme (multiple instance learning) gibi tekniklerin önemini artırmaktadır. Diğer yandan, güvenlikle ilgili önemli olaylar gerçek

hayatta nadir görüldüğünden, bu sınıflar için büyük hacimli veriler toplamak zordur. Bunu çözmek için ses olaylarına ait temiz kayıtları arka plan üzerine bindirerek sentetik veri üretimiyle sınıf dengesizliği giderilmeye çalışılır; ancak sentetik veriler gerçek kayıtlardaki ayrıntıları her zaman yansıtmayabilir. Kısacası, yüksek kaliteli ve çeşitli eğitim verisinin sınırlı olması, güçlü SED modellerinin eğitimi için önemli bir darboğaz oluşturmaktadır. Bu nedenle, daha az etiketli veriyle daha fazla başarı sağlayabilecek yöntemlerin geliştirilmesi büyük önem taşımaktadır (Cai ve ark, 2025).

SED sistemlerinde teknik açıdan önemli bir diğer husus ses kanalı yapılandırmasıdır. Düşük maliyetli sensörlerde yaygın olan mono-kanal ses kayıtları, yalnızca tek bir ses akışını kaydeder ve ortam bilgi içermez. Bu da ses kaynaklarının yerini belirleme kapasitesini sınırlar. Buna karşılık, çok kanallı ses (örneğin stereo ya da mikrofon dizileri), ortam hakkında ipuçları sağlayarak hem tespit doğruluğunu artırabilir hem de ses kaynağının konumlandırılmasını mümkün kılar. Bu sistemlerde, belirli yönlerden gelen sesleri güçlendirmek amacıyla ışın biçimlendirme (beamforming) gibi mekânsal sinyal işleme teknikleri kullanılmaktadır. Ancak çok kanallı sistemler, daha karmaşık donanım ve işlem gücü gerektirdiğinden, maliyet hassasiyeti olan veya büyük ölçekli uygulamalarda tek kanallı çözümler daha uygulanabilir olmaktadır (Griffin ve ark., 2011).

3.3. Ses Sinyal İşleme

Temelde ses, hava gibi bir ortamda akustik dalga şeklinde yayılan fiziksel bir titreşimdir. Bir mikrofon, bu hava basıncı titreşimlerini sürekli zamana ait bir ses temsili olan analog elektrik sinyaline dönüştürür. Bu sesi bir bilgisayarda depolamak veya işlemek için, analog sinyalin örnekleme (sampling) yoluyla dijital forma dönüştürülmesi gerekir. Örnekleme, sürekli sinyalin genliğinin belirli aralıklarla ölçülmesi işlemidir ve bu sayede ayrık zamanlı (discrete-time) bir sinyal elde edilir. Örnekleme frekansı (ya da hızı), saniyede kaç örnek alındığını belirler (örneğin 16 kHz veya 44.1 kHz gibi). Zaman uzayı (dalga formu) temsili, sinyalin zaman ekseninde genlik (ya da basınç) cinsinden ifade edilmesidir. Zaman uzayındaki bu dalga formu, ses olaylarını doğrudan yansıtsa da (örneğin bir adım sesi, genlikte kısa süreli bir yükselme olarak görünür), sesin frekans içeriğini açık şekilde göstermez.

Zaman uzayına ait öznitelikler (örneğin genlik zarfı ya da sıfır geçiş oranı) çoğu zaman karmaşık ses olaylarını ayırt etmekte yetersiz kalır. Bu durum, özellikle birden fazla ses kaynağının üst üste geldiği ya da geçici (transient) olayların çok kısa sürdüğü gerçekçi ses ortamlarında daha belirgindir. Zaman uzayı dalga formu, tüm frekansların karışımı olduğu için oldukça karmaşık yapılar içerebilir. Bu nedenle, adım sesi gibi belirli ses olaylarını analiz edebilmek ve tespit edebilmek için sinyalin genellikle frekans uzayında ya da zaman-frekans uzayında incelenmesi gerekir; bu temsillerde sinyalin spektral içeriği görünür hâle gelir (Cakir, 2019).

Frekans içeriğini incelemek için temel araçlardan biri Fourier dönüşümüdür. Fourier kuramına göre, zaman uzayındaki herhangi bir sinyal, farklı frekanslara, genliklere ve fazlara sahip sinüsoidal bileşenlerin toplamı olarak temsil edilebilir. Ayrık Fourier Dönüşümü (DFT), ayrık zamanlı bir sinyali frekans bileşenlerine matematiksel olarak ayırmak için kullanılır. Pratikte ise bu işlemi verimli şekilde gerçekleştirmek için Hızlı Fourier Dönüşümü (FFT) algoritması tercih edilir. DFT işleminin çıktısı, sinyalin frekans alanındaki temsili olup, sinyal enerjisinin frekans bantları boyunca dağılımını gösterir.

Ancak klasik DFT'nin önemli bir sınırlaması vardır: bu dönüşüm sabit uzunlukta bir zaman penceresi üzerinde uygulanır ve hangi frekans bileşeninin sinyalin hangi zaman aralığında gerçekleştiği bilgisi elde edilemez. Bu durum, durağan sinyaller için kabul edilebilir olsa da, ortam sesleri gibi zamanla değişen (non-stationary) sinyaller için ciddi bir sorundur. Örneğin bir adım sesi geçici bir olaydır; enerjisi kısa bir süreye yoğunlaşmıştır ve üretildiği anda frekans içeriği hızla değişebilir. Uzun bir ses kaydında yalnızca tek bir DFT hesaplandığında, adım sesinin zaman bilgisi global spektrum içinde kaybolur.

Bu zaman bilgisinin kaybını önlemek için Kısa Zamanlı Fourier Dönüşümü (Short-Time Fourier Transform – STFT) kullanılır. STFT, sinyalin kısa ve genellikle üst üste binen parçalara bölünmesi ve her bir çerçeve (frame) için ayrı ayrı Fourier dönüşümünün uygulanmasıdır (Algermissen ve Hörnlein, 2021). STFT'de ses sinyali x[n], uzunluğu N olan çerçevelere ayrılır (genellikle 20–40 ms), her çerçeve arasında H örnek adım (örneğin %50 örtüşme için 10–20 ms) bırakılır. Her çerçeveye, spektral sızıntıyı azaltmak için Hamming penceresi gibi bir pencereleme fonksiyonu w[n] uygulanır. Burada m, zaman çerçevesini; k, frekans bandını temsil eder. X[m, k] ise, m-inci zaman çerçevesinde k-inci frekans bileşenine ait karmaşık spektrumu gösterir.

Pencerenin zaman ekseni boyunca kaydırılmasıyla, ardışık çerçeveler için X[m, k] hesaplanır ve böylece sesin frekans içeriğinin zamanla nasıl değiştiği elde edilir. Bu temsil, ses olaylarının zaman-frekans düzleminde izlenmesini mümkün kılar.

$$X[\mathrm{m},\mathrm{k}] = \sum_{n=0}^{N-1} x[mH+n]w[n]e^{-j\frac{2\pi}{N}kn} \tag{3.1}$$

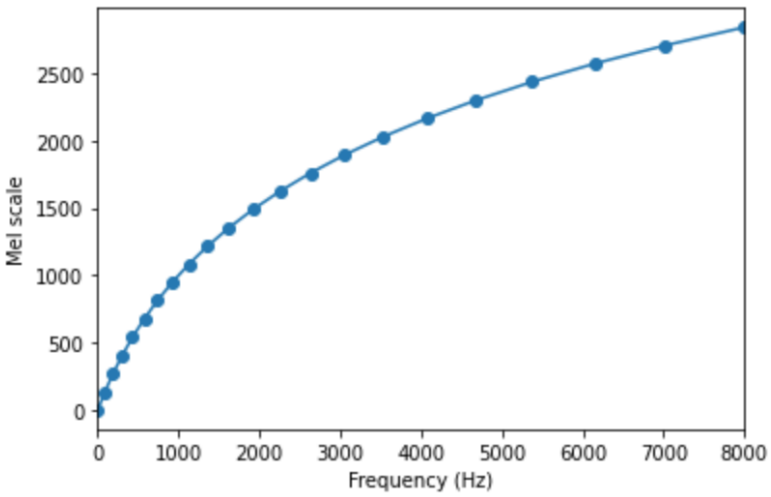
STFT (Kısa Zamanlı Fourier Dönüşümü) yönteminde, zaman ve frekans çözünürlüğü arasında dikkat çekici bir ödünleşim (trade-off) vardır ve bu durum doğrudan seçilen çerçeve (frame) uzunluğundan etkilenir. Kısa çerçeveler yüksek zaman çözünürlüğü sağlar ancak frekans çözünürlüğü düşer; uzun çerçeveler ise daha hassas frekans çözünürlüğü sunar fakat hızlı zaman değişimlerini bulanıklaştırır. Ses olaylarının tespiti için, genellikle olayların süresine uygun bir çerçeve boyutu tercih edilir. Örneğin adım sesleri milisaniyeler mertebesinde çok kısa süreli olaylar olduğundan, yüksek zaman çözünürlüğü tercih edilir. Algermissen ve Hörnlein (2021) adım sesi tespitine yönelik yaptığı bir çalışmada, 4096 örnek uzunluğunda bir FFT penceresi ve çok küçük bir adım boyutu (hop size) kullanılarak yaklaşık 2.5 milisaniyelik bir zaman çözünürlüğü elde edilmiştir. Bu tercih, 5–10 ms süren adım geçişlerinin doğru şekilde çözümlenmesini sağlamak amacıyla yapılmıştır.

STFT ile elde edilen doğrusal frekans spektrumu bilgilendirici olsa da, bu spektrumdan daha kompakt ve insan algısıyla daha uyumlu öznitelikler elde etmek için ek işlemler ya da dönüşümler yapılabilir. Ham spektrogram verisi genellikle yüksek boyutludur ve sınıflandırma görevleri için gereksiz veya yinelenen bilgiler içerebilir. Bu nedenle SED alanında yaygın olarak kullanılan iki temsil biçimi, mel spektrogramı ve MFCC'dir. Her ikisi de mel frekans ölçeğine dayanır (Cakir, 2019).

Mel ölçeği, insan işitme sistemini taklit etmeyi amaçlar. İnsanlar düşük frekanslı sesleri yüksek frekanslı seslere göre daha kolay ayırt edebilir. Mel ölçeğine dayalı filtreler, sinyal bileşenlerini doğrudan güçlendirmez ya da zayıflatmaz; bunun yerine, frekans spektrumunun insan algısı açısından en önemli bileşenlerini vurgulayan ağırlıklı bir temsil sunar. Kısaca mel ölçeği, iki ses arasındaki perde (pitch) farkını insan kulağının algılayış biçimine daha yakın şekilde ölçen bir sistemdir.

Mel ölçekli filtreler Şekil 3.1'de gösterilmiştir. Bu filtrelerin kapsadığı frekans aralıkları kullanılarak filtre bankaları tasarlanır. Bu filtre bankaları, ses sinyalinin

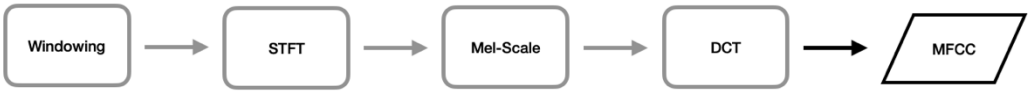
spektral içeriğini belirli frekans bantlarında çıkarmak için kullanılan, üst üste binen bant geçiren filtrelerden oluşur (Heller ve ark., 1999).



Şekil 3.1. Mel-scale.

MFCC'ler ilk olarak konuşma tanıma alanında Davis ve Mermelstein (1980) tarafından önerilmiş olup, sesin tını (timbral) dokusunu verimli ve etkili biçimde temsil edebilme özellikleri sayesinde ses analizi uygulamalarında yaygın olarak kullanılmaktadır. MFCC sonuçları yorumlanırken, her bir katsayının sinyalin genel temsilindeki katkısını dikkate almak önemlidir. Genellikle ilk birkaç katsayı sinyalin toplam enerjisini ve spektral eğimini (spectral slope) yansıtırken, daha yüksek sıradaki katsayılar sesin daha ayrıntılı spektral özelliklerini temsil eder.

Bir ses sinyalinden MFCC'lerin hesaplanması aşağıdaki adımları içerir: pencereleme (windowing), kısa zamanlı Fourier dönüşümü (STFT), mel ölçeğine dönüştürme ve son olarak doğrusal kosinüs dönüşümü (DCT). Bu süreç Şekil 3.2'de görselleştirilmiştir.

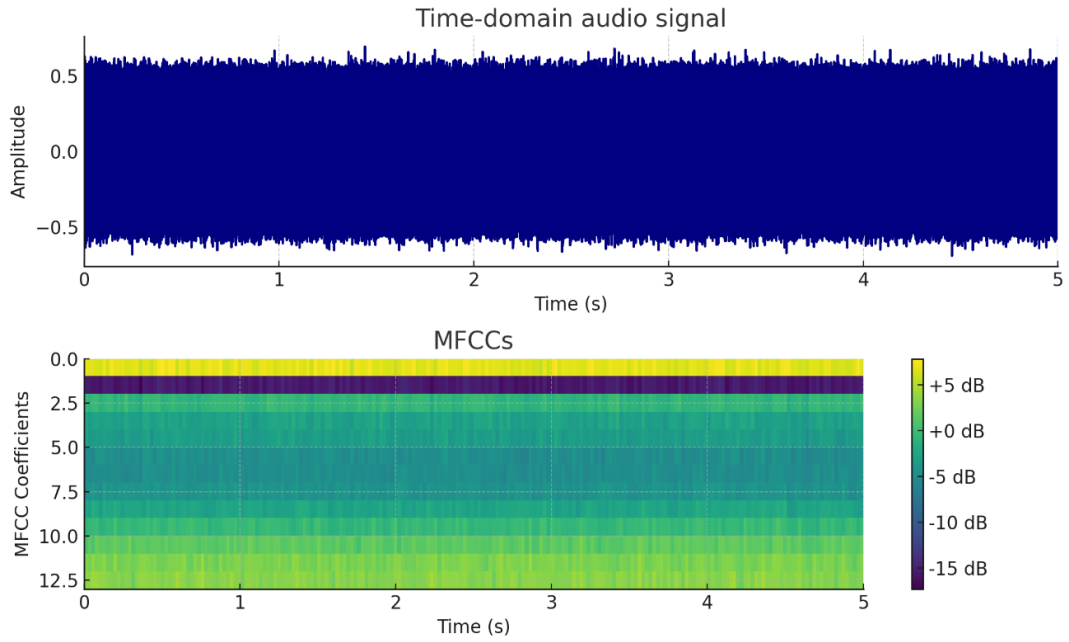


Şekil 3.2. MFCC özniteliği çıkartma adımları.

Her bir kareye bir pencereleme fonksiyonu (örneğin Hamming penceresi) uygulanır. Pencereleme işlemi, kare kenarlarını yumuşatarak ileride gerçekleştirilecek Fourier analizinde spektral sızıntıyı azaltmayı amaçlar. Ardından, STFT kullanılarak pencere uygulanmış karenin genlik spektrumu elde edilir. Bu spektrum, daha sonra mel filtre bankasından geçirilir. Frekans ekseni üzerinde, mel ölçeğine göre aralıklı

yerleştirilmiş üçgensel bant geçiren filtrelerden oluşan bir filtre kümesi tanımlanır. Her bir filtrenin çıktısı, ilgili mel bandındaki spektral enerjinin toplamı olarak elde edilir. Diğer bir ifadeyle, sonuç olarak bir dizi mel bandı enerjisi elde edilir. Bu enerjiler logaritmik dönüşüme tabi tutulduktan sonra, elde edilen log-mel vektörü DCT ile dönüştürülür. DCT, ayrık frekans üzerinde Fourier benzeri bir dönüşümdür ve log-mel spektrumunu kepstral uzayda yer alan bir dizi kepstral katsayıya dönüştürür. Elde edilen bu katsayılar, MFCC olarak adlandırılır (Algermissen ve Hörnlein, 2021).

Ardışık önişlem adımlarının ardından, zamana bağlı ham dalga formu (Şekil 3.3'ün üst grafiğinde gösterilmiştir), algısal temelli, düşük boyutlu ve büyük ölçüde dekorele edilmiş bir özellik kümesi olan MFCC matrisine dönüştürülür (Şekil 3.3'ün alt grafiğinde gösterilmiştir). Bu özellik kümesi, sağlam ses olayı sınıflandırması için en anlamlı spektral zarf bilgilerini korurken, yinelenen gürültü bileşenlerini eler (Davis ve Mermelstein, 1980; Çakır, 2019).



Şekil 3.3. İlk grafik, bir ses sinyalinin zaman alanındaki dalga biçimini göstermektedir. İkinci grafik bir ses sinyalinin MFCC'lerini göstermektedir.

3.4. Ses Olay Tespiti için Derin Öğrenme

Derin öğrenme (Deep Learning – DL), geleneksel yapay sinir ağı (YSA) paradigmasını genişleterek, çok sayıda doğrusal olmayan işlem katmanının (buradaki "derin" terimi buradan gelir) üst üste yerleştirilmesiyle, verilerden giderek soyutlaşan temsiller öğrenebilen hiyerarşik öznitelik çıkarıcılar oluşturur (Alpaydın, 2014). Daha önceki

"sığ" ağlar genellikle el ile çıkarılan özniteliklere ve bir veya iki gizli katmana dayanırken, modern DL modelleri onlarca hatta yüzlerce katmandan oluşabilir ve ham girdilerden doğrudan görevle ilişkili öznitelikleri otomatik olarak öğrenebilir (Goodfellow ve ark, 2016). Bu uçtan uca öznitelik öğrenme kapasitesi, bilgisayarlı görü, desen tanıma ve hem görüntü hem de ses sınıflandırma görevlerinde geleneksel makine öğrenimi tekniklerine göre belirgin performans artışları sağlamıştır (Krizhevsky ve ark, 2017; Diro ve Chilamkurti, 2017).

DL içindeki en etkili mimarilerden biri olan CNN'ler, iki temel ilkeyi kullanır: Yerel bağlantı, yani her nöron yalnızca önceki katmanın küçük bir alıcı alanını işler; ağırlık paylaşımı, yani aynı evrişim çekirdeği tüm giriş boyunca tekrar tekrar uygulanır. Bu özellikler parametre sayısını büyük ölçüde azaltır ve CNN'leri spektrogramlar veya görüntüler gibi yüksek boyutlu girdiler için oldukça uygun hale getirir (LeCun ve ark, 1998). Tipik bir CNN işleyişi, evrişim katmanlarını doğrusal olmayan aktivasyon fonksiyonları (örneğin ReLU), ortam havuzlama (örneğin max pooling) ve nihai sınıflandırmayı gerçekleştiren tam bağlantılı katmanlarla birleştirir. Grafik işlem birimi (Graphics Processing Unit – GPU) hızlandırmasıyla desteklenen bu derin evrişim yığınları, milyonlarca örnek üzerinde eğitilebilir ve daha önce görülmemiş veriler üzerinde genelleme yapabilir hale gelir—bu, büyük ölçekli ses olay algılama ve diğer büyük veri uygulamaları için temel bir özelliktir (Krizhevsky ve ark, 2017).

Denetimli DL iş akışlarında, kullanılabilir veri genellikle üç ayrı alt kümeye ayrılır. Eğitim veri kümesi, ağırlıkların öğrenilmesini sağlar ve gradyan tabanlı optimizasyonu yönlendirir. Doğrulama (validation) kümesi, eğitim sırasında periyodik olarak kullanılır ve öğrenme oranı, mini-yığın boyutu ve model derinliği gibi hiperparametrelerin ayarlanmasında kullanılır. Test veri kümesi ise tüm tasarım kararları verildikten sonra modele gösterilen, genelleme performansını tarafsız biçimde değerlendirmeye yarayan veri kümesidir (Maloof, 2006). Eğitim süreci, belirli sayıda eğitim döngüsü boyunca veriler üzerinde tekrar tekrar yapılır; bir eğitim döngüsü, eğitim veri kümesi boyunca yapılan tam bir geçiştir. Amaç, düşük hata oranı, kabul edilebilir eğitim süresi ve güçlü genelleme kabiliyeti sağlayan bir kayıp fonksiyonunu minimize eden ağırlık kümesini bulmaktır. Bu süreç genellikle dropout veya weight decay gibi tekniklerle desteklenir (Goodfellow ve ark, 2016).

Bir CNN'in hesaplama yolculuğu, uzamsal veya zamansal boyutlar ile kanal bilgilerini taşıyan tensör biçimli bir girişle başlar — örneğin 128 × 128 × 1 boyutunda bir log-

mel spektrogram ya da 224 × 224 × 3 boyutunda bir RGB görüntü (Goodfellow ve ark, 2016). Evrişim katmanı, bu tensör üzerinde küçük boyutlu (örneğin 3 × 3) öğrenilebilir çekirdekleri kaydırarak bir öznitelik haritası yığını üretir. Her çekirdek, tüm konumlarda aynı ağırlıkları kullanarak aynı örüntüyü konumdan bağımsız şekilde algılar. Bu sırada stride (adım uzunluğu) ve padding (sıfır dolgusu) gibi hiperparametreler, çıktı çözünürlüğünü ve kenar davranışlarını kontrol eder (LeCun, Bottou, Bengio ve Haffner, 1998). Çok sayıda bu tür katman üst üste yerleştirildiğinde, ağ giderek daha soyut temsiller inşa eder—erken katmanlarda kenar veya spektral desenler, daha derinlerdeyse adım seslerinin topuk vuruşu gibi sınıfa özgü örüntüler öğrenilir. Belirli aralıklarla uygulanan havuzlama katmanları, her örtüşmeyen pencere içerisindeki en yüksek aktivasyonu koruyarak boyutları düşürür, böylece hem hesaplama maliyeti azalır hem de konumsal küçük kaymalara karşı dayanıklılık sağlanır (Krizhevsky ve ark, 2017).

Evrişim ve havuzlama katmanları yerel örüntüleri özetledikten sonra, bu çok boyutlu aktivasyonlar düzleştirilerek tam bağlantılı katmanlara aktarılır. Bu katmanlar, tüm önceki aktivasyonlardan sinyal alarak bilgiyi bütünsel olarak işler. Aşırı öğrenmeyi (overfitting) önlemek için dropout katmanları, eğitim sırasında belirli bir orandaki (p) aktivasyonu sıfıra çevirerek modeli birden fazla yedek yola güvenmeye zorlar ve genelleme kabiliyetini artırır (Srivastava ve ark, 2014). Her doğrusal işlemden sonra doğrusal olmayan aktivasyon fonksiyonları uygulanır: ReLU (Denklem 3.2), gizli katmanlarda eğitimin hızlanmasını ve kaybolan gradyan probleminin önlenmesini sağlar; Leaky-ReLU veya PReLU gibi varyantlar ise negatif girişlerde küçük bir eğim bırakarak gradyan akışını devam ettirir (Nair ve Hinton, 2010; He ve ark, 2015). Çıkış katmanında ise, ikili sınıflandırma problemlerinde tek bir sigmoid (Denklem 3.3) nöron [0,1] aralığında olasılık üretirken, çok sınıflı görevlerde softmax (Denklem 3.4) çıkışı her sınıf için bir olasılık dağılımı sağlar, fonksiyonlar aşağıda sırasıyla verilmiştir.

$$f(x) = \max(0, x) \tag{3.2}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.3}$$

$$f(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{3.4}$$

Eğitim süreci, boyutu B olan mini-yığınlar (mini-batch) üzerinden yinelemeli olarak gerçekleştirilir. Daha büyük yığınlar, daha düzgün gradyan tahminleri sunarken aynı zamanda daha fazla GPU belleği gerektirir. Tüm eğitim örnekleri üzerinde yapılan bir tam geçiş, bir eğitim döngüsü olarak tanımlanır. Aşırı öğrenmeyi önlemek amacıyla ayrılmış bir doğrulama kümesi üzerinde erken durdurma (early stopping) kriterleri uygulanır.

Parametre güncellemeleri, Adam optimizasyon algoritması tarafından yönlendirilir. Adam, gradyanın birinci ve ikinci momentlerinin üstel hareketli ortalamalarını tutar ve bu ortalamaların yanlılıklarını düzelterek her ağırlık için uyarlanabilir bir öğrenme oranı hesaplar (Kingma ve ark, 2015). Genellikle 0.001 gibi bir başlangıç öğrenme oranı ile başlar ve her ağırlık için adım büyüklüğünü otomatik olarak ölçeklendirir. Bu sayede, klasik stokastik gradyan inişine kıyasla genellikle daha hızlı yakınsama sağlar.

Eğitimin hedefi, modelin tahminleri ile gerçek etiketler arasındaki farkı ölçen bir kayıp fonksiyonunu (loss function) minimize etmektir. İki sınıflı (binary) tespit problemlerinde standart olarak ikili çapraz-entropy (binary cross-entropy) kullanılır. Bu, N örnek için tanımlanır; burada y_i gerçek etiket, y_i' ise modelin tahmin ettiği olasılıktır (Denklem 3.5). Bu ifade, Bernoulli dağılımı için negatif log-olasılık (log-likelihood) olarak da yorumlanabilir: Modelin yaptığı hatalar ne kadar fazla, düzenlemeler de o kadar büyük olur. Ayrıca bu fonksiyon, gradyan tabanlı optimizasyon için düzgün bir yapı sunar ve tekil örnekler için sıkı konveks (strictly convex) özellik taşıdığı için mükemmel tahminler durumunda tekil bir minimuma sahiptir (Goodfellow ve ark, 2016).

Bu eğitim hattı sayesinde; tensörlerden çekirdeklere, havuzlamadan dropout'a, uyarlanabilir optimizasyondan ilkesel kayıp minimizasyonuna kadar, bir CNN modeli ham girdileri ayrıştırıcı iç temsillere ve nihayetinde doğru sınıf olasılıklarına dönüştürebilir.

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log y'_i + (1 - y_i)\log(1 - y'_i)] \tag{3.5}$$

Özetle, ham bir ses kaydı ilk olarak sayısallaştırılır ve ardından hızla değişen frekans içeriğinin analiz edilebilmesi için üst üste binen kısa çerçevelere (frame) bölünür (Cakir, 2019). Her bir çerçevenin frekans spektrumu, mel filtre bankasından geçirilir ve DCT ile sıkıştırılır. Bu işlem sonucunda, sinyalin insan algısına en uygun frekans bileşenlerini temsil eden kompakt bir MFCC vektörü elde edilir (Davis ve Mermelstein, 1980). MFCC dizisi, etiketlendikten sonra, denetimli öğrenme için uygun bir veri kümesi oluşur.

Bu öznitelik matrisleri, mini-yığınlar hâlinde gruplanarak, tekrarlayan spektral-zamansal desenleri öğrenebilecek şekilde yapılandırılmış derin CNN'e aktarılır. Eğitim sırasında, ağırlıklar geriye yayılım (backpropagation) yöntemiyle ikili çapraz-entropy kayıp fonksiyonu minimize edilerek güncellenir (Goodfellow ve ark, 2016). Eğitim süreci tamamlandıktan sonra, elde edilen model yeni bir ses kaydını işleyebilir ve hedef ses olayının (örneğin adım sesi) varlığını veya yokluğunu belirten zaman damgalı tahminler üretebilir. Böylece, uçtan uca bir ses olayı tespit hattı tamamlanmış olur.

Ses olayı tespiti için bir CNN modeli eğitilirken, modelin görülmemiş ses verilerine genelleme yapabilme yetisi kritik öneme sahiptir. Bu yetiyi güvenilir şekilde değerlendirmek için kullanılan sağlam tekniklerden biri k-katlı çapraz doğrulama (k-fold cross-validation) yöntemidir. Burada ses MFCC dizileri yaklaşık eşit büyüklükte k adet farklı parçaya (fold) ayrılır. Bu ayrım, döngüsel bir değerlendirme sürecinin temelini oluşturur.

K-katlı çapraz doğrulamanın temel mantığı, CNN modelinizi k kez eğitmek ve değerlendirmektir. Her yinelemede, farklı bir fold doğrulama (validation) veya test kümesi olarak seçilir – yani modelin o eğitim döngüsünde görmediği verileri temsil eder. Geriye kalan (k-1) fold, modelin eğitimi için kullanılır. Eğitimden sonra, modelin başarımı yalnızca elde tutulan doğrulama kümesinde değerlendirilir. Bu işlem, tüm fold'lar bir kez doğrulama kümesi olarak seçilene kadar tekrarlanır (Hastie ve ark, 2009).

Bu döngüsel değerlendirme sonucunda, her biri farklı bir doğrulama parçasına ait olmak üzere k farklı performans metriği elde edilir. CNN modelinin genelleme yeteneğini daha sağlam biçimde değerlendirebilmek için bu metrikler genellikle

aritmetik ortalama alınarak birleştirilir. Bu ortalama, modelin yeni ses olayları içeren kayıtlarda ne kadar iyi performans göstereceğine dair daha istikrarlı bir gösterge sunar.

K-katlı çapraz doğrulamanın önemli avantajlarından biri, özellikle sınırlı boyuttaki ses veri kümelerinde verilerin etkili şekilde kullanılmasıdır. Her bir veri noktası farklı iterasyonlarda hem eğitim hem de doğrulama sürecine katkı sağladığı için, tek bir eğitim-test ayrımına göre daha kapsamlı bir değerlendirme yapılabilir. Ayrıca bu yöntem, modelin eğitim verisine aşırı uyum sağlayıp (overfitting) yeni verilerde kötü performans göstermesi gibi durumları da tespit etmeye yardımcı olur. Eğer CNN modeli farklı iterasyonlardaki doğrulama parçalarında sürekli düşük performans gösteriyorsa, bu durum aşırı öğrenmenin güçlü bir göstergesi olabilir (Stone, 1974).

Genel olarak k-katlı çapraz doğrulama için tek bir evrensel formül olmasa da, temel hesaplama mantığı her bir doğrulama parçasından elde edilen performans metriklerinin ortalamasını almaktır. Eğer (P_i) her bir i. parçadaki başarı metriğini (örneğin doğruluk, F1 skoru) temsil ediyorsa ve i 1'den k'ya kadar değişiyorsa, modelin genel başarı metriği ($P_{Overall}$) aşağıdaki Denklem 3.6 ile gösterilebilir:

$$P_{Overall} = \frac{1}{k}\sum_{i=1}^{k} P_i \tag{3.6}$$

Bu ortalama, CNN modelinizin görülmemiş ses olayı verileri üzerindeki beklenen performansını daha sağlam ve güvenilir bir şekilde ölçmeye olanak tanır.

3.5. Değerlendirme Kriterleri

Bir ses olayı tespit sisteminin güvenilir bir şekilde değerlendirilmesi için gerçek sınıflandırmayla model sonuçlarının karşılaştırıldığı 2 × 2 boyutundaki karışıklık matrisi (confusion matrix) kullanılır. Tablo 3.1'de gösterilen bu matris, sistemin kararlarını özetler.

- Doğru pozitif (TP), modelin hedef bir olayı (örneğin bir adım sesi) doğru şekilde tespit ettiği durumu ifade eder.
- Yanlış pozitif (FP), aslında hedef olayın olmadığı bir durumda sistemin alarm vermesiyle oluşur.

- Doğru negatif (TN), modelin hedef ses olayı olmadığında alarm vermediği durumu gösterir.
- Yanlış negatif (FN) ise, modelin mevcut bir hedef olayı algılayamayıp atladığı durumu tanımlar (Sokolova ve Lapalme, 2009).

Tablo 3.1. Karışıklık Matrisi

	Predicted Positive	Predictive Negative
Gerçek Pozitif	Doğru Pozitif (TP)	Yanlış Negatif (FN)
Gerçek Negatif	Yanlış Pozitif (FP)	Doğru Negatif (TN)

Bu dört değerden çeşitli tamamlayıcı performans ölçütleri türetilir. Kesinlik (precision), Denklem 3.5'te gösterildiği üzere, "Bir adım sesi algılandığında, bunun gerçekten doğru olma olasılığı nedir?" sorusuna yanıt verir. Duyarlılık (recall) veya hassasiyet (sensitivity) ise, Denklem 3.6'da belirtildiği gibi, "Gerçek ses olaylarının ne kadarını tespit edebildik?" sorusunu sorar (Powers, 2020).

Karar eşiği (threshold) değiştirilerek yüksek kesinlik ile yüksek duyarlılık arasında denge kurulabilir. Bu nedenle, F1 skoru, bu iki ölçütü harmonik ortalama yoluyla birleştirir ve yalnızca her iki değerin de yüksek olduğu durumlarda maksimuma ulaşır.

Doğruluk (accuracy) ise genel başarı oranını sunar, fakat sınıf dengesizliğinin (örneğin negatif örneklerin çoğunluğu) söz konusu olduğu durumlarda yanıltıcı olabilir. Bu yüzden, performansın farklı boyutlarını daha doğru yansıtmak adına doğruluğa ek olarak kesinlik, duyarlılık ve F1 skoru birlikte rapor edilir (Sokolova ve Lapalme, 2009).

$$Kesinlik = \frac{TP}{TP + FP} \tag{3.5}$$

$$Duyarlılık = \frac{TP}{TP + FN} \tag{3.6}$$

$$F1 = 2 \times \frac{Presicion \times Recall}{Presicion + Recall} \tag{3.7}$$

Eğitim süreci boyunca ağ, yukarıda açıklanan ayrık (discrete) performans ölçütleriyle değil, sürekli türevlenebilir bir kayıp fonksiyonu (loss function) ile yönlendirilir. Bu çalışmada kullanılan ikili sınıflandırıcı, tahmin edilen olasılıklarla gerçek etiketler arasındaki farkı cezalandıran ikili çapraz entropi (binary cross-entropy) fonksiyonunu en aza indirmeye çalışır. Bu kayıp, her örnek için ayrı ayrı hesaplanır ve ardından gradyan inişi (gradient descent) yöntemiyle model ağırlıkları, beklenen hata değeri azalacak şekilde güncellenir (Goodfellow ve ark, 2016). Eğitim ve doğrulama süreçleri boyunca kayıp ve doğruluk eğrilerinin (loss and accuracy curves) görselleştirilmesi, modelin öğrenmeye devam edip etmediğini, yakınsayıp yakınsamadığını veya aşırı öğrenmeye (overfitting) başladığını –örneğin eğitim kaybı düşerken doğrulama kaybının yükselmesi durumunda– anlamayı sağlar.

Sınıflandırma raporları (classification reports), özellikle CNN gibi derin öğrenme algoritmalarıyla gerçekleştirilen SED görevlerinde, modelin başarımını kapsamlı bir biçimde değerlendirmek için kullanılan önemli araçlardandır. Bu raporlar genellikle kesinlik (precision), duyarlılık (recall), F1 skoru ve destek (support) gibi temel ölçütleri içerir. Böylece modelin her bir ses sınıfını ne derece başarılı ayırt ettiğine dair ayrıntılı bir anlayış sunar. Destek (support), veri kümesinde her sınıfa ait gerçek örnek sayısını belirtir ve bu bilgi, diğer metriklerin yorumlanmasına bağlam kazandırır.

Sınıflandırma raporlarının kullanımı, başlangıçta metin sınıflandırması ve görüntü işleme alanlarında yaygınlık kazanmış; sonrasında ses işleme ve ses sınıflandırma alanlarına da uyarlanmıştır. Bu tür raporlar, modelin farklı sınıflar üzerindeki performansını ayrıntılı şekilde inceleyerek karşılaştırma ve kıyaslama yapmaya olanak tanır.

Son olarak, modelin ikinci fazının ilk adımlarında (Bölüm 4.2'de ayrıntılı olarak açıklanacaktır), t-dağıtımlı stokastik komşuluk yerleştirme (t-distributed Stochastic Neighbor Embedding – t-SNE) algoritması kullanılmıştır. t-SNE, yüksek boyutlu uzaydaki örnekler arası benzerlikleri ortak olasılıklara dönüştürerek, noktaları iteratif olarak yerleştirir: Yakın komşular birbirine yakın, uzak olanlar ise daha ayrık konumlandırılır. Bu sayede yerel yapı korunurken genel sınıf kümelenmeleri görselleştirilir (van der Maaten ve Hinton, 2008).

t-SNE grafiğinde adım sesi ve adım sesi olmayan (veya kalitesiz/niteliksiz) örneklerin net biçimde ayrışması, henüz nicel metriklere geçmeden önce bile ön işleme ve

öğrenme süreçlerinin ayırt edici özellik uzayları (manifoldlar) oluşturduğuna dair nitel bir kanıt sunar.

Tüm bu istatistiksel ölçütler ve görsel teşhis araçları, modelin başarımını ve bilinmeyen seslere karşı genelleme yapabilme yetisini bütüncül olarak değerlendirmeyi sağlar.

4. ADIM SES OLAYI TESPİTİ

Ses olayları, insan kaynaklı ve doğal çok çeşitli etkinlikleri kapsar. Yaygın örnekler arasında insan tarafından üretilen sesler (konuşma, gülme, adım sesleri, kapı tıklamaları), nesneye ait sesler (cam kırılması, kapı açılıp kapanması, silah sesleri, alarm zilleri) ve ortam sesleri (yağmur, hayvan sesleri, motor gürültüsü) yer alır. Mülk güvenliği bağlamında, bazı ses olayları izinsiz girişleri veya acil durumları işaret edici nitelikleri nedeniyle özellikle önem arz eder. Örneğin, kısıtlı bir alanda duyulan adım sesleri, yetkisiz bir kişinin varlığına işaret edebilir; bir kapının gıcırdayarak açılması ya da camın kırılması, bir hırsızlık girişimini gösterebilir. Adım sesi tespiti, çevresel güvenlik açısından dikkat çeken bir konu haline gelmiştir. Her ne kadar geçmişte kamuya açık veri kümelerinde net adım sesi örnekleri yaygın olmasa da, 2022 yılında sunulan Real-Life Indoor Sound Event Dataset (ReaLISED) adlı veri kümesi, çok sayıda adım sesi örneğini içeren kapsamlı bir iç mekân ses verisi koleksiyonu sunmuştur. Bu tür güvenlikle ilgili seslere dayalı olarak eğitilen SED modelleri, sessiz bir ortamda bile bu tür ipuçlarını (örneğin yaklaşan adım sesleri, izinsiz giriş sesleri) algılayarak sistemin alarm üretmesini sağlayabilir.

Bu çalışmanın başlangıç noktası etiketlenmiş adım sesi kayıtlarını içeren bir veri kümesinin araştırılıp bulunması olmuştur. Bu ihtiyacı karşılayan ReaLISED veri kümesi, 2022 yılında yayımlanmış ve zamanla hizalanmış etiketlere sahip geniş bir iç mekân ses yelpazesi sunmuştur (Mohino-Herranz ve ark., 2022). Bölüm 3.3 ve 3.4'te açıklanan ön işleme ve modelleme adımlarımız bu veri kümesine uygulanmıştır. Başlangıç eğitim verileri ve karışıklık matrisi analizleri, ham veride iki temel zayıflığı ortaya koymuştur: Modelin adım sesi sınıfını öğrenmeye meyilli olmasına neden olan sınıf dengesizliği ve düşük kaliteli ya da hatalı etiketlenmiş ses kliplerinin varlığı. Bu nedenle, ikinci deneysel aşama veri kümesinin iyileştirilmesine odaklanmıştır: Azınlık sınıfa çeşitlilik kazandırmak amacıyla Epidemic Sound ses kütüphanesinden (Epidemic Sound, 2024) yüksek kaliteli adım sesi örnekleri seçilmiş; ayrıca belirsiz ya da gürültülü klipler, spektrogram görselleştirmesi yardımıyla elle taranarak veri kümesinden çıkarılmıştır. Bu şekilde yeniden dengelenmiş ve iyileştirilmiş veri kümesi, aynı CNN modelleme hattına yeniden sunulmuş; böylece daha net bir

öğrenme sinyali elde edilmiş ve hiperparametre ayarları ile model değerlendirmesi için daha kararlı bir temel sağlanmıştır.

4.1. Faz 1: ReaLISED Veri Kümesi

ReaLISED veri kümesi (Mohino-Herranz ve ark., 2022), 2022 yılında yayımlanmış olması, çeşitli ses olaylarını kapsaması ve adım sesi kayıtlarını içermesi bakımından değerli bir kaynak olarak öne çıkmaktadır. Bu veri kümesi, toplamda 2.479 ses dosyasından oluşmakta olup, bir saatten fazla kayıt süresi içermektedir. Kayıtlar; evler, apartman daireleri ve üniversite binaları gibi farklı ortamlarda yerleştirilmiş dört adet Olympus LS-100 ses kayıt cihazı kullanılarak alınmıştır. Tüm kayıtlar, 44.1 kHz örnekleme hızı ve 24-bit çözünürlük ile gerçekleştirilmiş olup, bu sayede 20 kHz'e kadar bant genişliğinde sinyallerin yakalanması ve 144 dB'ye kadar dinamik aralık desteği sağlanmıştır. Ayrıca kayıtlar stereo modda yapılmıştır; bu da ortam bilgilerin korunmasına olanak tanımaktadır. ReaLISED veri kümesinde yer alan ses olaylarına ilişkin sınıflar Tablo 4.1'de sunulmuştur.

Tablo 4.1. ReaLISED veri kümesindeki sınıflar ve ses dosyası sayıları.

Ses olayları	Dosya sayısı
Pencere	104
Musluk	140
Çamaşır makinesi	117
Adım sesi	119
Süpürge makinesi	166
Anahtar	108
Televizyon	143
Konuşma	190
Düşen obje	119
Mikrodalga	124
Duman dedektörü	142
Mobilyanın hareketi	152
Çekmece	158
Kapı	109
Bulaşık makinesi	130
Dolap	156
Yemek yapma	176
Çırpıcı	126

4.1.1. Veri kümesi

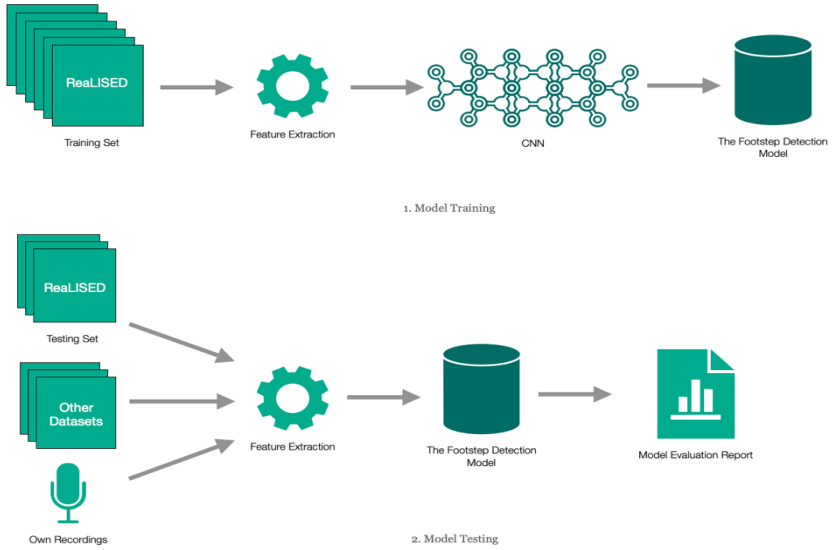
Bu çalışmada, mülkleri izinsiz girişlerden korumak amacıyla adım sesi olaylarının tespiti üzerine odaklanılmıştır. Televizyon veya kapı açılması gibi diğer ses olayları bu çalışmanın kapsamı dışında bırakılmıştır. Geliştirilen model, farklı ses olayları arasından yalnızca adım sesi olaylarını ayırt etmeye yönelik olarak tasarlanmıştır. Bu nedenle, adım sesi olayları "1" etiketiyle işaretlenmiş, diğer tüm ses olayları ise birlikte değerlendirilerek farklı bir sınıf altında toplanmıştır. Dosyalar üzerindeki etiket dağılımı Tablo 4.2'de verilmiştir.

Tablo 4.2. Faz 1'deki etiketlenmiş ses dosyalarının sayısı

Ses Olayları	Dosya Sayısı	Etiketler
Diğer ses olayları	2362	0
Adım sesi olayı	117	1

4.1.2. Metodoloji ve model

Mohino-Herranz ve ark. (2022), ReaLISED veri kümesindeki tüm etiketleri çok sınıflı çeşitli sınıflandırıcılar eğiterek incelemiştir. Bu çalışma ise daha dar bir yaklaşımla başlamaktadır: ilk deneyde, yalnızca adım sesi ile diğer tüm ses olaylarını ayırt eden tek bir CNN modeli geliştirilmiştir. Ağ, yalnızca ReaLISED veri kümesindeki ses klipleriyle eğitilmiştir; ancak modelin performansı daha sonra ek veri kümeleri ve sahada yapılan bu çalışma için kaydediler veriler ile test edilmiştir. Eğitim veya test işleminden önce, her ses dosyası bir önişleme hattı aracılığıyla özelliklere dönüştürülmüştür. Kayıpsız ses kodeki (Free Loses Audio Codec - FLAC) formatında saklanan ham sesler, librosa kütüphanesi (McFee ve ark., 2015) kullanılarak yüklenmiş ve MFCC çıkarılmıştır. Adım sesi içermeyen tüm ses olaylarına 0 etiketi, adım sesi içeren parçalara ise 1 etiketi atanmıştır; böylece model, çok sınıflı bir haritalama yerine ikili bir karar sınırı öğrenmektedir. CNN ağı eğitildikten ve model oluşturulduktan sonra, performansı diğer veri kümeleri, bu çalışma için kaydediler sesler ve ReaLISED veri kümesinin eğitilmemiş kısmı kullanılarak değerlendirilmiştir. İlk fazın yöntemi Şekil 4.1'de gösterilmiştir.



Şekil 4.1. İlk akış birinci fazın eğitim aşamasını, ikinci akış ise test aşamasını göstermektedir.

Tüm hesaplamalı denemeler, ücretsiz GPU'lar, kalıcı depolama alanı ve kod sürüm kontrolü sunan Kaggle platformu (Kaggle, 2025) üzerinde gerçekleştirilmiştir. Bu özellikler, deneylerin güvenilir şekilde tekrarlanabilmesini ve kolayca paylaşılabilmesini sağlayan bir ortam oluşturur. Hiperparametre arayışımızda, daha önce sinyal işleme açısından etkili olduğu belirlenen STFT çerçeve uzunluğu (n_fft) ve MFCC (n_mfcc) üzerine odaklanılmıştır. Bir grid search yöntemiyle farklı değer çiftleri test edilmiş ve doğrulama F1 skorunun en yüksek olduğu yapılandırma seçilmiştir. Bu taramanın sonuçları Tablo 4.3'te sunulmaktadır. Son modelde n_fft değeri 2048, n_mfcc değeri ise 20 olarak belirlenmiştir.

Tablo 4.3. Eğitim doğruluğu ve F1 skoru açısından MFCC'lerin oluşturulmasında kullanılan n_fft ve n_mfcc parametreleri.

n_fft	n_mfcc	Doğruluk Oranı (%)	F1 Değeri (%)
	20	97	98
	19	92	92
	18	94	94
	17	91	91
2048	16	94	94
	15	92	92
	14	92	92
	12	89	89
4096	20	95	95
1024	20	94	93
512	20	91	91
256	20	83	93

Ağın omurgası, üst üste yerleştirilmiş evrişimsel katmanlardan oluşmaktadır; aktivasyonları dengelemek ve aşırı öğrenmeyi engellemek amacıyla batch normalizasyonu ve dropout teknikleri uygulanmıştır. Tüm gizli katmanlarda ReLU aktivasyon fonksiyonu kullanılmış, çıkış katmanında ise sınıf olasılıklarını elde etmek için softmax fonksiyonu tercih edilmiştir. Optimizasyon işlemi, öğrenme oranı $1x10^{-4}$ olan Adam algoritmasıyla gerçekleştirilmiş ve kayıp fonksiyonu olarak ikili çapraz entropi (binary cross-entropy) kullanılmıştır. Eğitim süreci 50 eğitim döngüsü (epoch) boyunca sürdürülmüş ve sonuçlar Tablo 4.4'te sunulmuştur. Model, eğitim verisi üzerinde %99 doğruluk, doğrulama verisi üzerinde ise %89 doğruluk elde etmiştir.

Tablo 4.4. Faz 1 hiperparametreleri

Hiperparametreler	Değerler
Gizli Katman	64
Dropout	0.2
Batch boyutu	64
Öğrenme oranı	0.0001
Eğitim döngüsü (epoch)	50
Optimizasyon	Adam
Kayıp fonksiyonu	Binary crossentropy

4.1.3. Model değerlendirmesi

Eğitilmiş ağ, ilk olarak eğitim süreci boyunca tamamen görülmemiş olan bir test bölümü üzerinde değerlendirilmiştir. Bu test, modelin gerçek dünyadaki doğruluğunu önyargısız bir şekilde tahmin etmeye olanak tanımaktadır. Elde edilen sonuçlara ait karışıklık matrisi Tablo 4.5'te, precision (kesinlik), recall (duyarlılık) ve F1 skorları ise Tablo 4.6'da sunulmuştur. Genel istatistikler ilk bakışta kabul edilebilir görünse de, matris önemli bir dengesizliği ortaya koymaktadır: model, yanlış negatiflerden (false negative) daha fazla yanlış pozitif (false positive) üretmektedir. Bu durum, sınıflar arasında sürekli bir dengesizlik olduğuna işaret etmektedir.

Daha sonra, model farklı akustik koşullar altında kaydedilmiş iki yeni adım sesi örneğiyle test edildi. Her iki örnek de yanlış sınıflandırılmış olup, modelin ReaLISED dağılımının dışındaki verileri genelleme performansında bir eksiklik olduğunu göstermiştir. Kesinlik ve duyarlık değerleri arasındaki büyük fark, sınıf dengesizliği (class imbalance) nedeniyle duyarlık değerinin düşük kaldığını, buna karşın kesinlik değerinin göreceli olarak yüksek olduğunu teyit etmiştir (He ve Garcia, 2009). Bu

bulgular doğrultusunda, azınlık sınıfı olan adım sesi örneklerini çoğaltma kararı alınmış ve yüksek kaliteli yeni kayıtlar toplanarak ReaLISED verisiyle birleştirilmiş; bu sayede hem genelleme kabiliyetinin artırılması hem de sınıf dengesizliğinin giderilmesi hedeflenmiştir.

Tablo 4.5. Faz 1'e ait karışıklık matrisi

	Tahmin Edilen Pozitif	Tahmin Edilen Negatif
Gerçek Pozitif	470	2
Gerçek Negatif	4	20

Tablo 4.6. Faz 1'e ait sınıflandırma raporu

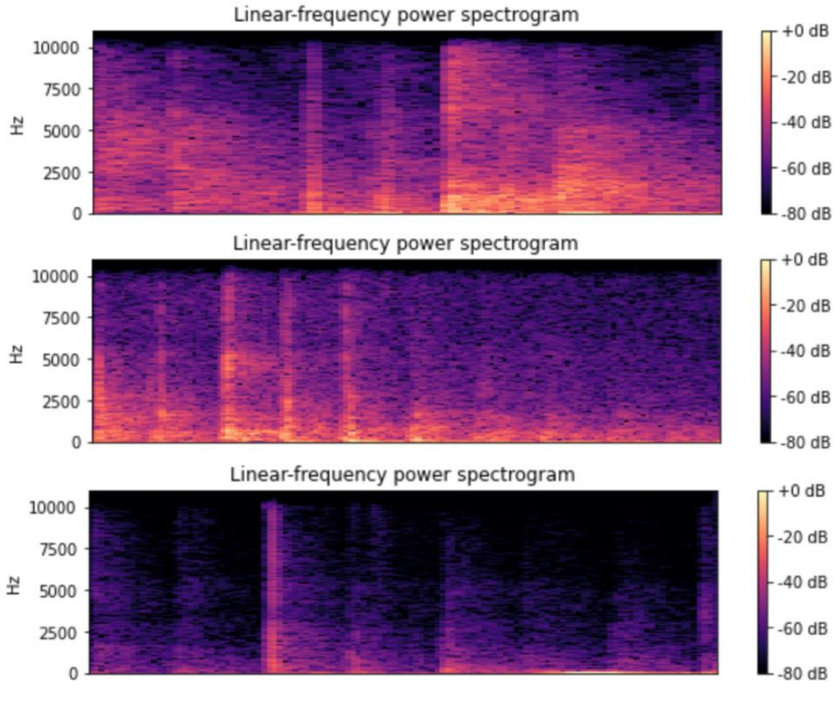
	Keskinlik	Duyarklık	F1 Skoru	Destek
0	0.99	1.00	0.99	472
1	0.91	0.83	0.87	24
Doğruluk			0.99	496
Makro Ort.	0.95	0.91	0.93	496
Ağırlıklı Ort.	0.99	0.99	0.99	496

4.2. Faz 2: Gözden Geçirilmiş ReaLISED ve Ek Veri

Analizlerin açıkça ortaya koyduğu üzere, veri kümesi sınıflar arasında ciddi bir dengesizlik göstermektedir. Bu tür dengesiz veri senaryolarında, sınıflandırıcılar genellikle çoğunluk sınıfını tahmin etmeye eğilimli hale gelir; çünkü eğitim örneklerinin büyük kısmı bu sınıfa aittir. Bu orantısız durum, modelin azınlık sınıflar üzerindeki tahminlere güvenilmez ve pratikte etkisiz hale getiren istenmeyen bir önyargı yaratır (He ve Garcia, 2009). Bu sorunu etkili bir şekilde ele almak ve sonraki aşamalara geçmeden önce, orijinal ReaLISED veri kümesini kapsamlı bir şekilde incelendi.

Bu detaylı inceleme sonucunda, veri kümesinde bazı dikkat çekici tutarsızlıklar ve etiketleme hataları tespit edildi. Bu hatalar arasında, "yürüme" olarak etiketlenmiş bazı ses olaylarının, aslında daha yakından incelendiğinde "koşma" kayıtları olduğu ortaya çıktı. Koşma olayları, yürümeden belirgin şekilde farklıdır; zamanlama açısından farklı desenler sergiler, daha yüksek enerji seviyelerine sahiptir ve spektral özellikleri de ayrışır. Ayrıca, merdivenleri hızla çıkan bireylerin bulunduğu bazı bölümler de gözlemlenmiştir ve bunlar da tipik adım sesi desenlerinden açık bir şekilde

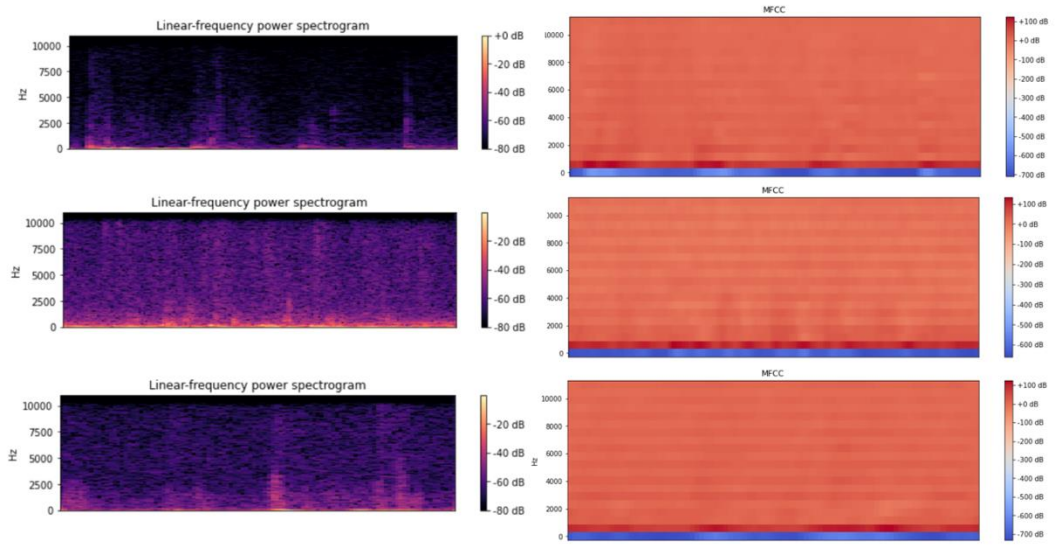
ayrışmaktadır. Şekil 4.2, ReaLISED veri kümesinde adım sesi olarak etiketlenmiş ama incelendiğinde koşma ses olayına ait olan ses sinyallerinden üç tanesini spektrogramları aracılığıyla, frekans ve zamansal desenlerdeki farklılıkları açıkça göstermektedir.



Şekil 4.2. Farklı frekans ve şiddet düzeylerine sahip koşma sesi sinyaline ait üç ayrı spektrogram.

Veri kümesinde gözlemlenen ikinci önemli tutarsızlık, dikkate değer derecede düşük ses kalitesine sahip örneklerin varlığıdır. Bu kayıtlar dikkatlice dinlendiğinde ve karşılık gelen spektrogramları titizlikle incelendiğinde, bazı ses olaylarının anlamlı özellik çıkarımı ve sınıflandırma görevleri için yeterince güçlü akustik özellikler içermediği açıkça görülmüştür. Bu tür kayıtların akustik netlikten yoksun oluşu, makine öğrenimi süreçlerinin güvenilirliğini ve etkinliğini olumsuz etkiler; zira bu tür düşük kaliteli sinyallerden çıkarılan özellikler genellikle gürültülü ve ayırt edici olmayan yapıdadır. Ayrıca, ayrıntılı inceleme sırasında tahta zemin gıcırtıları gibi arka plan parazitleri ve çok zayıf ses olayları da tespit edilmiştir; bu durum hedef sinyallerin kalitesini ve netliğini daha da düşürmektedir. Bu sorunlu kayıtlara ait örnekler, karşılık gelen spektrogramları ve MFCC özellikleriyle birlikte Şekil 4.3'te gösterilmiştir. Yapılan kapsamlı değerlendirme sonucunda, bu kalite eksikliklerini taşıyan tüm ses dosyaları veri kümesinden sistematik olarak çıkarılmıştır. Bu iyileştirme süreci

sonucunda elde edilen Refined ReaLISED veri kümesi, 2360 adet adım sesi dışı sınıfa ait ses sinyali ve 24 adet adım sesi olayına özel olarak kategorize edilmiş ses sinyali içermektedir.



Şekil 4.3. Düşük kaliteli ses sinyaline ait üç adet spektrogram ve MFCC görselleştirmesi.

4.2.1. Veri kümesi

Ek adım sesi verileri, çeşitli kategorilerde profesyonelce kaydedilmiş yüksek kaliteli ses örnekleri sunan tanınmış bir kaynak olan Epidemic Sound web sitesinden (Epidemic Sound, 2024) elde edilmiştir. Yalnızca orijinal ReaLISED veri setine dayalı çalışmanın sınırlılıkları göz önüne alındıktan sonra, bu yeni yüksek kaliteli adım sesi kayıtları, ReaLISED içerisinden seçilen bazı ses olayı örnekleriyle birleştirilmiştir. Böylece oluşturulan ve dengeli yapıya sahip olan bu yeni koleksiyon, "Footsteps Sound Dataset" (Adım Sesleri Veri Seti) olarak adlandırılmıştır. Veri seti, 333 adet adım sesi ve 333 adet adım sesi dışı ses klibinden oluşmakta olup toplamda 666 ses dosyası içermektedir. ReaLISED içerisindeki ses örnekleri, adım sesi hariç 17 farklı sınıfı temsil eden kayıtlar arasından her sınıftan yaklaşık olarak eşit sayıda örnek içerecek biçimde manuel olarak seçilmiştir. Seçim sürecinde, sınıflar arası dengenin sağlanması öncelikli hedef olarak belirlenmiştir. Böylece, eğitim sürecinde modelin sınıflar arasında adil bir temsille karşılaşması ve öğrenme sürecinde belirli sınıflara aşırı uyum sağlamasının önüne geçilmesi amaçlanmıştır. Bu sayede, oluşturulan veri kümesinin sınıf temsiliyetinde denge sağlanarak, modelin genel performansını olumlu yönde etkileyen dengeli bir eğitim ortamı oluşturulmuştur.

Sınıf dağılımına ilişkin ayrıntılar Tablo 4.7'de sunulmuştur.

Tablo 4.7. Faz 2'deki etiketlenmiş ses dosyalarının sayısı

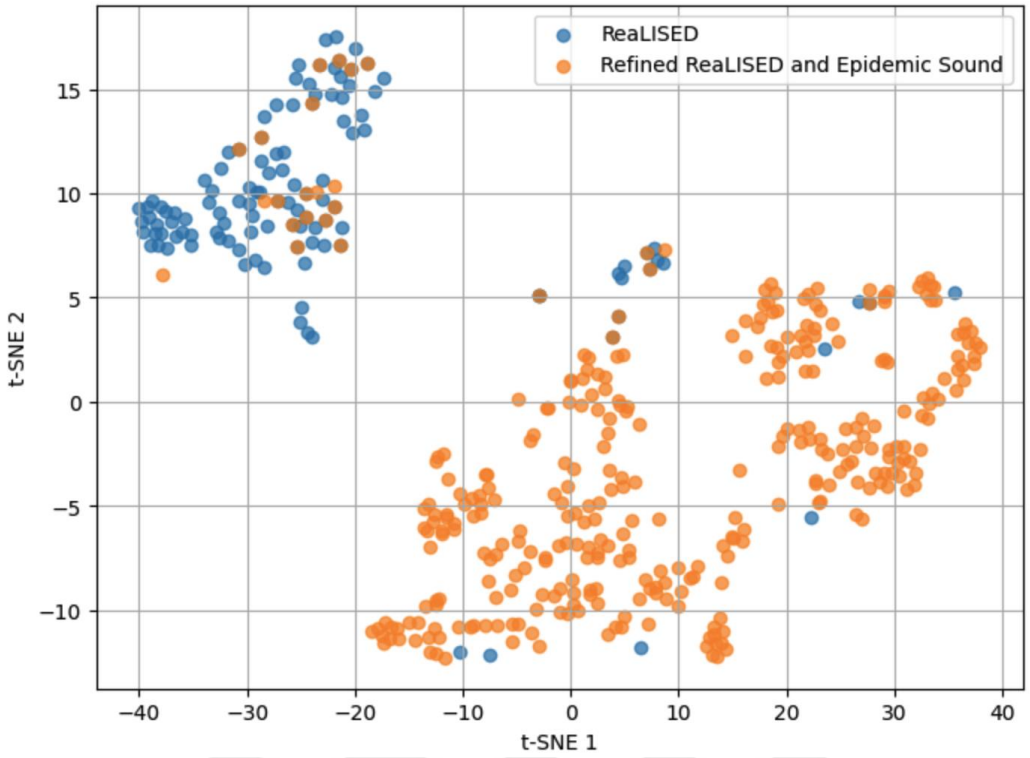
Ses Olayları	Dosya Sayısı	Etiketler
Diğer ses olayları	333	0
Adım sesi olayı	333	1

Şekil 4.4, yüksek kaliteli ses örneklerinin veri setimize entegre edilmesinin ardından ortaya çıkan olumlu etkileri görsel olarak ortaya koymaktadır. Bu iyileştirmeleri açıkça gösterebilmek için, karmaşık ve yüksek boyutlu veri setlerini görselleştirmek amacıyla özel olarak tasarlanmış doğrusal olmayan bir boyut indirgeme algoritması olan t-SNE yöntemi kullanılmıştır. İlgili şekilde, turuncu veri noktaları orijinal ReaLISED veri setinden alınan örnekleri, mavi noktalar ise Epidemic Sound kaynağından entegre edilen yeni örnekleri temsil etmektedir.

İyileştirme öncesinde, orijinal veri setindeki hedef sınıfa ait örnekler dar bir kümelenme göstermekteydi; bu durum, sınırlı çeşitlilik, tekrar eden örnekler ve düşük kaliteli ya da gürültülü kayıtların varlığına işaret etmektedir. Ancak Epidemic Sound verilerinin entegre edilmesiyle birlikte, bu kümenin gözle görülür şekilde genişlediği ve daha yapılandırılmış, zengin bir özellik dağılımının ortaya çıktığı görülmektedir.

Bu genişleme, veri setini yapay olarak dengelemeye yönelik yüzeysel bir müdahale değil; aksine, belirsiz, gürültülü ve tekrarlayan ses örneklerinin yüksek kaliteli verilerle değiştirilmesinin pratikteki önemini ortaya koymaktadır. Bu değişim süreci, makine öğrenimi modellerinin genelleme yeteneğini güçlendirerek, öğrenilen kalıpların daha temsil edici ve sağlam olmasını sağlamaktadır. Ayrıca, daha kaliteli örneklerin dahil edilmesiyle birlikte, veri setindeki güvenilir adım sesi sınıfına ait verilerin oranı %20,5'ten %100'e çıkarılmıştır. Bu kayda değer artış, verinin güvenilirliğini ve doğru özellik öğrenimi ile sınıflandırma görevlerine uygunluğunu büyük ölçüde artırmıştır.

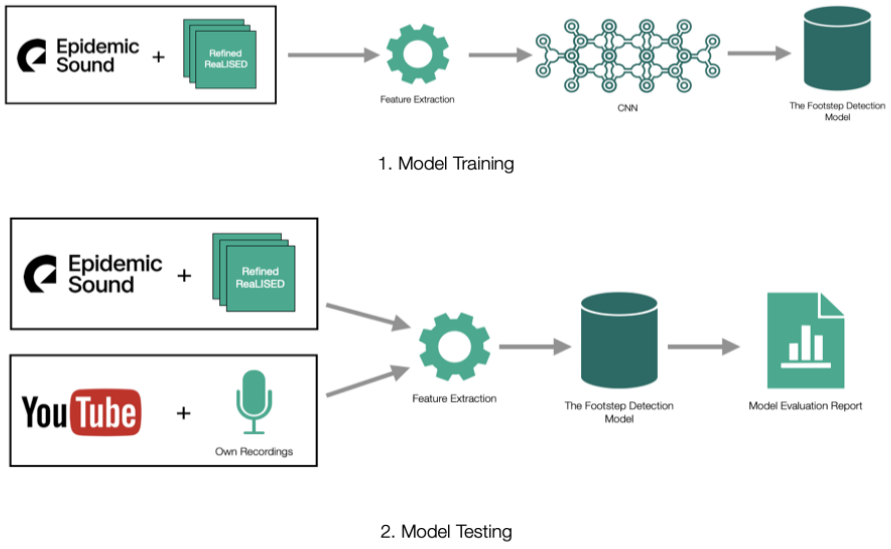
Sonuç olarak, ikinci faz, veriyi daha kaliteli hale getirerek, sınıflar arasında denge sağlayarak ve belirsiz örnekleri elimine ederek modelin performansını iyileştirmiştir. Bu da adım seslerini doğru biçimde algılayan ve tanıyan daha güvenilir ve etkili bir sistemin ortaya çıkmasını sağlamıştır.



Şekil 4.4. Faz 1 ve Faz 2 veri kümelerindeki hedef sınıfın t-SNE görselleştirmesi

4.2.2. Metodoloji ve model

CNN modelinin değerlendirilmesi ve ardından gelen performans analizi, yeni oluşturulan veri setinin model tarafından daha önce eğitim aşamasında hiç görülmemiş ayrı bir alt kümesi kullanılarak gerçekleştirilmiştir. Bu yaklaşım, değerlendirme sürecinin önyargısız ve güvenilir olmasını sağlamıştır. Bu veri setine ek olarak, değerlendirme aşamasında kendi kaydedilmiş adım sesi örnekleri ile çevresel ve kayıt gürültüleri içeren, YouTube gibi açık erişimli platformlardan temin edilen çeşitli ses klipleri de kullanılmıştır. İkinci fazın metodolojisi Şekil 4.5'te gösterilmiştir.

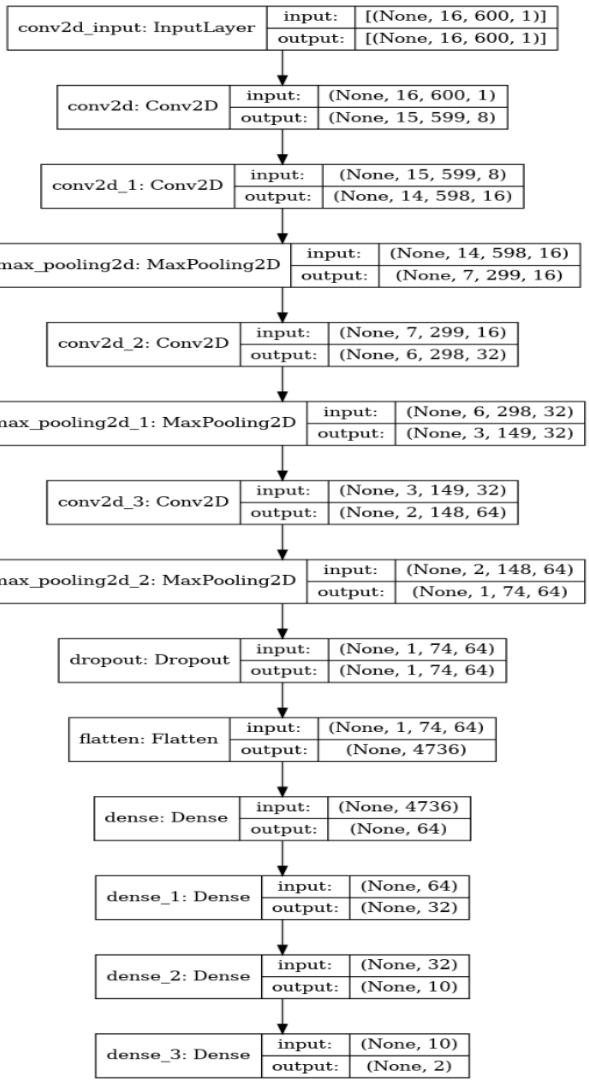


Şekil 4.5. İlk akış, ikinci fazın eğitim aşamasını; ikinci akış ise test aşamasını göstermektedir.

İlk deney aşamasında uygulanan prosedürün aksine, bu aşamadaki ses sinyalleri, öznitelik çıkarımından önce bir normalizasyon işlemine tabi tutulmuştur. Özellikle, normalizasyon işlemi sırasında librosa.util.normalize fonksiyonu kullanılmış; bu fonksiyon, ses sinyallerinin genliklerini maksimum mutlak değer olan 1.0'a ölçeklendirmiştir. Bu genlik normalizasyonu, giriş verilerinde bir biçimlilik sağlayarak istikrarlı öznitelik çıkarımını mümkün kılmakta ve sinyallerin orantısız şekilde ağır basmasının önüne geçerek sinir ağının daha verimli bir şekilde eğitilmesine katkıda bulunmaktadır. Önceki veri kümesinde olduğu gibi, bu yeniden düzenlenmiş ve genişletilmiş veri kümesi de 18 farklı iç mekân ses olayını içermektedir.

Derin sinir ağlarının geliştirilmesi ve eğitilmesi sürecinde, mimarinin karmaşıklığı ve yapısal derinliği, aşırı öğrenme riskine yol açan faktörlerdir (Goodfellow ve ark, 2016). Önceki aşamada karşılaşılan aşırı öğrenme problemleri, bu sorunu azaltmak amacıyla mimari sadeleştirmelere yol açmıştır. Bu nedenle, ağ mimarisi, Şekil 4.6'da gösterildiği üzere, fazla sayıda olan MaxPooling2D katmanlarının sayısı azaltılmıştır. Bu katmanların kaldırılmasındaki amaç, modelin yapısal karmaşıklığını azaltarak, daha önce görülmemiş veriler üzerinde genelleme yeteneğini artırmaktır. MFCC öznitelikleri çıkarılırken Tablo 4.3'de gösterildiği gibi n_fft = 2048 ve n_mfcc = 20 olarak belirlenmiştir. Ancak CNN giriş boyutunun (16, 600) olmasının sebebi, MFCC

özniteliklerinin sadece ilk 16 katsayısının kullanılmasıdır. Bu işlem, veri boyutunu azaltmak ve modelin eğitim performansını optimize etmek amacıyla yapılmıştır. Zaman boyutu ise ses dosyasının uzunluğu ve örnekleme parametrelerine göre 600 frame'e karşılık gelmektedir. Dört konvolüsyonel katmanda 10.904 ve onu izleyen dört tam bağlı katmanda 305.600 olmak üzere toplamda 316.504 öğrenilebilir parametre içermektedir.



Şekil 4.6. Ses olaylarını algılamaya yönelik CNN modelinin mimarisi.

Revize edilen modelin eğitim aşamasında, optimizasyon algoritması olarak Adam tercih edilmiş; kategorik çapraz entropi (categorical cross-entropy) kayıp fonksiyonu kullanılmıştır. Literatürde önerilen değer aralıkları dikkate alınarak ve çok sayıda deneysel deneme sonucunda hiperparametreler belirlenmiştir. Öğrenme oranı (0.0005), modelin doğruluk düzeyini artırırken overfitting riskini minimize etmek amacıyla, kademeli olarak yapılan testlerde elde edilen en dengeli sonuçları sağlayan

değer olarak seçilmiştir. Dropout oranı, eğitim sırasında aşırı öğrenmenin önlenmesi için yaygın olarak kullanılan 0.2 seviyesinde sabitlenmiştir. Batch boyutu (64) ve eğitim döngüsü (200), eğitim süresinin verimliliği ile modelin genel performansı arasında optimal dengeyi sağlamak amacıyla denenmiş ve gözlenen en iyi sonuçlara göre belirlenmiştir. Gizli katman sayısı ve yapısı ise, çeşitli yapılandırmalar denenerek modelin doğruluk ve genelleme başarısı temelinde optimize edilmiştir. Tablo 4.8'de listelenen bu parametrelerin belirlenmesinde hem deneysel sonuçlar hem de ilgili alan literatürü yönlendirici olmuştur.

Çıkış katmanı, sınıflandırma probleminin ikili (binary) yapısına uygun olarak iki nöron ile yapılandırılmıştır. Her bir nöron, hedef dizisi olan y içerisinde tek-sıcak kodlama (one-hot encoding) yöntemiyle temsil edilen iki olası sınıftan birine karşılık gelmektedir. Bu yapı, one-hot kodlu etiketlerle çalışmaya uygun olan kategorik çapraz entropi kayıp fonksiyonu ile tam uyum sağlamaktadır. Sınıf olasılıklarını açıkça ayırarak bu yapı, eğitim sürecinin etkinliğini artırmakta ve çapraz entropi hesaplamalarına dayalı doğru gradyan tahminleri sayesinde modelin daha hızlı yakınsamasını sağlamaktadır.

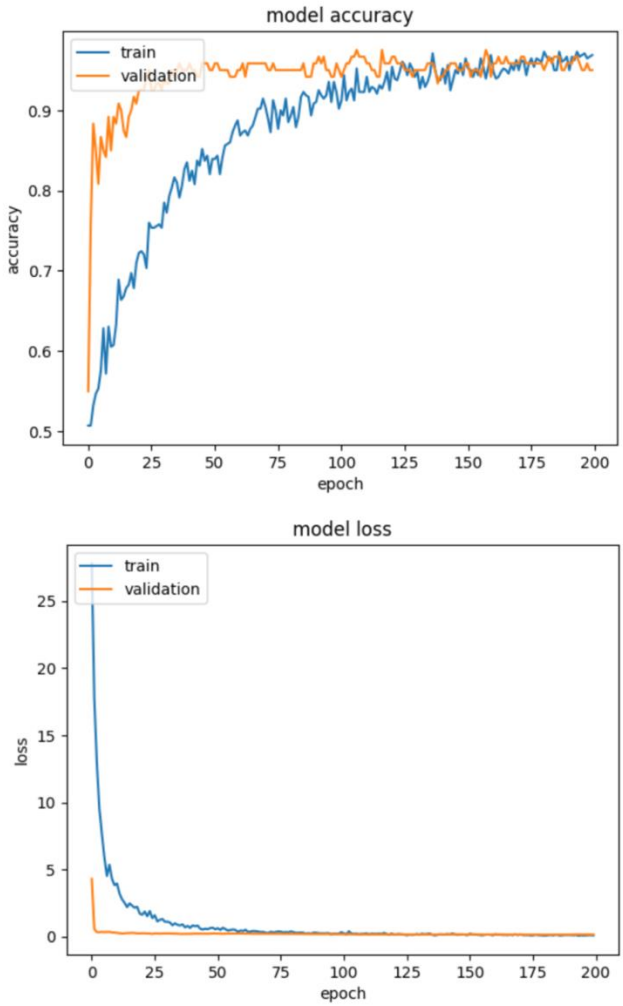
Tablo 4.8. Faz 2 hiperparametreleri

Hiperparametreler	Değerler
Gizli Katman	64
Dropout	0.2
Batch boyutu	64
Öğrenme oranı	0.0005
Eğitim döngüsü (epoch)	200
Optimizasyon	Adam
Kayıp fonksiyonu	Categorical crossentropy

4.2.3. Model değerlendirmesi

Eğitim sonuçları, yüksek doğruluk (accuracy) metrikleriyle desteklenen önemli model performansı iyileştirmelerine işaret etmektedir. Özellikle, eğitim doğruluğu yaklaşık %98.8 seviyesine ulaşırken, doğrulama doğruluğu ise etkileyici bir şekilde %97 olarak gerçekleşmiştir. Eğitim süreci boyunca doğruluk ve kayıp değerlerindeki değişimi görsel olarak sunan grafikler Şekil 4.7'de yer almakta olup, bu grafikler istikrarlı bir

eğitim süreci ve iyileştirilmiş CNN modelinin güçlü genelleme yeteneklerini açıkça ortaya koymaktadır.



Şekil 4.7. CNN modeli eğitiminin doğruluk ve kayıp grafikleri.

Modelin genelleme performansı, eğitim aşamasında kasıtlı olarak dışarıda tutulan, yeni oluşturulmuş veri kümesinden görülmemiş bir alt küme kullanılarak titizlikle değerlendirildi. Bu değerlendirmenin sonuçlarını açıkça göstermek amacıyla, karışıklık matrisi ve ayrıntılı sınıflandırma raporu sırasıyla Tablo 4.9 ve Tablo 4.10'da sunulmuştur. Ayrıca, performans değerlendirmemizin nesnelliğini ve sağlamlığını sağlamak amacıyla, büyük ölçüde YouTube gibi herkese açık kaynaklardan derlenen 85 ses örneğinden oluşan ek bir test seti oluşturulmuştur. Bu kamuya açık dosyalar, doğası gereği gerçekçi gürültü koşulları içermekte olup, modeli zorlayacak sağlam bir test ortamı sağlamıştır. Bu bağımsız test seti üzerinde yapılan değerlendirme sonucunda, model yüksek doğruluk göstermiş ve olay algılamada yaklaşık %1 gibi etkileyici derecede düşük bir hata oranı elde etmiştir.

Tablo 4.9. Faz 2'e ait karışıklık matrisi

	Tahmin Edilen Pozitif	Tahmin Edilen Negatif
Gerçek Pozitif	36	1
Gerçek Negatif	0	30

Tablo 4.10. Faz 2'e ait sınıflandırma raporu

	Keskinlik	Duyarklık	F1 Skoru	Destek
0	1.00	0.97	0.99	37
1	0.97	1.00	0.98	30
Doğruluk			0.99	67
Makro Ort.	0.98	0.99	0.98	67
Ağırlıklı Ort.	0.99	0.99	0.99	67

Modelimizin tahminlerinin güvenilirliğini ve tekrarlanabilirliğini daha da doğrulamak amacıyla K-Fold Cross-Validation tekniği uygulanmıştır (Szeghalmy ve Fazekas, 2023). Bu süreç, 5 katlı çapraz doğrulamanın 10 kez tekrarlanmasıyla toplamda 50 bağımsız eğitim ve test döngüsünün gerçekleştirilmesini içermektedir. Her bir iterasyonda, ilgili performans metrikleri dikkatle kaydedilmiş, analiz edilmiş ve karşılaştırılmıştır. Elde edilen F1-skorları 0.905 ile 0.992 arasında değişmiş, ortalama F1-skoru 0.960 ve standart sapması 0.021 olarak hesaplanmıştır. Bu sonuçlar, Tablo 5.1 ve Tablo 5.2'de özetlenmiş ve karşılaştırmalı analizler için kullanılmıştır. Bu kapsamlı çapraz doğrulama sürecinde elde edilen tutarlılık, modelin aşırı öğrenmeye karşı dayanıklı olduğunu göstermekte ve gerçekçi ses senaryolarına genellenebilirliğini güvenilir bir şekilde ortaya koymaktadır.

5. TARTIŞMA VE SONUÇ

5.1. Tartışma

Literatürde yer alan çeşitli önceki çalışmalar, adım sesi verilerini kullanarak bireylerin kimliğini belirlemeye odaklanmıştır; bu çalışmaların amacı, çevresel sesler arasında adım seslerini ayırt etmekten ziyade, yalnızca adım sesleri üzerinden kişilerin tanımlanması olmuştur. Buna karşın, bu çalışmada sunulan araştırma, çok çeşitli akustik ortamlarda adım sesi olaylarını tespit etmeye yöneliktir ve bu sesleri diğer iç mekân seslerinden etkin bir şekilde ayırt etmeyi hedeflemektedir.

Geliştirilen modelin tespit performansını tarafsız şekilde değerlendirebilmek amacıyla, 85 farklı ses örneğinden oluşan dikkatle hazırlanmış bir değerlendirme veri seti oluşturulmuştur. Bu ses klipleri, özellikle YouTube gibi tanınmış erişime açık platformlardan seçilmiş ve titizlikle derlenmiştir. Bu örneklerin farklı kaynaklardan gelmesi nedeniyle veri seti; çeşitli gürültü desenleri ve akustik koşulları doğal olarak içermektedir. Bu durum, değerlendirme sürecini daha zorlu hale getirmesine rağmen gerçek dünya senaryolarını daha iyi temsil edilmesini sağlamıştır. Bu karmaşıklığa rağmen, geliştirilen model adım sesi olaylarını doğru bir şekilde tespit etmede yaklaşık %1 gibi oldukça düşük bir hata oranı ile dikkate değer bir doğruluk seviyesi sergilemiştir. Bu kadar çeşitli ses girdilerine rağmen düşük hata oranına ulaşılması, yalnızca modelin etkinliğini değil, aynı zamanda farklı ses özellikleri ve ortam koşullarına karşı güçlü bir genelleme yeteneğine sahip olduğunu da ortaya koymaktadır.

Bu bulguları daha sağlam bir şekilde desteklemek amacıyla model, beş katlı ve on tekrar içeren K-Fold Cross-Validation yöntemi ile titizlikle doğrulanmıştır. Bu doğrulama stratejisi, 50 bağımsız eğitim ve değerlendirme döngüsünden oluşmuş ve F1 skorları 0.905 ile 0.992 arasında değişmiştir; ortalama F1 skoru ise 0.960 olarak hesaplanmıştır. Bu kapsamlı çapraz doğrulama süreci, aşırı öğrenme riskinin azaltılmasında önemli bir rol oynamış ve modelin farklı veri alt kümelerine karşı genelleme kabiliyetini güçlü şekilde teyit etmiştir. Ayrıca, test aşamasında kullanılan ses kliplerinin farklı akustik özellikleri ve karmaşıklıkları yansıtacak şekilde seçilmiş

olması, geliştirilen CNN modelinin sağlamlığını ve pratikteki uygulanabilirliğini daha da güçlendirmiştir.

Modelin CNN tabanlı yaklaşımının güçlü yönlerini nesnel biçimde değerlendirmek ve açık bir kıyas ölçütü sunmak amacıyla, ilave karşılaştırmalı deneyler gerçekleştirilmiştir. Bu deneylerde, farklı popüler makine öğrenimi algoritmaları ve alternatif öznitelik çıkarım yöntemleri kullanılmıştır. Bu yaklaşımın iki temel amacı olmuştur: İlki, önerilen CNN modeline yönelik bir taban referans (baseline) sağlamak; ikincisi ise, kullanılan veri setinin ve seçilen öznitelik çıkarım yönteminin bu sınıflandırma problemi için gerçekten uygun olup olmadığını test etmektir. Karşılaştırmalı deneylerin ilk aşaması, MFCC özniteliklerinin, Başlıca Bileşenler Analizi (PCA) ve Mel Spektrogram gibi alternatif yöntemlere kıyasla üstünlüğünü doğrulamaya odaklanmıştır. Ayrıca, Tablo 4.3'te ayrıntılı biçimde sunulan farklı MFCC parametre seçimlerinin etkisi de incelenmiştir. PCA, Mel Spektrogram ve çeşitli MFCC yapılandırmaları kullanılarak özel olarak geliştirilen CNN modelinin performansı sistematik biçimde karşılaştırılmış ve sonuçlar Tablo 5.1'de açıkça özetlenmiştir. Bu kapsamlı karşılaştırma süreci, seçilen MFCC özniteliklerinin ve parametre ayarlarının sağlamlığını, güvenilirliğini ve belirgin avantajlarını etkili biçimde ortaya koymuştur.

Tablo 5.1. Farklı öznitelikler ile karşılaştırmalar

Yöntem	Kesinlik	Duyarlık	F1 Skoru
CNN- PCA	0.92	0.907	0.911
CNN- Mel Spektogram	0.925	0.943	0.944
CNN- MFCCs	0.951	0.952	0.954
Bu çalışmadaki model ve öznitelikler	0.959	0.968	0.960

İkinci deneysel değerlendirme serisinde, özel olarak geliştirdiğimiz CNN modeli; ses işleme ve sınıflandırma alanında yaygın olarak tanınan birçok ileri düzey derin öğrenme mimarisiyle sistematik olarak karşılaştırılmıştır. Bu mimariler arasında ResNet50, ResNet101, EfficientNetB0, EfficientNetB4 ve Konvolüsyonel Tekrarlayan Sinir Ağı (CRNN) yer almaktadır. Özellikle CRNN mimarisi, ses sinyallerinin doğasında bulunan zamansal ve mekânsal bağımlılıkları etkin biçimde modelleyebilme kabiliyeti sayesinde son yıllarda büyük ilgi görmüştür.

Kapsamlı bir karşılaştırmalı analiz gerçekleştirebilmek için, performans değerlendirmesinde Doğruluk (Precision), Duyarlılık (Recall) ve F1-skoru gibi temel metrikler kullanılmıştır. Bu değerlendirmelerden elde edilen ayrıntılı sonuçlar Tablo 5.2'de açıkça özetlenmiştir ve önerilen CNN mimarimizin istikrarlı ve dikkat çekici bir üstünlük sergilediğini ortaya koymuştur.

Her ne kadar CRNN ve ResNet50 modelleri bazı durumlarda benzer ya da zaman zaman yaklaşan performanslar göstermiş olsa da, özel CNN modelimiz tüm değerlendirme metriklerinde genel olarak daha yüksek skorlar üretmiştir. Bu ayrıntılı analiz, ses tabanlı sınıflandırma görevlerinde karşılaşılan zorlukları etkin bir şekilde ele alma konusunda CNN modelimizin belirgin avantajını ve rekabetçi gücünü net bir şekilde ortaya koymaktadır.

Tablo 5.2. Modern mimariler ile karşılaştırma

Yöntem	Kesinlik	Duyarlık	F1 Skor
ResNet50	0.961	0.955	0.959
ResNet101	0.721	0.78	0.754
EfficientNetB0	0.512	0.537	0.526
EfficientNetB4	0.923	0.931	0.93
CRNN	0.963	0.952	0.955
Bu çalışmadaki model	0.959	0.968	0.960

Ayrıca, literatürdeki çalışmaların büyük bir kısmı adım sesi analizi yoluyla bireylerin tanınmasına (kişi tanıma) odaklanmış olsa da, konu ve yöntem açısından çalışmamıza en yakın araştırmaları temsil etmektedir. Bu çalışmaların adım sesi temelli kişi tanımlama yaklaşımı, çalışmamızla bazı benzerlikler taşısa da, bu yöntemlerin mülk güvenliğini sağlamada yeterince etkili olmadığı özellikle vurgulanmalıdır. Bu sınırlılık büyük ölçüde, kişi tanımlamaya yönelik veri kümelerinin genellikle az sayıda bireyden (örneğin Itai ve ark. (2008) çalışmasında 10 kişi, Algermissen ve Hörnlein (2021) çalışmasında 5 kişi) oluşmasından kaynaklanmaktadır. Gerçekçi güvenlik senaryolarında yetkisiz bir kişinin adımlarının yetkili kişilerin adım seslerinden ayırt edilebilmesi için, kişi tanıma modellerinin çok daha geniş ve çeşitli bir birey grubuna ait ses örnekleriyle eğitilmiş olması gerekmektedir. Dolayısıyla bu tür çalışmaların, mülk güvenliği amacıyla doğrudan uygulanabilirliği sınırlıdır.

Bu nedenle, küçük gruplar üzerinde eğitilen kişi tanımlama odaklı modellerin, daha büyük ve karmaşık ortamlarda, tanıdık olmayan bireyleri de içeren durumlara genellenebilir olduğu varsayımı geçerli değildir. Buna ilaveten Bölüm 4.2'de ReaLISED veri kümesindeki "walking" sınıfına ilişkin veri kalitesi sorunlarına değinilmişti. Özellikle "walking" sınıfındaki ses örneklerinin birçoğu, yalnızca adım sesi değil; koşma, merdiven çıkma, ahşap zeminin gıcırdaması gibi belirsiz ve alakasız sesleri de içermektedir. Buna karşılık, bu çalışmadaki yaklaşım veri kümesini titizlikle filtreleyerek, alakasız veya gürültülü örnekleri çıkarmış ve böylece sadece adım sesi tespiti açısından daha doğru, güvenilir ve geçerli değerlendirme sonuçları elde etmiştir.

Son olarak, teorik bulgularımızı ve değerlendirmelerimizi pratik bir bağlama dönüştürmek amacıyla, gerçekçi bir mülk güvenliği senaryosu önerilmiştir. Bu senaryoda, geliştirilen ses tabanlı tespit sistemi, olağandışı veya şüpheli bir faaliyet algıladığında ev sahibine otomatik olarak anlık bildirim göndermektedir. Bu sayede kullanıcılar sürekli olarak bilgilendirilmekte ve zamanında müdahale imkânı sağlanmaktadır. Ayrıca, önerilen sistem mimarisi uzaktan kontrol özelliklerini desteklemekte, böylece kullanıcıların sistemi kolayca yönetip özelleştirmelerine olanak tanımaktadır. Tanımlanan bu yapı, uç birimlerde (edge computing) ya da bulut tabanlı sistemlerde uygulanabilecek esneklikte ve uyarlanabilirliktedir. Ancak, bu çerçevenin temelinde yer alan ve sistemin kritik işlevselliğini oluşturan unsur, adım sesi tespitinin güvenilirliğidir ve bu yetenek çalışmada ayrıntılı şekilde doğrulanmıştır.

Bu senaryoda, mülkün farklı noktalarına stratejik olarak yerleştirilen birden fazla mikrofon sensörü, ortamdaki akustik verileri sürekli olarak toplar ve bu verileri merkezi bir hub cihazına iletir. Hub cihazı; gelen ses sinyallerini anlık olarak işler ve adım sesi tespit modeline aktarır. Model, anlık olarak adım sesi tespitine dair sonuç verir. Hub, bu sonuçları bir ağ yönlendiricisi aracılığıyla özel bir kontrol sistemine aktarır. Bu kontrol sistemi, tespit algoritmasından bağımsız olarak tasarlanmış olup, esasen kullanıcı arayüzlerinden ve bildirimler gibi çeşitli yapılandırma ayarlarından oluşur. Bu ayarlar arasında kullanıcıların aktif ve pasif zaman aralıklarını belirleyebilmesi de bulunmaktadır. Kullanıcının belirlediği aktif zaman dilimlerinde hub cihazından gelen anlık durumlar değerlendirilir. Bu bildirimlerin ev sahibine gecikmeden iletilmesiyle, zamanında müdahale ve güvenliğin sağlanması hedeflenmektedir. Özetle, ses sensörleri (mikrofonlar) sadece ses verisini algılar ve

hub cihazına aktarırlar. Hub cihazı bu ses sinyallerini toplar, işler, adım sesi tespit modeliyle değerlendirir ve sonucu kontrol merkezine aktarır. Kontrol merkezi, ev sahibinin bazı ayarlamaları yapmasına imkan verirken bildirim merkezi olarak çalışır ve ev sahibine bildirim gönderir.

Adım sesi tespit sistemleri, yalnızca mülk güvenliği değil, aynı zamanda çok çeşitli disiplinlerde pratik uygulamalara sahiptir. Örneğin, koruyucu sağlık alanında bu teknoloji; yaşlı bireylerin evde takibi, olası düşme veya hareket anormalliklerinin zamanında tespiti gibi senaryolarda kritik bir rol oynayabilir. Akıllı bina ve ev otomasyon sistemlerinde ise bireyin varlığını sessiz ve enerji verimli bir şekilde algılayarak aydınlatma, ısıtma veya güvenlik sistemlerinin dinamik olarak kontrol edilmesine olanak tanır. Askerî ve savunma uygulamalarında, görsel sistemlerin yetersiz kaldığı durumlarda yaklaşan kişi tespiti için düşük enerji tüketimli, pasif bir gözetleme aracı olarak görev alabilir. Benzer şekilde, iş yeri güvenliğinde, yetkisiz saatlerde gerçekleşen hareketlerin algılanması ya da güvenlik personelinin devriye davranışlarının takibi gibi durumlarda ses tabanlı çözümler, operasyonel verimliliği ve güvenliği artırabilecek önemli bir tamamlayıcı unsur olarak öne çıkmaktadır.

Bu sistemin pratikte uygulanabilirliği için kullanılacak hub cihazının hem donanımsal hem de yazılımsal olarak belirli yeteneklere sahip olması gereklidir. Donanımsal olarak, Raspberry Pi 4 veya NVIDIA Jetson Nano gibi en az 4 GB RAM'e sahip, tercihen GPU hızlandırması bulunan bir mikrodenetleyici önerilir. MFCC çıkarımı ve model çalıştırma işlemlerini gerçek zamanlı olacak şekilde 1-2 saniye içinde tamamlayabilmelidir. Böyle bir sistem, uç hesaplama (edge computing) prensiplerine dayalı olarak yerel işlem gücüyle yüksek gecikme riskini azaltır ve ağ trafiğini minimumda tutar. Sistemde kullanılacak mikrofon sensörlerinin yüksek hassasiyete ve düşük taban gürültüsüne sahip olması kritik öneme sahiptir. Bu sensörlerin dijital çıkışlı olması, ses sinyallerinin doğrudan sayısal olarak işlenmesine olanak tanıyarak veri kaybını azaltır. Minimum 16 bit çözünürlükte ve 44.1 kHz veya daha yüksek örnekleme hızına sahip mikrofonlar, insan adımlarının akustik detaylarını yeterli doğrulukla yakalayabilmek için tercih edilmelidir. Ayrıca, geniş alanlarda adım seslerini doğru biçimde algılayabilmek amacıyla yönlü (directional) mikrofonlar veya çoklu mikrofon dizilimi (mikrofon array) kullanılması önerilmektedir. Bu yaklaşım, ses kaynağının yönünü tespit edebilme kapasitesi kazandırarak yanlış sınıflandırmaları azaltır ve tespit doğruluğunu artırır. Sistem bileşenleri arasındaki iletişimin kesintisiz

sağlanabilmesi için TCP/IP protokolünü destekleyen, güvenilir bir ağ yönlendiricisi kullanılmalıdır. Mikrofonlardan gelen veriler önce hub'a, oradan ise sonuçları kontrol sistemine iletilirken düşük gecikme ve yüksek güvenlik hedeflenmelidir. Kontrol sistemi, kullanıcı arayüzlerini barındıran bir web sunucusu ve bildirim altyapısı ile desteklenmelidir. Bu amaçla, Node.js tabanlı bir arka uç yazılımı ile Firebase veya benzeri bir servis kullanılarak mobil cihazlara anlık bildirim gönderimi gerçekleştirilebilir. Bu yapı sayesinde kullanıcılar, tanımladıkları aktif zaman dilimlerinde gelen olası adım seslerine dair bildirimleri gecikmeden alabilir ve duruma hızlı müdahale edebilir.

5.2. Kısıtlar ve Sonuç

5.2.1. Kısıtlar

Bu çalışmada yapılan deneylerde elde edilen etkileyici sonuçlara rağmen, özellikle veri toplama süreci ve önerilen modelin sonuçlarının yorumlanmasıyla ilgili bazı kısıtların dikkatle ele alınması gerekmektedir. Bu çalışmanın başlıca kısıtlardan biri, kullanılan ses örneklerinin doğası ve özelliklerinden kaynaklanmaktadır. Özellikle, eğitim ve değerlendirme amacıyla kullanılan çevresel ses kayıtları ağırlıklı olarak kapalı alanlardan elde edilmiştir. Kapalı ortamlar, genellikle arka plan gürültüsünün daha az olduğu ve öngörülemeyen akustik olayların daha seyrek gerçekleştiği akustik koşullara sahiptir. Bu durum, geliştirilen modelin dış ortam koşullarında —örneğin rüzgâr, trafik gürültüsü ya da çeşitli yankı etkileri gibi değişken çevresel faktörlerin bulunduğu durumlarda— henüz yeterince değerlendirilmediği veya doğrulanmadığı anlamına gelmektedir. Tespit sistemimizin gerçek dünya koşullarında daha geniş bir yelpazeye uygulanabilirliğini sağlamak adına, gelecekteki çalışmaların veri kümesini dış ortam seslerini de içerecek şekilde genişletmesi ve modelin bu daha zorlu akustik ortamlardaki genelleme yeteneğini değerlendirmesi gerekmektedir.

Bir diğer kısıt ise, veri kümesinde kullanılan adım sesi örneklerinin çeşitliliğiyle ilgilidir; zira bu örnekler, belirgin ve nispeten daha net adım sesi olaylarını temsil etmektedir. Bu çalışma, farklı ayakkabı türleri ile zemin materyalleri arasındaki etkileşimlerin ses üzerindeki etkilerini dikkate almamıştır. Oysa bu tür etkileşimler, adım sesinin akustik özelliklerini önemli ölçüde değiştirebilir ve örneğin adım sesinin boğuk ya da zayıf duyulmasına neden olabilir. Örneğin, bir bireyin tahta zemin, halı, beton ya da çim üzerinde yürümesi veya yumuşak ya da sert tabanlı ayakkabı giymesi,

ortaya çıkan adım sesi imzasını büyük ölçüde etkileyebilir. Ayakkabı tabanı ve zemin yüzeyi arasındaki bu etkileşimlerin çeşitliliği, mevcut modelin doğrulama kapsamını sınırlandırmaktadır. Dolayısıyla, adım sesi sinyallerinin çeşitli yüzeyler ve ayakkabı türleri nedeniyle bozulduğu, maskelendiği ya da önemli ölçüde değiştiği durumlarda modelin etkinliği belirsizliğini korumaktadır. Bu nedenle, gelecekteki çalışmalar bu etkenlerin etkisini sistematik olarak incelemeli ve farklı yüzey türleri ile ayakkabı malzemelerini kapsayan çeşitli ses örnekleri içermelidir. Bu kısıtların giderilmesi, adım sesi tespit modelinin genel dayanıklılığını, güvenilirliğini ve gerçek dünya uygulamaları açısından kullanılabilirliğini önemli ölçüde artıracaktır.

Bir önceki bölümde sunulan pratikte uygulanabilecek sistemin karşılaşabileceği başlıca kısıtlar arasında yüksek ortam gürültüsü ve veri güvenliği sayılabilir. Gürültülü ortamlarda yanlış pozitif sonuçlar artabilirken, sürekli dinleme yapan mikrofonlar veri mahremiyeti açısından risk oluşturabilir; bu nedenle şifreleme ve güvenlik önlemleri şarttır. Bununla birlikte sistem, modüler yapısı sayesinde çok sayıda mikrofona kolayca ölçeklenebilir ve farklı ortamlarda (örneğin okul, hastane, otopark, fabrika gibi) uygulanabilirlik açısından değerlendirilebilir. Böylece sistem, hem bireysel hem kurumsal güvenlik senaryoları için genişletilebilir bir çerçeve sunar.

5.2.2. Sonuç

Bu çalışma, geleneksel görsel tabanlı güvenlik sistemlerine alternatif veya tamamlayıcı olarak ses tabanlı yaklaşımların uygulanabilirliği ve etkinliği konusunda güçlü kanıtlar sunmaktadır. Önerilen ses odaklı yöntem, görsel gözetim teknolojilerine kıyasla daha basit ve kompakt kurulum gereksinimleri ile birlikte, donanım maliyetlerinin önemli ölçüde azaltılması gibi önemli pratik avantajlar sağlamaktadır. Özellikle, geliştirilen ses olay tespit modeli, Derin Öğrenme tekniklerinden yararlanan CNN mimarisi üzerine kurulmuş olup, yaklaşık %98 sınıflandırma doğruluğu ve yaklaşık %1 gibi son derece düşük bir hata oranı ile dikkat çekici bir başarı elde etmiştir. Bu performans ölçümleri, 17 farklı çevresel ses kategorisini kapsayan çeşitli veri kümeleri üzerinde gerçekleştirilen titiz değerlendirmeler sonucunda elde edilmiştir ve modelin ayırt etme kapasitesi ile güvenilir sınıflandırma yeteneğini ortaya koymaktadır. Ayrıca, önerilen CNN modelinin dayanıklılığı ve güvenilirliği, çoklu yinelemelerle gerçekleştirilen Repeated Stratified K-Fold Cross-Validation yöntemi ile kapsamlı biçimde doğrulanmıştır. Bu titiz doğrulama süreci sonucunda, model ortalama 0.960 F1-Skoru ile istikrarlı ve yüksek performans sergilemiştir. Bu

tür kapsamlı çapraz doğrulama metodolojisi, modelin aşırı öğrenme (overfitting) eğilimini azaltarak, önerilen ses tabanlı tespit yaklaşımının daha önce karşılaşılmamış ve çeşitli akustik ortamlara başarıyla genellenebileceğine dair güçlü ampirik kanıtlar sunmaktadır. Ayrıca, güvenilir ve yaygın olarak bilinen platformlardan elde edilen açık kaynaklı ses veri kümelerinin modele dâhil edilmesi, çalışmanın şeffaflığını, yeniden üretilebilirliğini ve gerçek koşullara uygulanabilirliğini büyük ölçüde artırmıştır. Bu açık kaynak ses kayıtlarının çeşitliliği, modelin karmaşık, gürültülü ya da daha önce görülmemiş ses örnekleriyle karşılaştığında da yüksek performans göstermesini sağlayarak pratik geçerliliğini pekiştirmiştir.

Sonuç olarak, çalışmanın bulguları, ses tabanlı tespit sistemlerinin geleneksel görsel gözetim teknolojilerini ya tamamlayabileceğini ya da doğrudan onların yerini alabileceğini güçlü bir şekilde ortaya koymaktadır. Görsel gözetim sistemlerinin yüksek maliyet, görüş alanı kısıtlamaları ve sabotaja karşı savunmasızlık gibi yaygın zorluklarını başarılı şekilde ele alan ses tabanlı modelimiz, mülk güvenliğini artırmak için son derece pratik, maliyet etkin ve güvenilir bir çözüm olarak öne çıkmaktadır. Bu sonuçlar, ses tabanlı güvenlik çözümleri üzerine yapılacak daha fazla araştırma ve geliştirme çalışmalarına güçlü bir motivasyon sağlamakta olup, farklı alanlardaki güvenlik uygulamaları için daha geniş çapta benimsenmenin önünü açmaktadır.

KAYNAKLAR

Algermissen, S., ve Hörnlein, M. (2021). Person identification by footstep sound using convolutional neural networks. Applied Mechanics, 2(2), 257–273. https://doi.org/10.3390/applmech2020016

Al-Khalli, N., Alateeq, S., Almansour, M., Alhassoun, Y., Ibrahim, A. B., & Alshebeili, S. A. (2023). Real-time detection of intruders using an acoustic sensor and internet-of-things computing. Sensors, 23(13), 5792. https://doi.org/10.3390/s23135792

Alpaydin, E. (2014). Introduction to machine learning (3rd ed.). MIT Press.

Ball, S. J. (2022). Urban and Regional Planning and Real Estate Development. Routledge.

Bonet-Solà, D., ve Alsina-Pagès, R. M. (2021). A comparative survey of feature extraction and machine learning methods in diverse acoustic environments. Sensors, 21(4), 1274. https://doi.org/10.3390/s21041274

Bowles, S., ve Choi, J. K. (2019). The Neolithic Agricultural Revolution and the Origins of Private Property. Journal of Political Economy, 127(5), 2186–2228. https://doi.org/10.1086/701789

Cai, F. S., Philipson, D., & Syed, S. (2010, December 13). A step-by-step approach to footstep detection. Stanford University CS229: Machine Learning Course Projects. Retrieved May 11, 2025, from https://cs229.stanford.edu/proj2010/CaiPhilipsonSyed-FootstepDetection.pdf

Cai, P., Song, Y., Jiang, N., Gu, Q., & McLoughlin, I. (2025, April). Prototype based masked audio model for self-supervised learning of sound event detection. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE. https://doi.org/10.48550/arXiv.2409.17656

Cakir, E. (2019). Deep neural networks for sound event detection [Doctoral dissertation]. Tampere University.

Cakir, E., Heittola, T., Huttunen, H., & Virtanen, T. (2015). Polyphonic sound event detection using multi-label deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(1), 66–79. https://doi.org/10.1109/IJCNN.2015.7280624

Costin, A. (2016). Security of CCTV and video surveillance systems: Threats, vulnerabilities, attacks, and mitigations. In Proceedings of the 6th International Workshop on Trustworthy Embedded Devices (TrustED '16) (pp. 45–54). Association for Computing Machinery. https://doi.org/10.1145/2995289.2995290

Davis, S., ve Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

DeLoney, C. (2008). Person identification and gender recognition from footstep sound using modulation analysis.

Dh, S. Y., ve Ansari, A. (2020). Autonomous vehicles camera blinding attack detection using sequence modelling and predictive analytics. SAE Technical Paper, 2020-01-0123. https://doi.org/10.4271/2020-01-0719

Diro, A. A., ve Chilamkurti, N. (2017). Distributed attack detection scheme using deep learning approach for Internet of Things. Future Generation Computer Systems, 82, 761–768. https://doi.org/10.1016/j.future.2017.08.043

ENISA. (2024). ENISA Threat Landscape 2024. European Union Agency for Cybersecurity. Retrieved from https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024 on May 7, 2025.

Fennelly, L. J. (2017). Effective physical security (5th ed.). Butterworth-Heinemann.

Gencoglu, O., Virtanen, T., & Huttunen, H. (2014). Recognition of acoustic events using deep neural networks. Proceedings of the 22nd European Signal Processing Conference (EUSIPCO 2014), 506–510. https://ieeexplore.ieee.org/document/6952140

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 6645–6649). IEEE. https://doi.org/10.1109/ICASSP.2013.6638947

Griffin, A., Hirvonen, T., Tzagkarakis, C., Mouchtaris, A., & Tsakalides, P. (2011). Single-channel and multi-channel sinusoidal audio coding using compressed sensing. IEEE Transactions on Audio, Speech, and Language Processing, 19(5), 1382–1395. https://doi.org/10.1109/TASL.2010.2090656

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

He, H., ve Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. IEEE International Conference on Computer Vision (ICCV 2015) Proceedings, 1026–1034. https://doi.org/10.1109/ICCV.2015.123

Heittola, T. (2013). Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1), 1–13. https://doi.org/10.1186/1687-4722-2013-1

Heller, P. N., Karp, T., & Nguyen, T. Q. (1999). A general formulation of modulated filter banks. IEEE Transactions on Signal Processing, 47(4), 986–1002. https://doi.org/10.1109/78.752597

Hourri, S., ve Kharroubi, J. (2020). A deep learning approach for speaker recognition. International Journal of Speech Technology, 23(1), 123–131. https://doi.org/10.1007/s10772-019-09665-y

Itai, A., ve Yasukawa, H. (2008, May). Footstep classification using simple speech recognition technique. In 2008 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 3234–3237). IEEE. https://doi.org/10.1109/ISCAS.2008.4542147

Kaggle. (2025). Kaggle: Your home for data science. Retrieved May 16, 2025, from https://www.kaggle.com

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. https://doi.org/10.1145/3065386

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

Liu, X., ve Kong, H. (2021, June). Research on footstep recognition method based on HMM and MFCC. In Proceedings of the 2021 IEEE 5th Information Technology and Networking, Electronic and Automation Control Conference (ITNEC)(pp. 157–160). IEEE. https://doi.org/10.1109/ITNEC52019.2021.9587306

Maloof, M. A. (2006). Machine learning and data mining for computer security: Methods and applications. Springer.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In S. Doré, J. K. VanderPlas, & G. Varoquaux (Eds.), Proceedings of the 14th Python in Science Conference (SciPy 2015, pp. 18–25). Austin, TX. https://doi.org/10.25080/Majora-7b98e3ed-003

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. Information Fusion, 99, 101869. https://doi.org/10.48550/arXiv.2305.00359

Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021). Sound event detection: A tutorial. IEEE Signal Processing Magazine, 38(5), 67-83. https://doi.org/10.1109/MSP.2021.3090678

Mohino-Herranz, I., García-Gómez, J., Aguilar-Ortega, M., Utrilla-Manso, M., Gil-Pita, R., & Rosa-Zurera, M. (2022). Introducing the ReaLISED dataset for sound event classification. Electronics, 11(12), 1811. https://doi.org/10.3390/electronics11121811

Mottakin, K., Guo, J., & Song, Z. (2024). Cover More with Less: Eliminating Blind Spots for Surveillance Camera via Passive WiFi Sensing. In Proceedings of the 30th IEEE International Conference on Parallel and Distributed Systems (ICPADS 2024) (pp. 84–91). https://doi.org/10.1109/ICPADS63350.2024.00021

Nair, V., ve Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (pp. 807–814).

Nakadai, K., Fujii, Y., & Sugano, S. (2013, July). Footstep detection and classification using distributed microphones. In 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS) (pp. 1–4). IEEE. https://doi.org/10.1109/WIAMIS.2013.6616127

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061. https://doi.org/10.48550/arXiv.2010.16061

Radkau, A. (2008). Nature and Power: A Global History of the Environment. Cambridge University Press.

Savills Research. (2022). Global real estate report 2022. Savills. Retrieved from https://www.savills.com on May 6, 2025.

Savills World Research. (2016). The world's real estate assets 2016. Savills. Retrieved from https://www.savills.com/ on May 6, 2025.

Shams, J., Nassi, B., Koda, S., Shabtai, A., & Elovici, Y. (2025). A privacy enhancing technique to evade detection by street video cameras without using adversarial accessories. arXiv. https://doi.org/10.48550/arXiv.2501.15653

Sokolova, M., ve Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 111–147.

Summoogum, K., ve Palaniappan, R. (2021, November). Acoustic based footstep detection in pervasive healthcare. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 4974–4977). IEEE. https://doi.org/10.1109/EMBC46164.2021.9630125

Szeghalmy, S., ve Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors, 23(4), 2333. https://doi.org/10.3390/s23042333

van der Maaten, L., ve Hinton, G. (2008). Visualizing data using t-SNE. In Proceedings of the 25th International Conference on Machine Learning (pp. 259–268). Association for Computing Machinery.

Zieger, C., Brutti, A., & Svaizer, P. (2009, September). Acoustic based surveillance system for intrusion detection. In Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009)(pp. 314–319). IEEE. https://doi.org/10.1109/AVSS.2009.49

ÖZGEÇMİŞ

Ad-Soyad : Furkan Yusuf YAVUZ

ÖĞRENİM DURUMU:

- **Lisans :** 2015, Yıldız Teknik Üniversitesi, Elektrik Mühendisliği
- **Yükseklisans :** 2018, İstanbul Şehir Üniversitesi, Bilgi Güvenliği Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

- 2015-2018 yılları arasında TÜBİTAK BİLGEM'de araştırmacı olarak çalıştı.
- 2018-2019 yılları arasında BMW Group'da yazılım test otomasyon mühendisi olarak çalıştı.
- 2020-2024 yılları arasında Recognyte'da yazılım test mühendisi olarak çalıştı.

TEZDEN TÜRETİLEN ESERLER:

- Yavuz, F. Y., & Yumusak, N. (2025). Utilising Footstep Sound Event Detection by Using CNN Techniques for Assuring Property Security. IEEE Access.