NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2024; 13(4), 1207-1215
Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences
Araştırma makalesi / Research article
www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

TinyML tabanlı görsel işitsel anahtar kelime tespiti

TinyML based audio visual keyword detection

Mehmet Tosun[1,*], Hamit Erdem[2]

[1,2] *Başkent Üniversitesi, Elektrik & Elektronik Mühendisliği Bölümü, 06790, Ankara Türkiye*

Öz

Anahtar kelime tespiti (AKT), makine öğreniminin kullanıldığı alanlardan birisidir. Amacı, ses veya görüntü verisinden belirli kelime veya objenin otomatik tespit edilmesidir. Taşınabilir yapay zekâ uygulamalarının artmasıyla beraber, bu alanda da uygulamalar artmaktadır. Özellikle AKT uygulamalarının etkinliğini artırmak için hibrit sistemler (ses ve görüntünün birlikte kullanımı) üzerinde çalışma yapılmaktadır. Bu sistem ile birlikle iki farklı kanaldan algılanan ses ve görüntü komutlarının birleştirilmesi amaçlanmaktadır. Bilgisayar (PC) ortamında görsel işitsel AKT üzerinde birçok çalışma yapılmış ve iyi sonuçlar elde edilmiştir. Diğer taraftan derin öğrenme uygulamalarını düşük kapasiteli işlemciler üzerinde gerçekleştirmek için TinyML (Düşük Kapasiteli Makine Öğrenmesi) kapsamında çalışmalar yapılmaktadır. Bu uygulamalarda, derin öğrenmeye yönelik geliştirilen modelin parametrelerini azaltarak (nicelleştirme, kırpma) sıradan mikrodenetleyici üzerinde uygulama imkânı oluşturmaktadır. Bu çalışmada ses ve görüntü verisi kullanılarak, TinyML alanında AKT uygulaması önerilmiştir. Önerilen hibrit modelin eğitiminde öncelikle ses ve görüntü modelleri Edge Impulse yazılım ortamında ayrı ayrı eğitilmiştir. Geliştirilen MobileNetV2 ve CNN tabanlı modeller ESP32-CAM ve Arduino Nano BLE geliştirme kitlerine yüklenerek, denenmiştir. Daha sonra modeller doğrusal ağırlıklı birleştirme metodu ile birleştirilerek denenmiştir. Sistemin başarısı standart ölçütlere göre test edilmiştir. Deneysel sonuçlarda doğruluk ölçütüne göre, sadece ses tabanlı AKT başarısı %85, sadece görüntü tabanlı AKT başarısı %85 olurken, görsel işitsel hibrit uygulamasında sınıflandırma başarısı %90 civarında olmuştur.

Anahtar kelimeler: TinyML, Makine öğrenimi, Anahtar kelime tespiti, Evrişimsel sinir ağları, Edge impulse

Abstract

Keyword detection (KWD) is one of the areas where machine learning is used. Its purpose is the automatic detection of specific words or objects from audio or image data. As portable artificial intelligence applications become more prevalent, the number of applications in this field is also growing. In particular, hybrid systems (the use of audio and video together) are being studied to increase the effectiveness of KWD applications. The system aims to combine audio and visual commands detected through two different channels. Extensive work has been done on audiovisual keyword detection in a computer environment, yielding good results. On the other hand, efforts are being made within the scope of TinyML (Low-Power Machine Learning) to implement deep learning applications on low-capacity processors. In these applications, reducing the parameters of the deep learning model (quantization, pruning) makes it possible to implement the model on ordinary microcontrollers. In this study, a keyword detection application in the field of TinyML is proposed using audio and visual data. In the training of the proposed hybrid model, the audio and visual models were first trained separately in the Edge Impulse software environment. Developed MobileNetV2 and CNN-based models were loaded onto ESP32-CAM and Arduino Nano BLE development kits and tested. Subsequently, the models were combined using a linear weighted fusion method and tested. In the experimental results, according to the accuracy criterion, the success rate of the audio-based KWD was 85%, the success rate of the image-based KWD was 85%, while the classification success in the audiovisual hybrid application was around 90%.

Keywords: TinyML, Machine learning, Keyword spotting, Convolutional neural network, Edge impulse

1 Giriş

Günümüzün hızla dijitalleşen dünyasında yapay zekâ ve derin öğrenme, teknoloji alanında devrim niteliğinde ilerlemeler sağlayarak pek çok sektörde köklü değişikliklere yol açmaktadır. Yapay zekâ, karmaşık problemleri çözme, büyük veri setlerinden anlamlı sonuçlar çıkarma ve insan benzeri kararlar alabilme yetenekleriyle öne çıkarken, derin öğrenme bu yetenekleri daha da ileriye taşıyan bir alt dal olarak kendini göstermektedir [1]. Bu alandaki ilerlemeler, birçok sektörde devrim yaratma potansiyeline sahiptir. Özellikle son yıllarda, derin öğrenme algoritmalarındaki iyileşmeler sayesinde, makine öğrenmesi modelleri giderek daha insan benzeri zekâ sergilemeye başlamıştır.

Makine öğrenimi, bilgisayar sistemlerinin, açık bir programlama olmadan öğrenme ve performanslarını iyileştirme yeteneği olarak tanımlanabilir [2]. Bu teknoloji, büyük miktardaki veriden kalıplar ve eğilimler tespit ederek, sistemlerin çeşitli görevleri daha etkin bir şekilde yerine getirmesine olanak sağlar. Son yıllarda, makine öğrenmesi

* Sorumlu yazar / Corresponding author, e-posta / e-mail: mehmet_tsn.96@outlook.com (M. Tosun)
Geliş / Recieved: 11.05.2024 Kabul / Accepted: 30.07.2024 Yayımlanma / Published: 15.10.2024
doi: 10.28948/ngmuh.1482481

yöntemleri hızla gelişmiş ve birçok yeni uygulama alanı ortaya çıkmıştır. Bunlardan biri de TinyML'dir.

TinyML, makine öğrenmesi modellerinin mikro-kontrolörler dahil olmak üzere düşük işlemci gücüne ve hafızaya sahip cihazlarda çalıştırılmasını ifade eder [3]. Genellikle 1 miliwatt veya daha az enerji tüketimi olan cihazlarda kullanılan bu modeller, her yerde bulunan cihazları ve fiziksel dünyayı daha akıllı hale getirebilir [4]. Literatürde yeni bir yaklaşım olan TinyML teknolojisi birçok araştırmacı tarafından da incelenmektedir.

Altayeb ve arkadaşları (2023), çalışmalarında Edge Impulse platformunu kullanarak bir SiPM mikro radyasyon sensörü ve TinyML ile otomatik bir gama radyasyon algılama ve tanımlama sistemi tasarlamışlardır. Çalışmalarında kullanılan model, gerçek gama kaynağı verileri kullanılarak eğitilmiştir. Model olarak CNN kullanılmıştır. Yapılan testler sonucunda gerçek zamanlı işlemede %90 doğruluk elde etmişlerdir [5]. Mansoureh (2021), çalışmasında çamaşır makinesinin arıza tespiti için TinyML kullanmıştır. Donanım olarak Arduino Nano 33 BLE kullanmıştır ve Arduino bir çamaşır makinesine montaj edilmiştir. Kartta bulunan ivme sensörü çamaşır makinesi verileri ile arıza tespit çalışması yapmıştır. Çalışmada eğitim modeli olarak CNN kullanılmıştır. Çalışmasında %92'lik başarı elde etmiştir [6]. GRAU (2021), çalışmasında TinyML teknolojisine genel bir bakış yapıp, Arduino 33 Ble ile anahtar kelime tespiti yapmıştır. Çalışmasında yes ve no işitsel verileri ile anahtar kelime tespiti üzerinde çalışmıştır. Eğitim modeli olarak CNN kullanmıştır. Sonuç olarak modelde %94 eğitim doğruluğu elde edilmiştir [7].

TinyML teknolojisinin kullanım alanlarından birisi de anahtar kelime tespit işlemidir. AKT, konuşulan cümleleri veya metin parçalarını analiz ederek belirli anahtar kelimeleri otomatik olarak algılayan bir uygulamadır [8]. Önceki çalışmalar genellikle işitsel tabanlı anahtar kelime tespiti üzerine odaklanmıştır [9,10]. Ancak bu yaklaşımlar gürültülü ortamlarda başarı sağlamada zorluklar yaşatabilmektedir.

Görsel işitsel anahtar kelime tespiti hem görüntü hem de işitsel verilerini birlikte analiz ederek, sadece işitsel verilerine dayanan geleneksel anahtar kelime tespitinden daha yüksek doğruluk ve güvenilirlik sağlamaktadır. Görsel işitsel anahtar kelime tespiti uygulamaları bilgisayar tabanlı başarılı uygulamalar olmasına rağmen, mikrodenetleyici tabanlı TinyML uygulamalarında yapılmamıştır.

Bu çalışmada TinyML kullanarak görsel işitsel anahtar kelime tespit çalışması yapılmıştır. Çalışmada sırasıyla aşağıdaki işlemler uygulanmıştır.

- Ses ve görsel ifadeler için ayrı ayrı veri seti kullanılmıştır. Ses verileri için Yes-No, görsel veriler için Açık El- Kapalı El veri seti kullanılmıştır.
- Ses ve görüntü modelleri Edge Impulse uygulama programında ayrı ayrı eğitilerek, işlemcilere yüklenmiş ve test edilmiştir.
- Hibrit uygulamanın başarısını izlemek için PC ortamında GUI geliştirilmiştir. Ses ve görüntü tabanlı AKT sonuçlarının birleştirilmesinde doğrusal ağırlıklı birleştirme yapılmıştır.

2 Materyal ve metot

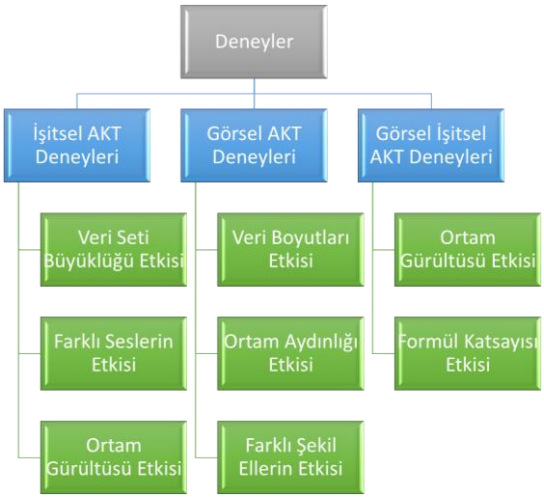
Önerilen çalışmada AKT uygulamasının etkinliğini artırmak için AKT uygulamasında başarılı olan hibrit uygulamanın, TinyML uygulaması olarak gerçekleşmesi üzerinde bir çalışma yapılmıştır. TinyML uygulamalarına uygun olacak şekilde veri setleri kullanılmıştır. Gerçekleştirilen derin öğrenme tabanlı uygulamada TinyML'e uygun model ve donanımlar kullanılmıştır. Modeller Edge Impulse uygulama programında eğitilip, optimizasyon ve kırpma yapılarak donanımlarda çalışmaya hazır hale getirilmiştir.

Şekil 1'de AKT modellerine yapılan deneyler gösterilmektedir. Deneylerdeki amaç görsel AKT ve işitsel AKT modellerinin tek başına değişen fiziksel koşullardaki performansını gözlemlemektir. Daha sonrasında hibrit modelinde performansı test edilip sonuçları karşılaştırarak hibrit modelin başarısını gözlemlemektir.

İşitsel AKT deneyleri 3 aşamada yapılmıştır. Öncelikle veri seti büyüklüğüne göre modelin doğruluk performansı ve aşırı öğrenme (overfitting) durumu gözlemlenmiştir. Daha sonrasında veri setinde olmayan insan sesleriyle model test edilmiştir. Son olarak gürültülü ortamlarda işitsel AKT'nin başarısı gözlemlenmiştir.

Görsel AKT deneylerinde ilk olarak görsel verilerin boyutlarının bellek kullanımı, doğruluk gibi parametre etkisi gözlemlenmiştir. Daha sonrasında ortam aydınlığının modelin performansına etkisi gözlemlenmiştir ve son olarak veri setinde bulunmayan farklı el şekillerinin model performansına etkisi gözlemlenmiştir.

Görsel işitsel AKT deneylerinde ise Denklem (5) 'de verilen doğrusal birleştirme yöntemiyle oluşturulan hibrit modelin formül katsayısı ve gürültülü ortamdaki performansı gözlemlenmiştir.

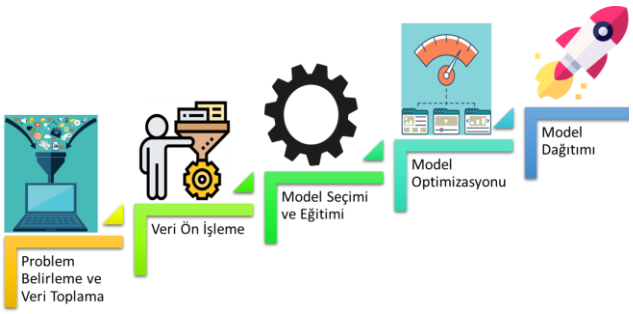


Şekil 1. Çalışmada yapılan deneyler

2.1 TinyML

Geleneksel makine öğrenmesi algoritmalarının aksine, TinyML modelleri çok daha az bellek ve işlem gücü kullanarak çalışabilmektedir. TinyML modellerini cihazlarda etkin bir şekilde çalıştırmak için tasarlanmış birçok platform bulunmaktadır.

Edge Impulse, TinyML uygulamalarını geliştirmek için kullanılan popüler bir platformdur. Bu platform, kullanıcılara veri toplama, model eğitimi, optimizasyon ve cihaza derleme gibi tüm TinyML geliştirme aşamalarını tek bir çatı altında sunar. Bu sayede, geliştiriciler kolayca TinyML uygulamaları oluşturabilir ve ürettikleri ürünlere akıllı özellikleri entegre edebilirler. Şekil 2'de TinyML akış şeması gösterilmiştir.



Şekil 2. TinyML ile model oluşturmadaki akış şeması

TinyML'de eğitim süreci, büyük ölçekli makine öğrenimi uygulamalarından farklıdır. Eğitim işlemi genellikle yüksek performanslı bir bilgisayarda gerçekleştirir ve daha sonra bu eğitilmiş model, bir mikrodenetleyici gibi bir hedef cihazda çalıştırılmak üzere optimize edilir [11].

TinyML, düşük güç tüketimi, sınırlı bellek ve işlem kapasitesi gibi zorluklarla karşı karşıya kalmaktadır. Bu nedenle, TinyML'de optimizasyon süreci çok önemlidir. Optimizasyon süreci, TinyML modelinin performansını ve verimliliğini artırmak için yapılan bir dizi adımdan oluşur.

Nicelleştirme, TinyML'de genellikle modelin boyutunu küçültmek ve hızını artırmak için kullanılan bir optimizasyon tekniğidir [12]. Spesifik olarak, modelin ağırlıklarının hassasiyetini azaltarak ve daha az dahili hafıza kullanarak çalışmasına olanak tanır. Büyük veri merkezlerinde çalışırken, model ağırlıkları genellikle 32 bitlik kayan nokta değerleri (float32) olarak saklanır. Ancak bu, donanımsal olarak düşük cihazlarda (örneğin cep telefonu veya mikrokontrolcüde) gereğinden fazla hassasiyet olabilir ve gereksiz yer kaplar. Nicelleştirme, bu ağırlıkları daha düşük hassasiyetli bir biçime (örneğin int8 veya uint8) dönüştürerek modelin hafızada kapladığı alanı önemli ölçüde azaltabilir.

2.2 ESP32-CAM modülü ile görsel anahtar kelime tespiti

Görsel anahtar kelime tespiti, belirli görsel desenleri veya nesneleri tanımlama sürecidir. Görsel AKT işleminde kamera sürekli olarak görüntü yakalar ve bu görüntüler model tarafından analiz edilir. Model, belirlenen nesneyi tespit ettğinde bir sinyal göndererek tespiti sağlar.

Bu çalışmada görsel verilerden yararlanarak AKT işlemi ESP32-CAM modülü ile yapılmıştır. ESP32-CAM, bir ESP32 WiFi modülü içeren ve dahili bir video kamera barındıran bir geliştirme kartıdır.

2.2.1 Görsel anahtar kelime tespitinde veri toplama

Klasik bir makine öğreniminde olduğu gibi TinyML ile de bir model oluştururken ilk yapılan işlem modelin veri setini hazırlamaktır. Bu çalışmada görsel veri olarak Hand Gesture Recognition Dataset kullanılmıştır [13]. Bu veri setinde 20 farklı el hareketinden 24000 görüntü mevcuttur. Her sınıf için 900 eğitim verisi 300 test verisi bulunmaktadır. Veriler %75 eğitim, %25 test verileri olacak şekilde ayrılmıştır. Bu çalışmada, 2 farklı el durumu (açık el, kapalı el) ile tespit yapılmıştır. Şekil 3'de çalışmada kullanılan el hareketleri gösterilmiştir. İlerleyen bölümlerde açıklandığı gibi açık el hareketi YES sesi ile kapalı el hareketi ise NO sesi ile birlikte kullanılmıştır.

(a) (b)

Şekil. 3. Veri setinde kullanılan el hareketleri (a) açık el hareketi ve (b) kapalı el hareketi

2.2.2 Görsel anahtar kelime tespitinde dürtü tasarımı

TinyML uygulamalarında veri seti oluşturma işleminden sonra dürtü tasarımı yapılmaktadır. Dürtü tasarımı, veri setinden alınan verileri işleyip özellikler çıkarmak ve yeni verileri sınıflandırmak için bir makine öğrenimi akışı oluşturmaya yarar. Dürtü tasarımında Edge impulse platformunda açıklanan bloklardaki parametreler girilir, daha sonrasında veri setindeki veriler siyah skalaya çevrilir. Siyah skalaya çevirme işlemi modelin daha hafif olmasını sağlar.

Bu çalışmada Transfer Learning tekniği ile veriler işlenmiştir. Eğitim mimarisi olarak MobilenetV2 kullanılmıştır. MobileNets, mobil ve gömülü cihazlarda hafif ve hızlı derin öğrenme modeli olarak kullanılmak üzere tasarlanmış bir mimaridir. Bu mimari, geleneksel derin sinir ağlarının (DNN) hesaplama yoğunluğunu azaltarak ve parametre sayısını düşürerek, daha küçük boyutlu ve daha hızlı çalışan modeller elde etmeyi hedefler. MobileNet, "depthwise separable convolution" olarak bilinen bir yapı kullanır [14].

Modelin dürtü tasarımı ve ardından eğitim mimarisi ile eğitimi sonucunda bir karışıklık matrisi (confusion matrix) verilmektedir. Şekil 4'de görsel AKT için karışıklık matrisi verilmiştir.

		GERÇEK	
		AÇIK EL	KAPALI EL
TAHMİN	AÇIK EL	100%	0%
	KAPALI EL	0%	100%
	F1 SKORU	1.00	0

Şekil 4. Görsel AKT eğitimi sonrası edge ımpulse'da hesaplanmış karışıklık matrisi

F1 skoru, bir modelinin performansını ölçen bir metriktir. Hassasiyet ve geri çağırma parametrelerinin harmonik ortalaması olarak hesaplanır. Denklem (1) hassasiyet hesaplamasını göstermektedir. Hassasiyet, modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma oranını ifade etmektedir.

$$\textbf{Hassasiyet} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}} \quad (1)$$

Geri çağırma, gerçek pozitif örneklerin ne kadarının doğru bir şekilde tanındığını ifade etmektedir. Denklem (2) geri çağırma hesaplamasını göstermektedir.

$$\textbf{Geri Çağırma} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \quad (2)$$

Denklem (1) ve Denklem (2)'deki TP, gerçek pozitif örneklerin doğru bir şekilde pozitif olarak tahmin edilmesi anlamına gelmektedir. FP, gerçek negatif örneklerin yanlış bir şekilde pozitif olarak tahmin edilmesini göstermektedir. FN, gerçek pozitif örneklerin yanlış bir şekilde negatif olarak tahmin edilmesini göstermektedir. F1 skoru, Denklem (3)'de gösterilmiştir.

$$\textbf{F1} = 2 \times \frac{\textbf{Hassasiyet} \times \textbf{Geri Çağırma}}{\textbf{Hassasiyet} + \textbf{Geri Çağırma}} \quad (3)$$

F1 skoru, 0 ile 1 arasında bir değer alır. 1, en iyi performansı temsil ederken, 0 en kötü performansı temsil eder.

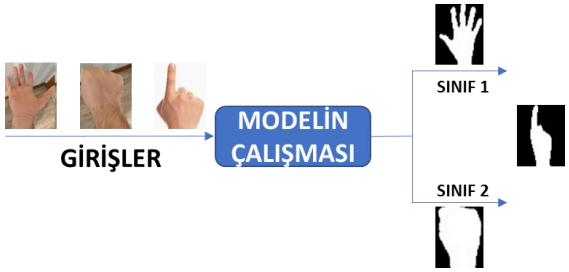
2.2.3 *Görsel anahtar kelime tespitinde dağıtım*

Dağıtım işlemi, modelin cihazlarda yerel olarak çalıştırmasını sağlamak için gerekli olan adımları içerir. Edge Impulse, makine öğrenim modellerinin boyutunu ve hesaplama maliyetini azaltmak için niceleme tekniğini kullanır. Tablo 1'de niceleme tekniği ile optimize edilmiş ve edilmemiş modelin farkları gösterilmektedir.

Tablo 1. Görsel AKT'de optimizasyon sonucu edge ımpulse tarafından hesaplanan değerler

Teknik	RAM (kByte)	Flash RAM (kByte)	Gecikme Süresi (ms)
Niceleme Uygulanmış	170.4	584.5	809
Niceleme Uygulanmamış	236.7	1600	1704

Model gerçek zamanlı çalışarak el hareketlerini algılamaktadır. Modelin çalışması Şekil 5'de gösterilmiştir. Modelin çalışmasında 2 sınıf bulunmaktadır. Açık el ve kapalı el hareketi olmayan diğer nesneler sınıflandırılmamaktadır. Görsel AKT'de bilinmeyen sınıfı yapılmamıştır. Bunun sebebi 3 sınıf olması durumunda bellek kullanımı artarak ESP32-CAM modeli çalıştıramamıştır.



Şekil 5. Görsel AKT modelinin el hareketlerini sınıflandırması

2.3 *Arduino nano BLE ile işitsel anahtar kelime tespiti*

İşitsel anahtar kelime tespiti, ses girdisinden belirli anahtar kelimelerin otomatik olarak tanınması işlemidir. Bu çalışmada Edge Impulse'da eğitilen işitsel AKT modeli Arduino Nano BLE Sense'de çalıştırılmıştır. Arduino BLE Sense, Bluetooth Low Energy (BLE) bağlantısı ve dahili olarak MP34DT05 model bir mikrofon içermektedir. Arduino Nano BLE sense, bir 32-bit Arm cortex-M4 işlemci barındırmaktadır.

2.3.1 *İşitsel anahtar kelime tespitinde veri toplama*

İşitsel AKT'de 3 farklı veri seti kullanılmıştır. İlk veri seti, Google tarafından hazırlanmış olan Yes-No veri setidir [15]. Veri seti yüzlerce farklı kişiden alınan ses kayıtlarından oluşmaktadır. Her bir kelime 1 saniyelik verilerle kaydedilmiştir. Veri seti 342 makalede referans gösterilmiştir [16]. Tablo 2'de veri setinin farklı modellerdeki doğruluk yüzdeleri gösterilmiştir. Bu modeller bilgisayar ortamında çalıştırılmış ve AKT uygulamalarında kullanılmıştır. Tablonun son satırında ise bu çalışmadaki Edge Impulse tarafından hesaplanan değeri gösterilmektedir.

Tablo 2. Farklı çalışmalarda kullanılan yes-no veri setinin doğruluk değerleri

Model	Doğruluk (%)
TripletLoss-res15 [17]	98,56
BC-ResNet-8 [18]	98
Wav2KWS [19]	97,9
Res 8 [20]	97,8
KWT-3 [21]	97,49
CNN	95

Doğruluk, makine öğreniminde modelin doğru tahmin ettiği örneklerin oranını ifade etmektedir. Doğruluk, Denklem (4)'de gösterilmiştir.

$$\textbf{Doğruluk} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}} \quad (4)$$

Denklemdeki TN, gerçek negatif örneklerin doğru bir şekilde negatif olarak tahmin edilmesini ifade etmektedir. TP, FP ve FN ifadeleri Denklem (1) ve Denklem (2)'de açıklanan ifadelerdir.

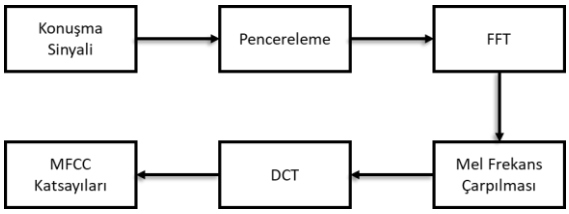
İkinci veri seti ise Microsoft tarafından hazırlanmış olan MS-SNSD veri setidir [22]. MS-SNSD veri seti, arka plan gürültüsünü bastırmak için derin sinir ağları modellerini eğitmek amacıyla hazırlanmıştır. Bu veri seti çalışmadaki gürültü sınıfında kullanılmıştır.

Üçüncü veri seti ise bu çalışma için Edge Impulse platformunda hazırlanmış olan Open Close veri setidir. Çalışmada Arduino Nano geliştirme kartında bulunan MP34DT05 model mikrofon ile 16 kHz örnekleme hızında ses veri seti hazırlanmıştır. Her bir kelime 1 saniye olacak şekilde kaydedilmiştir.

2.3.2 *İşitsel anahtar kelime tespitinde dürtü tasarımı*

İşitsel AKT'de dürtü tasarımı adımı, görsel AKT'da olduğu gibidir. İşitsel AKT işleminde, ses sinyalleri dijital bir yapıya dönüştürülüp işlenmektedir.

Mel-Frekansı Kepstrum Katsayıları (MFCC), ses sinyallerinin spektral özelliklerini temsil etmek için kullanılan ve insan kulağının frekans algılamasını taklit eden bir gösterim şeklidir. Bu, MFCC'yi konuşma tanıma, ses sentezi, hoparlör doğrulama ve müzik bilgi alma gibi birçok alanda en iyi özelliklerden biri haline getirir. MFCC literatürde de anahtar kelime tespit çalışmalarında kullanılmaktadır [23].



Şekil 6. MFCC yapısının blok şeması

Şekil 6'da gösterilen blok şemasında, öncelikle ses sinyalleri analogdan dijitale dönüştürülür. Örnekleme hızı ve bit derinliği belirlenir. Ses sinyali kare pencereler halinde bölünür. Pencerelerin uzunluğu ve örtüşme miktarı belirlenir. Daha sonrasında her pencereye FFT (Hızlı Fourier Dönüşümü) uygulanır. FFT'nin çıktısı, frekans bileşenlerinin genliklerini ve frekanslarını içeren bir spektrumdur. Spektrum, insan işitme sistemine daha yakın bir frekans temsili elde etmek için Mel ölçeği kullanılarak yeniden örnekleme edilir. Mel ölçekli spektrumun DCT (Discrete Cosine Transform)'si hesaplanır. DCT'nin ilk birkaç katsayısı, MFCC'leri oluşturur. En önemli MFCC'ler seçilir. Seçilen MFCC'ler, ses sinyalinin sınıfını tahmin etmek için kullanılır.

İşitsel anahtar kelime tespitinde, MFCC ile ses özellikleri çıkarılıp eğitim modeli olarak tek katmanlı CNN yapısı kullanılmıştır. Eğitim sonucunda karışıklık matrisi Şekil 7'de gösterilmiştir.

		GERÇEK		
		NO	GÜRÜLTÜ	YES
TAHMİN	NO	99.2%	0%	0.8%
	GÜRÜLTÜ	1.4%	98.2%	0.5%
	YES	0.4%	1.6%	98.0%
	F1 SKORU	0.99	0.98	0.98

Şekil 7. İşitsel AKT eğitimi sonrası edge ımpulse'da hesaplanmış karışıklık matrisi

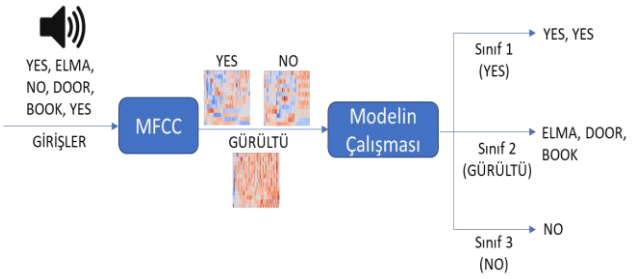
2.3.3 *İşitsel anahtar kelime tespitinde dağıtım*

Dağıtım işleminde Edge Impulse bir kütüphane dosyası olarak model çıktısını vermektedir. Bu kütüphane Arduino IDE programına eklenerek kod cihaza yüklenir. Görsel AKT'de olduğu gibi işitsel AKT için niceleme işlemi Tablo 3'de gösterilmiştir.

Tablo 3. İşitsel AKT'de optimizasyon sonucu edge ımpulse tarafından hesaplanan değerler

Teknik	RAM (kByte)	Flash RAM (kByte)	Gecikme Süresi (ms)
Niceleme Uygulanmış	13	35.7	340
Niceleme Uygulanmamış	28.5	39.4	981

Model gerçek zamanlı çalışarak sesleri sınıflandırmaktadır. Ses sınıflandırmada 3 sınıf bulunmaktadır. Ses verileri görsel verilere göre daha az yer kapladığı için bu model 3 farklı sınıfı sınıflandırmaktadır. Modelin çalışması Şekil 8'de gösterilmiştir.



Şekil 8. İşitsel AKT modelinin sesleri sınıflandırması

2.4 *Görsel işitsel anahtar kelime tespiti*

Görsel-işitsel anahtar kelime tespiti hem görsel hem de işitsel verilerden anahtar kelimeleri otomatik olarak çıkarma ve tanımlama işlemidir. Gürültülü ortamlarda anahtar kelime tespiti yapılması gereken durumlarda kullanılmaktadır. Görsel işitsel anahtar kelime tespiti, makine öğreniminde önemli ve zorlu bir araştırma konusudur. Bu hibrit modeller her iki veri kaynağının sunduğu özellikleri kullanarak daha güçlü ve güvenilir tahminler yapabilir.

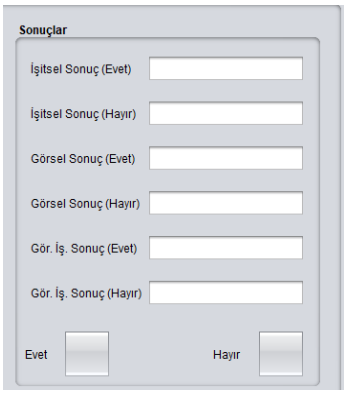
Xu ve arkadaşları (2021), uzak alanlarda kelime tespitinin çevresel sağlamlığını iyileştirmek için hem ses hem de görsel bilgiden yararlanarak bir model geliştirmişlerdir. Çalışmada ilk başta çoklu mikrofon sesini ele alıp ses özelliklerini çıkarmak için WPE (Weighted Prediction Error) kullanılmıştır. Görsel verilerin özellik çıkarımı için ROI (Region of Interest) metodu kullanılmıştır. Ardından çok katmanlı CNN ile modeller eğitilmiştir. Modeller doğrusal ağırlıklı birleştirme yapıldıktan sonra %91 doğruluk elde edilmiştir [24].

Bu makelede, Xu ve arkadaşlarının yapmış olduğu çalışmayı referans alınarak ses için Yes-No ve görüntü için Hand Gesture Recognition Dataset kullanılarak önerilen formül üzerinden modeller birleştirilmiştir. Ayrıca Xu ve arkadaşları bu çalışmayı bilgisayar ortamında yapmışken mevcut çalışma Arduino Nano BLE sense ve ESP32-CAM modüllerinde yapılmıştır.

Donanımlardan ayrı ayrı alınan sonuçlar Denklem (5) kullanılarak doğrusal ağırlıklı birleştirme yapılmıştır.

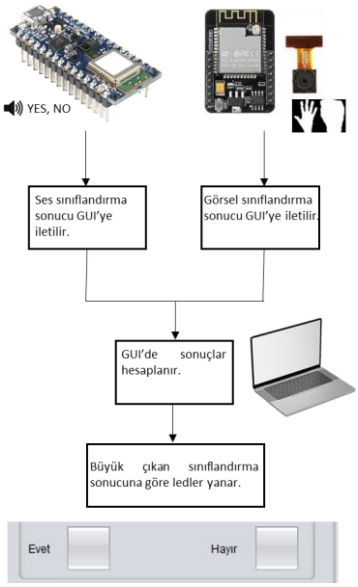
$$\mathbf{y} = (\mathbf{w_a} \times \mathbf{y_a}) + (\mathbf{w_b} \times \mathbf{y_b}) \quad (5)$$

Denklemdeki w_a ve w_b ifadeleri ağırlık değerleridir. Ağırlık değerleri olarak referans çalışmada w_a için 0.7 ve w_b için 0.3 önerilmiştir. Ayrıca y_a ses modelinin doğruluğunu ve y_b görsel modelin doğrulunu temsil etmektedir. Model sonuçları, bu çalışma için tasarlanmış JAVA tabanlı bir grafiksel arayüze aktarılmaktadır. Donanımlar ve bilgisayar arasındaki haberleşme UART ile yapılmaktadır. Şekil 9'da grafiksel arayüz gösterilmektedir.



Şekil 9. Java tabanlı tasarlanan grafiksel arayüz

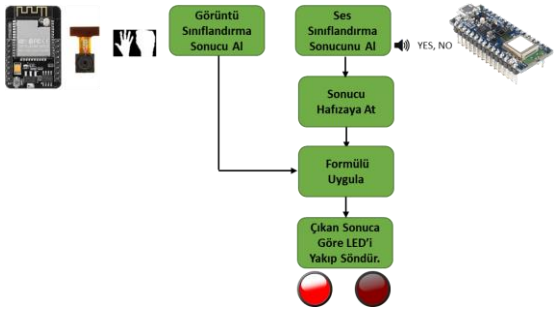
Görsel AKT sonucu 809 milisaniye aralıkla gelmektedir. İşitsel AKT sonucu ise 340 milisaniye aralıkla gelmektedir. Bu yüzden verilerin sınıflandırılması çıkarım zamanı büyük olan sisteme göre yapılmıştır. Bu işlemde öncelikle ses verisi geldiğinde hafızaya atılıp görsel sonuç beklenmektedir. Görsel sonuç geldiğinde ise sonuçlar Denklem (5) ile hesaplanmıştır. Büyük çıkan doğruluk sonucuna göre ilgili led yanmaktadır.



Şekil 10. GUI ortamında görsel işitsel anahtar kelime tespiti modelinin çalıştırılması

Denklem (5) değeri ilgili sınıfta 0.7 (%70) üzeri olduğunda ise ilgili sınıfın ledi ile sonuç gösterilmiştir.

TinyML uygulamaları düşük güç tüketimine ve düşük maliyetli sistemlerde çalışmayı hedeflemektedir. Çalışmanın bilgisayar ortamından bağımsız bir şekilde de çalışması sağlanmıştır. Burada da öncelikle ses sınıflandırma sonucu alınıp bir hafızaya atılmıştır. Daha sonrasında görsel sınıflandırma sonucuna göre ESP32-CAM modülündeki GPIO14 pini aktif olmaktadır. İlgili sınıflandırma doğruluğu %70 (0,7) üzerinde olduğunda GPIO14 pininden voltaj çıkışı verilmektedir. ESP32-CAM'de açık el hareketi algılandığında GPIO14 pininden 24V çıkış vermektedir. GPIO14 pini Arduino'daki A0 pinine bağlıdır. A0 pini burayı sürekli olarak okumaktadır. Ses sonucu "Yes" olarak algılanıp A0 pininden de 24V okunduğunda Arduino Nano BLE Sense'deki dahili led yanmaktadır. Ses sonucu "No" olarak algılanıp A0 pininden de 0V okunduğunda Arduino Nano BLE Sense'deki dahili led sönmektedir.



Şekil 11. Donanım ortamında görsel işitsel anahtar kelime tespiti modelinin çalıştırılması

3 Bulgular ve tartışma

Bu bölümde, Edge Impulse platformunda geliştirilen ve Arduino IDE ortamında cihazlara yüklenen TinyML modellerinin ayrı ayrı AKT deneyleri ve daha sonrasında tasarlanmış olan GUI üzerinde görsel işitsel AKT deneyleri açıklanmıştır.

3.1 İşitsel AKT deneyleri

İşitsel AKT deneyleri sonucunda gürültülü bir ortamda %85 doğruluk elde edilmiştir.

3.1.1 İşitsel AKT'de veri seti büyüklüğü etkisi

Yapılan çalışmada, değişen veri seti büyüklüğü referans alınarak, doğruluk değeri ve Arduino geliştirme kartı üzerinde yapılan testlerin deneysel doğruluğu gösterilmiştir.

Veri seti olarak Edge Impulse sitesinde Arduino'daki dahili mikrofon ile hazırlanan Open-Close ve hazır olarak alınan Yes-No ile deneyler ayrı ayrı yapılmıştır. Test sütunundaki değer Arduino'dan alınan 20 örnek sonucunda alınan doğruların yüzdesidir. Tablo 4'de Open-Close veri setinin büyüklüğünün test ve doğruluğa etkisi gösterilmektedir. Tablo 5'de Yes-No veri setinin büyüklüğünün test ve doğruluğa etkisi gösterilmektedir.

Tablo 4. Open-Close veri setinin büyüklüğünün doğruluğa etkisi

Eğitim Verisi (Dakika)	Test Verisi (Dakika)	Doğruluk (%)	Test (%)
5	1.25	100	35 (20/8)
10	2.5	100	60 (20/12)
20	5	100	75 (20/15)

Tablo 5. Yes-No veri seti büyüklüğünün doğruluğa etkisi

Eğitim Verisi (Dakika)	Test Verisi (Dakika)	Doğruluk (%)	Test (%)
5	1.25	93.9	55 (20/11)
10	2.5	96.4	80 (20/16)
20	5	98.5	95 (20/19)

Open-Close veri seti 3 farklı insan sesi kullanarak hazırlanmıştır. Yes-No veri seti ise daha profesyonel olarak yüzlerce kişi tarafından Google destekli hazırlanmıştır. Deneylerde Open-Close veri setinin yeterli dengede hazırlanamadığı ve aşırı öğrenme durumu oluşturduğu gözlemlenmiştir. Diğer deneyler daha dengeli ve daha yüksek doğruluk değeri elde ettiği için Yes-No veri seti ile yapılmıştır.

3.1.2 İşitsel AKT'de farklı insan seslerinin etkisi

Bir işitsel AKT modelinin yüksek performans kriterlerinden birisi de farklı seslerde benzer performansı göstermesidir. Tablo 6'da 10 farklı insan sesi ile modelin test edilmesini göstermektedir.

Tablo 6. Farklı İnsan Seslerinin Test Doğruluğuna Etkisi

Cinsiyet	Yaş	Test (%)
Erkek	28	95 (20/19)
Kadın	51	90 (20/18)
Erkek	51	85 (20/17)
Erkek	30	95 (20/19)
Erkek	40	90 (20/18)
Erkek	23	90 (20/18)
Kadın	32	95 (20/19)
Erkek	28	90 (20/18)
Kadın	21	95 (20/19)
Kadın	22	95 (20/19)

Görüldüğü gibi sonuçlar benzer çıkarak modelin farklı seslerde de benzer performans verdiği gözlemlenmiştir.

3.1.3 İşitsel AKT'de ortam gürültüsü etkisi

Ortam gürültüleri genel olarak modelin performansını bozan etkilerdir. Yüksek performans gösteren bir modelin ortam gürültüsünde de başarılı bir performans göstermelidir.

Tablo 7. Ortam Gürültüsünün Test Doğruluğuna Etkisi

Ortam Gürültüsü	Test (%)
Beyaz Gürültü	95 (20/19)
Siyah Gürültü	95 (20/19)
Kahve Dükkânı Arka Plan Gürültüsü	85 (20/17)
Mutfak Arka Plan Gürültüsü	90 (20/18)
Boş Laf Gürültüsü	85 (20/17)
Komşu Sesi Gürültüsü	85 (20/17)

Tablo 7'de görüldüğü gibi arka plan gürültüleri doğruluk değerini düşürmüştür. Arka planda insan sesi olan gürültü değerleri diğer gürültülere göre doğruluğu daha azaltmıştır.

3.2 Görsel AKT deneyleri

Görsel AKT deneyleri sonucunda yeterli aydınlığa sahip bir odada %85 doğruluk elde edilmiştir.

3.2.1 Görsel AKT'de veri boyutları etkisi

Görsel verilerin boyutları, model doğruluğu, RAM kullanımı ve çıkarım zamanını etkileyebilir. Büyük boyutlu veriler genellikle daha fazla bilgi sağlar, ancak modelin eğitimi ve dağıtımı için daha fazla bellek ve işlem gücü gerektirebilir.

Tablo 8. Görsel verilerin boyutlarının doğruluk, ram ve çıkarım zamanına etkisi

Genişlik (mm)	Boy (mm)	Doğruluk (%)	RAM (kByte)	Çıkarım Zamanı (ms)
32	32	95	583.4	394
48	48	100	584.5	588
64	64	100	586.5	645
96	96	100	590.6	2231
160	160	100	721.6	4084

Tablo 8'de görüldüğü gibi 48x48 boyutta en iyi sonucu vermiştir. En iyi sonuç en yüksek doğruluk, en düşük RAM kullanımı ve en düşük çıkarım zamanına göre belirlenmiştir. Edge Impulse görüntü verileri için en düşük 32x32 boyutuna izin vermektedir.

3.2.2 Görsel AKT'de ortam aydınlığı etkisi

Ortam aydınlığı görsel AKT'de performansı etkileyen faktörlerden biridir. Ortamın yeterli düzeyde aydınlık olması test performansını olumlu yönde etkilemektedir. Tablo 9'da yeterli ve yetersiz ışıklandırılmış bir odada ki görsel AKT performansı gösterilmiştir. Alınan 20 örnek sonucunda doğru sınıflandırmaya göre yüzde belirlenmiştir.

Tablo 9. Görsel AKT'de ortam aydınlığının test doğruluğuna etkisi

Ortam Aydınlığı	Test (%)
Yeterli Işıklandırılmış Oda	85 (20/17)
Yetersiz Işıklandırılmış Oda	70 (20/14)

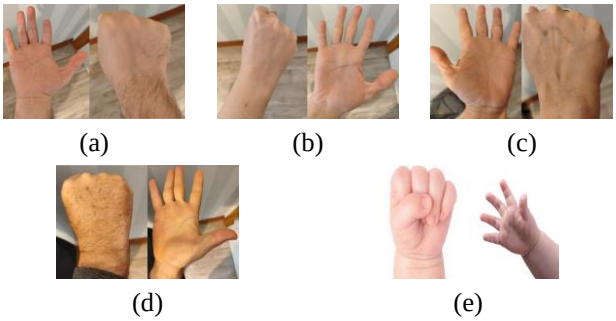
Testler aynı odada yapılmıştır. Yeterli ışıklandırılma, öğle vakti lamba açık bir oda iken, yetersiz ışıklandırılma, akşamüstü lamba açık olmayan bir odadır. Test sütunu, yapılan 20 hareket sonucu doğruluk yüzdesidir.

3.2.3 Görsel AKT'de farklı şekildeki ellerin etkisi

Başarılı bir AKT'de beklenilen bir başka durum, farklı boyut ve şekillerdeki ellerde de benzer performansı göstermesidir. Tablo 9'da Şekil 12'deki (a) eli ile deney yapılmıştır. Tablo 10'da ise Şekil 12'deki tüm eller ile deneyler yapılmıştır.

Tablo 10. Görsel AKT'de farklı şekildeki ellerin performansı

El Tipi	Test Performansı (%)
(A)	85 (20/17)
(D)	80 (20/16)
(B)	85 (20/17)
(C)	85 (20/17)
(E)	60 (20/12)



(a) (b) (c) (d) (e)

Şekil 12. Görsel AKT testlerinde kullanılan eller

Tablo 10'da görüldüğü gibi bebek eli haricinde sonuçlar benzer çıkmıştır. Bebek eli ile yapılan deneylerde test yüzdesi diğer deneylerden düşük çıkmıştır. Bunun sebebi kullanılan veri setinin yetişkin bir el ile hazırlanmasıdır.

3.3 Görsel işitsel AKT deneyleri

Bu bölümde Şekil 9'daki GUI ile görsel işitsel AKT deneyleri gerçekleştirilmiştir. Deneylerdeki değerlendirme kriteri, görsel işitsel AKT performansının işitsel AKT performansından daha yüksek olmalıdır.

3.3.1 Görsel işitsel AKT'de ortam aydınlığının etkisi

Tablo 11'de farklı ortam şartlarına göre görsel işitsel kelime tespit uygulaması yapılmıştır. Gürültülü ortam, komşu sesi gürültüsü açılmış bir odadır. Yeterli ışıklandırılmış ortam, öğle vakti lamba açık bir odayı temsil ederken, yetersiz ışıklandırılmış ortam akşamüstü lamba kapalı bir odayı temsil etmektedir. Testlerde bazı durumlarda işitsel ve görsel sınıflandırmalar ters yapılmıştır. Örneğin AÇ hareketi yaparken NO sesi söylenmiş veya KAPAT hareketi yaparken YES sesi söylenerek gözlemlenmiştir. Denklem (5) değerleri olarak w_a için 0.7 ve w_b için 0.3 kullanılmıştır.

Tablo 11. Görsel İşitsel AKT'de Ortam Aydınlığının Etkisi

Ortam	Görsel İşitsel Algılama (%)	İşitsel Algılama (%)
Yeterli Işık / Sessiz	95	95
Yeterli Işık / Gürültülü	90	85
Yetersiz Işık / Sessiz	95	95
Yetersiz Işık / Gürültülü	80	85

Testler sonucunda yeterli ışık alan gürültülü bir ortamda görsel işitsel algılama işitsel algılama doğruluğundan yüksek çıkmıştır.

3.3.2 Görsel işitsel AKT'de formül katsayılarının etkisi

Referans çalışmada katsayı değerleri w_a için 0.7 ve w_b için 0.3 kullanılmıştır. Yapılan deneylerde w_a katsayısının daha yüksek olması durumunda daha iyi sonuç verildiği gözlemlenmiştir. Tablo 12'de w_a katsayısının daha yüksek olduğu diğer durumlar gözlemlenmiştir. Ortam olarak gürültülü ve yeterli ışık alan bir ortam tercih edilmiştir.
Yapılan deney sonucunda referans çalışmanın önerdiği katsayının en yüksek doğruluk sonucunu verdiği gözlemlenmiştir.

Tablo 12. Görsel işitsel AKT'de farklı katsayıların etkisi

Katsayılar (Denklem5 Ağırlıkları)	Görsel İşitsel Algılama (%)	İşitsel Algılama (%)
w_a =0,6- w_b =0,4	85	85
w_a =0,7- w_b =0,3	90	85
w_a =0,8- w_b =0,2	85	85
w_a =0,9- w_b =0,1	85	85

4 Sonuçlar

Bu çalışma, TinyML ile görsel ve işitsel AKT sonuçlarının entegrasyonunu ele alarak, görsel işitsel anahtar kelime tespiti konusunda önemli bir katkı sağlamıştır.

Öncelikle görsel ve işitsel AKT modelleri Edge Impulse platformunda eğitilip cihazlarda ayrı ayrı test edilmiştir. Görsel AKT yeterli ışık alan bir odada %85 doğruluk elde ederken, işitsel AKT gürültülü bir ortamda %85 doğruluk elde edilmiştir. Daha sonrasında tasarlanan GUI'de ses ve görüntü komutlarını eş zamanlı olarak alarak, doğrusal ağırlıklı birleştirme yapılmıştır. Görsel işitsel anahtar kelime tespiti ile gürültülü bir ortamda %90 doğruluk sonucu elde edilmiştir.

Çalışma, Xu ve arkadaşlarının yapmış olduğu çalışmayı referans alarak farklı veri seti ve eğitim modeli ile gömülü cihazlarda uygulanmıştır. Referans çalışmada hazırlanan veri seti ve kullanılan modeller ile %91 oranında bir başarı sağlamışlardır. Sonuç olarak, modelin küçülmesi ve daha az kapasite kullanmasına rağmen doğruluk sonuçları benzer çıkmıştır. Önerilen yöntemin gerçek zamanlı sistemlerde düşük güç tüketimi ve düşük bellek gereksinimi gibi pratik avantajları, endüstrideki kullanılabilirliğini artırmaktadır.

Çalışmamız, düşük güçlü cihazlarda gerçek zamanlı anahtar kelime tespiti için TinyML'nin potansiyelini göstermektedir. Ancak, gelecekteki çalışmalar için birkaç önemli husus bulunmaktadır. Görsel işitsel yöntemlerin birleştirilmesinin en önemli kullanım alanı robot-insan ve insan makine iletişimi olacaktır. TinyML açısından bu uygulamanın geliştirilmesi küçük robotlar ve taşınabilir cihazlar ve insan iletişimi olabilir. Ayrıca bu çalışmada görsel ve işitsel veriler veri setlerinden ayrı ayrı alınarak fotoğraf ve ses tabanlı çalışmaktadır. İleri çalışmalarda bir videodan otomatik olarak görsel ve işitsel veriler alınarak model çalıştırılabilir.

Çıkar çatışması

Yazarlar çıkar çatışması olmadığını beyan etmektedir.

Benzerlik oranı (iThenticate): %5

Kaynaklar

[1] J. Tian, The human resources development applications of machine learning in the view of artificial ıntelligence. IEEE 3rd International Conference, 39-43, 2020. https://doi.org/10.1109/CCET50901.2020.9213113.

[2] M. Rusci and T. Tuytelaars, On-device customization of tiny deep learning models for keyword spotting with few examples. IEEE Micro, 43(6), 50-57, 2023. https://doi.org/10.1109/MM.2023.3311826.

[3] Y. Abadade, A. Temouden, H. Bamoumen, N. Benamar, Y. Chtouki and A. S. Hafid, A comprehensive survey on TinyML. IEEE Access, 11, 96892-96922, 2023. https://doi.org/10.1109/ACCESS.2023.3294111.

[4] P. Warden and D. Situnayake, TinyML machine learning with TensorFlow lite on arduino and ultra-low-power microcontrollers. O'Reilly Media, 2019.
[5] M. Altayeb, M. Zennaro and E. Pietrosemoli, TinyML gamma radiation classifier. Nuclear Engineering and Technology, 55(2), 443-451, 2023. https://doi.org/10.1016/j.net.2022.09.032.
[6] M. Lord, TinyML, anomaly detection. Masters Thesis, California State University, Computer Science, Northridge, USA, 2021.
[7] M. Monfort Grau, TinyML from basic to advanced applications. Bachelor Thesis, Universitat Politècnica de Catalunya, Facultat d'Informàtica de Barcelona, Spain, 2021.
[8] S. Sadhu and P. K. Ghosh, Low resource point process models for keyword spotting using unsupervised online learning. 25th European Signal Processing Conference, 538-542, 2017. https://doi.org/10.23919/eusipco.2017.8081265.
[9] Z. Tang, L. Chen, B. Wu, D. Yu and D. Manocha, Improving reverberant speech training using diffuse acoustic simulation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6969-6973, 2020. https://doi.org/10.48550/arXiv.1907.03988.
[10] J. M. Phillips and J. M. Conrad, Robotic system control using embedded machine learning and speech recognition. 19th International Conference on Smart Communities, Improving Quality of Life Using ICT, IoT and AI (HONET), 214-218, 2022. https://doi.org/10.1109/HONET56683.2022.10019106.
[11] H. Han and J. Siebert, TinyML: A systematic review and synthesis of existing research. International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 269-274, 2022. https://doi.org/10.1109/ICAIIC54071.2022.9722636.
[12] N. S. Huynh, S. De La Cruz and A. Perez-Pons, Denial-of Service (DoS) Attack Detection Using Edge Machine Learning. International Conference on Machine Learning and Applications (ICMLA), 1741-1745, 2023. https://doi.org/10.1109/ICMLA58977.2023.00264.
[13] H. Andrew, Z. Menglong, C. Bo, K. Dmitry, W. Weijun, W. Tobias, A. Marco and A. Hartwig, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Computer Vision and Pattern Recognition, 2017. https://doi.org/10.48550/arXiv.1704.04861.
[14] Kaggle, Hand Gesture Recognition Dataset. https://www.kaggle.com/datasets/aryarishabh/hand-gesture-recognition-dataset, Accessed 14 January 2024.
[15] W. Pete, Speech commands: A dataset for limited-vocabulary speech recognition. Computation and Language, 2018. https://doi.org/10.48550/arXiv.1804.03209.
[16] Papers with code, Speech Commands, https://paperswithcode.com/dataset/speech-commands, Accessed 2 February 2024.
[17] V. Roman and M. Nikolay, Learning efficient representations for keyword spotting with triplet loss. 23rd International Conference SPECOM, 2021. https://doi.org/10.1007/978-3-030-87802-3_69.
[18] B. Kim, S. Chang, J. Lee and D. Sung, Broadcasted Residual Learning for Efficient Keyword Spotting. Proceedings of INTERSPEECH, 2021. https://doi.org/10.48550/arXiv.2106.04140.
[19] D. Seo, H.-S. Oh and Y. Jung, Wav2KWS: Transfer Learning From Speech Representations for Keyword Spotting. IEEE Access, 9, 80682-80691, 2021. https://doi.org/10.1109/ACCESS.2021.3078715.
[20] R. Tang, J. Lee, A. Razi, J. Cambre, I. Bicking, J. Kaye and J. Lin, Howl: A Deployed, Open-Source Wake Word Detection System. Computation and Language, 2020, https://doi.org/10.48550/arXiv.2008.09606.
[21] A. Berg, M. O'Connor and M. Tairum Cruz, Keyword Transformer: A Self-Attention Model for Keyword Spotting. Interspeech, 4249-4253, 2021, https://doi.org/10.21437/Interspeech.2021-1286.
[22] C. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan and J. Gehrke, A scalable noisy speech dataset and online subjective test framework. InterSpeech, 2019. https://doi.org/10.48550/arXiv.1909.08050.
[23] A. Mahmood and U. Köse, Speech recognition based on convolutional neural networks and MFCC algorithm. Advances in Artificial Intelligence Reseach, 1(1), 6–12, 2021.
[24] Y. Xu, J. Sun, Y. Han, S. Zhao, C. Mei, T. Guo, S. Zhou, C. Xie, W. Zou, X. Li, S. Zhou, C. Xie, W. Zou and X. Li, Audio-Visual Wake Word Spotting System For MISP Challenge 2021. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 9246-9250, 2021, https://doi.org/10.48550/arXiv.2204.08686.