ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

LIFELONG LEARNING FOR AUDITORY SCENE ANALYSIS

Ph.D. THESIS

Barış BAYRAM

Department of Computer Engineering

Computer Engineering Programme

JULY 2022

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

LIFELONG LEARNING FOR AUDITORY SCENE ANALYSIS

Ph.D. THESIS

Barış BAYRAM
(504152503)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. Gökhan İNCE

JULY 2022

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

İŞİTSEL SAHNE ANALİZİ İÇİN HAYAT BOYU ÖĞRENME

DOKTORA TEZİ

Barış BAYRAM
(504152503)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Gökhan İNCE

TEMMUZ 2022

Barış BAYRAM, a Ph.D. student of ITU Graduate School student ID 504152503, successfully defended the dissertation entitled "LIFELONG LEARNING FOR AUDITORY SCENE ANALYSIS", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Assoc. Prof. Dr. Gökhan İNCE** ....
Istanbul Technical University

Jury Members : **Assoc. Prof. Dr. Sanem SARIEL UZER** ....
Istanbul Technical University

Assoc. Prof. Dr. Emre UĞUR ....
Boğaziçi University

Assoc. Prof. Dr. Nazım Kemal ÜRE ....
Istanbul Technical University

Assoc. Prof. Dr. Sinan KALKAN ....
Middle East Technical University

Date of Submission : 3 June 2022
Date of Defense : 4 July 2022

To my dearest Mischa,

FOREWORD

Firstly, I would like to thank my advisor Assoc. Prof. Dr. Gökhan İNCE for his excellent facilities and advice during this thesis. Also, I would like to thank the members of my thesis committee Assoc. Prof. Dr. Sanem SARIEL and Assoc. Prof. Dr. Emre UGUR for their valuable time, examinations and suggestions. Finally, I would like to thank my family who always supported me in my whole education life.

July 2022

Barış BAYRAM
(Computer Engineer, M.Sc.)

TABLE OF CONTENTS

Page

ABBREVIATIONS

AAD	: Acoustic Anomaly Detection
AdaBoost	: Adaptive Boosting
AE	: Autoencoder
AER	: Acoustic Event Recognition
AND	: Acoustic Novelty Detection
ASA	: Auditory Scene Analysis
ASC	: Acoustic Scene Classification
AUC	: Area Under Curve
CAE	: Convolutional Autoencoder
CNN	: Convolutional Neural Network
Conv-LSTMAE	: Convolutional Long Short Term Memory Autoencoder
DBN	: Deep Belief Network
DNNs	: Deep Neural Networks
DOA	: Direction Of Arrival
ESR	: Environmental Sound Recognition
ESC	: Environmental Sound Classification
FFT	: Fast Fourier Transform
FPR	: False Positive Rate
F-TDNN	: Factorized Time-Delay Neural Network
GEVD-MUSIC	: Generalized Eigen-Value Decomposition based Multiple SIgnal Classification
GMM	: Gaussian Mixture Model
HMM	: Hidden Markov Model
iCaRL	: Incremental Classifier and Representation Learning
ICL	: Incremental Class-Learning
iForest	: Isolation Forest
KL	: Kullback-Leibler
kNN	: k-Nearest Neighbour
LPC	: Linear Predictive Codes
LPCC	: Linear Predictive Cepstral Coefficients
LSP	: Line Spectral Pairs
LSTM	: Long Short-Term Memory
LwF	: Learning without Forgetting
MFCC	: Mel-Frequency Cepstral Coefficient
MLP	: Multilayer Perceptron
mPFC	: medial Prefrontal Cortex
MUSIC	: Multiple SIgnal Classification
OCSVM	: One-Class Support Vector Machine
PLP	: Perceptual Linear Prediction
PNN	: Progressive Neural Network
ResNet	: Residual Neural Network
RF	: Random Forest

ROC-AUC	: Receiver Operating Characteristic Area Under Curve
SAI	: Stabilized Auditory Image
SIF	: Spectrogram Image Feature
SSL	: Sound Source Localization
STFT	: Short-Time Fourier Transform
SVM	: Support Vector Machine
TDNN	: Time-Delay Neural Network
TNR	: True Negative Rate
TPR	: True Positive Rate
US8K	: UrbanSound8K
VAE	: Variational Autoencoder
VGG	: Visual Geometry Group
VSA	: Visual Scene Analysis
ZCR	: Zero-Crossing Rate

SYMBOLS

X	:	A set of audio features
C	:	A set of Mel-Frequency Cepstral Coefficients
e	:	Acoustic event class
ε	:	Constant value
$\mathbf{\Sigma}_i)$	:	Covariance Matrix
D	:	Deep audio features
D	:	Discriminant capability
p	:	Distribution of the i^{th} feature unit in the feature vector of an event class, e
q	:	Distribution of the i^{th} feature unit of all the feature vectors
T_{L_e}	:	Event-wise likelihood threshold
I	:	Feature set of a segmented pattern
f_{and}	:	Function of a novelty detector
H_{and}	:	Hidden Markov Model for Acoustic Anomaly Detection
H_e	:	Hidden Markov Model for the event class, e
s	:	Hidden state
k	:	Index of the window
$\boldsymbol{\mu}_i$	:	Mean Vector
I	:	Mel-spectrogram
$N(\dot{)}$	:	Multivariate Gaussian distribution
G_i	:	Novelty score for the audio sample, i
N_s	:	Number of spectrograms
N_i	:	Number of windows from the audio sample, i
N_i'	:	Number of windows selected from the audio sample, i
b	:	Observation probability
o	:	Observation state
r_j	:	Prior probability of the j^{th} event class
S_i	:	Raw audio signal from an acoustic event sample, i
E_R	:	Reconstruction error
λ_e	:	Set of HMM parameters for event, e
X_e	:	Set of the audio feature vectors of event, e
υ_i^k	:	Sum of the sample values from a raw signal
T_S	:	Suspicion threshold
t	:	Time
T	:	Time dimension
A	:	Transition matrix
a_{ij}	:	Transition probability from the state, s_i to the state, s_j
W	:	Window on a raw signal

I^a	**:** Actual spectrogram
f_{icl}	**:** Function of a recognizer for Incremental Class-Learning
y_i	**:** Label of an acoustic event sample, i
I^r	**:** Reconstructed spectrogram
^{T}	**:** Transpose operator
J	**:** Youden's index

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIFELONG LEARNING FOR AUDITORY SCENE ANALYSIS

SUMMARY

Due to the evolution in artificial intelligence and machine learning with the recent advancements in sensor technology, scene analysis is getting more attention for automatic sensing and understanding of dynamic environments including various targets, non-target objects, and noises. The sensory information stemming from the environments can be efficiently analyzed to infer the events, activities, and related objects. However, many issues encountered in the real world, exist that prevent robust sensing and information processing required for important real-time tasks in real dynamic environments with background noises, overlapping targets, high processing complexity, and so on. Automatic scene analysis is a major aspect of the process of collecting and extracting useful knowledge of objects and events to analyze scenes in terms of the places, situations, events, and activities. A significant number of scene analysis studies have mainly focused on visual processing approaches for the analysis of objects in various environments.

Auditory Scene Analysis (ASA) has been used in various real-world tasks, which relies on the perception and analysis of stationary and non-stationary sounds from environmental events and activities, background noises, human voices, and other sound sources. In realistic environments, the dynamic spatio-temporal nature and complexity of environmental sounds, and the existence of novel events may eventually deteriorate the performance of ASA. Therefore, ASA in real-world environments is a difficult task and has not been extensively investigated.

Lifelong learning that is progressively becoming a more crucial task in artificial intelligence is a continuous learning process in acquiring and adapting knowledge from dynamic environments. In this thesis, the task of lifelong ASA for Acoustic Event Recognition (AER) with Acoustic Novelty Detection (AND) is addressed to detect novel acoustic events, recognize known events, and learn in a self-learning manner. The problem is investigated by identifying and tackling various issues that may or will affect the ASA in a real-world learning environment. The main issues of lifelong learning in realistic environments are (i) existence of novel acoustic classes, (ii) existence of unlabeled data, (iii) cost of annotation, (iv) lack of adequate data for novel classes, (v) imbalanced data between classes, (vi) forgetting of previous data, and (vii) lack of memory for storing all the data and (viii) computational power for lifelong learning.

In dynamic acoustic environments, the lifelong ASA for intelligent systems, agents, or robots in real-time is still an open issue. Also, recent deep learning methods for ASA have not been investigated yet while avoiding the issues of real-world environments. In

this thesis, two approaches regarding the main issues, 1) a real-time ASA approach and 2) a deep learning-based ASA approach are investigated, which are able to recognize acoustic events, detect the novel events and then learn by the AER and AND models. However, both lifelong learning approaches have certain issues; which are the lack of incremental learning capability in the real-time approach for the ASA in a realistic environment, and the computational time in the deep learning-based approach. One of the main differences between the approaches is that the real-time ASA is applied to the streaming signal. Thus, each salient sound source in an acoustic scene is identified and localized by a Sound Source Localization (SSL) method to robustly perform the source-specific analysis of its signal. In addition to the SSL method, a segmentation technique is employed to segment variable-length time-series audio patterns of acoustic activities from the streaming signal to efficiently analyze the events and scenes. Moreover, the approaches differ in their audio features used for AER and AND taking into account the requirements of the algorithms and real-time processing.

The first approach for lifelong learning in ASA based on a multilayered Hidden Markov Model (HMM) comprises five main steps: (1) SSL used for detection and location monitoring of the most salient sound source in a scene to perform source-specific analysis, (2) segmentation of time-series audio patterns on the streaming signal, (3) feature extraction from the segmented patterns and construction of a feature set for each pattern, (4) AER in a semi-supervised manner performed by class-specific HMMs associated with known events, (5) AND carried out using a single HMM for all the known events, from the outputs of AER module, and (6) lifelong self-learning (Chapter 4). In the step of lifelong learning, the updates of the models are realized, in which after recognizing an event, the HMM is retrained using more likely knowledge selected among all the previous and new knowledge of the event, and for a new acoustic event recently detected, a class-specific model is generated and the AND model is retrained. In Chapter 4, the offline and real-time experiments are given in detail, which are conducted using streaming signals from a real domestic environment. In the experiments, it is demonstrated that for real-time ASA, HMM for modeling the time-series audio patterns from the streaming signal is the most efficient algorithm for the AER and AND.

In Chapter 5, the steps of the other proposed lifelong learning approach which is a deep learning-based approach for ASA in offline mode are explained, which are: (1) raw acoustic signal pre-processing, (2) extraction of low-level, time-varying spectral representation (spectrogram), and deep audio features, (3) AND, (4) acoustic signal augmentation, (5) AER, and (6) Incremental Class-Learning (ICL) of the audio features of the novel events. The self-learning on different types of audio features extracted from the acoustic signals of various events occurs without human supervision. For the extraction of deep audio representations, in addition to Visual Geometry Group (VGG) and Residual Neural Network (ResNet), Factorized Time-Delay Neural Network (FTDNN) and TDNN-based Long Short-Term Memory (TDNN-LSTM) networks are pre-trained using a large-scale audio dataset called Google AudioSet. As the input of the networks, Mel-Frequency Cepstral Coefficient (MFCC) and raw signals are used by F-TDNN and TDNN-LSTM, respectively, and Mel-spectrograms are taken by the VGG and ResNet. The performances of ICL with AND using Mel-spectrograms, and deep features with TDNNs, VGG, and ResNet are validated on benchmark audio datasets such as ESC-10, ESC-50, UrbanSound8K (US8K), and an audio dataset collected in a real domestic environment, also used with the proposed real-time ASA approach. The results demonstrate that the FearNet algorithm with the VGG-16 features is a more promising algorithm for incremental learning of new acoustic classes in the audio domain, and the GMM algorithm provided the best AND performances in various AND scenarios. Moreover, for the ICL with AND, the FearNet integrated with a GMM exhibits the effectiveness of scene analysis in a real-world acoustic environment to deal with novel events and recognition of unlabeled data using the ICL-based AER. The efficient performances in the experiments of AND and AER tasks are observed also using F-TDNN, and iCaRL has the ICL performances with VGG close to the performances of FearNet in ESC-10 and Domestic datasets.

İŞİTSEL SAHNE ANALİZİ İÇİN HAYAT BOYU ÖĞRENME

ÖZET

Sensör teknolojisindeki son gelişmeler ile yapay zekâ ve makine öğrenmesinde gerçekleşen evrim niteliğinde değişimlerden dolayı sahne analizi, durağan ve durağan olmayan sahnelerde bulunan çeşitli hedefler, hedef olmayan nesneler ve gürültüler dâhil olmak üzere akustik olayların ve aktivitelerin dinamik olarak algılanması ve anlaşılması için son zamanlarda daha yoğun bir ilgi görmektedir. Akustik sensörlerin çeşitli çevrelerde yakaladığı veriler, o sahnedeki olayları, aktiviteleri ve bunlarda yer alan nesnelerden yüksek başarımlı bir şekilde anlam çıkarmak için analiz etmek mümkündür. Bununla birlikte, gerçek dünyadaki dinamik ortamlar, arka plan sesleri ve gürültüleri, birbiriyle örtüşen hedef ses kaynakları, hesaplama için yüksek zaman ve alan gerekliliği vb. gibi sahne analizinde gerçek zamanda hayat boyu öğrenme için verimli algılamayı ve veri işlemeyi olumsuz etkileyen birçok sorun mevcuttur. Otomatik sahne analizi konusu, bu tip dinamik olarak değişen sahnelerde bulunan ortamları, durumları, olayları ve aktiviteleri analiz etmek için nesneler ve olaylar hakkında faydalı bilgiler toplama ve tespit etme sürecinin önemli bir yönüdür. Sahne analizi çalışmalarında çeşitli ortamlardaki objelerin analizi için çoğunlukla bilgisayarda görü (computer vision) alanındaki problemlere odaklanılmıştır.

Sahne analizi gerektiren farklı görevlerde işitsel yetilerin avantajları, İşitsel Sahne Analizi (Auditory Scene Analysis, ASA) araştırma konusu altında çevresel olaylardan, arka plan seslerinden ve insan seslerinden gelen durağan ve durağan olmayan seslerin dinamik olarak algılanmasına ve analizine dayanan çeşitli çalışmalarda ele alınmıştır. Gerçekçi akustik ortamlardaki seslerin, durağan olmayan uzamsal-zamansal doğası ve daha önce gözlemlenmemiş farklı olayların ortaya çıkma ihtimali nedeniyle işitsel sahne analizi performansı zamanla olumsuz olarak etkilenmektedir. Bu nedenle, gerçek dünya ortamlarında işitsel veriler ile sahne analizi derinlemesine incelenmemiş zor bir problemdir.

Yapay zekâ alanında her geçen gün önem kazanan hayat boyu öğrenme, dinamik ortamlarda elde edilen verilerin hayat boyu öğrenildiği bir süreci ifade etmektedir. Bu tezde, dinamik işitsel sahnelerde ortaya çıkabilecek yeni olayları tespit etmek, bu olayları modele öğretmek ve öğrenilen olaylara ait yeni verileri tanımak için işitsel sahne analizi yönteminde Akustik Yenilik Tespiti (Acoustic Novelty Detection, AND) ve Akustik Olay Tanıma (Acoustic Event Recognition, AER) yetileri için gerekli algoritmalar ile hayat boyu öğrenme (lifelong learning) yaklaşımları sunulmaktadır. Bu yaklaşımlar ile yapay zekanın, işitsel sahnedeki sesleri insan müdahalesi olmadan hayat boyu öğrenmesi amaçlamaktadır. Bu tez kapsamında gerçek dünya öğrenme

probleminde, hayat boyu öğrenme tabanlı sahne analizi performansını etkileyebileceği öngörülen ve bu doğrultuda araştırmalar ve geliştirmeler yapılan temel sorunlar şunlardır; (i) yeni ses sınıflarının varlığı, (ii) ortamda yeni ve daha önce öğrenilmiş ses sınıflarına ait etiket bilgisi olmayan büyük miktarda verinin varlığı, (iii) makine öğrenmesi alanındaki en maliyetli iş olan veri etiketlemenin zorluğu, (iv) yeni sınıfların tespit edildiği anda yeterli miktarda verisi olmama ihtimali, (v) sınıflar arası etiketli veri dengesizliği, (vi) kullanılan tanıma ve yeni olay tespit modellerinin daha önceden öğrendiği verileri ve sınıfları yıkıcı unutabilmesi (catastrophic forgetting) ve (7) sahne analizinde hayat boyu öğrenme amacıyla elde edilen verilerin depolanması için ihtiyaç duyulan alan ile (8) bu yöntemler için gereken hesaplama gücü yetersizliğidir.

Dinamik olarak değişen ve durağan olmayan akustik ortamlarda, akıllı bir sisteme, makineye veya araca kazandırmak için gerçek zamanda hayat boyu öğrenme yetisinin geliştirilmesi hala araştırılmakta olan bir problemdir. Ek olarak, gerçek dünya ortamlarının sorunları söz konusuyken ASA için en güncel derin öğrenme tabanlı algoritmalarından ve ses işleme yöntemlerinden faydalanılabilir. Bu tezde işitsel sahne analizinde de karşılaşılabilecek bu temel sorunların ele alındığı hayat boyu öğrenme için (1) gerçek zamanlı ve (2) derin öğrenme tabanlı olarak 2 farklı yaklaşım, araştırılmış, geliştirilmiş ve tartışılmıştır. Ancak, her iki hayat boyu öğrenme yaklaşımının belirli temel sorunları vardır. Bunlar gerçekçi işitsel ortamlarda sahne analizi için gerçek zamanlı yaklaşımda artan öğrenme (incremental learning) yeteneğinin uygulanabilirliği ve derin öğrenme tabanlı yaklaşımda hesaplama süresidir. Bu çözümlerin temel farklılıklarından biri, gerçek zamanlı ASA'nın akış sinyali üzerinde gerçekleştirilmesidir. Bu nedenle, sinyalleri verimli bir şekilde analiz etmek için belirgin ve güçlü ses kaynaklarının bir Ses Kaynağı Lokalizasyonu (Sound Source Localization, SSL) yöntemi ile gerçek zamanda tespit edilip ve lokalize edilmesi sağlanmıştır. Aynı zamanda, gerçek zamanlı ASA için, olayları ve sahneleri verimli bir şekilde analiz edebilmek amacıyla mikrofon ile elde edilen akan sinyaller ile mevcut akustik aktivitelere ait değişken uzunluktaki zaman serisi ses örüntülerini ayrıştırmak için bir segmentasyon yöntemi gereklidir. Bu gerçek zamanlı ve derin öğrenme tabanlı çözümler, algoritmaların ve gerçek zamanlı işlemenin gereksinimlerine bağlı olarak AER ve AND için kullanılan ses öznitelikleri türlerinde de farklılık gösterir.

Çok katmanlı Gizli Markov Modeli (multilayered Hidden Markov Model, multilayered HMM) tabanlı gerçek zamanda ASA için ilk önerilen hayat boyu öğrenme yöntemi beş ana adımdan oluşur: (1) Kaynağa özel analiz gerçekleştirmek için bir sahnedeki en belirgin ses kaynağının tespiti ve konumunun izlenmesi için kullanılan bir SSL adımı, (2) akan sinyal üzerinde zaman serisi ses örüntülerin segmentasyonu, (3) bu örüntülerden öznitelik çıkarma ve her bir örüntüye karşılık bir özellik setinin oluşturulması, (4) AER, her bir bilinen sınıfa özgü eğitilmiş HMM'ler tarafından gerçekleştirilen yarı denetimli (semi-supervised) olay tespiti adımı, (5) yeni sınıfların tespiti amacıyla AER adımında bulunan HMM'lerin çıktıları ile tüm sınıflar için yaratılan bir adet HMM algoritmasının kullanıldığı AND adımı, (6) insan müdahalesi olmadan hayat boyu kendi kendine öğrenme. Hayat boyu öğrenme adımında, AER adımını içeren ve çok katmanlı HMM'in alt katmanında kullanılmak için yakın zamanda tespit edilmiş yeni bir akustik olay sınıfına ait kısıtlı miktardaki öznitelik setleri kullanılarak yeni bir HMM eğitilmesi ve bilinen bir olayın tanınmasından sonra bu sınıfın alt katmanda bulunan modeli, bu işitsel olay sınıfına ait tüm eski ve yeni öznitelik setleri arasından seçilen bu sınıfı en iyi temsil edebileceği düşünülen en değerli öznitelik setleri ile yeniden eğitilmesi gerçekleştirilmektedir. Aynı zamanda, bu hayat boyu öğrenme adımında, üst katmanda bulunan HMM, yeni bir sınıfın tespiti ve bilinen bir sınıfın tanınmasından sonra HMM çıktılarına göre belirlenen en ideal öznitelikler ile yeniden eğitilmesi sağlanmaktadır. Bölüm 4'de, sırasıyla tespit edilen bir dizi ses örneği ve gerçek bir ev ortamından yerleştirilen 4-kanallı mikrofon dizisi ile elde edilen ses sinyalleri kullanılarak yürütülen gerçek zamanlı olan deneyler ayrıntılı olarak paylaşılmıştır. Birçok deney ile geleneksel makine öğrenmesi algoritmalarının da kullanıldığı hayat boyu öğrenme ile gerçek zamanlı ASA için akan sinyalden zaman serisi ses örüntülerini çıkarmak ve modellemek amacıyla önerilen çok katmanlı HMM'nin AER ve AND adımlarını yerine getirmek için en ideal algoritma olduğu gösterilmiştir.

Bölüm 5'te, ASA için derin öğrenme tabanlı hayat boyu öğrenme yöntemindeki adımlar detaylı açıklanmıştır. Bu adımlar sırasıyla: (1) ham akustik sinyal üzerinde ön işleme, (2) düşük-seviyeli geleneksel, zamanla değişen spektral temsil (spektrogram) ve derin ses temsil özniteliklerinin çıkarılması için işitsel öznitelik çıkarma adımı (3) yeni ses sınıflarının tespiti için benzer bir AND adımı, (4) yeni sınıfa ait kısıtlı miktardaki verinin sayısını çoğaltmak için akustik sinyal artırım adımı, (5) bilinen sınıfların tespitinin gerçekleştiği AER adımı ve (6) yeni olayların işitsel öznitelikleri ile Artımlı Sınıf Öğrenimi (Incremental Class-Learning, ICL) adımı. Geleneksel algoritmaların kullanıldığı gerçek zamanlı ASA yönteminde olduğu gibi çeşitli işitsel olayların akustik sinyallerinden çıkarılan farklı özelliklerdeki ses öznitelikler ile bu derin ağ tabanlı artımlı hayat boyu öğrenme de, insan denetimi olmadan gerçekleştirilmektedir. Bu öznitelikler arasında bulunan derin ses özniteliklerinin çıkarılması için Görsel Geometri Grubu (Visual Geometry Group, VGG) ve Artık Sinir Ağı (Residual Neural Network, ResNet), Faktörize Zaman Gecikmeli Sinir Ağı (Factorized Time-Delay Neural Network, F-TDNN) ve TDNN tabanlı Uzun Kısa Süreli Bellek (TDNN based Long Short-Term Memory, TDNN-LSTM) ağları, çok büyük ölçekli bir ses veri seti olan Google AudioSet kullanılarak eğitilmiştir. Bu derin sinir ağlarından, VGG ve ResNet için Mel spektrogramları, F-TDNN için Mel-Frekans Cepstral Katsayısı (Mel-Frequency Cepstral Coefficient, MFCC) öznitelikleri ve TDNN-LSTM için ham ses sinyalleri girdi olarak işlenip bu derin ses özniteliklerini üretilmektedir. ESC-10, ESC-50, UrbanSound8K (US8K) ve tez kapsamında gerçek zamanlı ASA deneyleri için toplanılmış ev ortamındaki sesleri içeren veri setinden çıkarılan bu derin öznitelikler ve Mel-spektrogram ile çeşitli AND ve ICL tabanlı AER modellerinin performansları karşılaştırılıp değerlendirilmiştir. Bu elde edilen sonuçlar, 16 katmanlı VGG'nin derin öznitelikleri üzerinde Gauss Karışım Modeli (Gaussian Mixture Model, GMM) algoritmasının çeşitli AND senaryolarında en iyi performansları sağladığı ve FearNet algoritmasının, işitsel problemlerde yeni sınıfların öğrenilmesi ve tanıması için daha umut verici olduğu gösterilmiştir. Ayrıca bu deneylerde AND ile ICL için, FearNet ve GMM'nin entegre edilmesi, yeni olayların tespit edilmesi, bu yeni olayların öğrenilmesi ve ortamda bulunan etiketlenmemiş verilerin tanınması için sunulan yaklaşımlar ile sahne analizinin güvenilirliği ve verimliliği gösterilmiştir. Bunun yanında, F-TDNN derin özniteliği ile tatmin edici AND ve ICL tabanlı AER performansları ve iCaRL algoritması ile de bu veri setlerinde etkin ICL performansları elde edilmiştir.

1. INTRODUCTION

Due to significant advancements in artificial intelligence and recent breakthroughs in machine learning, traditional and deep learning algorithms have been developed for automatic scene analysis which is rather a challenging task in different domains (e.g. computer vision and audition). Scene analysis is a major research topic of sensing, understanding, and interpreting relationships of objects and events in a scene. For various purposes, Auditory Scene Analysis (ASA) is a developing research area that takes advantage of more mature fields of machine learning, signal processing, and other modalities such as computer vision and robotics (Figure 1.1). The ASA works have generally focused on different tasks such as recognition of urban and natural events [1], [2], recognition of domestic events [3], interaction with objects [4], animal life monitoring [5], surveillance [6], etc. Scene analysis in acoustic environments including several sources making concurrent sounds for detecting target sounds and understanding the relationships between the events, objects, actions, and humans is also known as the cocktail party problem [7]. The problem is one of the most challenging problems faced in auditory perception and computational ASA discussed in many studies. However, the real-world environments have streaming data such as acoustic signals from unseen classes, and a large scale of unlabeled data. Therefore, an ASA system should be able to recognize acoustic events and detect and then learn novel events.

The tasks of Acoustic Event Recognition (AER) and Acoustic Novelty Detection (AND) have not received as much attention as speech recognition which is still considered to be the primary goal in the auditory domain. Scene analysis with AER and AND on audio signals has been carried out regarding different application domains such as surveillance [6], elderly human monitoring [8], home automation [3], and robotics [9]. In real-world acoustic scenes, environmental sounds have spectral characteristics with a broad flat spectrum unlike speech signals and musical sounds [10]. Therefore, for learning the non-stationary and dynamic nature of the

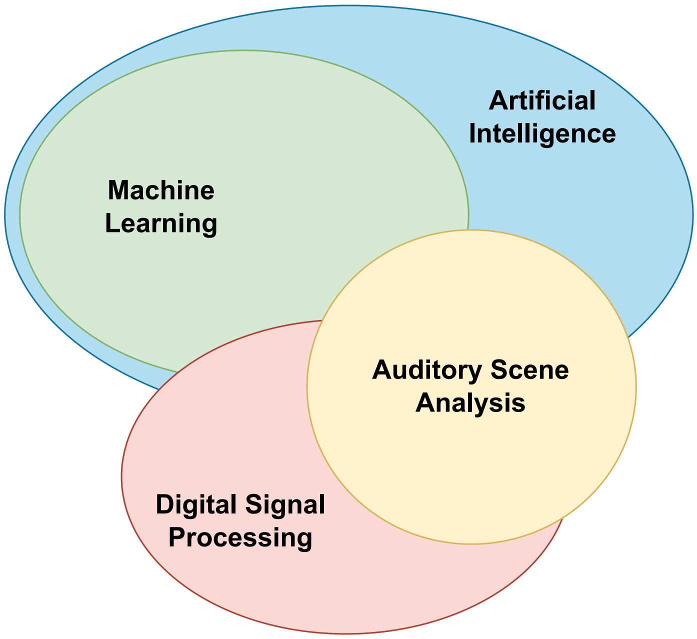

Figure 1.1 : Auditory Scene Analysis (ASA).

environmental and event sounds, several works have focused on the development of AER, AND, and Acoustic Anomaly Detection (AAD) algorithms for alleviating the difficulties of the sounds. The differences of environmental sounds from speech signals adversely affect the recognition of events in the environments [11]. Various types of audio features including low-level features (e.g., zero-crossing rate and short-time energy) and complex high-dimensional features (e.g. Mel-Frequency Cepstral Coefficient (MFCC) [12], Mel-spectrograms [13], gammatone-spectrograms [14], and wavelet-based features [15]) have been investigated for these audio tasks.

One of the most popular tasks in the computer vision domain that has not been deeply investigated yet in the auditory domain is human-like lifelong learning also known as continual, online, evolutionary, streaming, sequential, or incremental learning. Traditional machine learning algorithms are not scalable, thus not able to fit the model with large-scale data [16] compared to deep learning algorithms. On the other hand, traditional Deep Neural Networks (DNNs) perform poorly on data of unseen classes [17] and may be inefficient to incrementally learn in real-time due to the computational time and inadequate volume of data for generation of DNN models [18]. Moreover, the DNNs are generally prone to "catastrophic forgetting" [19, 20], in which previously seen instances, classes, or tasks may be forgotten. Traditional algorithms can be exploited by learning from scratch to address the forgetting problem by retraining the model to maintain accurate prediction performances.

In the last few years, Incremental Class Learning (ICL) using deep learning-based approaches has attracted considerable attention. The aim of ICL is to sequentially learn new classes and/or tasks without forgetting and re-using previously learned data [21] in a different domain using architectural, regularization, and rehearsal strategies, in combination or independently. In literature, most of the ICL studies have investigated catastrophic forgetting problem in computer vision tasks (image classification [22], semantic segmentation [23], image classification in a number of isolated tasks [24]). Incremental learning of images from a new class or task was aimed, thus, the ICL algorithms were developed using the images of different objects/scenes, and have not been evaluated using different types of sensory data such as audio signals individually or in combination with visual images like a hybrid ICL method. Only the works of Eunjeong et al. [25] and Yu et al. [26] have dealt with incremental learning in the auditory domain to achieve better Acoustic Event Recognition (AER) performances. Besides AER, several tasks such as speech recognition, voice detection, AAD, AND, etc. can be enhanced by the capability of incremental learning without forgetting.

In real-world tasks, for supervised traditional and deep learning-based algorithms, a large amount of annotated data is required, but the data specific to the problem may not exist due to the nature of streaming data and the expense of annotation. Also, in real dynamic environments, novel acoustic events or unknown noise signals may occur that exacerbate AER performance. Therefore, to improve the ASA performances in the acoustic tasks and to achieve AER in acoustic scenes with novel event classes, the novel acoustic classes should be detected and adapted into the algorithms without catastrophic forgetting to enhance the recognition capability and to provide the continuity of the scene analysis process. After detecting a novel acoustic event, only a few data are adapted into the AER model by retraining the model using a subset of previous data, or incrementally learned by an ICL-based AER model without access to the previous data. Furthermore, the AND model is re-generated using the scarce data in a semi-supervised manner. However, the scarcity is an essential challenge also in acoustic tasks, therefore, audio signal augmentation is investigated to increase the number of audio samples by time-stretching of the signals [27].

1.1 Problem Statement and Technical Issues

In traditional tasks of machine learning, statistical and algorithmic models are trained once using entire data with a static setting in which an assumption is made that there is only data sampled from a stationary distribution (a sine qua non-condition) and a fixed number of classes. Also, the data used for training and evaluation is entirely labeled, and the existence of unlabeled data is ignored. However, in real-world problems, analyzing and learning from dynamically changing environments including non-independent and identically distributed sampled data and the possibility of new objects/classes emergence is a natural and continual process. The data of the environments are often processed in a semi-supervised manner [28, 29] due to the existence of a large number of unlabeled data [30] and the lack of ground truth labels [31], and the emergence of new acoustic classes [32]. The annotation of the data requires too much time and high cost, and it may be the most expensive work in machine learning tasks, which is the main reason why lifelong learning is an absolute necessity. On the other hand, new classes in the unlabeled data cause more challenges for lifelong learning. Lifelong learning is the paradigm that combines multiple tasks and algorithms to tackle the problem of learning in such dynamic environments especially regarding the tasks of adaptation of new data and classes, preserving the learned knowledge, detection of novel classes, and memory and computational time limitations.

One of the lifelong learning solutions for the evolutionary process is achieved by saving all data or little portion of the data to combine with the novel data for re-generation of the learning model. The streaming data in real-world in a semi-supervised manner, because only a few acoustic data are labeled and there is the lack of supervision and ground truth labels for new data. For lifelong learning based ASA with low computational cost, it is important to find a reliable and stable solution, in which the AND and AER algorithms are affected as little as possible when new knowledge of a known class or a completely new class is being learned. Hence, the proposed approach returns a prediction among three outputs which are “known”, “unknown”, or “uncertain”. In real environment, comprehensive experiments were conducted to demonstrate the efficiency of the approach with low computational cost. However, the optimization of computational time and memory needs to be taken into account in the lifelong learning tasks, while continually learning and preserving the previously learned knowledge. Moreover, the streaming audio signal may even not be stored or accessible any longer and there may not be sufficient computational power to retrain the model. Therefore, the continual learning approach should be able to sequentially insert a new class/task under these challenges while tackling the catastrophic forgetting problem [19] which refers to the phenomenon of a neural network, and disentangling the complexity.

1.2 Overview of Proposed Approaches

In this thesis, lifelong learning-based AER with AND is proposed for AER in acoustic scenes with unknown events using traditional methods for lifelong learning with low computational complexity [33] and deep learning algorithms for lifelong learning approach [34]. For the approaches, various algorithms are evaluated on different types of audio features in terms of the performances of AER and AND in several benchmark audio datasets, and the ASA with low computational complexity is assessed on the streaming audio signal from a real domestic environment. In this thesis, a lifelong learning approach for ASA is presented which is reliably capable of 1) starting learning with one class and a small subset of the data, 2) detection of new classes over an undefined period of time, and 3) sequentially learning of new knowledge in time. This approach includes four main components: (1) segmentation of audio patterns from streaming audio signal, (2) a AND model that is a detector of novel events in the environment, (3) an AER model for each known event to predict on the known events, and (4) adapting new data into the AND model and AER models in a semi-supervised manner. The ASA with self-learning on streaming data can be used by intelligent systems, agents, robots, etc., which would robustly analyze auditory scenes through time.

For the proposed deep learning based lifelong learning for ASA [34], audio features composed of well-known low-level features are assessed in terms of computational time and effectiveness for the detection of novel events and recognition of known events. High-level audio features such as spectrograms of the sounds, and deep audio representations based on the spectrograms are preferred to feed deep learning models used for AER, ICL, and AND. To extract the deep audio representations from traditional audio features, spectrograms, or raw signals, in this thesis, various deep networks pre-trained using a large scale of audio datasets are exploited, and the networks are evaluated in different benchmark datasets. In addition, different types of Time-Delay Neural Networks (TDNNs) based networks used to recognize speech, speech emotion, and speaker and to detect voice activity are sufficient to capture complex temporal characteristics of sounds with transient, intermittent, and continuous temporal structures. Therefore, Factorized TDNN (F-TDNN) [35] and TDNN with Long-Short Term Memory (TDNN-LSTM) [36] are employed to extract the deep representations from MFCCs, traditional audio features, and raw signals, respectively, to demonstrate the use of TDNNs in AER task and the use of the acoustic data for extraction of deep features instead of spectrograms. ResNet and VGG which are the most used networks for feature extraction from audio spectrograms and TDNNs are pre-trained on AudioSet [37], including a large amount of audio data and then tested with different benchmark datasets.

In the low computational cost approach for ASA [33], a multilayered HMM composed of a lower HMM for each known event and an upper HMM for detection of unknown events is developed in which the AND model is employed in a semi-supervised manner. The performance of AND is compared with One-Class Support Vector Machine (OCSVM) [38] and regular Hidden Markov Model (HMM). For deep learning-based ASA, an ICL model is integrated with an AND model in which the most appropriate AND algorithm is heuristically found after conducting several experiments using the following state-of-the-art methods in a semi-supervised manner: stacked Autoencoder (AE) [39], Variational AE (VAE) [40], k-Nearest Neighbour (kNN) [41], Gaussian Mixture Model (GMM) [42], OCSVM [38], and Isolation Forest (iForest) [43]. To achieve ICL, Learning without Forgetting (LwF) [44], Incremental Classifier and Representation Learning (iCaRL) [45], and FearNet [46] were employed with five types of audio features: Mel-spectrograms, and deep features from TDNN, TDNN-LSTM, ResNet, and VGG.

The low computational cost approach for ASA [33] is developed regarding the requirement of computations with low complexity composed of the steps; (1) Sound Source Localization (SSL) used for detection and location monitoring of the most salient sound source in a scene to perform source-specific analysis, (2) segmentation of variable-length time-series audio patterns on the streaming signal, (3) feature extraction from the segmented patterns and construction of a feature set for each pattern, (4) AER in semi-supervised manner performed by a number of AER models including a class-specific model for each known event, (5) AND carried out using a model for all the known events using the outputs of AER module, and (6) lifelong self-learning in which for a recognized event, a model is retrained using more likely knowledge selected among all the previous and new knowledge of the event, and for a new acoustic event recently detected, a class-specific model is generated and the AND model is retrained. Using the capabilities, it is planned to achieve self-learning-based ASA in real dynamically changing environments. The most appropriate type of audio features for AER and AND is determined and utilized for the ASA process. However, in real environments, polyphonic sounds exist which make it harder to analyze the activities. Thus, the segmentation of audio patterns is performed on the incoming signals to obtain multiple segments from the sound of acoustic activity. The SSL and segmentation steps of the ASA will improve the recognition of the events and adaptation of the knowledge from known or unknown events, to the current AER and AND models.

Furthermore, the proposed ICL based ASA approach [34] comprises the following steps: (1) pre-processing, (2) extraction of audio features, (3) AND, (4) augmentation of audio signals, (5) AER, and (6) ICL of new acoustic event classes. The SSL and segmentation steps are required for analyzing streaming signals, and due to the use of source-specific audio samples, the localization of the sources and segmentation of the audio patterns are not required, thus, these are not included in the deep learning based approach. Besides these steps, there is a pre-processing step in which a window slides on the signal to eliminate the segments including silence instead of acoustic activities. In this approach, the ASA process is invoked using only one randomly selected event class, and a number of samples from all the unknown events is sequentially analyzed. The event instance recently detected as a new class may have only a limited number of audio features due to the size of the selected windows. Thus, an augmentation technique is applied to the window to increase the amount of data used in the ICL step. Moreover, the evaluation of different types of audio features and several algorithms of ICL and AND is carried out in the experiments.

1.3 Contribution of the Thesis

The contributions of this thesis can be divided into two main parts for each approach low computational cost lifelong learning and deep learning-based lifelong learning. For the lifelong learning with a low computational cost for ASA, the contributions are as follows;

1. detection and adaptation of novel acoustic events with low computational complexity for lifelong learning,
2. ASA with self-learning on the audio features without human supervision,
3. integration of the novelty detection with AER in a single framework for AER in the scenes with unknown acoustic events, and
4. the realization of lifelong learning using an audio dataset collected in a real domestic environment in the scope of this thesis.

To our knowledge, this work is the first lifelong learning study on streaming audio signals collected from real environments. The effectiveness and applicability of the proposed approach are demonstrated in a set of comprehensive experiments using domestic sounds recorded in the real world.

Furthermore, the contributions of the proposed deep learning-based lifelong learning for ASA can be listed as follows;

1. the use of ICL algorithms using acoustic signals in an AER task,
2. the integration of ICL algorithm based AER with the novelty detection method,
3. extraction of deep audio representations with the pre-trained ResNet, VGG, F-TDNN, and TDNN-LSTM,

4. augmentation of audio signals to increase the number of audio samples belonging to recently detected novel events, to be used for ICL and the update of the AND algorithm,

To the best of our knowledge, the contributions regarding ICL with AND in acoustic tasks and pre-training of several networks for ICL and AND appear in this thesis for the first time in the literature. The proposed approach for ICL with AND was evaluated on the benchmark audio datasets Environmental Sound Classification with 10 classes (ESC-10), ESC with 50 classes (ESC-50) [47], UrbanSound8K (US8K) [1], and the audio dataset (Domestic) collected by the authors using a microphone array located 1 m from the sound sources.

1.4 Thesis Structure

The rest of the thesis is organized as follows: Chapter 2 is a discussion of the related work in traditional and deep learning-based algorithms for AER and AND, low-level audio features and deep audio features using various DNNs for ASA, and the algorithms for low computational cost lifelong learning and ICL based lifelong learning. The steps of the proposed traditional and deep learning-based lifelong learning approaches are given in Chapter 4 and Chapter 5, respectively. The implementation details of the techniques and algorithms used for all of the tasks included in these approaches and the experiments are provided at the end of the related Chapter 4 and Chapter 5. The performances of AER, AND, and ICL algorithms with various types of audio features used in the experiments are discussed in Chapter 6. Finally, the conclusion and future work for lifelong learning based ASA are detailed in Chapter 7.

2. LITERATURE REVIEW

For the lifelong learning problem in ASA, various steps are required to learn in a lifelong manner for efficient analysis of events. In this chapter, the works for the steps which are preprocessing techniques, audio features, AER, AND and lifelong learning are explained.

2.1 Audio Preprocessing

In the study of [48], the augmentation of spectrograms extracted from the sounds in the ESC-50 dataset was carried out to generate a CNN-based ensemble method for AER. Pandeya et al. investigated different augmentation techniques such as random time-stretching, pitch-shifting, inserting noises of different ranges and dynamic range compression in order to augment the raw audio signals in the datasets of domestic cat sounds [49]. Also, in [50], the pitch-shifting and time-stretching methods were applied to augment spectrograms for use as inputs to a CNN model. The augmentation of the raw audio signals and the spectrograms of the signals was performed on natural animal sounds to improve performance of the classification of animal sounds [51]. A few audio data are obtained when a new event class is detected, which is required to incrementally learn the new class by the AER model and to update AND model. Therefore, in this thesis, augmentation of an audio signal is applied to increase the number of audio features. Moreover, a segmentation task based on a energy-based detector exploited in the proposed ASA approach with low computational complexity is one of the preprocessing tasks to detect change points with high energy values in the streaming audio signal. The segmentation step is important for the elimination of the segments without at least 2 seconds to analyse the signal with lower computational time than the analysis of the entire streaming signal. An overview of various techniques for audio segmentation has been presented in the work [52]. Using onset and offset thresholds, each frame which is called a "super-frame" is selected, since the frames may include more discriminative information representing the event [53]. Eventually, it is shown by several works that a super-frame-wise recognition is achieved with higher accuracy. Therefore, AER is performed on these interleaved super-frames. Also, for estimation of the threshold values is achieved by manual or a rule for dynamically determining the thresholds.

Another task of ASA is Sound Source Localization (SSL) to localize the sound sources which may be belonging to target and non-target events and background noises. The location information can be useful for the analysis of source-specific signals. Various techniques have been developed to estimate the Direction Of Arrival (DOA) from the emitted signal captured by a microphone array [54]. The SSL approaches based on beamforming [55] and Multiple SIgnal Classification (MUSIC) [56] are widely used and investigated in different types of environments. MUSIC, used in this thesis, is robust and efficient for real-time SSL, which is a subspace-based method to derive a steering vector in order to detect peaks with low computational cost. Another MUSIC-based SSL approach for dynamic environments is Generalized Eigen-Value Decomposition based MUSIC (GEVD-MUSIC) [57] which uses a noise correlation matrix to suppress the noises and performs noise-robust localization in real-time. For ASA, an SSL technique can be deployed to localize dominant sound sources in the scene, to monitor the localized sources, and to analyze the source. In the works [58, 59], Bian et al. proposed an SSL approach to monitor and infer domestic activities. Moreover, the combination of the tasks for sound recognition and SSL based on MUSIC is employed using Gaussian Mixture Model (GMM) with outlier rejection for ASA [60]. In this thesis, the SSL is utilized as a step of the ASA approach to achieve source-specific analysis.

2.2 Audio Feature Extraction

Traditional and deep learning algorithms have been thoroughly applied to different types of audio features for the acoustic tasks of Acoustic Scene Analysis (ASA) such as Acoustic Scene Classification (ASC), Acoustic Event Recognition (AER), Environmental Sound Recognition (ESR), Acoustic Novelty Detection (AND), and Acoustic Anomaly Detection (AAD). Various low-level audio features, time-frequency representation, and deep audio features have been used in the tasks. Besides the acoustic tasks, lifelong learning has been rarely discussed in the auditory domain in the literature.

Different types of low-level audio features can be divided into time and frequency-domain features considering type of the acoustic problem. In the early works, Zero-Crossing Rate (ZCR) [61], an important time-domain feature has been mostly used for the classification of vocal and non-vocal sounds to measure the number of times when the sign changes among the successive samples. The ZCR feature is exploited to determine the most dominant frequency of an audio frame, which provides important information to discriminate voices from unvoiced speech [62] and to be combined with energy features for the separation of music signals from speech signals [63]. Also, the ZCR feature is combined with other features such as Mel-Frequency Cepstral Coefficient (MFCC) and discrete wavelet transform to improve the AER performance [64].

The MFCC feature which is the most well-known audio feature used in speech-related problems has been exploited in various AER works to describe cepstral characteristics of the sounds. The features are computed using discrete cosine transform on the log-energy values of the Mel-scaling filter-bank [65]. Also, its first and second-time derivatives $\Delta MFCCs$, $\Delta^2 MFCCs$ [66] are utilized for AER tasks in different acoustic environments. Other low-level audio features are utilized to support MFCC for AER in many works. A feature set composed of the combination of MFCC and the spectral features which are spectral centroid, spectral spread, and spectral flatness is employed in an HMM-based AER approach for an acoustical surveillance task [67]. In the work [68], the performance of MFCC feature representation combined with the MPEG-7 low-level descriptors and wavelet packets is evaluated for AER. Moreover, MFCCs were used as the input of a DNN in the study of Kataria et al. [69] to construct x-vector embeddings for speaker verification tasks. Also, other cepstral-based features such as Linear Predictive Codes (LPC) and Linear Predictive Cepstral Coefficients (LPCC) were utilized to characterize audio contents [70], the sounds from acoustic events [61], the signals of pure and vocal music [71] and for the sound classification. In this thesis, the MFCC feature is used for low computational cost lifelong learning as the main feature. Also, in deep learning based ASA approach, the feature is exploited for the extraction of deep audio representations by a pre-trained deep network.

MPEG-7 audio descriptors have been utilized as features for the classification of environmental sounds, music sounds [72], and speaker recognition [73]. Also, MPEG-7 based ESR was proposed in [74] in which the spectral descriptors including spectral envelope, spectral centroid, spectral spread, spectral flatness, etc. are used. The spectral features extracted from the acoustic events are as follows: the spectral centroid is a measure of the shape of the power spectrum; the spectral spread gives the information about the spectral shape by taking the root-mean-square deviation of the spectrum from its centroid; and the spectral flatness measures the flatness of the power spectrum. The other frequency-domain spectral feature is roll-off [75] for the measurement of the frequency, which takes 95% of the power spectrum and indicates the skewness of the spectral shape, and pitch ratio calculated on the pitches estimated in each frame. The roll-off spectral features are used to discriminate the speech from non-speech signals [76].

Time-frequency representations of sounds such as Mel-spectrograms are the most widely used feature with DNNs in the past years. In a DNN based work for AER, a Convolutional Neural Network (CNN)-like network, ResNet was utilized on Mel-spectrograms [77], in which as the input of the ResNet three-channel (RGB) images were used, so the spectrograms were modeled like RGB images. Also, Müller et al. [78] present a study of the use of deep audio representations extracted by ResNet and VGG from audio images such as Mel-spectrograms for the detection of abnormal sounds. The spectrograms which are the 2D audio representations were employed to generate a model of Deep Belief Network (DBN) to extract high-level feature representations [79]. In a transfer learning fashion, a DNN trained using a large-scale dataset of a different particular task that may not be in the auditory domain was exploited to extract deep audio representations for the sounds of a different dataset [27]. In addition, Time Delay Neural Networks (TDNNs) were exploited using different audio features or raw signals in order to employ the recognition of speech [80], emotion [81], speaker, or voice activity. However, the facilities of TDNNs have not been investigated in the AER problem using event sounds in different acoustic environments. In this thesis, for AER, AND and ICL tasks of the proposed deep learning-based ASA framework, the effectiveness and the efficiency of transfer learning were demonstrated with TDNN based DNNs that are pre-trained on a large scale audio dataset to extract the embedding of acoustic events.

Feature selection is applied to different types of low-level audio features in order to reduce the number of features and to estimate the most discriminative features for AER and AAD. A feature selection method with a Kullback-Leibler (KL) distance-based Adaptive Boosting (AdaBoost) model was carried out to estimate the most discriminative audio features [82]. In this thesis, the KL distance is investigated to determine the most important features to improve and speed up the training, update, and prediction of HMMs for AER and AND with low computational cost.

2.3 Acoustic Event Recognition

For recognition of acoustic events, various traditional algorithms such as HMM, Gaussian Mixture Model (GMM), Dynamic Time Warping (DTW), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), k-Nearest Neighbor (kNN), etc. and deep learning-based algorithms such as Long Short-Term Memory (LSTM), CNN, etc. have been investigated using different types of audio features [83].

2.3.1 Traditional algorithms

In several studies [84–86], for AER in a time-series manner, Hidden Markov Model (HMM) was a widely used algorithm with MFCC features for speech recognition in the past many years [87]. In the work of Do et al. [88] in which elderly life monitoring is carried out using an SVM based AER platform to improve their life quality. Using the platform, human voice and activities are reliably localized and recognized using MFCC features in a domestic environment. In the study [89], an AER approach based on a multiple instance learning classifier was presented to recognize the events in the scenes like a crowded restaurant, busy street, pool hall, play area of kids, etc. The approach is a maximum margin early event detector both operating in the feature and landmarking spaces used for measuring similarities of the temporal sequences of MFCC features. In another AER work [68], GMM and HMM with MFCC, MPEG-7 low level descriptors and wavelet features are employed for automatic recognition of different sounds from speech, music, pets, doorbell, baby crying, explosion, gunshot, and laughter is carried out in a domestic environment. Also, the ASA with low-cost computations is desirable in various robotic tasks to be integrated into robot audition. In the study of [9], the SVM and MLP algorithms were exploited to analyze the acoustic signals in a robotic task to infer the physical and functional properties of the objects, which are the material type and the event that occurred while the robot interacts with the objects. Another study focused on the use of kNN based approach for the deployment of ASA in a robotic manipulation task in which five manipulation behaviors performed by the robot on 36 objects, which are grasping, shaking, dropping, pushing, and tapping are recognized [90]. Phan and Mertins [91] proposed an event-wise recognition system using different SVM models, and demonstrated the SVM model with Chi-square and histogram intersection kernels. Feki et. al. [92] proposed a HMM based framework to recognize environmental sounds, consisting of three stages; (1) preprocessing for detecting the silence by segmenting the audio stream, (2) automatically classifying sounds based on time frequency analysis into speech, music and urban, and (3) implementation of encapsulating of binary classifiers. Also, Lafay et. al. [93] presented a model for simulating environmental acoustic scenes in order to detect events with stronger background noise by the classifiers, 2-layer HMM, RF, SVM and Non-negative Matrix Factorization. Another HMM based approach [94] was developed to recognize the environmental sounds using Local Discriminant Bases (LDB) technique to estimate the discriminatory time-frequency subspace as features.

2.3.2 Deep learning algorithms

In recent years, Deep Neural Networks (DNNs) for the acoustic tasks have not received as much attention as computer vision-related tasks. One of the most recent deep learning problems, which is Incremental Class-Learning (ICL) mostly handled in many studies in the computer vision domain has been covered in the auditory domain for the first time in this thesis. This chapter describes several ASA works focused on the use of various machine learning algorithms with different types of audio features for AER, AND, and ICL.

Two-dimensional time-frequency representations extracted from the sounds, which are Stabilized Auditory Image (SAI) [95], Spectrogram Image Feature (SIF) with SVM, and with a deep neural network studied in [96], were investigated for AER. Many works have demonstrated that the DNNs substantially increase the performances of acoustic tasks. For the enhancement of the AER performance, DNN-based approaches have been presented for different acoustic tasks such as noise reduction [97], and dereverberation and beamforming [98].

2.4 Acoustic Novelty Detection

The novelty detection problem has been investigated for different purposes in various domains. The problem was initially investigated on computer vision tasks such as image processing, video surveillance where a large amount of data is available. Also, in the other early works, the novelty detection was achieved in text mining and network intrusion [7]. The techniques are divided into probabilistic, distance-based, reconstruction-based and domain-based techniques.

2.4.1 Probabilistic algorithms

The probabilistic novelty detection algorithms aim to estimate generative probability density function for the data of known classes. Using the algorithms, the training data is assumed to be represented by the density function, and the novelty detection is achieved with a threshold to define the boundaries of the known in the data space. One of the popular probabilistic algorithms for novelty detection is Gaussian Mixture Model (GMM) which is based on the Expectation Maximization (EM) technique [99] to fit a number of Gaussian distributions. The EM aims to maximize the likelihood value. Another probabilistic method is Hidden Markov Model (HMM) which is a sequential algorithm based on graphical models to estimate state sequences for sequential data in successive order [100, 101]. In this thesis, for the novelty detection steps of the proposed approaches, the probabilistic algorithms which are different types of GMMs and left-to-right HMMs are investigated regarding the number of states.

2.4.2 Domain-based algorithms

One-class algorithms have been developed to learn the domain by taking into account normal classes, which are known as domain-based algorithms. For the novelty detection, the one-class novelty detection algorithms have been employed using known classes in different domains. One of the famous algorithms is One-Class Support Vector Machine (OCSVM) which is based on a kernel function to compute a hyperplane with a maximum margin in a feature space, to separate the sets from the origin [102, 103].

2.4.3 Reconstruction-based algorithms

Besides the traditional one-class methods, many studies have focused on unsupervised DNNs generated by the data of known/normal classes for the detection of novel and anomaly sounds in the environment [40, 104, 105]. Mostly, for this purpose, unsupervised deep networks such as Recurrent Neural Network-based Autoencoders on a benchmark speech dataset [105], and as in our work, Convolutional Long Short Term Memory Autoencoders (Conv-LSTMAE) and Convolutional Autoencoders (CAE) on the sounds of the different types of manufacturing processes [106] have been employed on spectrograms of the sounds. Using the traditional networks, the adaptation of the recently detected class into the novelty detection model consumes too much time and is not proper for ASA with low-cost computations. Sequential AEs have been investigated for AAD in industrial environments [106]. In the other AE-based work [107], Nguyen et al. presented a semi-supervised Variational Autoencoder (VAE) method for the detection of novel anomalous classes. Only a few studies have been conducted in a few-shot learning manner to detect new acoustic event classes. To detect rare sounds in acoustic environments with background noises, a few-shot learning-based solution is presented [108].

2.5 Lifelong Learning

Lifelong learning is a learning paradigm in which continually learning new knowledge is carried out while retaining the previously learned knowledge. In real-world environments, the capability of continual learning is essential for intelligent systems and agents in lifelong time. For lifelong learning, brain-like algorithms have been developed to achieve many tasks in robotics [109], self-driving [110], monitoring for defect recognition [111], unmanned aerial vehicles with IoT devices [112], natural language processing, etc. A few solutions have been proposed to achieve incremental HMM algorithm without retraining depending on the online Baum-Welch and EM techniques [101], but none of them has been assessed in a real-world task with real streaming data, and so the works may not provide suitable AER performances as much as retraining itself. The major problem of incremental learning with deep neural networks is catastrophic forgetting [19] extensively investigated in several studies [113].

After learning a new class, the parameters of the new model should not deviate too far from the configurations of the previous model. The base strategies of ICL are architectural, regularization, and rehearsal used in combination or independently for the development of ICL algorithms in several studies. The main aim is to learn new classes incrementally while avoiding the catastrophic forgetting problem in the deep learning-based incremental learning works, and so different types of deep networks have been proposed for this aim. One of the first and most used ICL algorithms based on a data-driven regularization technique is Learning without Forgetting (LwF) [44], which focuses on transferring knowledge for tackling the forgetting problem. Another popular ICL algorithm called "Incremental Classifier and Representation Learning" (iCaRL), combining rehearsal and regularization strategies, was proposed in [45] using the nearest exemplar algorithm and a fixed size external memory for previous exemplars. Therefore, a nearest-exemplar algorithm is employed for classification and substantial changes are prevented using the memory. In many studies, the efficient performances of the iCaRL algorithm have been presented using different computer vision tasks, however, the algorithm needs to store the samples of known classes. Moreover, a memory-based generative approach known as FearNet [46], based on a brain-inspired, dual-memory system achieves continual learning of new classes by storing and replaying detailed statistics about known classes instead of the previous knowledge. For its incremental learning process, three short- and long-term memory networks and a decision network are exploited to choose the activated network [46].

In a few studies, the lifelong learning paradigm has been discussed in the audio domain. Most of the recent studies for novel object detection and incremental learning have used deep learning methods in the field of computer vision. A few recent works have focused on incremental learning using CNN for AER. In [25, 26], the performances of incremental learning were evaluated using Mel-spectrograms from one-second audio files. However, the detection of novel classes in the algorithms for incremental learning has not been addressed in the ICL studies. Furthermore, the works did not investigate widely used ICL algorithms and make analyses of different audio features. To the best of our knowledge, the combination of ICL and AND has not yet been explored in the audio domain.

The lifelong learning problem has been widely investigated in studies using various traditional and deep learning-based algorithms. Only a few works have focused on real-time or near real-time learning on the streaming data known as *streaming learning* that may belong to a novel class, therefore, in these works, an incremental approach was proposed in which a recognition method is integrated with a novelty detection algorithm. Husam et al. [114] presents a Parzen window kernel density estimator-based incremental approach with novelty detection for gesture recognition on streaming data in real-time. The approach is able to detect novel gesture classes and adapt the new classes to the recognition model. In a relevant work [115], an algorithm was developed to achieve ICL without forgetting new classes. However, the label information of new classes is required for the algorithm like in the other ICL methods, in which there is not a novelty detector. Also, ICL was carried out in a few-shot learning manner using Attention Attractor Network which is a meta-learning method [116].

In various studies, a lifelong learning approach has been extensively utilized using different machine learning algorithms. However, only a few works have been developed to combine the tasks of recognition and novelty detection. For incremental learning and novelty detection, one approach has been proposed based on the Parzen window kernel density estimator using gesture data including novel classes [114]. Another work [117] for gesture recognition and detection of novel gesture classes has been presented using an ensemble method in a semi-supervised learning manner. In general, most online learning approaches require a high computational cost and a lack of detection and learning of new classes in real-time or near real-time. However, learning is a continual process based on experience throughout life, and a variety of tasks in real-world problems from diverse domains require a continual learning capability to enhance the performance of the algorithms. The other ICL approach using CNN was proposed in [115] in which unknown classes are detected and incrementally learned by preserving the previous knowledge of the known classes. An incremental few-shot learning approach based on Attention Attractor Network, a meta-learning method has been presented to incrementally achieve few-shot learning without retraining the data [118].

3. PROBLEM DEFINITION AND THEORETICAL BACKGROUND

In this chapter, the problem of lifelong learning for ASA is discussed, and the main issues in scenarios for low computational cost and deep learning-based lifelong learning which are taken into account in this thesis are explained.

3.1 Problem Definition

In real acoustic environments, unseen acoustic classes may exist, and different audio patterns of a known class may be obtained. The ordinary situations affect the scene analysis and decrease the performance of the sound recognition. Therefore, an intelligent system for ASA needs a lifelong learning capability to learn novel events and instances in a long-term. This capability is essential to adapt to changing environmental acoustic conditions.

Traditional methods used for various tasks of audition aim to process streaming audio signals from a microphone. However, the lifelong learning problem has not been extensively investigated using the methods to improve the task performance. Also, deep learning methods used for the tasks have different problems for lifelong learning.

3.2 Issues in Lifelong Learning

The ability of human-like lifelong learning is an important problem in machine learning. However, some of the main issues of lifelong learning in realistic environments covered in this thesis are; (i) the existence of novel acoustic classes, (ii) the existence of unlabeled data, (iii) the cost of annotation, (iv) lack of adequate data for novel classes, (v) imbalanced data between classes, and (vi) requirement of large memory for storing all the data and (vii) computational power for lifelong learning.

3.2.1 Issues in scenarios for low computational cost lifelong learning

Besides the main issues, for lifelong learning with low computational complexity, there are also theoretical and practical issues affecting the sensing of real acoustic scenes and processing of the data captured from the scenes. Most state-of-the-art algorithms have no online learning capability, so the adaptation of new data is performed by retraining the model using a large dataset. Thus, the retraining of the model has computational and memory costs. For some algorithms such as Hidden Markov Model (HMM), the theoretical solutions for the capability exist in the literature, but in practice, it provides poor performances when using real data. The algorithms which are able to model the audio features in a frame-wise manner and to incrementally learn such as Support Vector Machine (SVM), Gradient Boosting Trees (GBTs), Stochastic Gradient Descent (SGD) based algorithms, etc. tend to overfit and are not able to adapt new classes into the model which is aimed in this thesis. Also, the incremental algorithms are not capable of learning for a long time due to the overfitting problem, however, the lifelong learning for ASA is aimed to learn during the lifetime of a scene. In practice, retraining of the frame-wise algorithms can take extremely large space and high computational time to complete, which is impossible for real-time or near real-time lifelong learning. Also, the other problem of lifelong learning is the space complexity which scales with the size of features belonging to the detected new class or recognized known class. Therefore, a deep learning-based ASA is also proposed in the scope of this thesis, which is able to incrementally learn without saving and retraining the old data. However, there are also a number of issues in deep learning-based lifelong learning for ASA as discussed in the next section.

3.2.2 Issues in deep learning based lifelong learning

Due to the issues of lifelong learning with low-cost computations, in this thesis, a deep learning-based approach is presented, and its advantages over the low computational cost approach are discussed. The main issue of the deep learning-based approach is that the algorithms based on batch calculations are not able to learn from streaming data, so the approach is not applicable to real-time or near real-time tasks of intelligent systems, robots, agents, etc.

In this thesis, regarding the issues and the gap in the literature, the questions are stated for the research of lifelong learning in the audio domain which are;

- Can acoustic events be recognized in scenes with unknown events?
- Can the AER capability be conducted in a lifelong manner without human supervision?
- Can the lifelong learning be performed in real-time?
- Can the lifelong learning be performed using deep learning algorithms?
- What are the most promising audio features and algorithms for the lifelong learning capability?

3.3 Scene Analysis

Scene analysis which is described in visual processing tasks in the early study [119] as analyzing all the visible properties together (edges, surface textures, colors, e.g.) depends on several tasks such as detection, localization, recognition, and tracking of events, objects, humans, actions, etc. for a variety of purposes in dynamically changing noisy environments. Nowadays, due to different kinds of sensors such as microphone arrays, cameras, tactile sensors, laser scanners, etc., scenes where the events take place can be monitored and analyzed to achieve the tasks. Moreover, depending on the sensors to perceive the scenes, different kinds of modalities have been individually or jointly developed in order to extract for analyzing useful cues about the targets. In the real world, many issues exist for Visual Scene Analysis (VSA) that prevent accurate perception and learning in dynamic environments which are not affecting ASA as much as VSA such as unexpected illumination changes in the scene, occlusion of multiple targets, not being in the field of camera view, a high time and space complexity for processing, and so on. Therefore, audition capability has important advantages over VSA which is an intensively investigated research field.

3.3.1 Auditory scene analysis

Analyzing acoustical scenes in a realistic environment is one of the major problems of audition [120], since the environments involve a mixture of sounds with multiple events including target and non-target actions, voices of several humans, background noises, etc. To analyze the scenes for understanding the relationships between the events, objects, and humans with each other, and labeling everything correctly is also known as cocktail party problem [7]. For example, an intelligent device like a household robot with an auditory perception ability should also be capable of learning and reasoning about everyday objects and events not covered even in its field of view, e.g. to understand the baby is crying in a different room while the television is turned on and the doorbell is ringing. Under such a complex domestic situation, the cognitive system should be well trained with the features of the sound of crying which is the main target overlapped with the coexisting sounds, and noises. Thus, in an ASA approach, recognition ability requires preprocessing techniques such as noise suppression, Sound Source Localization (SSL), and Sound Source Separation (SSS) capabilities in complex environments with overlapping sounds, since the context of an event including not only identity but also time and location information has an important role in the scene analysis [121]. Furthermore, it is important to design a learning model required for the cognitive aspects of scene analysis. The recent learning ability is considered a lifelong learning ability. An incremental learning paradigm allows a number of new instances in a known class and also a new class of sounds is analyzed and inserted into the generated model, and so this process continues in a lifelong manner.

Several types of research on ASA have been achieved with methods applied to different kinds of feature sets extracted from auditory signals. Also, for capturing the signals, single and multi-channel sensor systems have been developed. An audition capability for scene analysis has the main steps which are preprocessing, SSL, AER, AND, and lifelong learning.

3.4 Sound Source Localization

Several SSL methods have been proposed for single sources, multiple sources, and only speech or only non-speech sound source localization. The spatial feature provides much information about the time, the place of the event, and the place of the other sounds interfering with the target sound, which are essential for acoustic event detection and recognition. The famous approaches for SSL are based on Multiple SIgnal Classification (MUSIC) [56], working in real-time to detect the peak of the spatial spectrum. MUSIC is a subspace-based method used by deriving a steering vector to detect reliable peaks with low computational cost.

The SSL approach for dynamic environments which is exploited in this thesis is Generalized Eigen-Value Decomposition based MUSIC(GEVD-MUSIC) [57] which performs noise-robust localization in real-time. Also, it uses a noise correlation matrix to suppress the noises. The GEVD-MUSIC has a few static parameters. One is the number of sources to be detected and localized at the same time, and the most salient sound source can be detected by setting the parameter to 1. Also, a power threshold and an interval threshold in angle can be specified depending on the acoustic scene and the target events in the scene. If a source in the scene has the highest power bigger than the threshold, and the difference between its current azimuth position and the previous position is smaller than the interval threshold, the sound is detected as belonging to the same source, and so the tracking of the source is carried out using these parameters.

3.5 Audio Feature Extraction

Various types of low-level audio features have been employed individually or in combinations. In the ASA studies, several audio features such as which are mostly Zero-Crossing Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCCs), spectral energies, the entropy of energies, chroma vector, chroma deviation, spectral centroid, flux, roll-off, spectrum image, etc. have been analyzed to estimate the features representing the characteristics of the sound. Therefore, it is important to find the most discriminative type of audio features for describing the signals from each event, human, object, and so on. In this thesis, different time-domain features, frequency-domain features, time-frequency representations, and deep audio features are analyzed for the AER, AND, and lifelong learning tasks.

3.5.1 Time-domain features

In this thesis, the AER performances are evaluated using a combination of different audio features also including the time-domain feature of Zero-Crossing Rate (ZCR) [61] which is an important time-domain feature mostly used for the classification of vocal and non-vocal sounds by indicating the most dominant frequency of a frame. It is a measure of the number of times when the sign changes between successive samples along with a signal. Also, the time feature has been mostly utilized to discriminate voices from unvoiced speech [62] and to model the music. In the audio feature analysis experiments of the proposed low computational cost approach, the features are evaluated by combining them with the other low-level features.

3.5.2 Frequency-domain features

The low-level spectral descriptors are frequency-domain features such as the spectral centroid, the spectral spread, and the spectral flatness [72]. These descriptors are also used to extract features from the acoustic events in which the spectral centroid is a measure of the shape of the power spectrum; the spectral spread gives information about the spectral shape by taking the root-mean-square deviation of the spectrum from its centroid; and the spectral flatness measures the flatness of the power spectrum. Another frequency-domain feature is spectral roll-off [75] used to measure the frequency, which takes 95% of the power spectrum and indicates the skewness of the spectral shape, and pitch ratio calculated on the pitches estimated in each frame to discriminate the speech from non-speech signals [76].

In the frequency-domain audio features, the MFCC feature which is the most famous cepstral feature employed especially for speech processing applications represents the shape of the spectrum with a few coefficients of the Mel cepstrum. The Mel scale is defined to represent the relations between pitch and frequency by inspiring the non-linear human auditory system. The Discrete Cosine Transform (DCT) is utilized to decorrelate the log Mel spectrograms [96] to obtain a compressed representation of the filter banks as the MFCC features.

3.5.3 Time-frequency representations

The spectrograms which have received increasing attention in recent years are a time-frequency image representation of a sound inspired by the human auditory system. In many studies, it is demonstrated that the feature is more suitable for speech and event recognition in which the Mel-spectrograms of the audio data are utilized as the input of DNNs. For the extraction of the 2-D representations, Short-Time Fourier transforms (STFT) and Mel-filters are applied to an audio frame/segment to be transformed into a Mel-spectrogram. The STFT is a well-known Fourier-related transform technique by setting a window function with a constant size to determine the sinusoidal frequency and phase content of a local cross-section of a time-varying signal [122].

3.5.4 Deep audio features

The F-TDNN, a deeper network than TDNN includes four more channels, and a weight matrix of each TDNN layer in F-TDNN is factorized by multiplying two smaller matrices to decrease the number of parameters in the layers [123, 124]. The TDNN-type network has a factorized architecture instead of singular value decomposition in a traditional DNN, to reduce and fine-tune the parameters after the reduction. On the other hand, TDNN-LSTM architecture is composed of LSTM layers replaced with two TDNN layers.

3.6 Acoustic Novelty Detection

The traditional one-class algorithms (OCSVM, AEs, and iForest) have been evaluated in the experiment of the novelty detection in the real-time ASA. Also, the extreme GBT algorithms employed in a variety of machine learning tasks are fast and sufficient for multi-class classification problems. Therefore, for real-time AND, a semi-supervised variant of the ensemble model proposed in [125] is generated using the audio features to compare with the multilayered HMM-based approach. The algorithm is based on unsupervised outlier detection algorithms to improve data representation, greedy selection for the selection of useful representations, and then an XGBoost classifier on the improved feature space. Moreover, iForest [126] another widely used algorithm for outlier detection is exploited in the proposed low computational cost approach. The algorithm randomly and recursively partitions the input data to construct a forest composed of multiple trees for isolating the novel/abnormal data.

Besides the stacked AE (CAE for Mel-spectrogram), VAE, kNN, GMM, iForest, and OCSVM algorithms were exploited for the AND task. A one-class kNN model is generated only using the feature sets of known classes for novelty detection. The algorithm aims to estimate the nearest feature set in all the sets of known classes to an audio feature set sample [41]. If the sample's distance to the nearest set is bigger than the distance of the sample's nearest set's nearest set, the sample will be identified as a novel [127]. In addition, a GMM, a parametric probability density function abbreviated as a weighted sum of Gaussian component densities is employed for the detection of novel acoustic events [42]. On the other hand, iForest and OCSVM algorithms were investigated using audio features from the streaming signal in the low computational cost approach. In the ASA with ICL based AER approach, the algorithms were evaluated on the spectrogram and deep audio features.

3.6.1 Gaussian mixture model

To capture comprehensive features of known classes, a single Gaussian Mixture Model (GMM) known as universal GMM (uGMM) trained only by the known data can be employed for the novelty detection problem. The GMM can be defined as follows:

$$p(X_i) = \sum_{m=1}^{M} w_m \mathcal{N}(X_i|\theta_m) \quad m = 1,2,3,.....M \tag{3.1}$$

where $\mathcal{N}$ is the normal density with the parameter set of $\theta_m \in \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, X_i is an instance of a known class, and w_m represents the weights of the mixture, m with the constraint, $\sum_{m=1}^{M} w_m = 1$. The GMM algorithm aims to model density, and the parameters are optimized by the Maximum Likelihood (ML) criterion with the Expectation Maximization (EM) technique.

3.6.2 One-class support vector machine

One of the famous novelty detection algorithms is One-Class Support Vector Machines (OCSVMs) which learns the data of known classes to discriminate the instances of the classes amongst all objects, and so the technique is able to detect the instances of a new class, which are different from the instances learned by the OCSVM model. Based on the kernel, the algorithm maps the data into a higher dimensional feature space and separates the data on the space by an optimum hyperplane from the origin with maximum margin.

3.6.3 Hidden Markov model

Hidden Markov Model (HMM) is the most generative algorithm used for speech recognition, sound recognition, music genre recognition, etc. For novelty detection, the HMM has been investigated in a few works in which a left-right HMM called universal HMM (uHMM) is generated with the normal data [33,42]. A threshold value is estimated considering the likelihood values produced by the uHMM. In the lifelong learning approach with low computational cost proposed in this thesis, an uHMM is employed for AND.

3.6.4 Isolation forest

The other novelty detection algorithm is Isolation Forest (iForest) [126] which is able to perform even when the training data is contaminated with outliers. The algorithm aims to isolate the outliers/anomalies/novelties using random forests instead of profiling the normal data [128]. In the audio domain, iForest has been investigated in a few studies [34, 43].

3.6.5 Nearest neighbour

One of the famous distance metric-based algorithms which has been commonly utilized for different classification problems is Nearest Neighbour (NN) used for the novelty detection in the audio data. The search algorithm depends on a distance metric (Euclidean, weighted Euclidean, e.g.) in which the distance between an instance and its nearest neighbors is computed to detect global anomalies/outliers/novelties instead of local ones in either normal data or known classes [129]. For Acoustic Novelty

Detection (AND), novel audio features not similar to the known features are detected. Also, the algorithm is dependent on a threshold value, thus, it is in practice not easy to select an appropriate threshold, if required.

3.6.6 Autoencoders

The AEs which are unsupervised DNNs contain an encoder and a decoder, which aim to efficiently reconstruct the input data if the data belongs to one of the known classes. Thus, the AE models trained with the features of known event classes will reconstruct novel data with a high reconstruction error. The reconstruction error values used as the novelty scores in the AND step are compared with a dynamic threshold to detect the audio sample not reconstructed well as a novel class. For deep learning-based ASA, a stacked AE with multiple AEs in a stacked form, and a VAE which is a deep generative network with a Gaussian mixture-like model for learning statistics of the known events are evaluated for the AND task.

The AE-based unsupervised DNN algorithms contain an encoder and a decoder, which aim to efficiently reconstruct the input data if the data is belonging to the known classes. Thus, the AE models trained with the features of known event classes will reconstruct novel data with a high reconstruction error. The reconstruction error values used as the novelty scores in the AND step are compared with a dynamic threshold to detect the audio sample not reconstructed well as a novel class. For deep learning-based ASA, a stacked AE with multiple AEs in a stacked form, and a VAE which is a deep generative network with a Gaussian mixture-like model for learning statistics of the known events are evaluated for the AND task. Moreover, to use spectrograms as the input to the AE, a Convolutional AE (CAE) (Fig. 3.1 (a)) in our work [106] on detection of acoustic anomalies, is employed in a stacked manner. Also, in this work, the Convolutional Long Short-Term Memory AE (Conv-LSTMAE) demonstrated in (Fig. 3.1 (b)) was utilized for sequentially reconstructing the streaming signal from a stationary sound source in a manufacturing process to detect abnormal sound activities [106]. However, the Conv-LSTMAE is useful for novelty detection on stationary streaming signals instead of non-stationary sounds from different types of sources.

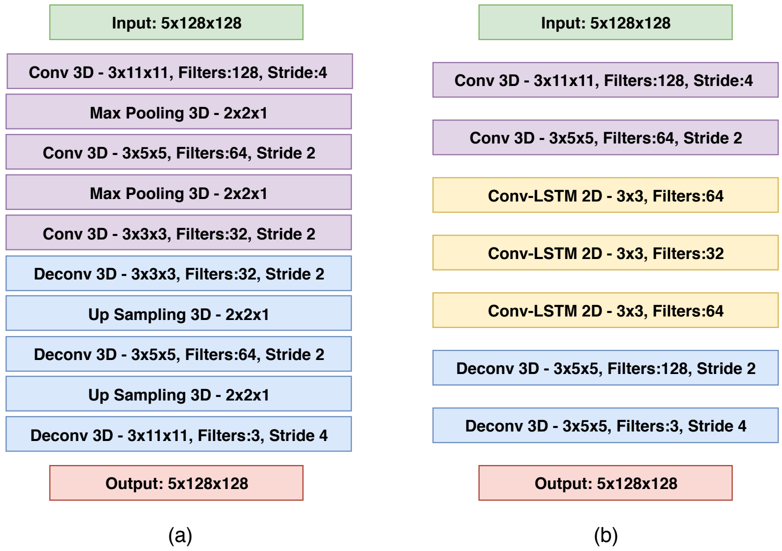

Figure 3.1 : The sequential AEs used for real-time AND: (a) the architecture of CAE based on convolutional layers, max pooling, and up sampling layers, and (b) the architecture of Conv-LSTMAE composed of 2D LSTM layers based on Recurrent Neural Network and convolutional layers.

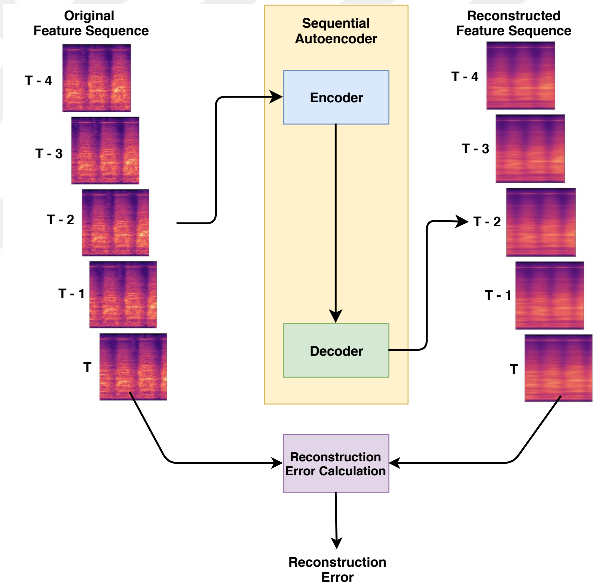

Figure 3.2 : The reconstruction and the reconstruction error calculation stages

Calculation of Reconstruction Error. To estimate the reconstruction error, Euclidean distance is calculated for each actual spectrogram and the reconstructed spectrogram in the corresponding sequences. The average of the distance between the sequences gives the reconstruction error [130] as in Eq. (3.3).

$$E_R = \frac{1}{N_s}\sum_{i=1}^{N_s}(X_i^a X_i^r)^T(X_i^a X_i^r), \qquad (3.2)$$

$$E_R = \frac{1}{N_s} \sum_{i=1}^{N_s} (I_i^a I_i^r)^T (I_i^a I_i^r), \tag{3.3}$$

where E_R is the reconstruction error, N_s denotes the number of spectrogram features in the sequence, I^a is the actual spectrogram, I^r is the reconstructed spectrogram, ^{T} is the transpose operator, and i represents the index of the spectrogram feature belonging to the sequence. The reconstruction error is checked against the threshold obtained in the testing step of the model generation stage using Youden's index, J, as in Eq. (3.4). This value is applied to each threshold present in the Receiver Operating Characteristic Area Under Curve (ROC-AUC) curve estimated using the Euclidean distance values [131] in order to estimate the best threshold with the maximum J value as follows;

$$J = Sensitivity + Specificity - 1, \tag{3.4}$$

where *Sensitivity* stands for True Positive Rate (TPR) and *Specificity* stands for True Negative Rate (TNR). The aim of Youden's index is to maximize the difference between TPR and False Positive Rate (FPR). Thus, the system will set the optimal threshold value by maximizing the anomaly detection accuracy and separability rate.

Furthermore, using AEs, the lower and upper threshold values of the decision interval are dynamically determined with the errors of the known events. When detecting a new event class and then updating the AE model, the thresholds are re-computed by all the error values of known events including the new one. In the reconstruction stage (Fig. 3.2), the spectrograms of an audio sample are reconstructed in the encoder and decoder layers of the CAE. The AE model produces a reconstruction and the reconstruction error is computed for the spectrogram of each selected window. The average of all the errors is taken to be used as the novelty score of the audio samples.

3.7 Lifelong learning

Various traditional algorithms using different types of audio features have been proposed for Acoustic Event Recognition (AER). Most recognition models for acoustic tasks employed in a frame-wise or time-series manner are not able to incrementally learn new data or classes. Therefore, in this thesis, one of the time-series AER models, the Hidden Markov Model (HMM) is investigated and modified for the low computational cost approach for lifelong learning due to the complexity time of the analysis of streaming audio signals. Also, for the deep learning based approach, the deep networks capable of Incremental Class-Learning (ICL) are evaluated for acoustic tasks. In this section, the background information is given for these algorithms.

3.7.1 Low computational cost lifelong learning

The lifelong learning which is a paradigm for learning new incoming data in real-time is also known as streaming learning, online learning, and incremental learning. However, due to the issues of real-time or near real-time processing, the algorithms have been developed depending on a domain, task, or an exact number of classes. In the audio domain, the detection and recognition of acoustic classes have not been well studied in environments with new incoming classes on streaming data.

3.7.2 Deep learning based lifelong learning

Traditional neural networks may tend to forget the previously learned knowledge while incrementally learning new data. The bottleneck is known as catastrophic forgetting which is related to the plasticity-stability dilemma [132]. If a deep network is too plastic it causes forgetting, but if the network is too stable to decrease the plasticity, the new data may not be efficiently learned.

The task of incremental learning is confronted with two major challenges which are to update the model using the data of a new class and to retain a subset of the old data. The weights of the network are updated by backpropagation and stochastic gradient descent techniques with a loss function during the learning. When new knowledge is being learned, the update may substantially affect the weights optimized by the previous knowledge, known as catastrophic forgetting [19] which is the main challenge of incremental learning. The stability-plasticity dilemma [132] which is an issue of the learning models, is a constraint for deep networks not to suffer from the forgetting problem. According to the dilemma, a deep network needs to be plastic to learn new knowledge, and stable to retain the existing knowledge. For the learning of new knowledge, class, or task, the weights of the network may considerably change, which will cause forgetting. To overcome the changes, the weights need to be kept stable, but the level of stability will affect the learning [133]. For the development of deep networks capable of incremental learning, the human's learning capability is inspired to overcome the forgetting problem. In real-world continuous environments, humans are able to continuously acquire, learn and accumulate new knowledge and tasks without forgetting the previous ones in their lifetime.

3.7.2.1 Strategies for mitigating catastrophic forgetting

To overcome the bottleneck forgetting problem, three types of strategies are adopted which are architectural strategy (3.7.2.1), regularization strategy (3.7.2.1), and rehearsal strategy (3.7.2.1). These three strategies are not contradicting each other. Therefore, like in several studies, multiple strategies can be employed cooperatively.

Architectural strategy

The main idea of the architectural strategy is to learn new classes or new knowledge of known classes while maintaining the stability of the network. Therefore, it aims to determine the most promising model among multiple different models trained for each new class. The strategy is firstly employed with traditional machine learning techniques such as Support Vector Machine (SVM) combined with the Learn++ method to train multiple SVM classifiers using different training sets for incremental learning. However, an important challenge of the methods is the imbalanced data of a new class. Also, when learning a new class, the number of parameters in the network will enhance and its architecture may require manual adjustment, which is not viable in a real-world scenario with a lack of human supervision and information about the classes. The ICL algorithms based on an architectural strategy have not been exploited on the audio data in literature.

Regularization strategy

In the regularization strategies, an additional regularization term is added to the loss function to control the update of parameters of the network. The goal is to enhance the stability of the network to mitigate forgetting. By the use of the term, the performance is maintained considering outputs of the old and new models while learning a new class or task (multiple classes). Therefore, storing the old knowledge is not required for the regularization strategy-based methods.

The forgetting is mainly observed when a new model is specialized to the new class while updating the weights of the old model trained on the known classes. Thus, measuring the importance of the weights is carried out to prevent the knowledge of the old classes by limiting the learning rate of the loss function, known as the weight regularization strategy. Moreover, knowledge distillation is another type of regularization strategy, that aims to transfer knowledge from one model to a new one to make the new model consistent while learning a new class or task.

Rehearsal and pseudo-rehearsal strategies

In the earlier study of [20], rehearsal and pseudo-rehearsal strategies are explained by taking into account the problem of catastrophic forgetting. For rehearsal strategy, there are different methods which are *straightforward rehearsal* that aims to store and reuse the recent samples, *random rehearsal* using an equal probability for the selection of samples in all the existing samples, and *sweep rehearsal* that re-selects new samples in each epoch of the training process.

3.7.2.2 Learning without forgetting

The earliest ICL algorithm based on knowledge distillation is Learning without Forgetting (LwF) [44] used for ICL tasks in the visual domain, in which the data of a new class is predicted using the output of the old model, and all the predictions are exploited for the update of the model. The LwF algorithm is based on a regularization strategy to add a term into the loss function of the algorithm in order to reduce the effect of forgetting. The term provides the knowledge distillation to get the output of the network while learning a new class, close to the original network output.

3.7.2.3 Incremental classifier and representation learning

Incremental Classifier and Representation Learning (iCaRL) [45] depending on a combined strategy of knowledge distillation and rehearsal was developed for ICL by retaining a number of samples for each class. For the ICL based AER, the algorithm classifying the events by the nearest exemplar method is also evaluated in terms of prediction performances while incrementally learning new classes. The remarkable performance of iCaRL has been demonstrated in various visual ICL tasks, but it has not been investigated in the audio domain besides our ICL work.

3.7.2.4 FearNet

For the ICL task, FearNet [46] which is an ICL algorithm based only on a pseudo-rehearsal strategy [20] was proposed by inspiring from mammalian brains. The algorithm covers a dual-memory system to learn new samples retained in the short-term memory, using a hippocampal network, and progressively consolidate the samples in the long-term memory of the system with medial Prefrontal Cortex (mPFC) network. Also, one more network is included in the algorithm for deciding on the use of the hippocampal or mPFC network for a sample.

4. LOW COMPUTATIONAL COST LIFELONG LEARNING FOR AUDITORY SCENE ANALYSIS

In this chapter, the approach for lifelong learning with low computational cost for Auditory Scene Analysis (ASA) is presented. The steps of the approach are explained in detail. In the experiments, the algorithms for the steps are evaluated in a domestic acoustic environment.

4.1 Overview of Proposed Lifelong Learning Approach

The proposed Hidden Markov Model (HMM) based approach (Figure 4.1) for low computational cost lifelong learning in ASA is composed of five main modules: (1) preprocessing, (2) extraction of audio features, (3) AER, (4) AND, and (5) lifelong learning. Each module aims to decrease the computational time while preserving suitable AER and AND performances. The modules of the approach demonstrated in Figure 4.2 include the following steps: (1) preprocessing for eliminating the windows with no acoustic activity, (2) Sound Source Localization (SSL) used for detection and location monitoring of the most salient sound source in a scene to perform source-specific analysis, (3) segmentation of audio patterns on the streaming signal, (4) feature extraction from the segmented patterns and construction of a feature set for each pattern, (5) Acoustic Event Recognition (AER) in semi-supervised manner, performed by a number of AER models including a class-specific model for each known events, (6) Acoustic Novelty Detection (AND) carried out using a model for all the known events with the outputs of AER module, and (7) lifelong self-learning in which for a recognized event, a model is retrained using more likely knowledge selected among all the previous and new knowledge of the event, and for a new acoustic event recently detected, a class-specific model is generated and the AND model is retrained (Figure 4.2). The approach uses a two-level thresholding mechanism in which an event-wise likelihood threshold for AER and a suspicion threshold for AND are estimated by all the known events: the likelihood threshold, T_{L_e} compared with the outputs of the AER models, and the suspicion threshold, T_S compared with the output of the AND model if the feature set is not recognized. The AER is carried out in a semi-supervised manner to pseudo-label audio patterns in the environment without a supervisor, and to update the AER model of the pseudo-label event class. All steps are evaluated using various audio samples and streaming signals regarding the requirements of processing and analysis with low computational complexity. landscape [1-3]

Using the approach, an agent or an intelligent system will be able to identify the most salient sound source in the scene, recognize the known event or detect a new event class, and learn knowledge from a known or new source by updating the AER model or generating an event-specific AER model. Furthermore, the system includes a supporting module for an offline supervised initial model generation to begin the self-learning by one or multiple initial AER models, and only one AND model. The set of models may be generated using one, few, or many audio patterns of each randomly selected known event.

For the evaluation of the novelty detection step in the proposed approach, state-of-the-art anomaly/novelty detection algorithms including a universal HMM, One-Class SVM (OCSVM), Isolation Forest, one-class Gradient Boosting Trees (GBT), simple Autoencoder (AE), and Variational Autoencoder (VAE) are exploited to decide the algorithm with efficient and reliable detection performances. The initial AND model in the ASA process is generated using the initial set of audio samples and acoustic events used also for AER models with respect to a preferred scenario. Such algorithms provide a novelty score computed with the similarity of a new feature set of a recently segmented pattern and the features of the known patterns. In the thresholding mechanism, the score is compared with the suspicion threshold to decide whether the new set as belonging to a novel class or not.

4.2 Audio Preprocessing

The steps for the preprocessing in the proposed approach are SSL and audio pattern segmentation for the improvement of the lifelong learning approach with low computational cost. The SSL step aims to provide the scene analysis of the streaming signal of a localized source. Also, a segmentation technique is utilized to segment the

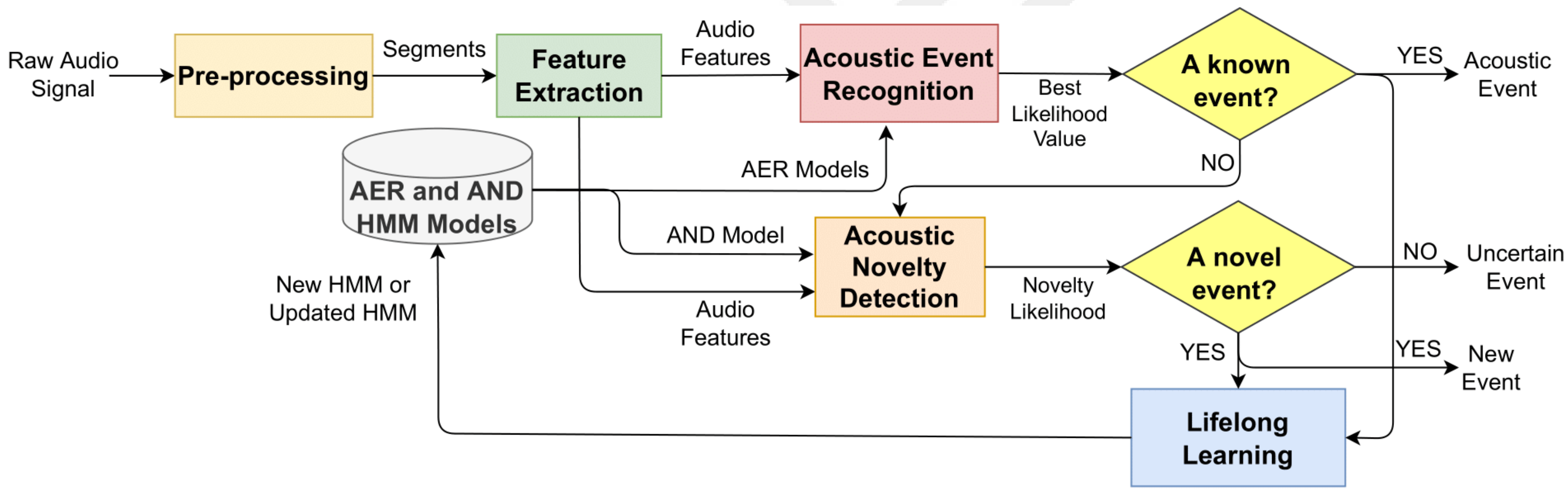

Figure 4.1 : The overview of the proposed framework for the low computational cost lifelong learning for

important parts of the signal to remove the noises and non-target patterns, and it helps to decrease the computational time of the analysis.

4.2.1 Sound source localization

For ASA, source-specific analysis is carried out to localize and track the most salient source, in which the segmentation and feature extraction steps are being performed on the captured signal of the source. In addition to the tracking of the source, the localization of sound sources may help to discriminate the salient sources from background noises or non-target sources that may exist in a dynamically-changing environment. Therefore, the location information of the source is important for reliable ASA and ignoring non-salient sounds. Also, depending on the type of acoustic scene, the spatial information may give essential cues for the recognition of events.

In the SSL step, Generalized Eigen-Value Decomposition based Multiple SIgnal Classification (GEVD-MUSIC) [57], mostly used for robot audition tasks, which performs noise-robust and fast localization technique is employed to quickly obtain the azimuth angle of multiple sound sources. A salient sound source can be monitored over its location to capture its signal for the realization of source-specific ASA. When a new source is detected, a new ASA process begins to analyze the signal from the new source.

4.2.2 Audio pattern segmentation

In real acoustic environments, multiple sound sources may overlap, and so many separation techniques have been proposed to solve the problem. However, the performance and the use of the techniques with low computational overhead are not satisfactory. The class-based HMM algorithm suitable for the source-specific AER is able to model variable-length time-series data, but may not be efficient for modeling too long sequences [134]. Therefore, segmentation of the time-series on a streaming signal may provide useful inputs to this algorithm and help the analysis of the overlapping sounds individually. In the pre-processing step of the proposed approach for ASA, a fast segmentation technique is performed to divide the signal from a salient source into variable-length segments. It aims to automatically determine its beginning and ending frames according to the energy features, called onset and offset frames. By a sliding window, the extracted energy features are compared with

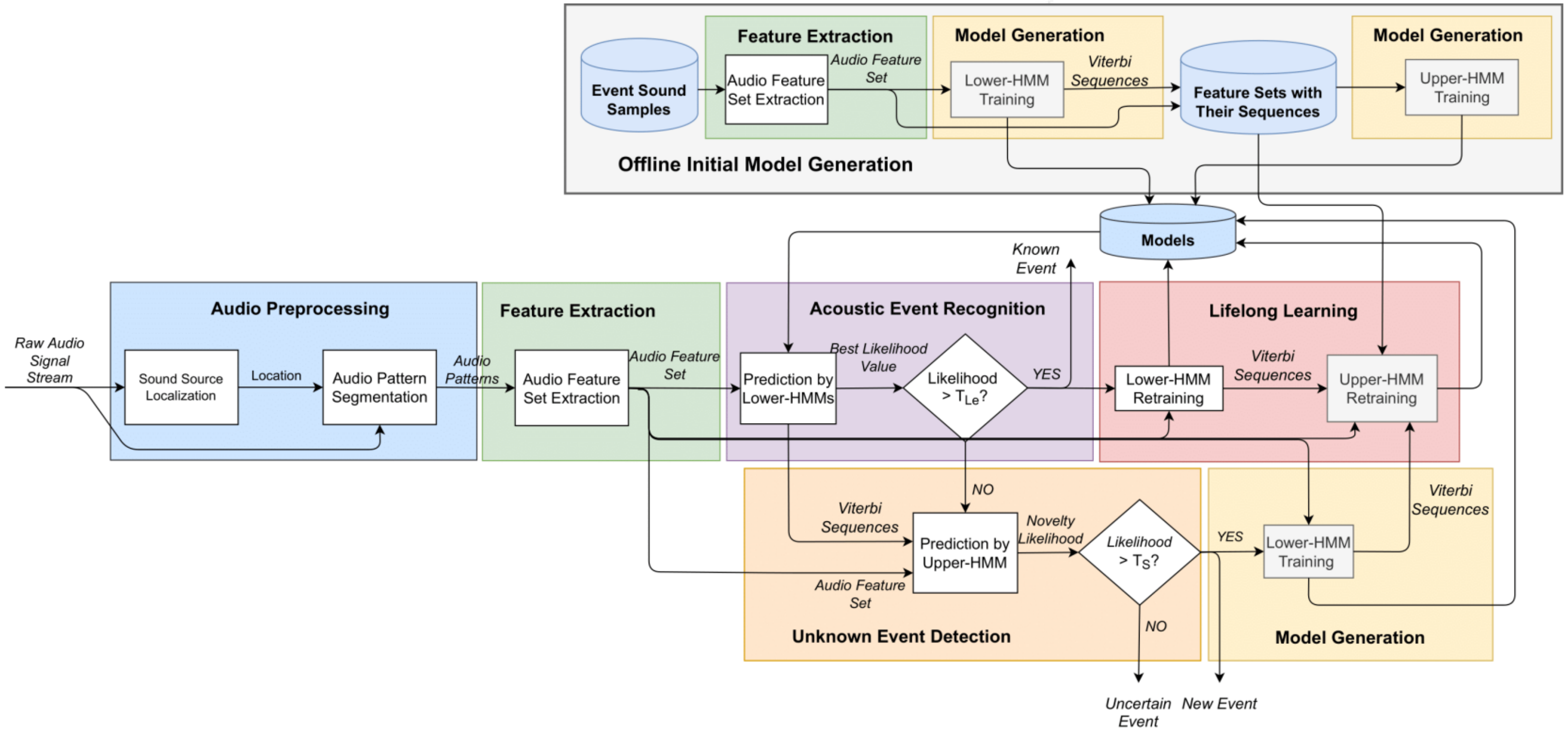

Figure 4.2 : The overview of the proposed framework for the lifelong learning including the tasks in the

a dynamic threshold to estimate only the onset frame, and then the offset which is the last frame before another onset frame, or a silent frame arrives is selected.

4.3 Audio Feature Extraction

One of the factors to develop a robust learning model for ASA is a cautious feature extraction task. In the audio feature extraction step of the ASA approach, audio feature sets are provided not only to AER, AND, and lifelong learning steps but also to the pattern segmentation task. For low computational cost lifelong learning, different types of audio features are evaluated in terms of the prediction performances in AER and AND tasks, and space and time complexities. The pattern segmentation task uses energy features to detect the boundaries of the pattern over the streaming signal. For AER and AND, the investigation of spectral, cepstral, and temporal audio features is realized to estimate the most suitable feature or combined features. One of the major factors for efficient AER, AND, and lifelong learning is a cautious extraction of the audio features.

Furthermore, in the feature extraction step, a feature set in which each feature vector corresponds to a time point is constructed for each pattern after the segmentation. The set is transmitted first to the lower layer of the multilayered HMM for AER, and if it is not recognized, it is analyzed in the upper layer for AND. Therefore, the most appropriate type of audio feature should be found considering its effectiveness not only in the AER task but also AND task. For this purpose, the cepstral features, Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Codes (LPC), Line Spectral Pairs (LSP), and Perceptual Linear Prediction (PLP) are analyzed for AER. Also, time-domain features such as energy and Zero Crossing Rate (ZCR), frequency-domain features such as spectral centroid, spectral flatness, spectral roll-off, and a combination of various types of features are evaluated.

4.3.1 Feature selection

To speed up the update of the HMMs in the upper and lower layers, Kullback-Leibler (KL) distance (1) based feature selection is employed on the determined audio features to reduce the computational time of training and update while preserving the AER and

AND performance as follows:

$$D(p_{ij}||q_i) = \int p_{ij} \log \frac{p_{ij}}{q_i}, \tag{4.1}$$

where p_{ij} represents the distribution of the i^{th} feature unit in the feature vector of the j^{th} audio event class, and q_i is the distribution of the i^{th} feature unit in the entire feature vectors of all the known event classes. Thus, the discriminant capability of a unit is;

$$D(p_i) = \sum_j r_j D(p_{ij}||q_i), \tag{4.2}$$

where r_j denotes the prior probability of the j^{th} event class. The features with a high distance to all the others in the vector are selected for AER, AND, and lifelong self-learning tasks. The feature selection technique is exploited to remove redundant and similar feature units among different classes. It aims to attain similar AER performance with the feature vectors with a lower dimension to decrease the computational time required especially in the retraining process of the AER and AND models.

4.4 Lifelong Learning

For scene analysis in real acoustic environments, an ASA approach with the AND and AER models in which the unlabeled samples are pseudo-labeled and adapted into the recognition models, and new classes are detected and inserted into the set of models is developed for lifelong learning with low computational cost.

4.4.1 Multilayered HMM

The computation time of the HMM algorithm depends on the length of the sequence and the parameter set, which scales linearly and quadratically with the increase of the dimension of the feature vector and the number of states, respectively. Therefore, the selection of the most distinctive features and the optimal number of states and iterations is an essential task to control the required time and the AER and AND performances.

The multilayered HMM is composed of a lower layer for the AER task and an upper layer for AND explained in the previous sections. For each known event, an HMM is generated, and a common HMM is trained using the outputs of all the lower layer HMMs (Figure 4.3). Each HMM in the multilayered HMM aims to optimize an objective function with the parameter set, λ of the traditional HMM, which are transition matrix, A, observation probabilities, B and initial probabilities, π.

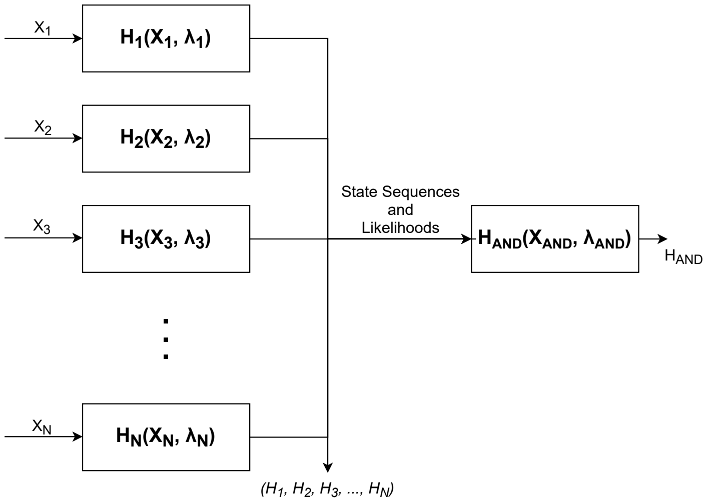

Figure 4.3 : The structure of the multilayered HMM.

An initial HMM, $H_e(X_e, \lambda_e)$, is generated for a randomly selected event, e to be used in the lower and upper layers of the multilayered HMM. The training for each model aims to maximize the objective function:

$$H_e(X_e, \lambda_e) = \arg\max_{\lambda_e} P(I|X_e, \lambda_e), \tag{4.3}$$

where X is a set of the audio feature vectors in which the size of X depends on the length of the audio sample, I denotes the feature set of a recent segmented pattern, and λ is the set of HMM parameters based on each event class, optimized with the initial training and retraining processes. A number of states are connected by transitions and the HMM aims to compute the likelihood value of the most likely state sequence, in which each feature vector in the set, X, is represented by a state, and the sequence corresponds to the entire set in the segment.

Each probability value, a_{ij} in A matrix that is the transition probability from the state, s_i to the state, s_j, is defined as $a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$ where s_i is the actual state at time, t. The probabilistic model is here a Markov chain in which an observable Markov model of the states and the chain depends on the transition probabilities between the states. Thus, the audio feature sequence from each segment is modeled using the transition probabilities for the AER model of each event class and AND model of all the known events.

The observation probability b_i of the state, s_i in B is calculated by a probability density function, $b_i(O) = P(o_t = u_k | q_t = s_i)$ where o_t denotes an observation state at the time t. The density function is a multivariate Gaussian distribution, $b_i(O) = P(O|q_t = s_i) = N(O; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

To estimate the optimal parameters for each HMM of AER and the HMM of AND, the EM algorithm is employed to maximize the likelihood value until convergence or reaching the maximum iteration value. The algorithm is composed of two steps; expectation, E-step, and maximization, M-step. In addition to the HMM parameters, the mean vectors and the covariance matrices used for the GMM are optimized during training/retraining.

The state-to-state transition probabilities are derived by the forward and backward processes called Forward/Backward algorithm term instead of the Expectation-Maximization algorithm. Also, the initial state probabilities are re-estimated by the forward variable, α, and backward variable, β. The variables are computed as $\alpha_t(i) = P(o_{1:t}, q_t = s_i | \lambda)$, which is the probability of a partial observation sequence from the beginning to time, t given the model λ.

The steps to compute the forward variables are as the following:

1. Initialization:

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N \tag{4.4}$$

2. Induction:

$$\alpha_{t+1}(j) = \Big(\sum_{i=1}^{N} \alpha_t(i) a_{ij}\Big) b_j(o_{t+1}), \quad 1 \leq t \leq T-1;\ 1 \leq j \leq N \tag{4.5}$$

3. Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{4.6}$$

Using Baum-Welch algorithm, the backward variable is defined as;

$$\beta_{t+1}(j) = P(o_{t+1}, o_{t+1}, ..., o_T | q_t = s_i, \lambda). \tag{4.7}$$

The probability of a Markov process in which the i^{th} feature vector of a set, s_i is at time, t, and the j^{th} vector as another state, s_j is at time, $t+1$, given the parameter set,

λ of an HMM and an observation sequence, O is defined as;

$$\xi_t(i,j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda). \tag{4.8}$$

The probability, ξ is computed using the forward and backward variables as follows;

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \tag{4.9}$$

The probability of being in the state, S_i at time, t given the observation sequence and the model is estimated as follows;

$$\gamma_t(i) = P(q_t = S_i | O, \lambda) = \sum_{j=1}^{N} \xi_t(i,j). \tag{4.10}$$

Consequently, the update of the parameters depending on the length of the pattern, p, so the length of the observation sequence, $o^{(p)}$, is iteratively achieved using the probability values, ξ, as follows:

For updating the initial probabilities, $\overline{\pi}$;

$$\overline{\pi}_i = \frac{\sum_p \gamma_1^{(p)}(i)}{\sum_p 1} \tag{4.11}$$

For updating the transition probabilities, $\overline{a}_{ij}$;

$$\overline{a}_{ij} = \frac{\sum_p \sum_{t=1}^{T^{(p)}-1} \xi_t^{(p)}(i,j)}{\sum_p \sum_{t=1}^{T^{(p)}-1} \gamma_t^{(p)}(i)} \tag{4.12}$$

For updating the observation probabilities, $\overline{b_j}(u_k)$;

$$\overline{b_j}(u_k) = \frac{\sum_p \sum_{t=1}^{T^{(p)}-1} [o_t^{(p)} = u_k] \gamma_t^{(p)}(j)}{\sum_p \sum_{t=1}^{T^{(p)}-1} \gamma_t^{(p)}(j)} \tag{4.13}$$

At each step, the new parameter set, $\lambda_e^{'}$ of an HMM belonging to an event for AER or all of the known events for AND is estimated, and this re-estimation loop continues over the forward and backward variables until satisfying the stopping criterium;

$$\log P(O|\lambda_e^{'}) - \log P(O|\lambda_e) < \varepsilon, \tag{4.14}$$

where ε denotes a constant value. The optimization process is applied to each event's HMM, H_e and the HMM of AND task, $H_a nd$.

4.4.2 Acoustic event recognition

The main problem of an ASA is the recognition of unlabeled sound sources of events and activities in acoustic scenes. The pseudo-labeling of the acoustic events is a process of semi-supervised learning in which unlabeled audio samples are utilized to improve the recognition model as well as the labeled data by predicting the labels of the samples. Therefore, in the proposed approach, the semi-supervised process is carried out to recognize the known event classes in the AER step in the lower layer of the multilayered HMM.

For each class, an HMM is generated and an event-wise likelihood threshold is computed for the model to both be used for the analysis of each unlabeled audio sample regarding its computed likelihood value. If the value is lower than the threshold of the most likely HMM, the sample is pseudo-labeled by the class of the HMM. Otherwise, the state sequence with the likelihood value computed by the most likely HMM is used as the input of the AND step. For each feature set, the successful estimation of the most likely state sequence with the greatest likelihood value using Viterbi is an important factor to build a robust AND model.

The event-wise thresholds associated with the known events are computed after fitting the HMMs. The audio samples of an event, used for the training are predicted by the HMM, and a likelihood value is estimated for each sample. The average of all the values is used as the likelihood threshold value of the event. After a new class is detected in the AND step, the event-wise threshold is determined with the data of the class, used in the novelty detection.

4.4.3 Acoustic novelty detection

In real environments, novel classes may also exist besides the unlabeled instances of known classes. The ASA tasks in such environments have the same challenges as real-world acoustic problems, so new acoustic events in the scenes should be detected and learned by analyzing the signals from the sources. In the AND step, a universe HMM is generated using the subset of the audio features of the known events. The model is exploited to predict the novelty of a segmented pattern by computing its novelty likelihood exactly as not belonging to any known events.

In the upper layer of the multilayered HMM, an initial HMM is generated using the outputs of the AER models in the lower layer. Using the Viterbi algorithm, an AER model for each event decodes each feature set to a state sequence with the greatest log-likelihood value to be used in the AND task. The entire state sequences of the existing feature sets by the associated AER models are exploited in the generation of HMM and prediction of the new incoming sequences for AND. For the detection of novelty, the likelihood value used as the novelty score of the sequence from the lower layer is computed by the HMM in the upper layer. Furthermore, the novelty score is compared with the suspicion threshold dynamically modified to predict the sequence as belonging to a novel class or uncertain. An example of novelty scores among a number of state sequences transmitted from the lower layer is demonstrated in Figure 4.4 where the novelty scores (likelihood values) of the unknown events are represented by the points (true-positive samples (blue points), the false-positive samples (green points), the patterns detected as "uncertain" (yellow points)) and the suspicion threshold is the red point over the whole data.

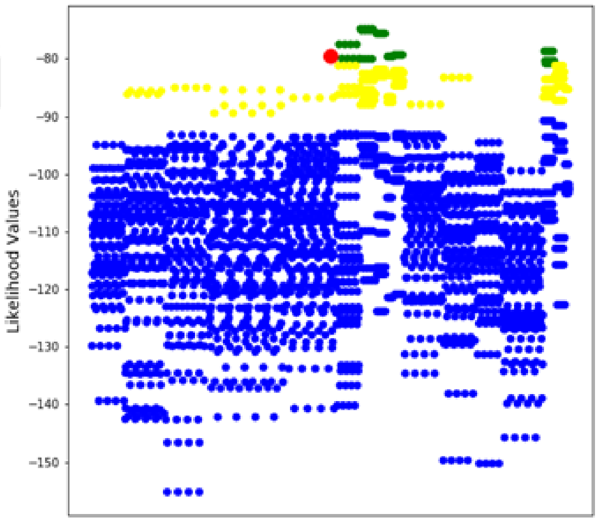

Figure 4.4 : The likelihoods of the unknown events that are the true-positive samples (blue points), the false-positive samples (green points), the patterns detected as "uncertain" (yellow points), and the suspicion threshold (red point) over the whole test data represented with the indices along the x-axis.

In addition to the "known" and "unknown" outputs of the approach, one more output, "uncertain" is included, which is required to improve the precision of the predictions and to avoid the wrong learning of the events. The performance of the AND task depends on the output of the AER step which are the state sequences and the log-likelihood values. Also, besides the output, the suspicion threshold directly affects the AND performance for the detection of new events and uncertain patterns. Therefore, the connection between the lower and upper layers of the approach is not only based on transmitting the data, but also the effective prediction performances.

4.4.3.1 Novelty score calculation

In the AND step, a novelty score for each segmented pattern from the audio samples is calculated depending on the type of algorithm and compared with the suspicion threshold to make a decision for the detection of an “unknown” or “uncertain” class. If the score is less than the threshold, the pattern is predicted as a novel. The distance to the hyperplane is used as the novelty scores derived from the OCSVM algorithm. Furthermore, the GBT algorithm provides a novelty score depending on the fitted feature sets of known classes, and the higher this value is, the more novel it is. Using iForest, the novelty score of a new feature set is computed by the negative of the average path length of the set overall trees in the forest.

4.5 Experiments

In this section, the experimental setup in which the software and hardware specifications are provided, evaluation metrics, and details of the dataset collected in a real domestic environment are explained in detail. Moreover, the experimental procedure carried out for the experiments of feature analysis, AND, AER, and lifelong learning is presented.

4.5.1 Experimental setup

For lifelong learning with a low computational cost for ASA, acoustic tasks for SSL and pre-processing were performed using HARK [135], an open-source robot audition software. For the audio feature extraction step, the Librosa python package was utilized for MFCC, spectral features, and energy, and for the extraction of LPCC, LSP, and PLP features, pymir which is also a python package was used. For the AER and AND tasks, HMM was implemented using the python package, hmmlearn, and the performances of algorithms for the tasks were evaluated and visualized by scikit-learn and matplotlib, respectively. Moreover, the recordings of the acoustic events were taken by a Kinect microphone array with a 4-channel, and the offline and online experiments were conducted with a computer with Intel Quad Core i7-3700K with 16 GB RAM on the recordings.

4.5.2 Experimental procedure

The offline experiments were conducted for feature analysis and the estimation of the optimal parameter settings for AER and AND models. The most appropriate feature types (Table 4.1) and parameters were found considering the requirements of ASA with low computational complexity in the online experiment. For the experiments, a number of audio recordings in a real domestic environment have been utilized, which are belonging to the acoustic event classes of *opening door* and *closing door*, *footsteps*, *kettle*, *vacuum cleaner*, *washing hands*, *dishwasher*, *toilet flush*, and *washing machine*. In the domestic dataset, for the offline experiments, 436 event sounds manually labeled (41 *opening door* and 50 *closing door*, 36 *footsteps*, 65 *kettle*, 55 *vacuum cleaner*, 24 *washing hands*, 85 *dishwasher*, 32 *toilet flush*, and 48 *washing machine* sounds) have been exploited. After applying the pattern segmentation method, 31905 audio patterns were obtained from the sound instances, consisting of 1981 door opening and 2711 door closing, 2416 footsteps, 2449 kettle, 3972 vacuum cleaner, 6008 washing hands, 5112 dishwasher, 3184 toilet flush, and 4072 washing machine audio patterns. For the experiments of feature analysis, parameter estimation, AER, and AND, the standard 5-fold cross-validation was performed in which the patterns were divided into a training set (70%) in which a few or all the patterns were used for the model generation and retraining, and a test set (30%) was only used for performance evaluation. A subset of the patterns was randomly selected, in which a few or all the patterns were for the generation of the AER and AND models and retraining, and the rest of the patterns were utilized for ASA in offline experiments. Thus, the ASA with the lifelong learning capability begins with an initial model using one or a few audio patterns and learns the rest in time concerning the predictions of AER and AND models such as “known”, “unknown” or “uncertain”.

Various types of audio features (4.1) were extracted from the audio signals with the sampling rate selected as 16kHz, and for the extraction MFCC, the parameters, the size of Fast Fourier Transform (FFT) as 512, and the window size as 256 were used. The

Table 4.1 : The audio features evaluated for AER.

Cepstral	Spectral	Temporal
MFCC	Flatness	ZCR
MFCC with Derivatives	Centroid	Energy
LPCC	Roll-off	
LSP	Flux	

number of states for HMMs of AER and AND tasks and the maximum iteration number used in the training of HMM were selected considering the length and dimension of the feature sets.

In the experiments for AER and AND tasks, the types of audio features and HMM parameters learned in the first experiment were utilized for the generation and the update of the models. For the prediction experiments, the sounds in the same collected domestic were exploited.

4.5.3 Evaluation metrics

For the evaluation of AER for determining the most appropriate type of audio features and parameters, the accuracy, precision, and recall were utilized, and the computational time was measured for retraining after prediction, so the average results were compared in which the average values of these metrics and measurement are computed according to the number of patterns. The performance of AND in terms of precision, recall, retraining time, and detection time including also the prediction time of the lower HMMs was exploited in the comparisons. Subsequently, the outcomes of the experiments were exploited to develop the models for the ASA experiment in the same domestic environment.

In the experiment of AAD with low computational cost, Receiver Operating Characteristic Area Under Curve (ROC-AUC) score is used as the evaluation metric to evaluate the AEs. The higher the ROC-AUC score is, the better the anomaly detector the model is. A ROC-AUC score close to 1.0 means that the AE model can efficiently reconstruct the Mel-spectrograms of known acoustic events and detect the anomaly events.

Table 4.2 : The AER performance using the cepstral features.

Features	Dimension	Avg. Accuracy	Avg. F1-Score	Avg. Retraining Time (sec.)
MFCC	13	0.881	0.716	2.524
MFCC with Derivatives	39	0.914	0.772	5.110
LPCC	20	0.711	0.389	4.171
LSP	255	0.761	0.471	12.820

4.6 Results

For the evaluation of the proposed real-time ASA approach, firstly, an offline procedure is performed to estimate the most useful feature types and optimal parameters to be used in the real-time experimental procedure.

4.6.1 Results of offline experiments

In this section, the results of offline experiments conducted for estimation of the most efficient audio feature set in terms of AER performance and computational time, and the AND algorithm with the respective optimal parameters providing the best performance are presented.

4.6.1.1 Results on audio feature analysis

In the experiment of feature analysis, various types of low-level audio features given in Table 4.1 were compared with respect to the prediction performance of AER and retraining time. The experiment aims to determine one of the cepstral feature types as the major part of the most efficient features and to estimate the promotive features in the spectral and temporal features. Therefore, the cepstral features (Table 4.2) and then other variations of features (Table 4.3) were analyzed by experimenting 10 times and the average values of the metrics and computational time were obtained. Furthermore, the updating process of the model by retraining is analyzed with the worst scenario in which the entire training set is utilized for the initialization.

In the analysis experiment of cepstral features, MFCCs and MFCCs with the derivatives have similar and efficient AER performances (Table 4.2). Due to the dimension of the feature set and so the computational time, using only MFCCs is

Table 4.3 : The performance of AER using different feature combinations of MFCCs with other low-level features.

Features	Dimension	Avg. Accuracy	Avg. F1-Score	Avg. Retraining Time (sec.)
MFCC	13	0.881	0.716	2.524
MFCC with Spectral Features	16	0.889	0.734	2.828
MFCC with Energy	14	0.879	0.708	2.580
MFCC with ZCR	14	0.883	0.721	2.644

Table 4.4 : AER performance and retraining time of both MFCC and selected coefficients of MFCC features.

Features	Avg. Accuracy	Avg. F1-Score	Prediction Time (sec.)	Avg. Retraining Time (sec.)
MFCC with 13 coefficients	0.881	0.716	0.0060	2.524
MFCC with selected 8 coefficients	0.849	0.670	0.0043	1.986
MFCC with selected 4 coefficients	0.811	0.577	0.0038	1.180

preferred in the subsequent experiments. As Table 4.3 illustrates, using the variations of features with 1-dimension, it is observed that the performance of MFCC either deteriorated or not improved remarkably.

In addition, the Kullback-Leibler distance-based feature selection is applied to the MFCCs with 13-dimension to select the most distinctive coefficients not redundant among different classes for the reduction of the computational time. The average AER accuracy, precision, and recall in the lower layer of the HMM, and the average computational time of the retraining of an HMM given in Table 4.4 were estimated by the realization of five experiments for the analysis of MFCC with 13 coefficients and MFCCs with selected coefficients.

4.6.1.2 Results of acoustic novelty detection

In the AND experiment, the aim is to estimate the effectiveness of the multilayered HMM by comparing it with the regular HMM and some state-of-the-art novelty detection methods in terms of precision and recall (Table 4.8). The solution for novelty detection which is the use of an HMM for each known event suffers when multiple

HMMs provide similar likelihood values when some event classes have sounds and so features looking like the data of other events. The similar sounds eventually deteriorate the precision and recall of the novelty detection and the false-positive rate increases. In the experiment, it is demonstrated that modeling of the most likely state sequences with likelihood values by a shared HMM increases the prediction performances and decreases the false-positive, and false-negative rates.

In Table 4.5, the multilayered HMM has the best AND performances, and the previously proposed HMMs with OCSVMs and HMMs with only one OCSVM like the proposed multilayered HMM has similar results obtained on the domestic sounds. However, especially AEs and Isolation Forest have unreliable novelty detection performances due to the inadequate number of examples used for the generation of the model. Also, the AEs generated using a few examples require too high computational time.

Table 4.5 : The Mean F1-Scores for AND.

	Mean F1-Scores			
Methods	**One Event**	**Two Events**	**Three Events**	**Five Events**
Multilayered HMM	**0.958**	**0.957**	**0.933**	**0.917**
HMMs with OCSVM	0.950	0.939	0.926	0.908
OCSVM	0.948	0.941	0.919	0.917
Isolation Forest	0.889	0.455	0.516	0.371
GBT	0.929	0.910	0.910	0.902
VAE	0.844	0.441	0.313	0.346
Simple AE	0.842	0.438	0.313	0.347

In Tables 4.6 and 4.7, the event-wise best and worst F1-scores of the methods trained in 1- and 3-shot manners, respectively providing suitable performances are given. The AEs have been evaluated in more than 30 sound samples for each known event, and much higher F1-scores are obtained. However, in real-time scenarios, learning with a few samples will be required, so the AEs are not suitable. The hybrid methods have suitable AND performances in these model generation conditions. Only using OCSVM requires a high number of instances to achieve the novelty detection task, and then get closer to the performance of the hybrid methods.

In addition to these event-wise results, AEs have provided the best performances on these events $door opening/closing$, and *washingmachine* with the F1-scores more than

Table 4.6 : Event-wise Average Best and Worst F1-Scores for 1-shot AND.

	Methods			
Events	**Multilayered HMM**	**HMMs with an OCSVM**	**OCSVM**	**GBT**
Washing Machine	0.999/0.998	0.999/0.997	0.997/0.991	0.998/0.992
Door Opening	0.994/0.991	0.995/0.977	0.995/0.912	0.995/0.877
Door Closing	0.980/0.968	0.979/0.963	0.966/0.893	0.982/0.914
Kettle	0.993/0.988	0.995/0.977	0.990/0.987	0.994/0.991
Footsteps	0.871/0.692	0.999/0.997	0.997/0.991	0.998/0.912
Vacuum Cleaner	0.978/0.952	0.981/0.967	0.963/0.961	0.981/0.972
Toilet Flush	0.891/0.859	0.897/0.773	0.861/0.692	0.810/0.752

Table 4.7 : Event-wise Average Best and Worst F1-Scores for 3-shot AND.

	Methods			
Events	**Multilayered HMM**	**HMMs with an OCSVM**	**OCSVM**	**GBT**
Washing Machine	0.999/0.998	0.999/0.997	0.997/0.993	0.998/0.993
Door Opening	0.994/0.990	0.995/0.983	0.995/0.908	0.995/0.893
Door Closing	0.980/0.971	0.980/0.971	0.978/0.891	0.984/0.944
Kettle	0.993/0.988	0.995/0.980	0.990/0.984	0.993/0.988
Footsteps	0.882/0.748	0.900/0.697	0.803/0.797	0.888/0.812
Vacuum Cleaner	0.974/0.966	0.974/0.967	0.965/0.963	0.978/0.973
Toilet Flush	0.903/0.861	0.895/0.724	0.864/0.783	0.817/0.778

0.90. After a long parameter tuning process on the structure of the network, the epoch number, learning rate, etc., the results of the best network are obtained. Using a few examples, well-known AEs may not provide suitable performances, and the existing works capable of few-shot learning and novelty detection are not developed on the acoustic data. Also, when more than one event is known, the AEs and Isolation Forest performances are substantially reduced.

Table 4.8 : The AND performances and detection times of a regular HMM and the multilayered HMM.

Number of Known Event	**Avg. Precision**	**Avg. Recall**	**Detection Time(sec.)**
Only HMM	0.799	0.494	0.0178
Multilayered HMM	0.889	0.818	0.0271

Furthermore, different AND scenarios are generated to evaluate the AND performances of multilayered HMM, in which are; (1) only one class is known, (2) three event classes are known, (3) five event classes are known, and (4) only one event class is unknown (eight events are known) where HMMs in the lower and upper layers of the multilayered HMM trained using few audio patterns segmented from only one sound sample. The computational times of model generation for the new class as a lower-layer HMM and retraining of the AND model by using a few patterns with respect to its likelihood value were employed for the measurement of time. The average values of the precision, recall, and computation time, are shown in Table 4.9.

Table 4.9 : The AND performance and detection time of the multilayered HMM in different scenarios.

Number of Known Event	Avg. Precision	Avg. Recall	Detection Time(sec.)
1 (only one event is known)	0.884	0.818	0.0103
3	0.859	0.792	0.0179
5	0.830	0.761	0.0210
8 (only one event is unknown)	0.819	0.745	0.0218

In the experiments for AND, the scenarios in which only one, three, or five events are known, or only one event is unknown are utilized to analyze the performance of the multilayered HMM. In Table 4.9, the results of the multilayered HMM algorithm are listed using these scenarios. The events are selected among dissimilar ones, and also events similar to the known ones are not included in the training of the AND model. In an example of a scenario, the likelihood values computed by the HMM using the unknown events (*dish machine*, *door closing*, and *vacuum cleaner*) are demonstrated in Figure 4.4, in which each point represents a segmented pattern of an "unknown" event, and the red point denotes the suspicion threshold. The correctly detected patterns are represented by the blue points, and the patterns detected as "uncertain", and "known" are represented by the green and yellow points, respectively. In the offline experiments of AND tasks, it is revealed that the number of sounds used for the training of HMMs directly indicates the threshold values.

4.6.2 Results of real-time experiments

In this section, the results of real-time lifelong learning experiments using the outcomes of the offline experiments about the audio feature set and the AND algorithm are explained in detail.

4.6.2.1 Experiments of real-time auditory scene analysis

In the experiment of real-time ASA, initial models for lower and upper layers of the multilayered HMM were generated with the parameter values using the select audio features in the previous experiments. Only a few audio patterns which are randomly selected of events randomly selected were utilized, and in the real domestic environment, the sounds of new events performed next to the microphone array were localized and analyzed by the approach. In the ASA process, the randomly selected events were *kettle*, *vacuum cleaner*, *washing machine*, and *dishwasher*, and unknown ones were *footsteps*, *toilet flush*, *washing hands*, and *opening* and *closing the doors*.

The prediction performances of AER and AND models in the experiment are shown in Table 4.10. The worst precision is observed while analyzing the *kettle* sounds due to the similarity with the *footstep*. Some patterns of *toilet flush* sound are mispredicted with the patterns from the sounds of *dishwasher* and *washing hands*. It is observed that the most distinctive sounds are from the *vacuum cleaner* and *closing the door*.

Table 4.10 : AER and AND performances and retraining time of lower and upper HMMs for different sounds.

Events	Avg. Precision for AER	Avg. Recall for AER	Avg. Precision for AND	Avg. Recall for AND
Door Opening	0.811	0.770	0.888	0.804
Door Closing	**0.874**	**0.841**	**0.910**	**0.864**
Footsteps	0.790	0.731	0.764	0.681
Kettle	0.757	0.689	0.784	0.599
Vacuum Cleaner	**0.889**	**0.814**	**0.929**	**0.871**
Toilet Flush	0.803	0.782	0.861	0.802
Washing Machine	0.819	0.800	0.844	0.814
Washing Hands	0.821	0.804	0.809	0.821
Dishwasher	0.796	0.764	0.840	0.790

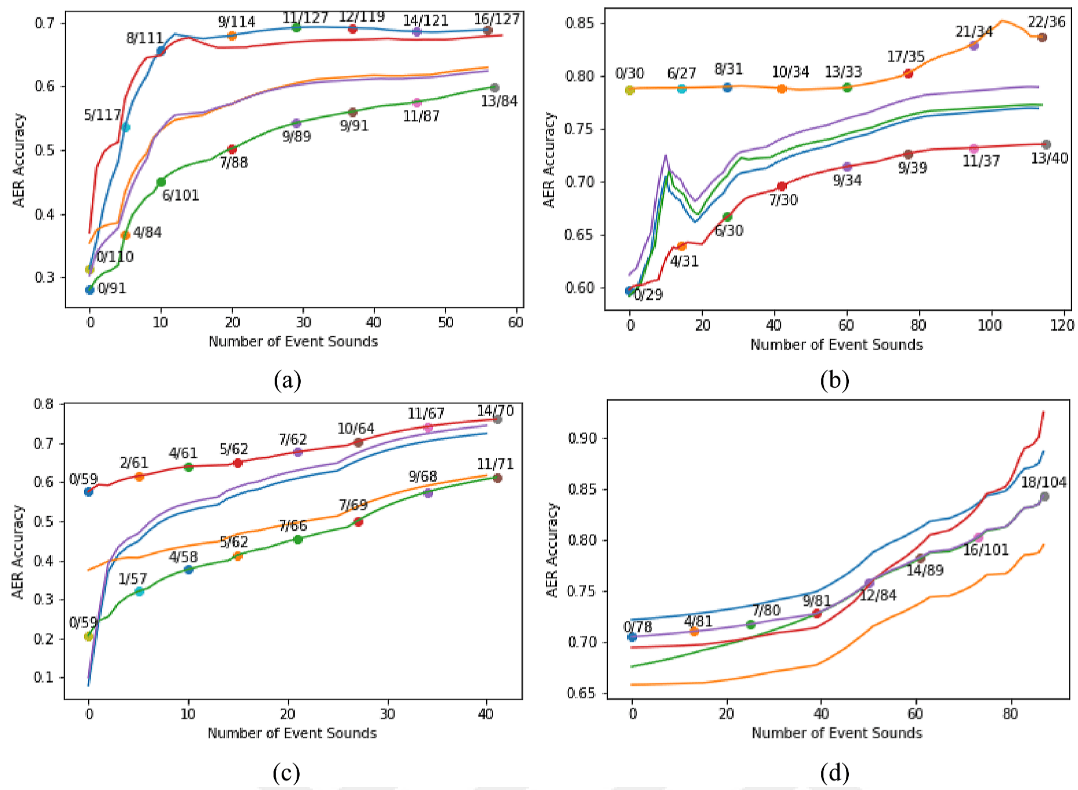

Figure 4.5 : The performances of AER in 4 different sessions of the lifelong learning experiment.

In the experiment of lifelong learning for ASA, four different learning processes were carried through, in which one event class is randomly selected to generate initialize lower and upper layer HMMs, and several event samples near the microphone array were performed for each event class. The performances of lifelong learning are demonstrated in Figure 4.5 in terms of accuracy changes after a new class is detected and adapted to the HMMs. Each panel in this figure belongs to a process and the x-axis values are the sum of the number of events sounds predicted as known and unknown events to be used in the update of the AER and AND models, respectively. The designated dots (x/y) along the lines represent the number of unknown events in the training (x) contributing to the numbers on the x-axis; and the number of known events in the test (y) contributing to the AER accuracy on the y-axis. The sounds predicted as “uncertain” were discarded. Thus, numbers of event sounds in those four arbitrary experiments with five different initial HMM parameters differ because of the number of uncertain patterns. The increase of AER accuracy over time is computed using only the predicted and actual labels of the known patterns in the test set is apparent in all sessions. Even though the learning begins with low accuracies (especially in

Figure 4.5(a) and Figure 4.5(c)), a substantial increase is achieved over time due to the retraining and AND. Different initial accuracy values obtained from a low number of known events may be observed in the same sessions such as in Figure 4.5(b) mainly due to the training of the model with only one sample. At the end of the process, 22 different classes were detected among 101 event sounds with 12901 audio patterns.

5. DEEP LEARNING BASED LIFELONG LEARNING FOR AUDITORY SCENE ANALYSIS

In this chapter, the deep learning-based approach for the lifelong learning problem for Auditory Scene Analysis (ASA) is proposed. The differences from the approach presented in Chapter 4 in the types of audio features and algorithms, regarding the issues of the real-time lifelong learning approach. The steps of the approach are explained in detail. In the experiments, the algorithms for the steps are evaluated in various benchmark datasets. Also, the performances of the approaches are compared in terms of the prediction performances for the AER and AND in a lifelong learning manner.

5.1 Overview of Proposed Deep Learning based Approach

In this section, the proposed approach for deep learning-based ASA with incremental learning is introduced in detail [34]. The approach has similar steps to the proposed real-time ASA approach except for the task required for streaming learning, which is audio pattern segmentation. The steps of the deep learning-based ASA are; (i) pre-processing, (ii) extraction of audio features, (iii) Acoustic Novelty Detection (AND), (iv) augmentation of audio signals, (v) Acoustic Event Recognition (AER), and (vi) Incremental Class-Learning (ICL) (Figure 5.1). The main goals in the AND, AER, and ICL tasks are to learn the function of

1. a semi-supervised detector for novel events, updated by retraining like in the real-time ASA approach and
2. a recognizer for acoustic events in scenes which is able to incrementally learn the new events detected in the AND step.

The deep learning-based ASA process differs from the HMM-based approach in the pipeline of the incoming data. After the feature extraction step, firstly, the novelty detector is employed instead of the recognizer like in the real-time ASA approach, since the two major steps are not connected in an algorithmic manner as in the multilayered HMM. The audio samples are not in a streaming manner, and a number of recordings are utilized in the deep learning-based incremental approach. The other differences are the use of only one model for each AER and AND task, and the update of the AER model is achieved by incremental learning instead of retraining. A thresholding-based decision mechanism is employed, in which an interval of two dynamic thresholds is used in the AND step to provide the same outputs in the previous approach, that are “known”, “unknown” and “uncertain”. There is no event-wise threshold for the AER task like in the HMM-based model.

5.2 Audio Preprocessing

In the ASA task, instead of audio pattern segmentation, windowing is crucial to construct fixed-size frames from the audio signal. The algorithms, especially Deep Neural Networks (DNNs), employed in the ICL-based AER and AND tasks require the frames as the input rather than variable-length time-series data like HMMs. Therefore, a sliding window through the signal which is the most popular windowing technique is applied to obtain the frames to be used in the feature extraction step. The length parameter of the sliding window technique is determined after conducting several analyses to preserve the feature stability and acoustic characteristics of events. Let S_i be the raw signal from an acoustic event sample, i, divided into overlapping windows (frames) by the sliding window with a constant temporal length (400 ms). $W_i^k \in R^T$, where k is the index of the window, and T is the time dimension which is equal to the size of the window. Each window, W_i^k is 50% overlapped by previous window, W_i^{k-1} and subsequent window, W_i^{k+1} where $k = 2, 3, \ldots, N_i$, and N_i is the number of windows obtained from the signal in the event recording, S_i.

The other task in the pre-processing step is a window-based segmentation method employed on the obtained windows to compute the ratio of silence and obtain the presence of sound events within all the 1-D windows, W_i^k. The ratio is computed by taking the sum of the samples in each window and then compared with a pre-defined silence threshold that is close to 0 determined by analyzing the absence of any sound

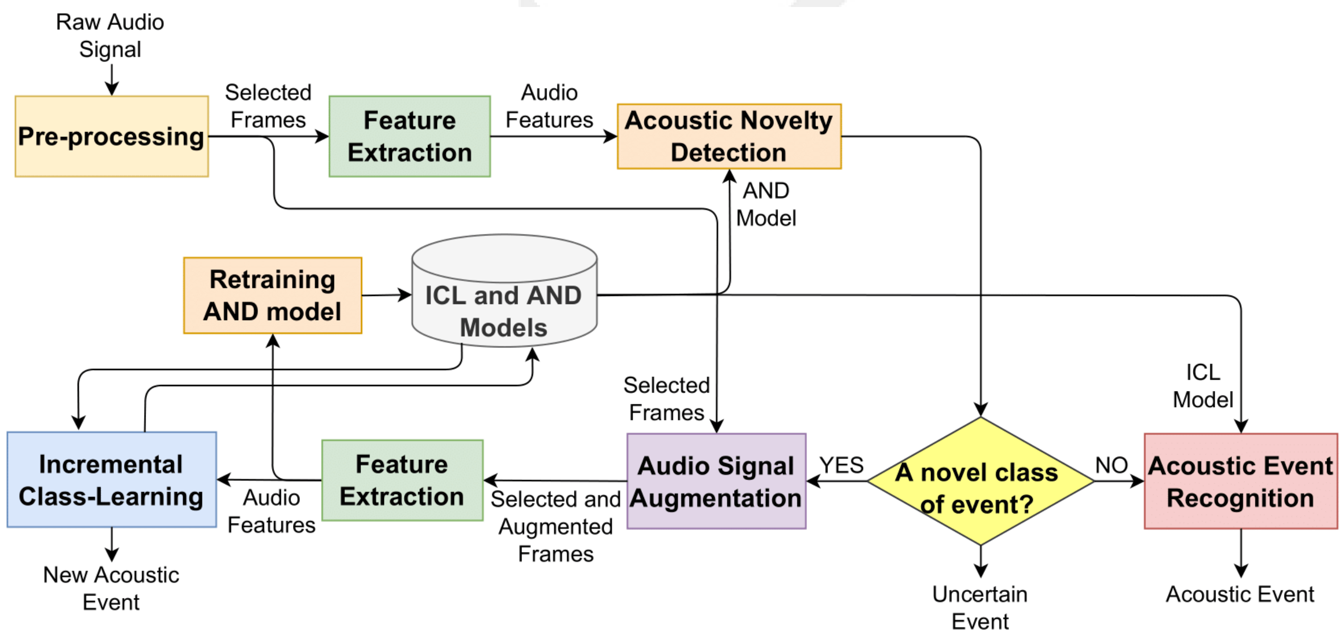

Figure 5.1 : The block diagram of the proposed approach for incremental class-learning with novelty

activity in the recording.

$$\upsilon_i^k = \sum_{t=1}^{T} W_i^k(t), \tag{5.1}$$

where t is the time index of the samples within the k^{th} window. For each window, the sum of the samples, υ_i^k, bigger than the silence threshold is selected to be used in the further tasks of ASA.

In addition to sliding window and window segmentation, the pre-processing includes the determination of the optimal values for the parameters which are an overlapping factor in windowing, the audio features, MFCC, the number of coefficients, Mel-spectrogram, the window size in short-time Fourier transform, and the hop size, and of the algorithms.

5.3 Audio Feature Extraction

The feature extraction step covers the analysis of five different types of audio features, $X_i^k \in \{I_i^k, C_i^k, W_i^k\}$, that are Mel-spectrograms and deep audio representations extracted using four pre-trained networks from one of Mel-spectrogram and MFCC. Each frame from the sliding window is exploited for the extraction of MFCCs, C_i^k and Mel-spectrograms, I_i^k to be used by a pre-trained model, and the algorithms of AER and AND, and two pre-trained models, respectively. One of the pre-trained models, TDNN-LSTM, uses the raw signal as its input for deep feature extraction. MFCC, C_i^k is a fixed number of d-dimensional feature vectors. Briefly, each type of audio feature, $X_i^k \in \{I_i^k, C_i^k, W_i^k\}$ is used to extract the deep audio features with 128 dimensions; D_i^k, by TDNN, TDNN-LSTM, VGG, and ResNet networks pre-trained with AudioSet, a large scale dataset.

As the inputs of F-TDNN and TDNN-LSTM, MFCCs, C_i^k and the windows including the part of the raw signal, W_i^k are utilized, respectively, instead of Mel-spectrograms like in the VGG and ResNet. Furthermore, Mel-spectrogram which is the most popular audio feature in the deep learning-based acoustic tasks is also used as one of the feature types in AND, AER, and ICL tasks. The feature analysis is achieved to estimate the most useful feature for the ASA task by conducting several experiments of AND and ICL-based AER in terms of the effectiveness of the performances of AND and ICL.

5.4 Lifelong Learning

The new acoustic events detected by the AND model are incrementally learned by the ICL model in a lifelong manner. Also, in the deep learning-based approach, novelty detection is achieved using a threshold, and the possible outcomes are similar to the real-time lifelong learning approach to preventing fast degradation in AER and AND performances.

5.4.1 Acoustic novelty detection

The AND step is one of the similar steps to the proposed real-time ASA approach, which aims to detect novel acoustic events. According to the novelty scores of the audio features and the pre-defined novelty threshold, the task provides three possible outcomes ("known", "unknown" or "uncertain") (Figure 5.1). The novelty score of an audio sample, that is too far from the decision threshold causes the "uncertain" prediction. The requirement of the "uncertain" outcome depends on the same reason to prevent the substantial decreases of the AER and AND performances due to the wrong predictions in these tasks, which would directly affect the incremental learning and so scene analysis.

For the novelty detector, a function, $f_{and} : X_i^{k:N_i'} \rightarrow G_i$, where N_i' is the number of selected windows aims to learn to compute a novelty score, G_i for each raw signal, S_i. The average of the scores for the features, $X_i^{k:N_i'}$ of the selected frames on the signal of the acoustic event recording is computed to obtain a novelty score for the recording, i. The function of the novelty detector is defined using the entire audio features of all the known events, and the model of the detector is updated by retraining if a recording is recently detected as a novel event.

In the case of that, the computed novelty score is in the decision interval, the audio sample is detected as "uncertain", but if the score is greater than the upper threshold of the interval, a novel event class will be detected. Otherwise, the features of the audio sample with the novelty score less than the lower threshold are presumed to be a "known" event and analyzed in the AER step for the recognition of its event class. To improve the precision of the ICL-based AER and AND, and prevent the propagation of the wrong learning, the thresholding mechanism is exploited in the ASA process.

The false positives of AND task may negatively affect the analyses in both the AND and AER steps and the incremental learning process.

However, the recently detected new event class may inevitably have a few features that may deteriorate the performance of scene analysis. Especially the deep learning-based models will probably be over-fit or forget the previously learned classes and features due to the sparse data. Therefore, in this framework, a few augmentation techniques on raw audio signals or audio features of new classes are investigated to overcome the sparsity problem by increasing the size of sparse data. It is observed that only the augmentation of audio signals by time-stretching with a number of stretching factors has improved the performances of the tasks in ASA. The techniques are employed to change the length of the signal while preserving the signal's spectral properties.

The ASA with ICL and AND is performed on the audio samples sequentially analyzed by the approach without human supervision, for label information of known and unknown events. Therefore, the novelty detection model is updated in a semi-supervised manner, regarding the detection output of the AND model. When a new class is detected, its knowledge is inserted into the previous data to retrain the model without evaluating the detection. The update of AND model is carried out as in the real-time ASA framework, by retraining the entire features of known classes including new features from a number of selected windows of the actual and augmented signals where the new features are also incrementally learned by the ICL based AER model.

For an efficient detector for the novel classes, the AND performances of various state-of-the-art algorithms are assessed to estimate the most robust and efficient algorithm, in which the one-class classification methods have been employed, that are stacked Autoencoder (AE), Variational AE (VAE), k-Nearest Neighbour (kNN), Gaussian Mixture Model (GMM), One-Class Support Vector Machine (OCSVM), and Isolation Forest (iForest). The decision interval is estimated for each algorithm by conducting several experiments and evaluations for each dataset.

Besides the stacked AE (CAE for Mel-spectrogram) and VAE, kNN, GMM, iForest, and OCSVM algorithms were exploited for the AND task. A one-class kNN model is generated only using the feature sets of known classes for novelty detection. The

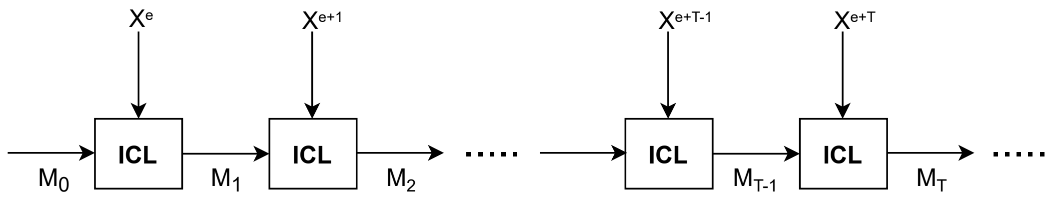

Figure 5.2 : The process of ICL of a new event class.

algorithm aims to estimate the nearest feature set in all the sets of known classes to an audio feature set sample [41]. If the sample's distance to the nearest set is bigger than the distance of the sample's nearest set's nearest set, the sample will be identified as a novel [127]. In addition, a GMM in which a parametric probability density function is abbreviated as a weighted sum of Gaussian component densities is employed for the detection of novel acoustic events [42]. On the other hand, iForest and OCSVM algorithms were investigated using audio features from the streaming signal in the real-time ASA approach. In the ASA with ICL-based AER approach, the algorithms were evaluated on the spectrogram and deep audio features.

5.4.1.1 Novelty score calculation

The AND algorithm computes a novelty score for each selected window of the audio samples. The average value of the scores from all the selected windows is utilized as the novelty score of the audio sample and compared with the novelty threshold value. The computation of novelty score by OCSVM and AEs explained in Section 4.4.3.1 is used in the AND experiments with the AEs.

5.4.2 Incremental class-learning for acoustic event recognition

A deep learning-based ASA is proposed in the scope of this thesis, which is able to incrementally learn without saving and retraining the old data since incremental learning is the only solution for continual learning from the streaming and ephemeral data. The feature set, X_{e+j} of the j^{th} new event,e, is incrementally adapted to the ICL model, M_t while retaining the knowledge of the previous event classes, as demonstrated in Figure 5.2.

Using an ICL algorithm, audio features from selected overlapping windows of the actual and augmented signal belonging to the new detected class are adapted to the ICL-based AER model with no human supervision at all.

Initially, an ICL model is trained in a supervised setup, including a number of audio features, $\{X_i^k, y_i^k\}_e$, of a randomly selected recording of a randomly selected acoustic event, e in which i denotes the index of the audio sample, and y is the event label. For the ICL task, the audio features from the samples of unknown events are sequentially and disjointedly learned while avoiding forgetting without retraining. Therefore, a function of the recognizer for ICL, $f_{icl} : X_i^k \rightarrow y_i^k$ is defined on the training features with the associated labels.

For the evaluation of the ICL task on the acoustic data, three ICL methods which are Learning without Forgetting (LwF), incremental classifier and representation learning (iCaRL), and FearNet based on one or more of the ICL strategies are implemented to find the most powerful algorithm avoiding the forgetting problem.

5.5 Experiments

This section outlines the experimental setup in which the software and hardware specifications are provided, and the experimental procedure with the explanation of the details for the experiments of feature analysis, AND, ICL, and lifelong learning. Also, the evaluation metrics and details of the datasets are presented.

5.5.1 Experimental setup

For the implementations of three novelty detection techniques (GMM, OCSVM, and iForest) and evaluation of all the AND and AER performances, Scikit-learn was used, which is a python package also used for the evaluation of the performances of the ASA approach with low computational cost lifelong learning. For the implementation of AEs, Keras which is one of the popular deep learning packages in python was used, and the pre-trained and ICL networks were implemented using PyTorch which is the most used Python package for these networks. The famous python package, Librosa, for audio analysis and signal processing tasks, utilized also in the real-time ASA was used for processing of audio signals and extraction of MFCCs and Mel-spectrograms. Specifically, the training of ICL algorithms and the pre-training of F-TDNN, TDNN-LSTM, VGG, and ResNet networks were run on a machine with Intel® Core™ i7-8700K CPU and NVIDIA's GeForce, GTX 1080Ti GPU.

5.5.2 Experimental procedure

The experiments of the deep learning-based ASA were conducted using four different datasets which are ESC-10, ESC-50 [47], UrbanSound8K (US8K) [1], and the domestic dataset collected in the scope of this thesis. The experiments cover the evaluation of the performances of (1) the algorithms with five types of audio features for the detection of novel acoustic classes, (2) the ICL algorithms using these types of features, and (3) ASA with the tasks of ICL and AND in which more efficient ICL and AND algorithms were employed with more distinctive feature types.

The ESC-50 dataset includes various types of 2000 sounds with 5 seconds duration with a sampling frequency of 44,100 Hz from 50 balanced classes (40 examples per class) in 5 sound categories (animal (*dog*, *rooster*, *pig*, *cow*, *frog*, *cat*, *hen*, *flying insects*, *sheep*, and *crow*), non-speech human (*crying baby*, *sneezing*, *clapping*, *breathing*, *coughing*, *footsteps*, *laughing*, *brushing teeth*, *snoring*, and *drinking/sipping*), urban/outdoor noises (*helicopter*, *chainsaw*, *siren*, *car horn*, *engine*, *train*, *church bells*, *airplane*, *fireworks*, and *hand saw*), indoor/domestic sounds (*door knock*, *mouse click*, *keyboard typing*, *door*, *wood creaks*, *can opening*, *washing machine*, *vacuum cleaner*, *clock alarm*, *clock tick*, and *glass breaking*), and natural (*rain*, *sea waves*, *crackling fire*, *crickets*, *chirping birds*, *water drops*, *wind*, *pouring water*, *toilet flush*, and *thunderstorm*)). The ESC-10 dataset with 10 acoustic classes is a subset of ESC-50. The classes of ESC-10 dataset in the categories are *dog barking*, *rain*, *sea waves*, *baby crying*, *clock ticking*, *person sneezing*, *helicopter*, *chainsaw*, *rooster*, and *fire crackling*. In the US8K dataset consists of 8732 sound clips with up to 4 seconds from 10 classes of the sounds captured in indoor and outdoor environments which are *air conditioner*, *car horn*, *children playing*, *dog bark*, *drilling*, *engine idling*, *gun shot*, *jackhammer*, *siren*, and *street music*. The last dataset includes the domestic audio samples of 436 short clips from the non-overlapping sounds with a sampling frequency of 44,100 Hz., of 10 domestic events, used also in the real-time ASA experiments. The events in the real domestic environment are *opening* and *closing doors*, *footsteps*, *taking a shower*, *kettle whistling*, *vacuum cleaner*, *cooking*, *dishwasher*, *toilet flushing*, *washing machine* with a duration between 1 and 12 seconds.

MFCCs with 20-dimension and Mel-spectrograms were extracted from each segmented window in the audio signal, in which the parameters for the extraction of the features were *window size* of 400 *ms*, *step size* of 200 *ms*, *overlap factor* of 50%, and a Fast Fourier Transform (*FFT*) *size* of 512.

For the experiments of AND and ICL tasks, the five types of audio features were utilized with the algorithms in order to determine useful feature types and algorithms. For this purpose, F-TDNN, TDNN-LSTM, VGG-16, and ResNet-34 models were pre-trained with a subset of the sounds in the AudioSet dataset, in which the subset has 40 classes of the environmental, urban, and domestic categories. The AudioSet dataset used in many studies in different acoustic tasks is composed of 5800 hours of video clips with an ontology of 527 types of sound events from YouTube. In AND and ICL-based AER tasks, Mel-spectrogram and four types of deep audio features extracted by VGG-16 and ResNet-34 using Mel-spectrograms, F-TDNN with MFCCs, and TDNN-LSTM from the raw signal of the audio samples in each of these datasets were utilized for training and test processes.

In the first experiment of the ICL-based ASA approach, several scenarios were used, in which different number pairs of known or unknown events exist for the generation of AND models and detection of novel classes. The algorithms (*stacked AE*, *VAE*, *kNN*, *GMM*, *iForest*, and *OCSVM*) were evaluated by analyzing the AND performances in the scenarios to estimate promising algorithms. In the scenarios for the $ESC-10$, $US8K$, and *Domestic* dataset, the number of known classes was set to 1, 3, 5, and 7, and the rest of the 10 classes were used as unknown. Furthermore, in the last scenario, only 1 class was unknown, and the rest of the event classes was known. The scenarios were different with *ESC*-50 dataset, in which 1, 5, 10, 20, 30, 40, and 45 event classes were known and the rest were unknown, and one where only 1 class was unknown (the rest of the 49 event classes were used as known). In the scenarios, for the AND experiment, all the combinations were considered for each dataset except ESC-50. For the dataset with 50 classes, the experiment was repeated 5 times for each scenario in which the known classes were randomly selected.

The ICL algorithms were evaluated with the feature types using these datasets in the experiment of the ICL task, and in the last experiment, the proposed ICL with AND based ASA approach was discussed and its performance was evaluated using the outputs of the first and second experiments, to estimate the useful feature types and more robust and efficient AND and ICL based AER algorithms. Moreover, in the second part of the last experiment for lifelong learning, the proposed ICL-based lifelong learning method is compared with the proposed real-time approach explained in Chapter 4.

5.5.3 Evaluation metrics

The performances of the AND algorithms were compared in terms of the average Area Under Curve (AUC) and F1-scores for each feature type in each dataset. The AUC metric is important to analyze the performances of binary and one-class classification algorithms, and the AUC is generated by plotting the rates of true positive and false positive. The F1-score is calculated by taking the harmonic mean of the precision and recall values of the predictions. For the evaluation of the performances of ICL-based AER and ICL with AND tasks, the average accuracies were utilized using a test set, which consists of randomly selected audio samples for each event class in each dataset. The average F1-scores of AER predictions after each class is incrementally learned are used to compare the five feature types. Moreover, for the evaluation of catastrophic forgetting in the ICL experiment, the Ω_{all} metric [133, 136] which normalizes an ICL performance (AER accuracy after learning a new class) by the offline performance of the ICL model trained from scratch on all training audio features as follows:

$$\Omega_{all} = \frac{1}{T}\sum_{t=1}^{T}\frac{\alpha_t}{\alpha_{offline,t}}, \quad t = 1, 2, ..., T \tag{5.2}$$

where α_t denotes the ICL performance of the algorithm at time t, $\alpha_{offline,t}$ is the ICL performance of the offline model, and T is the total number of testing processes. If the Ω_{all} value is equal to 1, the model does not catastrophically forget, therefore, it is expected to obtain values not close to 0.

5.6 Results

This section presents the results of the AND, ICL, and lifelong learning experiments.

5.6.1 Results of acoustic novelty detection

The experiment of AND task aims to estimate the algorithms with better performances, in which stacked AE, VAE, kNN, GMM, OCSVM, and iForest algorithms were employed using Mel-spectrograms, and deep features extracted by TDNN, TDNN-LSTM, VGG, and ResNet networks. In these AND scenarios, the average F1-scores for each feature type extracted from each dataset are listed in Tables 5.1–5.4. The deep features of VGG and F-TDNN through all the algorithms provided more efficient performances than the other feature types. The stacked AE, kNN, and OCSVM algorithms robustly detected novel classes in a few scenarios, but the best F1-scores and the most robust performances were obtained with the GMM algorithm, so the algorithm is selected to be used in the experiment of ICL with AND based ASA. The VGG and F-TDNN features provided more promising performances for AND, but after the analysis of the features also in the experiment of ICL-based AER tasks in Section 5.6.2, the satisfactory feature types will be determined.

In Figure 5.6a-d, the AUCs with the scores using the audio feature types are demonstrated in each dataset. The AUC performances were obtained in the less complicated scenario in which only one event is used as known. The VGG-16 deep features presented the most efficient AND performances in the scenario carried out using the audio samples of the Domestic dataset (Figure 5.6a) and the US8K dataset (Figure 5.6c). Moreover, in the ESC-10 and ESC-50 datasets, the best AUC scores observed using AND algorithms with the deep features by F-TDNN and VGG-16 were shown in Figure 5.6b,d. On the other hand, the ResNet-34 deep features were close to the performances of VGG-16 and F-TDNN features in some scenarios, but in the scenarios in which a high number of known classes exists. On average, the deep features extracted by ResNet-34 provided unreliable performances, and TDNN-LSTM features had the worst performances in all the scenarios in each dataset.

Table 5.1 : The Average F1/AUC Scores of AND on Domestic dataset.

Algorithm	Mel- Spectrogram	F-TDNN	TDNN-LSTM	ResNet-34	VGG-16
Stacked AE	89.7/92.1	91.7/94.0	86.1/88.3	90.1/92.9	92.8/95.0
VAE	83.5/86.2	84.1/85.2	80.7/81.7	87.3/90.3	88.9/91.1
kNN	86.0/88.2	91.1/94.7	81.5/83.8	88.8/90.4	94.7/**97.1**
GMM	92.7/94.4	**96.1/97.2**	86.5/86.7	92.3/94.2	**96.4/97.4**
OCSVM	86.3/90.6	86.1/91.2	80.1/87.5	85.1/91.7	91.4/94.9
iForest	78.4/81.6	77.4/84.1	74.9/76.8	80.7/86.1	83.1/88.1

Table 5.2 : The Average F1/AUC Scores of AND on ESC-10 dataset.

Algorithm	Mel- Spectrogram	F-TDNN	TDNN-LSTM	ResNet-34	VGG-16
Stacked AE	83.1/86.8	76.4/83.1	70.4/74.3	81.5/86.1	81.4/88.7
VAE	76.0/83.2	76.1/83.4	66.4/67.8	83.8/86.3	81.3/85.2
kNN	81.1/88.2	84.8/86.9	70.7/73.8	83.8/87.2	88.4/**89.1**
GMM	80.5/87.9	85.1/**88.9**	77.0/77.7	83.0/86.7	**89.0/89.1**
OCSVM	80.5/85.8	78.1/81.1	76.2/74.1	83.1/85.0	86.2/88.3
iForest	67.2/71.4	62.4/65.2	59.4/57.8	63.3/66.3	71.3/73.0

Table 5.3 : The Average F1/AUC Scores of AND on US8K dataset.

Algorithm	Mel- Spectrogram	F-TDNN	TDNN-LSTM	ResNet-34	VGG-16
Stacked AE	65.7/74.6	63.8/68.9	59.9/63.3	80.8/84.4	81.3/84.8
VAE	62.2/68.6	60.1/66.6	56.7/62.4	74.8/77.9	74.7/76.7
kNN	72.5/78.7	69.2/74.8	65.5/72.8	**83.5/87.1**	82.0/85.6
GMM	70.8/78.9	73.0/78.9	71.2/78.5	80.1/85.9	**85.1/87.7**
OCSVM	68.9/71.3	65.1/68.9	66.1/73.5	76.4/80.1	**84.1/87.3**
iForest	60.5/62.8	59.7/64.0	55.8/57.7	62.2/66.8	63.3/68.2

Table 5.4 : The Average F1/AUC Scores of AND on ESC-50 dataset.

Algorithm	Mel- Spectrogram	F-TDNN	TDNN-LSTM	ResNet-34	VGG-16
Stacked AE	68.9/69.8	65.5/67.6	58.1/60.4	68.9/70.1	**71.6/72.7**
VAE	53.4/58.9	60.9/64.4	59.9/62.2	67.4/68.8	66.3/69.7
kNN	67.9/70.7	66.6/68.3	60.1/63.8	70.9/71.8	70.4/71.1
GMM	71.0/73.8	68.1/69.8	59.5/64.7	**71.2/73.4**	**71.9/73.4**
OCSVM	68.9/71.4	64.3/66.8	58.8/60.0	65.7/68.8	68.1/69.4
iForest	52.2/54.2	56.1/58.2	48.1/52.8	56.2/58.1	59.2/60.1

In the AND scenario most complicated for the algorithms, in which only one event is unknown, and the rest is assumed as known classes, the performances obtained with VGG-16 and F-TDNN algorithms were more reliable, within most of the AND experiments. The algorithms with the audio features provided efficient performances are demonstrated in Figure 5.10a–d, within the scenario.

5.6.2 Results of incremental class-learning

In the experiment of ICL-based AER, each event class in the dataset was sequentially transmitted to the LwF, iCaRL, and FearNet algorithms to be incrementally learned. Therefore, the annotation information of the event classes is given to the models without the AND step, and each model tries to adapt the features of the class while avoiding the catastrophic forgetting problem. The ICL with AER performances were evaluated in terms of change in accuracy over the number of known classes, using a pre-determined test-set of multiple audio samples for all the events in each dataset. The samples in the test set were predicted by the ICL-based AER model, and the accuracy values were computed after each new event was incrementally learned.

The average change in accuracy values of LwF algorithm is demonstrated in Figure 5.11– 5.14. The performance was also examined in the same fashion in which a test set including 30% of the audio samples for each class was selected from each dataset to be used for the calculation of prediction accuracy after each new acoustic class was incrementally learned. The LwF algorithm provided the best performance using ResNet-34 deep features in our dataset, Domestic. The VGG features had similar accuracy values with ResNet deep features (Figure 5.12). Similar AER performances with the US8K and ESC-50 which are the most complicated datasets due to the number and characteristics of audio samples were obtained using the VGG-16, ResNet-34, and F-TDNN deep features (Figure 5.13, 5.14). The worst ICL-based AER performance was acquired using the audio feature representations of TDNN-LSTM in each dataset.

The performance of the iCaRL algorithm for ICL-based AER is shown in Figure 5.15– 5.18, in which the best performance for each dataset was observed with the deep features of VGG-16. Using the audio samples in ESC-50 including 50 classes, the ResNet-34 deep features provided similar AER performance to the VGG-16 deep features. However, Mel-spectrograms and the deep features from the TDNN-LSTM network had the worst performances with the iCaRL algorithm.

In Figure 5.19–5.22, the ICL-based AER performances of the FearNet algorithm are demonstrated on the dataset, in which the algorithm presented the best overall performances in each dataset. Moreover, the iCaRL algorithm provided efficient ICL

Table 5.5 : The Average Accuracy Values of the ICL Algorithms with the VGG/F-TDNN features in Five Experiments

Algorithm	Domestic	ESC-10	US8K	ESC-50
LwF	69.4/64.8	64.2/60.0	57.1/54.4	24.1/20.9
iCaRL	78.5/77.6	68.1/68.3	62.1/59.6	28.1/21.1
FearNet	**80.7/81.4**	**74.3/71.0**	**63.8/59.5**	**30.8/24.7**

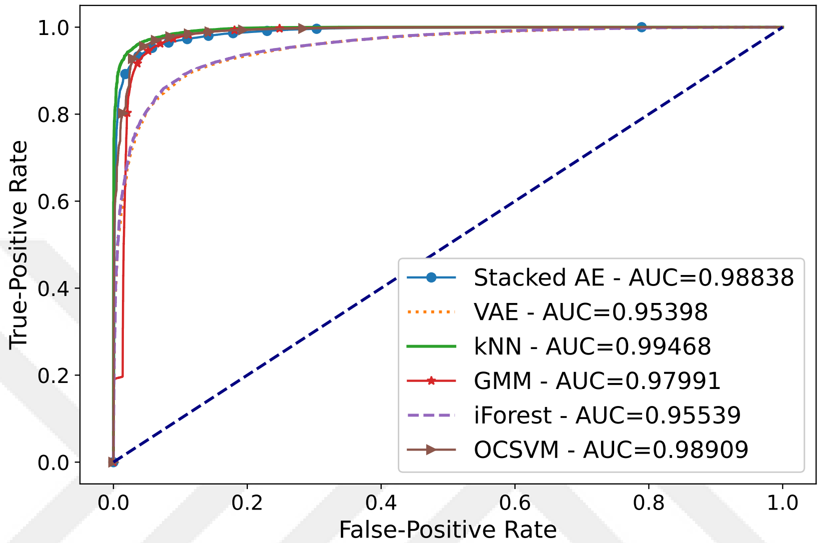

Figure 5.3 : The AUC curves of the best performances of methods with VGG-16 features obtained in the AND scenario in which only one event is known on Domestic.

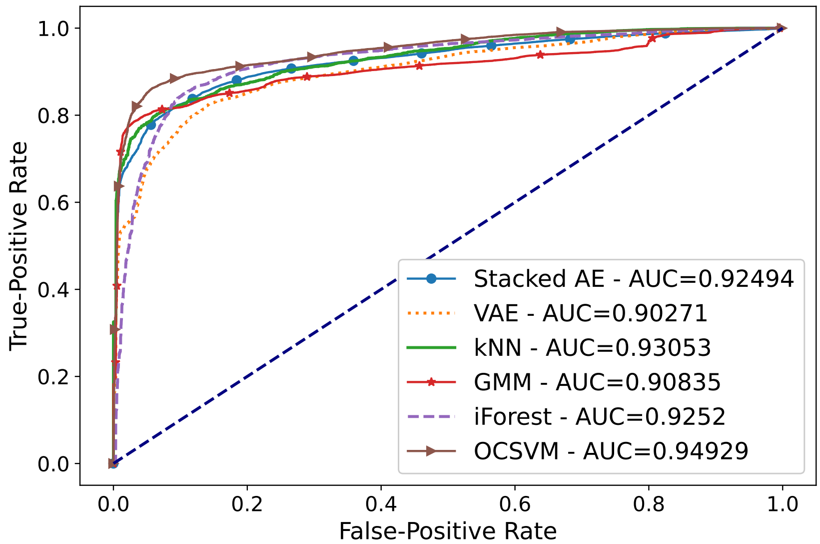

Figure 5.4 : The AUC curves of the best performances of methods with F-TDNN features obtained in the AND scenario in which only one event is known on ESC-10.

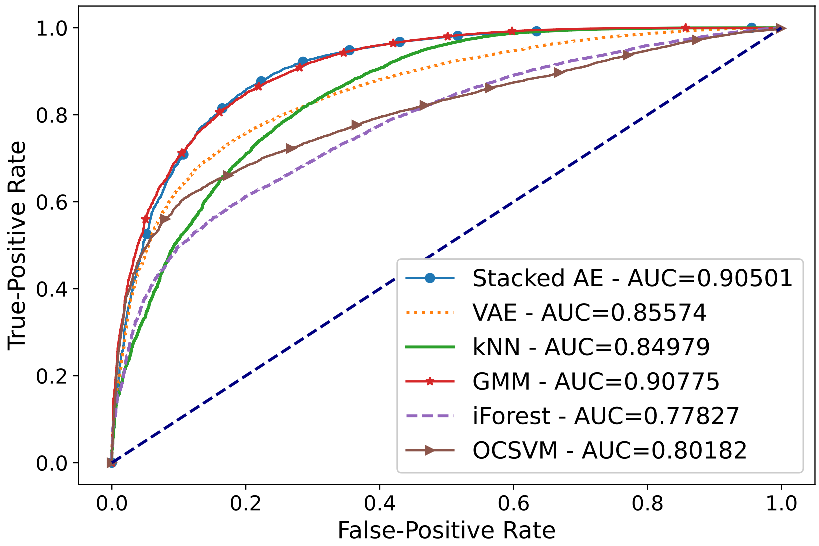

Figure 5.5 : The AUC curves of the best performances of methods with VGG-16 features obtained in the AND scenario in which only one event is known on US8K.

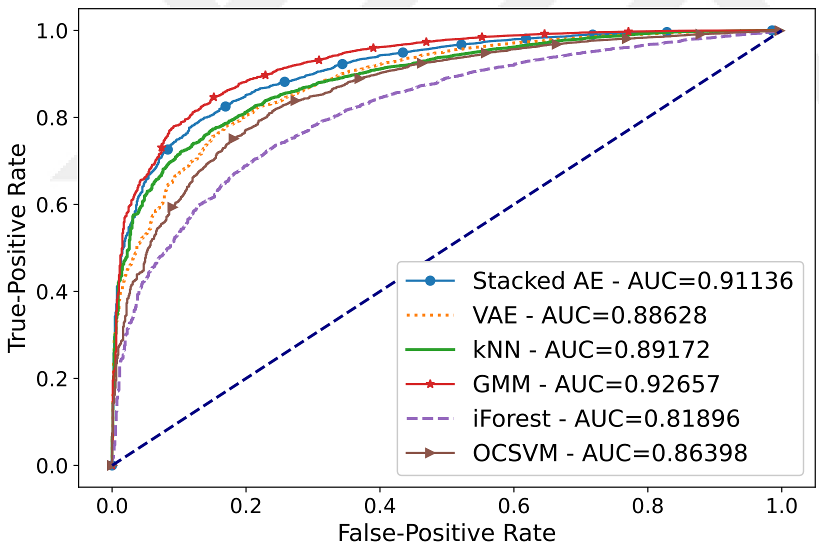

Figure 5.6 : The AUC curves of the best performances of methods with VGG-16 features obtained in the AND scenario in which only one event is known on ESC-50.

performances, but it is observed that in the acoustic domain, the LwF algorithm is not a robust and satisfactory algorithm, which had the worst performance due to the forgetting of the previous knowledge. Thus, in the experiment of ICL with AND-based ASA (Section 5.6.3), it is decided for a robust and efficient ASA to employ the

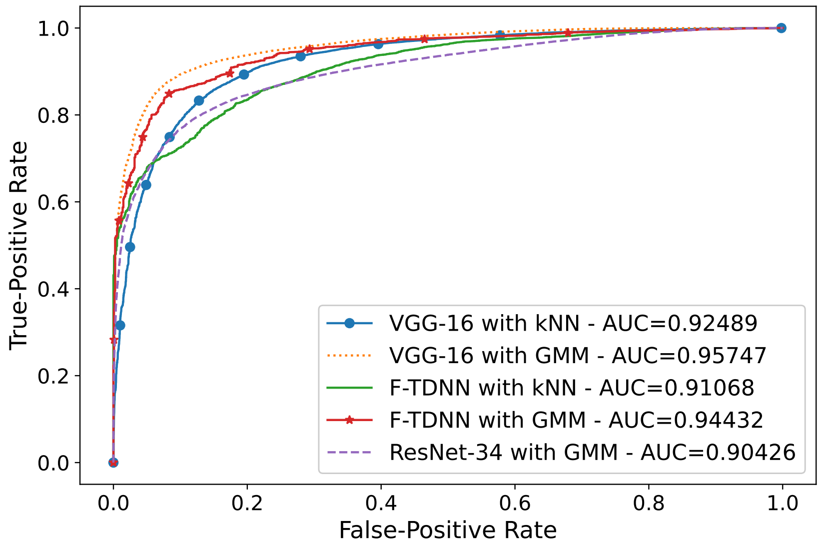

Figure 5.7 : The AUC curves of the best performances of the AND methods with VGG-16, F-TDNN, and ResNet-34 features in the most complicated AND scenario in which only one event is unknown on Domestic.

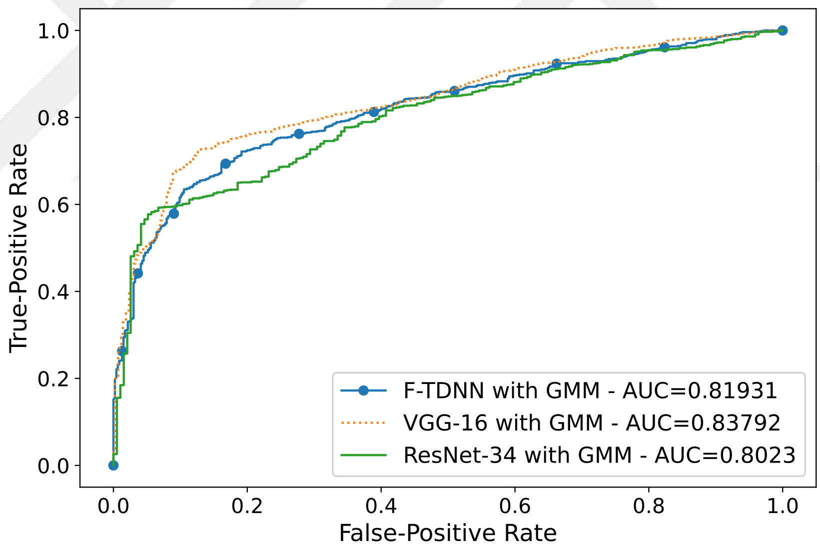

Figure 5.8 : The AUC curves of the best performances of the AND methods with VGG-16, F-TDNN, and ResNet-34 features in the most complicated AND scenario in which only one event is unknown on ESC-10.

integration of the FearNet or iCaRL algorithms for ICL with the GMM algorithm for AND using the deep features by VGG-16 and F-TDNN.

For the evaluation of the catastrophic forgetting problem, the Ω_{all} values of the ICL algorithms with VGG and F-TDNN features listed in Table. 5.6 are computed in which the AER accuracy of an ICL model after learning a new class is computed and

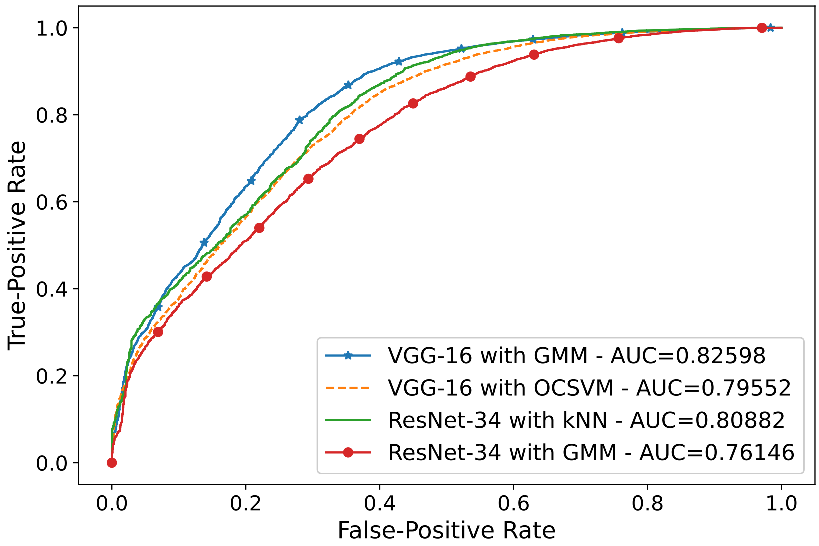

Figure 5.9 : The AUC curves of the best performances of the AND methods with VGG-16 and ResNet-34 features in the most complicated AND scenario in which only one event is unknown on US8K.

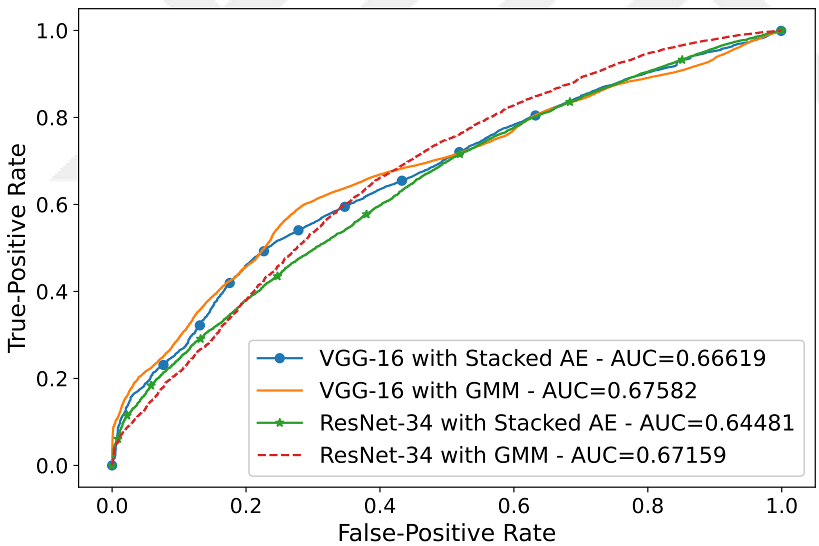

Figure 5.10 : The AUC curves of the best performances of the AND methods with VGG-16 and ResNet-34 features, in the most complicated AND scenario in which only one event is unknown on ESC-50.

normalized by using Eq. (5.2). It is observed that the iCaRL and FearNet algorithms were less affected by the forgetting problem than the LwF algorithm. In the Domestic and ESC-10 datasets, iCaRL and FearNet provided more than 0.7 and 0.8 values, respectively, in which the forgetting problem is observed by using the datasets at least.

Table 5.6 : The Average Ω_{all} Values of the ICL Algorithms and the Offline Algorithm with the VGG/F-TDNN features

Algorithm	Domestic	ESC-10	US8K	ESC-50
LwF	0.633/0.604	0.564/0.540	0.471/0.454	0.341/0.309
iCaRL	0.778/0.757	0.721/0.708	0.586/0.566	0.481/0.422
FearNet	0.840/0.795	0.802/0.790	0.738/0.695	0.580/0.524
Offline	0.927/0.878	0.886/0.773	0.812/0.786	0.744/0.697

5.6.3 Results of incremental class-learning with acoustic novelty detection

The experiment of the proposed ASA approach (Figure 5.1) for ICL of new acoustic classes recently detected without human supervision was conducted by using the most efficient ICL and AND algorithms with more distinctive audio features. Therefore, in deep features of VGG and F-TDNN in each dataset, an initial model is generated by a randomly selected audio sample of a randomly selected acoustic class. In the ASA process, each sample detected as a new class is incrementally learned by an ICL model, and more than one audio sample of an event may be detected as a new class. Therefore, at the end of the ASA, the total number of known classes may be bigger than the actual number of classes, because an acoustic class in a complicated dataset may be composed of completely different audio samples such as *dog barking*, *human voice*, *music*, etc. In the evaluation of the performance, the accuracy values were computed with respect to the pseudo-labels like sub-classes of the actual ones and the actual labels of base classes.

The average accuracy values of five different experiments for the proposed ASA approach are listed in Table 5.7. Using the largest dataset (ESC-50), each experiment was conducted by the same audio feature types and the ICL-based AER and AND models of the same algorithms, that are initially generated with different audio samples of a different event class. The accuracy values of the algorithms are demonstrated in Figure 5.23, and the most efficient ASA performance was obtained by FearNet using the VGG deep features like in the experiments of ICL (Section 5.6.2). At the end of the experiment, the ASA process started with only one event class used to generate the initial models, and a totally 214 acoustic event classes were detected by the GMM algorithm over time.

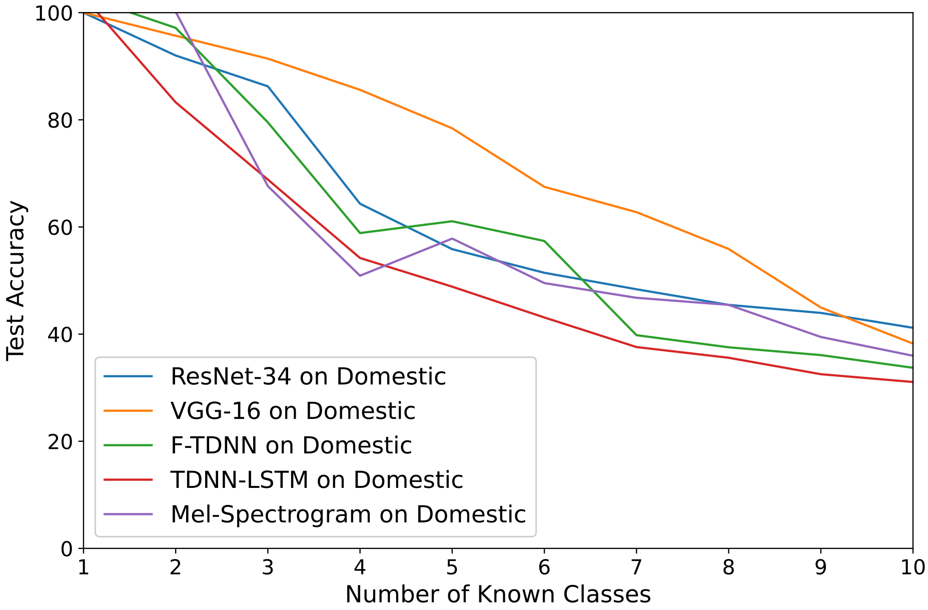

Figure 5.11 : The average change in accuracy while incrementally learning new classes by LwF on Domestic.

Table 5.8 : The Average Accuracy Values and Average Number of Detected Classes of the FearNet and iCaRL with GMM, and multilayered HMM in Five Experiments.

	Accuracy Values on VGG and Number of Detected Events			
Algorithm	**Domestic**	**ESC-10**	**US8K**	**ESC-50**
iCaRL	51.8/16	47.7/21	38.4/38	20.4/174
FearNet	51.6/26	**50.9**/21	**43.3**/38	**24.1**/174
Multileyered HMM	47.4/26	43.4/39	40.8/52	17.0/241

Table 5.7 : The Average Accuracy Values of the ICL Algorithms with GMM using VGG/F-TDNN Features and Average Number of Detected Classes in Three Experiments.

	Avg. Accuracy Values on VGG/F-TDNN/ Avg. Number of Detected Events			
Algorithm	**Domestic**	**ESC-10**	**US8K**	**ESC-50**
iCaRL	56.4/51.0/26	48.0/44.3/36	42.4/36.2/40	14.4/9.7/226
FearNet	**59.1/52.6**/26	**53.3/50.3**/36	**43.9/39.3**/40	**17.8/14.7**/226

In the second part of the experiment of lifelong learning, the deep learning method with satisfying ICL and AND models is compared with the proposed low computational cost approach for lifelong learning as explained in detail in Chapter 4, which is a multilayered HMM. During a lifelong learning process, the average changes in accuracy values of the proposed lifelong ASA methods using ICL algorithms with

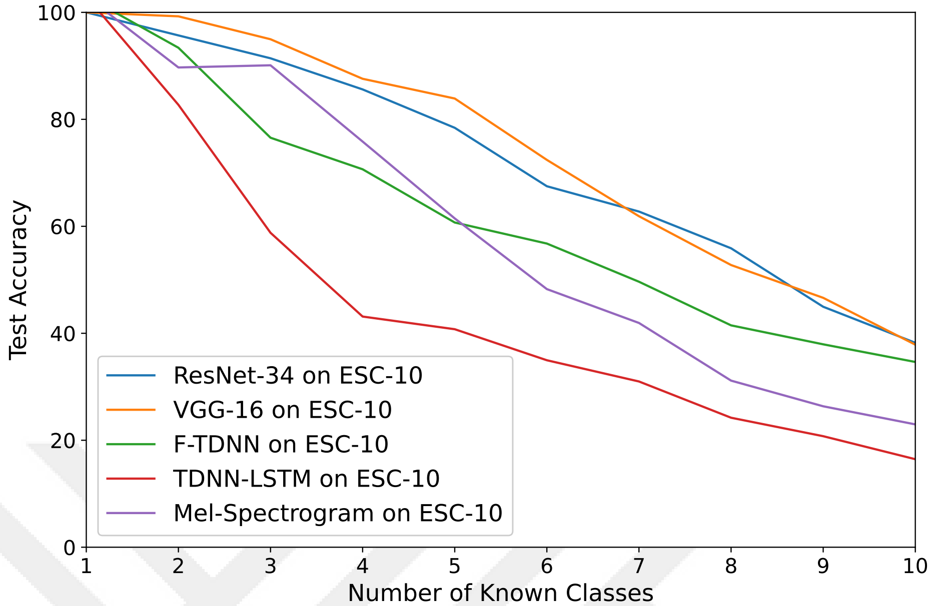

Figure 5.12 : The average change in accuracy while incrementally learning new classes by LwF on ESC-10.

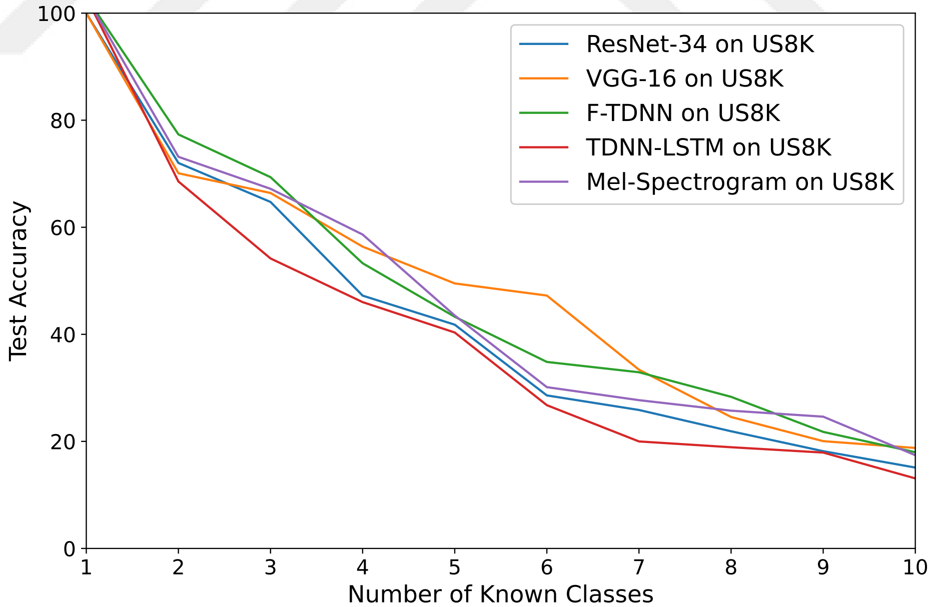

Figure 5.13 : The average change in accuracy while incrementally learning new classes by LwF on US8K.

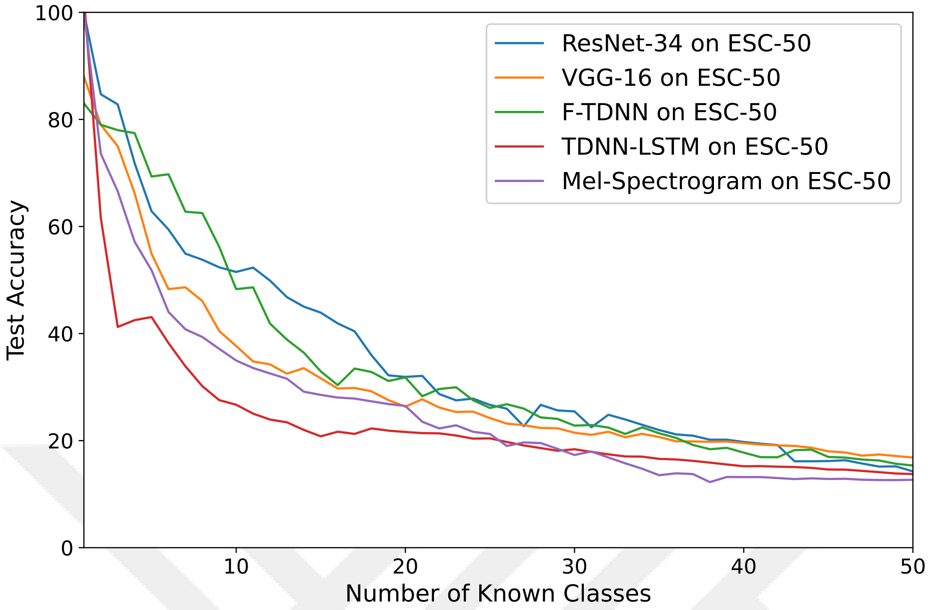

Figure 5.14 : The average change in accuracy while incrementally learning new classes by LwF on ESC-50.

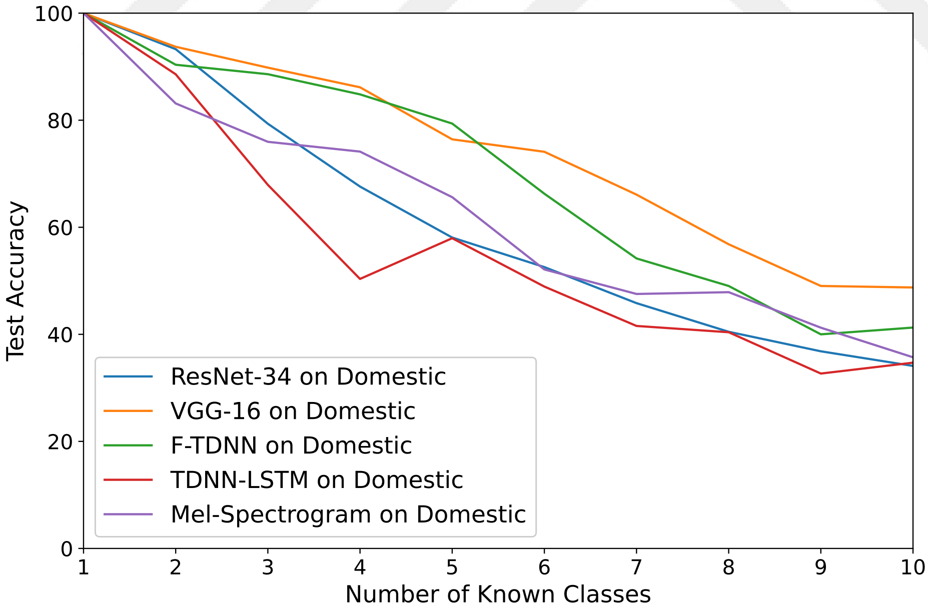

Figure 5.15 : The average change in accuracy while incrementally learning new classes by iCaRL on Domestic.

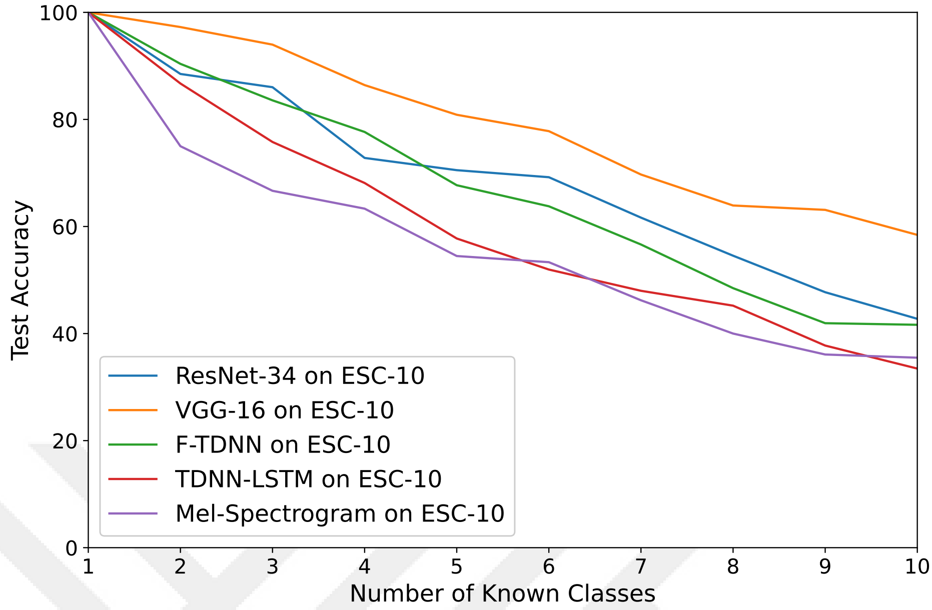

Figure 5.16 : The average change in accuracy while incrementally learning new classes by iCaRL on ESC-10.

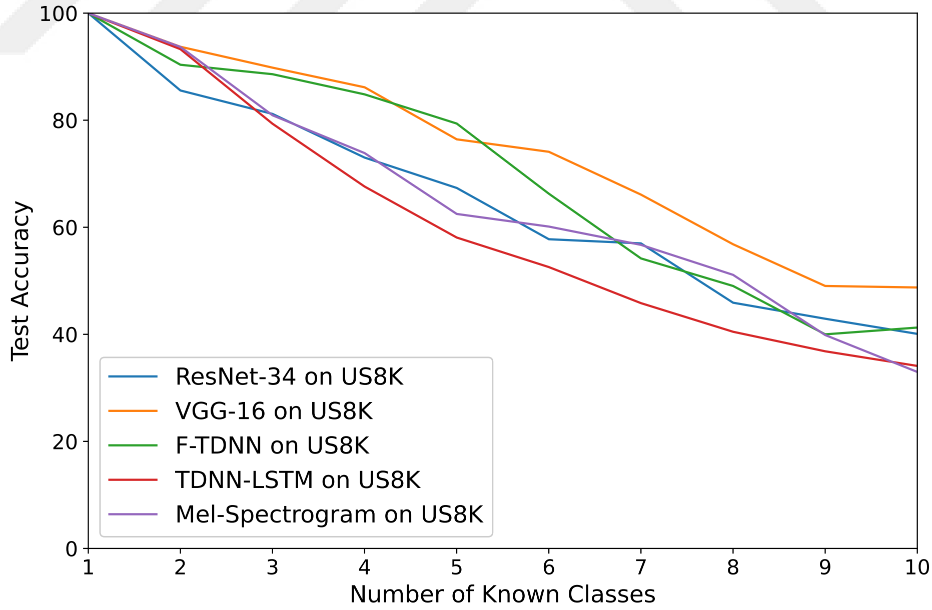

Figure 5.17 : The average change in accuracy while incrementally learning new classes by iCaRL on US8K.

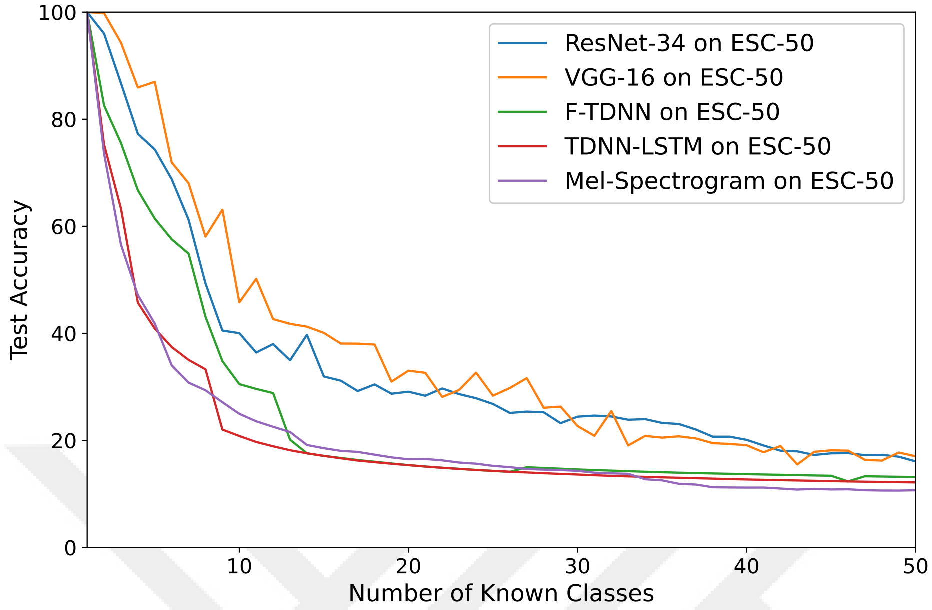

Figure 5.18 : The average change in accuracy while incrementally learning new classes by iCaRL on ESC-50.

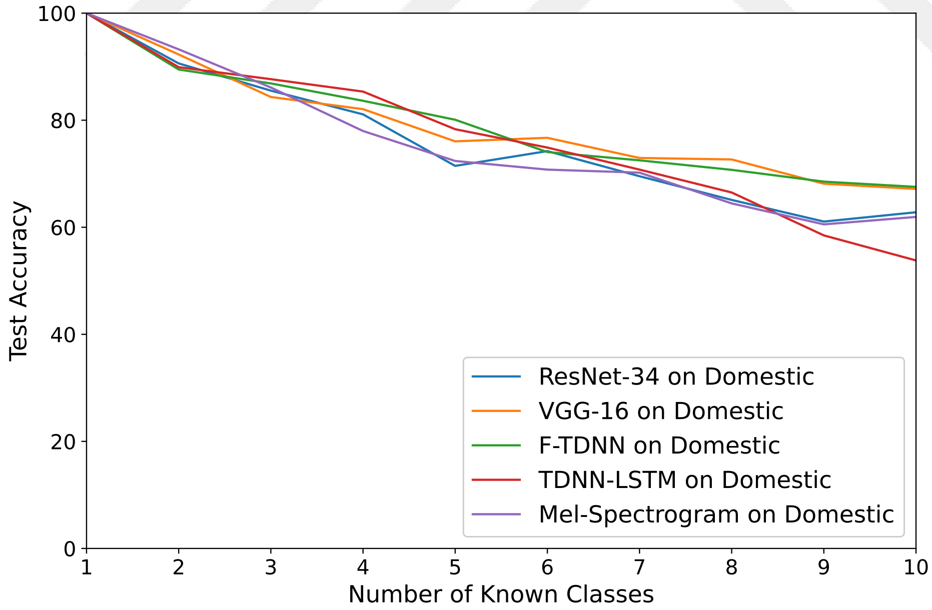

Figure 5.19 : The average change in accuracy while incrementally learning new classes by FearNet on Domestic.

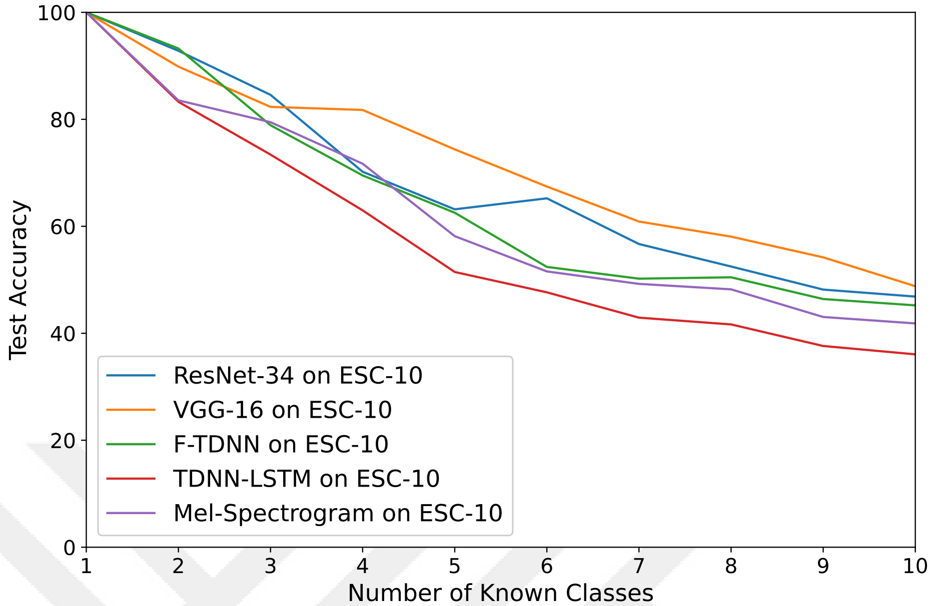

Figure 5.20 : The average change in accuracy while incrementally learning new classes by FearNet on ESC-10.

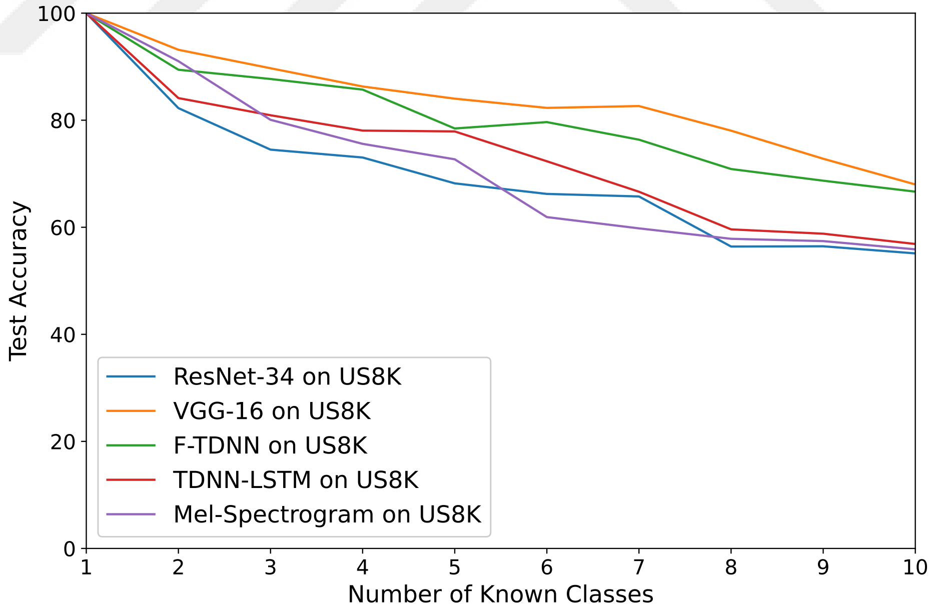

Figure 5.21 : The average change in accuracy while incrementally learning new classes by FearNet on US8K.

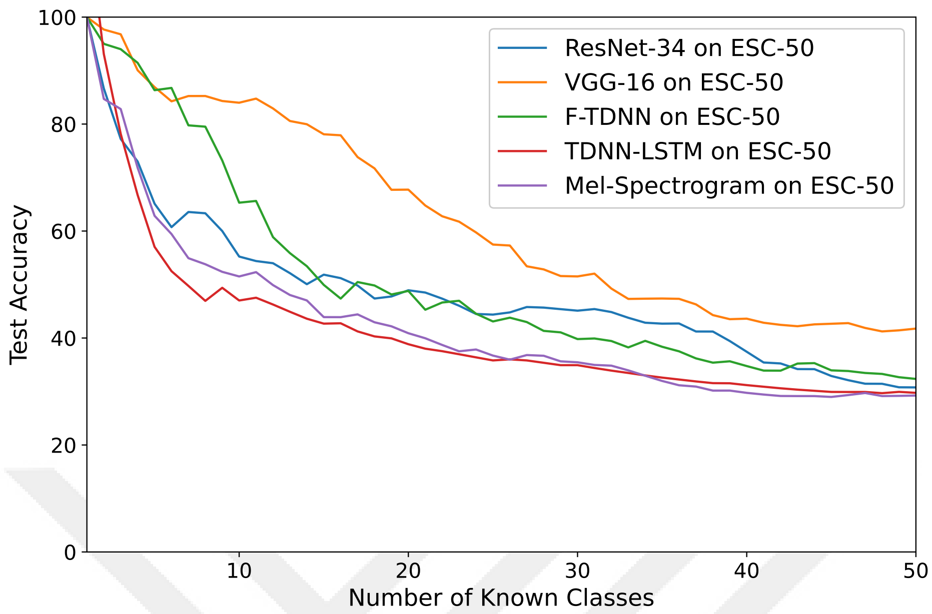

Figure 5.22 : The average change in accuracy while incrementally learning new classes by FearNet on ESC-50.

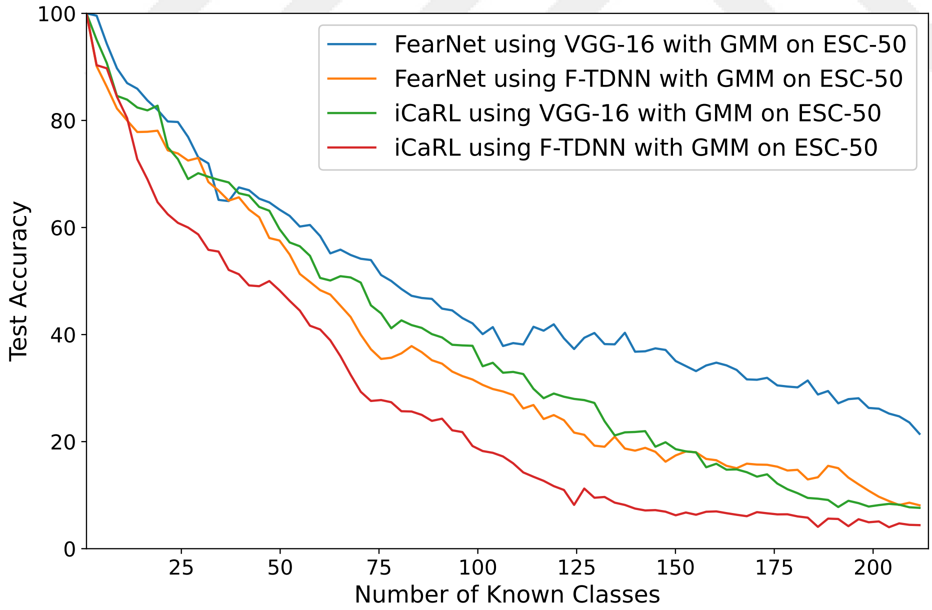

Figure 5.23 : The performances of FearNet and iCaRL with GMM on ESC-50 in which the deep features, VGG, and F-TDNN of new classes detected by GMM are incrementally learned.

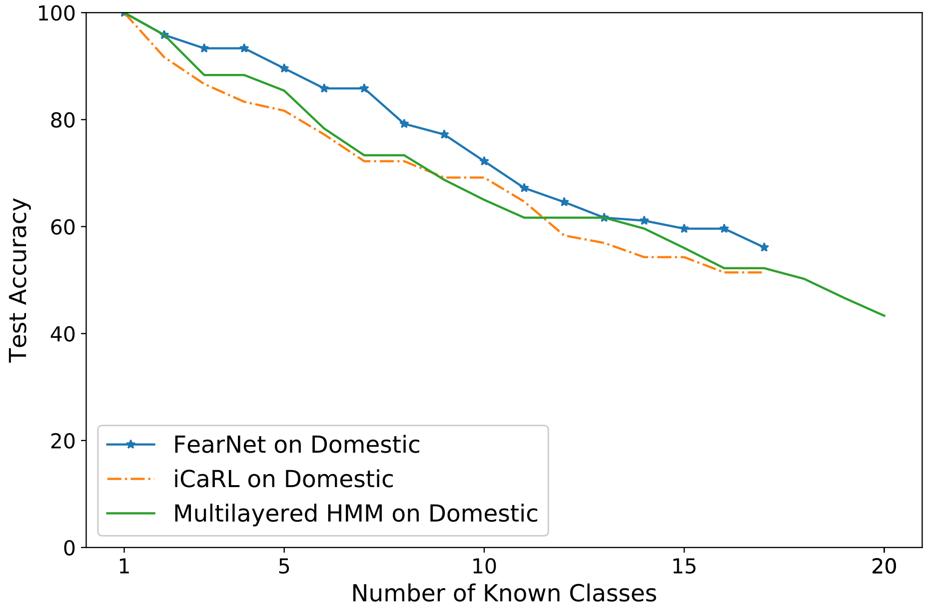

Figure 5.24 : The performances of FearNet with GMM, iCaRL with GMM, and multilayered HMM on Domestic in which VGG and MFCC features are used for the ICL models and the multilayered HMM, respectively.

GMM on VGG features and multilayered HMM on MFCC features are demonstrated in Figure 5.24– 5.27. Using the multilayered HMM, 38 pseudo-classes were detected, although only 21 new classes were detected in the AND step of the deep learning-based approach. Also, the AER performances of the ICL algorithms are better than the real-time approach, but the ICL algorithms are not usable in real-time.

In Table 5.8, the average F1-scores of AER performances of iCaRL with GMM, FearNet with GMM, and multilayered HMM are listed. In the experiment, an event class in the domestic dataset is randomly selected to generate initial models, and randomly selected samples of the other classes are sequentially fed into the models. The information about the event class of the samples is not used during the lifelong learning.

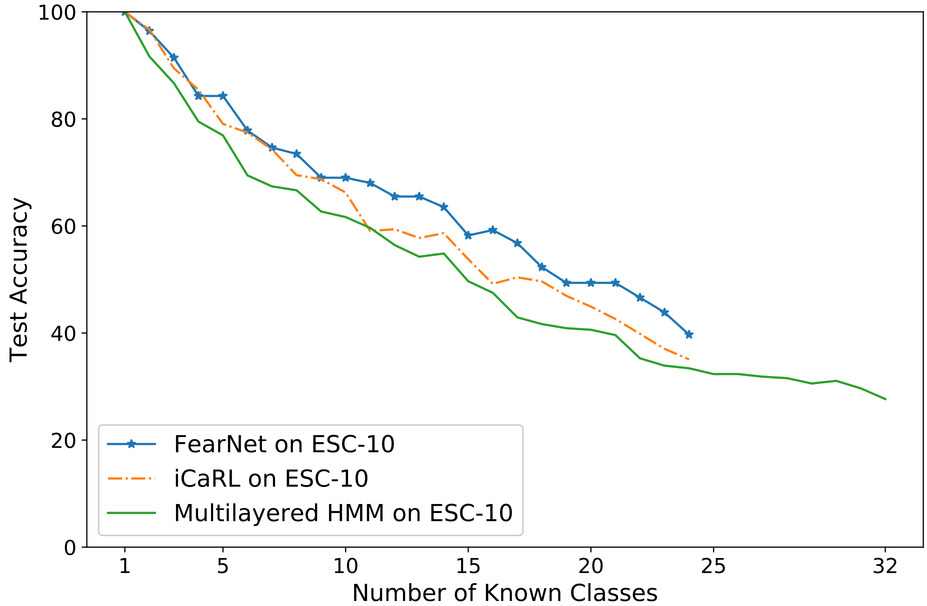

Figure 5.25 : The performances of FearNet with GMM, iCaRL with GMM, and multilayered HMM on ESC-10 in which VGG and MFCC features are used for the ICL models and the multilayered HMM, respectively.

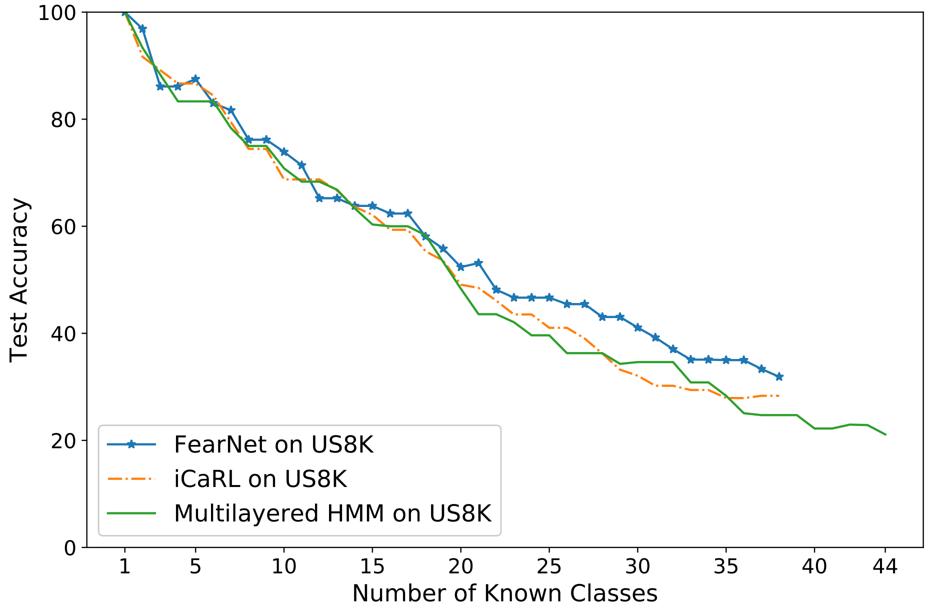

Figure 5.26 : The performances of FearNet with GMM, iCaRL with GMM, and multilayered HMM on US8K in which VGG and MFCC features are used for the ICL models and the multilayered HMM, respectively.

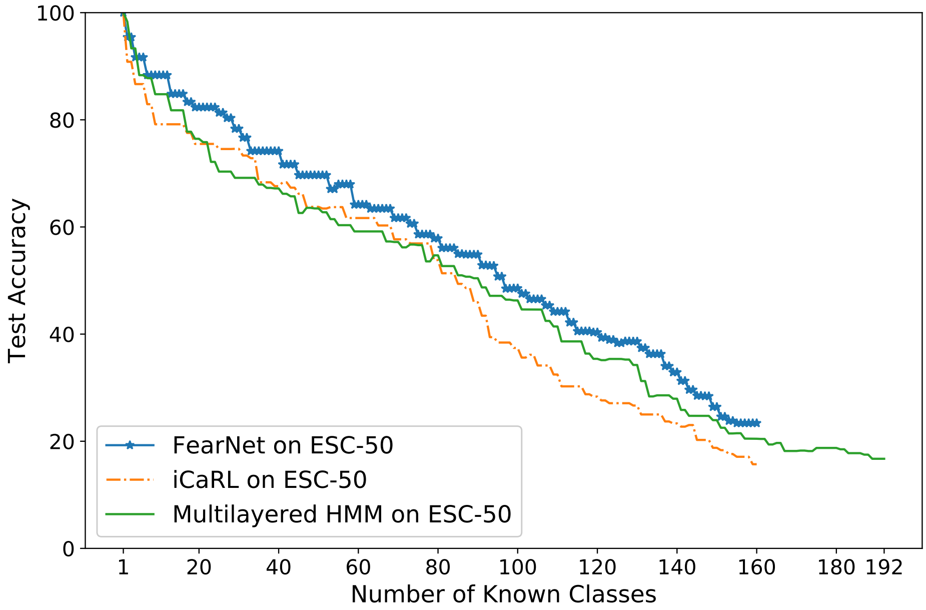

Figure 5.27 : The performances of FearNet with GMM, iCaRL with GMM, and multilayered HMM on ESC-50 in which VGG and MFCC features are used for the ICL models and the multilayered HMM, respectively.

6. DISCUSSION

This chapter gives a comprehensive discussion of the approaches and datasets including the comparative observations for the lifelong learning in Auditory Scene Analysis (ASA) and future work.

6.1 Observations

For processing the streaming signal, it is observed that the localization of the most salient sound source and segmentation of audio patterns are vital necessities to achieve the source-specific analysis in a time-series manner. For the adaptation of new knowledge from the sources, the online capabilities of the traditional algorithms such as SVM, GBT, and MLP provided unsatisfactory performances in real-world data in terms of AER prediction accuracy. As a result, the adaptation is performed by retraining the model using a limited number of useful feature sets for the known events, but all of the supervised algorithms are trained with the entire data of a fixed number of classes. Therefore, a class-specific time-series based algorithm like HMM was investigated to model the variable-length time-series patterns segmented from streaming audio signals for the tasks of AND, AER, and AER with AND in the ASA approach with low computational cost.

On the other hand, after conducting many experiments for novelty detection, the state-of-the-art probabilistic unsupervised method, GMM was estimated as the most reliable algorithm. The performances of GMM for novelty (outlier/anomaly) detection tasks have been demonstrated in various studies in different domains, so in this thesis, its satisfactory performances in varied datasets have been observed also in the auditory domain. Thus, in the experiment of ICL with AND-based ASA, GMM was integrated with an ICL algorithm. In the experiment of ICL algorithms using different audio feature types, the best performance was exhibited by the FearNet algorithm. In addition, the iCaRL algorithm which provided the best performances in several ICL works, presented robust performances in the experiments of ICL and ICL with

AND. One of the main challenges in incremental learning by deep neural networks, the catastrophic forgetting problem is not observed when using FearNet and iCaRL algorithms in each dataset. Therefore, it is proved in this thesis that ICL performances with acoustic data are comparable to the performances, especially in the computer vision domain, and so ICL is viable also in acoustic tasks. Therefore, ICL and ICL with AND-based ASA can be performed in various acoustic problems, in which incremental learning of new classes/tasks/domains will enhance the performance of the tasks such as bioacoustic [137], acoustic anomaly detection [106], or robot audition. Also, multi-modal frameworks (e.g., audio-visual recognition tasks [138]) including an audition modality can be efficiently achieved.

In the experiments of a low computational cost approach for lifelong learning, it is observed that the AND and AER performances were affected by similar classes in the domestic dataset collected in the scope of this thesis. Several patterns from the events classes of *toilet flush*, *dishwasher*, and *washing hands* are relatively similar, therefore the majority of the mispredictions in AND and AER were found due to these similarities of the classes. It is observed that the most distinctive sounds are from the *vacuum cleaner* and *closing the door* which are easily detected by the universal HMM in the lower layer and predicted by the HMMs in the upper layer for AER. Furthermore, in the experiments of the deep learning-based lifelong learning approach, a few observations were made about the sounds of the classes in the benchmark datasets. In ESC and US8K datasets, there are various environmental sounds with variations in background noise, which may be one of the reasons for the performance decrease in the AND and AER predictions, and the similarities between the classes were observed which degraded the performances of the AND and ICL based AER models. In ESC-50 dataset, the urban/outdoor sounds have many similar sounds especially from the classes of *washing machine*, *sea waves*, *airplane* and *train*. On the other hand, in the classes of US8K dataset, the classes of *air conditioner* and *drilling* include very similar sounds, which are misidentified in several AND scenarios and AER evaluation tests. Moreover, the ICL performance of the approach was affected less than the low computation cost approach by the similarities in the domestic dataset.

6.2 Future Work

For future work, an intelligent agent/robot endowed with the lifelong learning approach using low-cost computations will be used to analyze real-world scenes with unknown acoustic classes. Also, different intelligent and autonomous behaviors such as getting closer to and monitoring the events and tracking the moving sound sources in the scenes will be performed to improve the AER and AND performances. On the other hand, the performance of the proposed ICL-based ASA will be verified in real-time and evaluated on a larger audio dataset containing hundreds of classes. Regarding an AND algorithm updated by retraining in the proposed low computational cost and ICL-based lifelong learning approaches for ASA, an incremental novelty detection algorithm will be employed to be integrated with the AER algorithm.

7. CONCLUSION

In the real world, there is a large number of unlabeled data and unseen classes that may exist at any time in an acoustic environment/scene. Therefore, for both the proposed ASA approaches, the integration of Acoustic Event Recognition (AER) with Acoustic Novelty Detection (AND) in a lifelong manner is achieved in the audio domain for AER in acoustic scenes with novel event classes, considering different challenges and requirements. For low computational cost Auditory Scene Analysis (ASA), it is aimed to provide an efficient ASA capability on streaming data captured by an acoustic sensor in a real domestic environment, so the efficient and fast AER and AND algorithms were developed in a semi-supervised manner. A time-series based method for each task, AER and AND was employed for real-world scene analysis, and new knowledge about a known or a new class recently detected by the AND model was adapted into the models. Hence, a multilayered HMM for ASA with streaming learning was implemented to analyze audio features of the patterns segmented from the streaming signal. The approach uses a two-level thresholding mechanism for the prediction of either "known", "unknown", or "uncertain" for each audio pattern. To prolong the time of the ASA by decreasing the computational time, various tasks were accomplished, which are a selection of fast and efficient AER and AND algorithms, feature selection for dimension reduction, parameter tuning for faster training, and retraining of Hidden Markov Models (HMMs) using only a few numbers of patterns. A national journal about the ASA approach with low-cost computations for lifelong learning was published [33].

However, the ASA approach is based on the old-fashioned algorithm, HMM, and incapable of incremental learning, but the ASA process will not perform sufficiently in real-time or near real-time after a while due to the number of classes and the required computational time. Therefore, by using algorithms capable of incremental learning, the deep learning-based ASA approach was proposed and two international journals about the AND task of the approach and the ASA approach were published. In the experiments of the ASA approach, the integration of Incremental Class-Learning (ICL) based AER and AND, the most promising audio feature types, and the most efficient and effective ICL and AND algorithms were estimated in various analyses. Unlike the acoustic studies using Time Delay Neural Network (TDNN) to model the human voice, two TDNN-based networks, Factorized TDNN (F-TDNN) and TDNN based Long Term-Short Memory (TDNN-LSTM) have been exploited for the extraction of deep embedding vectors from Mel-Frequency Cepstral Coefficient (MFCC) and raw-signal, respectively, instead of spectrograms/images like in the other works. In the experiments, efficient and satisfactory AND and ICL performances are observed by the F-TDNN in each dataset, and TDNN-LSTM sometimes provided acceptable performances in the Domestic and ESC-10 datasets.

After the evaluations and analysis of the experiments, the ICL and AND algorithms can be efficiently employed in the auditory domain as much as in the computer vision tasks. In an acoustic scene, a new acoustic event class can be incrementally learned while avoiding catastrophic forgetting by the state-of-the-art ICL algorithms. However, according to the required inputs of the algorithms, for training, the networks require data in an image or embedding vector format. Therefore, to achieve the acoustic tasks using the networks, it is decided to change the input format of the state-of-the-art ICL and AND networks or to extract images or embedding vectors from the audio signal like in most deep learning-based acoustic works. Moreover, the embedding vectors are extracted using the images of the sounds. Therefore, the analysis of audio features for the algorithms and extraction of embedding vector using also the raw audio signal and most famous low-level audio features in addition to the spectrograms are within the scope of this thesis.

In this thesis, to achieve AER in acoustic scenes with novel events, the first work for the ICL of acoustic events was presented in [34], in which an ICL model is integrated with a novelty detection algorithm to incrementally learn new classes recently detected using various types of audio features. To evaluate several algorithms for ICL and Acoustic Novelty Detection (AND), the most famous audio benchmark datasets, ESC-10, ESC-50, and $US8K$, and a domestic audio dataset collected for this thesis, have been utilized in this presented work and thesis. For human- or animal-like lifelong learning, a novel ICL-based scene analysis approach was developed and proposed in this thesis, in which an AND capability was used to be integrated with the ICL module for acoustic tasks. In literature, the ICL without human supervision has not been well-investigated also in the tasks of other domains such as computer vision. The main aim is to incrementally learn the acoustic knowledge captured in an acoustic scene, which belongs to a known or unknown class. Due to the existence of unlabeled data and unseen classes, the AND and AER models in the approach were employed in a semi-supervised manner by retraining or incremental learning to pseudo-label and adapt the unlabeled data into the models.

The other challenge focused on in this thesis is the size of the data belonging to the new classes, which will be adapted into the models after the detection. Therefore, in this thesis, various augmentation techniques of the audio signal were investigated to increase the size of the data of the recently detected novel event class.

In the experiments of ICL with AND-based ASA, the performances of different ICL and AND algorithms were analyzed using five types of audio features, which are Mel-spectrogram and deep audio features of the pre-trained F-TDNN, TDNN-LSTM, Visual Geometry Group (VGG), and Residual Neural Network (ResNet) to estimate the robust feature types and algorithms for ICL and AND tasks. Moreover, the pre-training and use of TDNNs for extraction of audio embedding vectors were achieved for the first time in the work [34] and this thesis.

REFERENCES

[1] **Salamon, J.**, **Jacoby, C.**, and **Bello, J. P.** (2014). A dataset and taxonomy for urban sound research, *Proceedings of the 22nd ACM international conference on Multimedia*, pp.1041–1044.

[2] **Young, S. H.** and **Scanlon, M. V.** (2001). Robotic vehicle uses acoustic array for detection and localization in urban environments, *Unmanned Ground Vehicle Technology III*, volume4364, International Society for Optics and Photonics, pp.264–273.

[3] **Wang, J. C.**, **Lee, H. P.**, **Wang, J. F.**, and **Lin, C. B.** (2008). Robust environmental sound recognition for home automation, *IEEE transactions on automation science and engineering*, *5*(1), 25–31.

[4] **Sinapov, J.**, **Weimer, M.**, and **Stoytchev, A.** (2008). Interactive learning of the acoustic properties of objects by a robot, *Procceedings of the RSS Workshop on Robot Manipulation: Intelligence in Human Environments, Zurich, Switzerland.*

[5] **Lee, C. H.**, **Han, C. C.**, and **Chuang, C. C.** (2008). Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients, *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(8), 1541–1550.

[6] **Carletti, V.**, **Foggia, P.**, **Percannella, G.**, **Saggese, A.**, **Strisciuglio, N.**, and **Vento, M.** (2013). Audio surveillance using a bag of aural words classifier, *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, IEEE, pp.81–86.

[7] **Okuno, H. G.**, **Nakatani, T.**, and **Kawabata, T.** (1995). Cocktail-Party Effect with Computational Auditory Scene Analysis—Preliminary Report—, Advances in Human Factors/Ergonomics, volume 20, Elsevier, pp.503–508.

[8] **Naronglerdrit, P.** and **Mporas, I.** (2017). Recognition of Indoors Activity Sounds for Robot-Based Home Monitoring in Assisted Living Environments, *International Conference on Interactive Collaborative Robotics*, Springer, pp.153–161.

[9] **Saltali, I.**, **Sariel, S.**, and **Ince, G.** (2016). Scene analysis through auditory event monitoring, *Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents*, pp.1–6.

[10] **Rivenez12, M.**, **Gorea, A.**, **Pressnitzer, D.**, and **Drake, C.** (2003). The tolerance window for sequences of musical, environmental and artificial sounds, *Proceedings of the 7th International Conference on Music Perception and Cognition*, pp.17–21.

[11] **Chu, S.**, **Narayanan, S.**, and **Kuo, C. C. J.** (2011). Unstructured environmental audio: Representation, classification and modeling, Machine Audition: Principles, Algorithms and Systems, *IGI Global*, pp.1–21.

[12] **Wyse, L.** (2017). Audio spectrogram representations for processing with convolutional neural networks, *arXiv preprint arXiv:1706.09559*.

[13] **Piczak, K. J.** (2015). Environmental sound classification with convolutional neural networks, *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, pp.1–6.

[14] **Phan, H.**, **Chén, O. Y.**, **Pham, L.**, **Koch, P.**, **De Vos, M.**, **McLoughlin, I.**, and **Mertins, A.** (2019). Spatio-temporal attention pooling for audio scene classification, *arXiv preprint arXiv:1904.03543*.

[15] **Waldekar, S.** and **Saha, G.** (2018). Wavelet-based audio features for acoustic scene classification, *Tech. Rep., DCASE 2018 Challenge*.

[16] **Yu, F.**, **Fang, J.**, **Chen, B.**, and **Shao, Y.** (2021). An incremental learning based convolutional neural network model for large-scale and short-term traffic flow, *International Journal of Machine Learning and Computing*, *11*(2).

[17] **Pereira, R. M.**, **Bertolini, D.**, **Teixeira, L. O.**, **Silla Jr, C. N.**, and **Costa, Y. M.** (2020). COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios, *Computer Methods and Programs in Biomedicine*, *194*, 105532.

[18] **Shao, Y.**, **Zhao, Y.**, **Yu, F.**, **Zhu, H.**, and **Fang, J.** (2021). The Traffic Flow Prediction Method Using the Incremental Learning-Based CNN-LTSM Model: The Solution of Mobile Application, *Mobile Information Systems*, *2021*.

[19] **McCloskey, M.** and **Cohen, N. J.** (1989). Catastrophic interference in connectionist networks: The sequential learning problem, Psychology of learning and motivation, volume 24, Elsevier, pp.109–165.

[20] **Robins, A.** (1995). Catastrophic forgetting, rehearsal and pseudorehearsal, *Connection Science*, *7*(2), 123–146.

[21] **Vaila, R.**, **Chiasson, J.**, and **Saxena, V.** (2020). Continuous Learning in a Single-Incremental-Task Scenario with Spike Features, *International Conference on Neuromorphic Systems 2020*, pp.1–4.

[22] **Zhao, H.**, **Wang, H.**, **Fu, Y.**, **Wu, F.**, and **Li, X.** (2020). Memory Efficient Class-Incremental Learning for Image Classification, *arXiv preprint arXiv:2008.01411*.

[23] **Yu, L.**, **Liu, X.**, and **van de Weijer, J.** (2020). Self-Training for Class-Incremental Semantic Segmentation, *arXiv preprint arXiv:2012.03362*.

[24] **Maltoni, D.** and **Lomonaco, V.** (2019). Continuous learning in single-incremental-task scenarios, *Neural Networks*, *116*, 56–73.

[25] **Koh, E.**, **Saki, F.**, **Guo, Y.**, **Hung, C.Y.**, and **Visser, E.** (2020). Incremental Learning Algorithm For Sound Event Detection, *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp.1–6.

[26] **Wang, Y.**, **Bryan, N. J.**, **Cartwright, M.**, **Bello, J. P.**, and **Salamon, J.** (2021). Few-Shot Continual Learning for Audio Classification, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.321–325.

[27] **Palanisamy, K.**, **Singhania, D.**, and **Yao, A.** (2020). Rethinking cnn models for audio classification, *arXiv preprint arXiv:2007.11154*.

[28] **Zhu, X.**, **Goldberg, A. B.**, and **Khot, T.** (2009). Some new directions in graph-based semi-supervised learning, *2009 IEEE International Conference on Multimedia and Expo*, IEEE, pp.1504–1507.

[29] **Chapelle, O.**, **Scholkopf, B.**, and **Zien, A.** (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews], *IEEE Transactions on Neural Networks*, *20*(3), 542–542.

[30] **Mao, C. H.**, **Lee, H. M.**, **Parikh, D.**, **Chen, T.**, and **Huang, S. Y.** (2009). Semi-supervised co-training and active learning based approach for multi-view intrusion detection, *Proceedings of the 2009 ACM symposium on Applied Computing*, pp.2042–2048.

[31] **Abdel-Hakim, A. E.** and **Deabes, W.** (2020). Handling Missing Annotations in Supervised Learning Data, *arXiv preprint arXiv:2002.07113*.

[32] **Zhou, D. W.**, **Yang, Y.**, and **Zhan, D. C.** (2021). Learning to Classify With Incremental New Class, *IEEE Transactions on Neural Networks and Learning Systems*.

[33] **Bayram, B.**, **Duman, T. B.**, and **İnce, G.** (2020). Real time detection of acoustic anomalies in industrial processes using sequential autoencoders, *Expert Systems*, e12564.

[34] **Bayram, B.** and **İnce, G.** (2021). An Incremental Class-Learning Approach with Acoustic Novelty Detection for Acoustic Event Recognition, *Sensors*, *21*(19), 6622.

[35] **Zhou, S.** and **Beigi, H.** (2020). A transfer learning method for speech emotion recognition from automatic speech recognition, *arXiv preprint arXiv:2008.02863*.

[36] **Chen, C. P.**, **Zhang, S. Y.**, **Yeh, C. T.**, **Wang, J. C.**, **Wang, T.**, and **Huang, C. L.** (2019). Speaker characterization using tdnn-lstm based speaker embedding, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.6211–6215.

[37] **Gemmeke, J. F.**, **Ellis, D. P.**, **Freedman, D.**, **Jansen, A.**, **Lawrence, W.**, **Moore, R. C.**, ..., and **Ritter, M.** (2017). Audio set: An ontology and human-labeled dataset for audio events, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.776–780.

[38] **Popescu, M.** and **Mahnot, A.** (2009). Acoustic fall detection using one-class classifiers, *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, pp.3505–3508.

[39] **Meire, M.** and **Karsmakers, P.** (2019). Comparison of deep autoencoder architectures for real-time acoustic based anomaly detection in assets, *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 2, IEEE, pp.786–790.

[40] **Suefusa, K.**, **Nishida, T.**, **Purohit, H.**, **Tanabe, R.**, **Endo, T.**, and **Kawaguchi, Y.** (2020). Anomalous sound detection based on interpolation deep neural network, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.271–275.

[41] **Lakshmi, S. V.** and **Prabakaran, T. E.** (2014). Application of k-nearest neighbour classification method for intrusion detection in network data, *International Journal of Computer Applications*, *97*(7).

[42] **Ntalampiras, S.**, **Potamitis, I.**, and **Fakotakis, N.** (2011). Probabilistic novelty detection for acoustic surveillance under real-world conditions, *IEEE Transactions on Multimedia*, *13*(4), 713–719.

[43] **Antonini, M.**, **Vecchio, M.**, **Antonelli, F.**, **Ducange, P.**, and **Perera, C.** (2018). Smart audio sensors in the internet of things edge for anomaly detection, *IEEE Access*, *6*, 67594–67610.

[44] **Li, Z.** and **Hoiem, D.** (2017). Learning without forgetting, *IEEE transactions on pattern analysis and machine intelligence*, *40*(12), 2935–2947.

[45] **Rebuffi, S. A.**, **Kolesnikov, A.**, **Sperl, G.**, and **Lampert, C. H.** (2017). icarl: Incremental classifier and representation learning, *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp.2001–2010.

[46] **Kemker, R.** and **Kanan, C.** (2017). Fearnet: Brain-inspired model for incremental learning, *arXiv preprint arXiv:1711.10563*.

[47] **Piczak, K. J.** (2015). ESC: Dataset for environmental sound classification, *Proceedings of the 23rd ACM international conference on Multimedia*, pp.1015–1018.

[48] **Nanni, L.**, **Maguolo, G.**, **Brahnam, S.**, and **Paci, M.** (2020). An Ensemble of Convolutional Neural Networks for Audio Classification, *arXiv preprint arXiv:2007.07966*.

[49] **Pandeya, Y. R.**, **Kim, D.**, and **Lee, J.** (2018). Domestic cat sound classification using learned features from deep neural nets, *Applied Sciences*, *8*(10), 1949.

[50] **Schlüter, J.** and **Grill, T.** (2015). Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks., *ISMIR*, pp.121–126.

[51] **Nanni, L.**, **Maguolo, G.**, and **Paci, M.** (2020). Data augmentation approaches for improving animal audio classification, *Ecological Informatics*, *57*, 101084.

[52] **Theodorou, T.**, **Mporas, I.**, and **Fakotakis, N.** (2014). An overview of automatic audio segmentation, *International Journal of Information Technology and Computer Science (IJITCS)*, *6*(11), 1.

[53] **Shon, S.**, **Han, D. K.**, and **Ko, H.** (2013). Abnormal acoustic event localization based on selective frequency bin in high noise environment for audio surveillance, *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, IEEE, pp.87–92.

[54] **Asano, F.**, **Goto, M.**, **Itou, K.**, and **Asoh, H.** (2001). Real-time sound source localization and separation system and its application to automatic speech recognition, *Seventh European Conference on Speech Communication and Technology*.

[55] **Kushwaha, M.**, **Koutsoukos, X.**, **Volgyesi, P.**, and **Ledeczi, A.** (2009). Acoustic source localization and discrimination in urban environments, *2009 12th International Conference on Information Fusion*, IEEE, pp.1859–1866.

[56] **Schmidt, R.** (1986). Multiple emitter location and signal parameter estimation, *IEEE transactions on antennas and propagation*, *34*(3), 276–280.

[57] **Nakamura, K.**, **Nakadai, K.**, **Asano, F.**, **Hasegawa, Y.**, and **Tsujino, H.** (2009). Intelligent sound source localization for dynamic environments, *2009 IEEE/RSJ international conference on Intelligent Robots and Systems*, IEEE, pp.664–669.

[58] **Bian, X.**, **Abowd, G. D.**, and **Rehg, J. M.** (2004). *Using sound source localization to monitor and infer activities in the Home*, (technical report), Georgia Institute of Technology.

[59] **Bian, X.**, **Abowd, G. D.**, and **Rehg, J. M.** (2005). Using sound source localization in a home environment, *International Conference on Pervasive Computing*, Springer, pp.19–36.

[60] **Liu, Y. W.**, **Liang, H. M.**, **Lao, S. Y.**, **Wu, C. W.**, **Hao, H. K.**, **Kung, F. J.**, ..., and **Kang, S. C.** (2014). Developing “voice care”: Real-time methods for event recognition and localization based on acoustic cues, *2014*

IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, pp.1–6.

[61] **Atrey, P. K.**, **Maddage, N. C.**, and **Kankanhalli, M. S.** (2006). Audio based event detection for multimedia surveillance, *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, IEEE, pp.V–V.

[62] **Bachu, R.**, **Kopparthi, S.**, **Adapa, B.**, and **Barkana, B.** (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal, *American Society for Engineering Education (ASEE) zone conference proceedings*, American Society for Engineering Education, pp.1–7.

[63] **Panagiotakis, C.** and **Tziritas, G.** (2005). A speech/music discriminator based on RMS and zero-crossings, *IEEE Transactions on multimedia*, *7*(1), 155–166.

[64] **Vafeiadis, A.**, **Votis, K.**, **Giakoumis, D.**, **Tzovaras, D.**, **Chen, L.**, and **Hamzaoui, R.** (2017). Audio-based event recognition system for smart homes, *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, IEEE, pp.1–8.

[65] **Janse, P. V.**, **Magre, S. B.**, **Kurzekar, P. K.**, and **Deshmukh, R.** (2014). A comparative study between mfcc and dwt feature extraction technique, *International Journal of Engineering Research and Technology*, *3*(1), 3124–3127.

[66] **Logan, B.** (2000). Mel frequency cepstral coefficients for music modeling, *In International Symposium on Music Information Retrieval*, Citeseer.

[67] **Vozáriková, E.**, **Pleva, M.**, **Juhár, J.**, and **Cizmár, A.** (2011). Surveillance system based on the acoustic events detection, *Journal of Electrical and Electronics Engineering*, *4*(1), 255.

[68] **Ntalampiras, S.**, **Potamitis, I.**, and **Fakotakis, N.** (2010). A multidomain approach for automatic home environmental sound classification, *Eleventh Annual Conference of the International Speech Communication Association*.

[69] **Kataria, S.**, **Nidadavolu, P. S.**, **Villalba, J.**, **Chen, N.**, **Garcia-Perera, P.**, and **Dehak, N.** (2020). Feature enhancement with deep feature losses for speaker verification, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.7584–7588.

[70] **Li, D.**, **Sethi, I. K.**, **Dimitrova, N.**, and **McGee, T.** (2001). Classification of general audio data for content-based retrieval, *Pattern recognition letters*, *22*(5), 533–544.

[71] **Xu, C.**, **Maddage, N. C.**, and **Shao, X.** (2005). Automatic music classification and summarization, *IEEE transactions on speech and audio processing*, *13*(3), 441–450.

[72] **Wellhausen, J.** and **Höynck, M.** (2003). Audio thumbnailing using MPEG-7 low-level audio descriptors, *Internet Multimedia Management Systems IV*, volume5242, International Society for Optics and Photonics, pp.65–73.

[73] **Kim, H. G.**, **Moreau, N.**, and **Sikora, T.** (2004). Audio classification based on MPEG-7 spectral basis representations, *IEEE Transactions on Circuits and Systems for Video Technology*, *14*(5), 716–725.

[74] **Muhammad, G.** and **Alghathbar, K.** (2009). Environment recognition from audio using MPEG-7 features, *2009 Fourth International Conference on Embedded and Multimedia Computing*, IEEE, pp.1–6.

[75] **Rabaoui, A.**, **Lachiri, Z.**, and **Ellouze, N.** (2009). Using HMM-based classifier adapted to background noises with improved sounds features for audio surveillance application, *Int. J. Signal Process*, *3*, 535–545.

[76] **Abu-El-Quran, A. R.**, **Goubran, R. A.**, and **Chan, A. D.** (2006). Security monitoring using microphone arrays and audio classification, *IEEE Transactions on Instrumentation and Measurement*, *55*(4), 1025–1032.

[77] **Ciaburro, G.** and **Iannace, G.** (2020). Improving smart cities safety using sound events detection based on deep neural network algorithms, *Informatics*, volume 7, Multidisciplinary Digital Publishing Institute, p. 23.

[78] **Müller, R.**, **Illium, S.**, **Ritz, F.**, and **Schmid, K.** (2020). Analysis of Feature Representations for Anomalous Sound Detection, *arXiv preprint arXiv:2012.06282*.

[79] **Zhang, Z.**, **Xu, S.**, **Zhang, S.**, **Qiao, T.**, and **Cao, S.** (2020). Learning Frame Level Attention for Environmental Sound Classification, *arXiv preprint arXiv:2007.07241*.

[80] **Fathima, N.**, **Patel, T.**, **Mahima, C.**, and **Iyengar, A.** (2018). TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages., *INTERSPEECH*, pp.3197–3201.

[81] **Huang, J.**, **Tao, J.**, **Liu, B.**, **Lian, Z.**, and **Niu, M.** (2019). Efficient modeling of long temporal contexts for continuous emotion recognition, *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, pp.185–191.

[82] **Zhou, X.**, **Zhuang, X.**, **Liu, M.**, **Tang, H.**, **Hasegawa-Johnson, M.**, and **Huang, T.** (2007). HMM-based acoustic event detection with AdaBoost feature selection, Multimodal technologies for perception of humans, Springer, pp.345–353.

[83] **Chandrakala, S.** and **Jayalakshmi, S.** (2019). Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies, *ACM Computing Surveys (CSUR)*, *52*(3), 1–34.

[84] **Cowling, M.** and **Sitte, R.** (2003). Comparison of techniques for environmental sound recognition, *Pattern recognition letters*, *24*(15), 2895–2907.

[85] **Rabaoui, A.**, **Kadri, H.**, **Lachiri, Z.**, and **Ellouze, N.** (2008). Using robust features with multi-class SVMs to classify noisy sounds, *2008 3rd International Symposium on Communications, Control and Signal Processing*, IEEE, pp.594–599.

[86] **Stowell, D.**, **Giannoulis, D.**, **Benetos, E.**, **Lagrange, M.**, and **Plumbley, M. D.** (2015). Detection and classification of acoustic scenes and events, *IEEE Transactions on Multimedia*, *17*(10), 1733–1746.

[87] **Dong, R.**, **Hermann, D.**, **Cornu, E.**, and **Chau, E.** (2007). Low-power implementation of an HMM-based sound environment classification algorithm for hearing aid application, *2007 15th European Signal Processing Conference*, IEEE, pp.1635–1638.

[88] **Do, H. M.**, **Sheng, W.**, and **Liu, M.** (2015). An open platform of auditory perception for home service robots, *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp.6161–6166.

[89] **Sangnier, M.**, **Gauthier, J.**, and **Rakotomamonjy, A.** (2015). Early frame-based detection of acoustic scenes, *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, pp.1–5.

[90] **Sinapov, J.** and **Stoytchev, A.** (2009). From acoustic object recognition to object categorization by a humanoid robot, *Proc. of the RSS 2009 Workshop on Mobile Manipulation, Seattle, WA*.

[91] **Phan, H.** and **Mertins, A.** (2014). Exploring superframe co-occurrence for acoustic event recognition, *2014 22nd European Signal Processing Conference (EUSIPCO)*, IEEE, pp.631–635.

[92] **Feki, I.**, **Ammar, A. B.**, and **Alimi, A. M.** (2011). Audio stream analysis for environmental sound classification, *2011 International Conference on Multimedia Computing and Systems*, IEEE, pp.1–6.

[93] **Lafay, G.**, **Lagrange, M.**, **Rossignol, M.**, **Benetos, E.**, and **Roebel, A.** (2016). A morphological model for simulating acoustic scenes and its application to sound event detection, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(10), 1854–1864.

[94] **Su, F.**, **Yang, L.**, **Lu, T.**, and **Wang, G.** (2011). Environmental sound classification for scene recognition using local discriminant bases and HMM, *Proceedings of the 19th ACM international conference on Multimedia*, pp.1389–1392.

[95] **Walters, T. C.** (2011). *Auditory-based processing of communication sounds*, (Ph.D. Dissertation), University of Cambridge.

[96] **McLoughlin, I.**, **Zhang, H.**, **Xie, Z.**, **Song, Y.**, and **Xiao, W.** (2015). Robust sound event classification using deep neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(3), 540–552.

[97] **Zhou, Q.**, **Feng, Z.**, and **Benetos, E.** (2019). Adaptive noise reduction for sound event detection using subband-weighted NMF, *Sensors*, *19*(14), 3206.

[98] **Noh, K.** and **Chang, J. H.** (2020). Joint optimization of deep neural network-based dereverberation and beamforming for sound event detection in multi-channel environments, *Sensors*, *20*(7), 1883.

[99] **Dempster, A. P.**, **Laird, N. M.**, and **Rubin, D. B.** (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

[100] **Xing, Z.**, **Pei, J.**, and **Keogh, E.** (2010). A brief survey on sequence classification, *ACM Sigkdd Explorations Newsletter*, *12*(1), 40–48.

[101] **Khreich, W.**, **Granger, E.**, **Miri, A.**, and **Sabourin, R.** (2012). A survey of techniques for incremental learning of HMM parameters, *Information Sciences*, *197*, 105–130.

[102] **Schölkopf, B.**, **Williamson, R. C.**, **Smola, A.J.**, **Shawe-Taylor, J.**, and **Platt, J.C.** (1999). Support vector method for novelty detection, *NIPS*, volume 12, Citeseer, pp.582–588.

[103] **Yahaya, S. W.**, **Langensiepen, C.**, and **Lotfi, A.** (2018). Anomaly detection in activities of daily living using one-class support vector machine, *UK Workshop on Computational Intelligence*, Springer, pp.362–371.

[104] **Marchi, E.**, **Vesperini, F.**, **Squartini, S.**, and **Schuller, B.** (2017). Deep recurrent neural network-based autoencoders for acoustic novelty detection, *Computational intelligence and neuroscience*, *2017*.

[105] **Nguyen, D.**, **Kirsebom, O. S.**, **Frazão, F.**, **Fablet, R.**, and **Matwin, S.** (2019). Recurrent neural networks with stochastic layers for acoustic novelty detection, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.765–769.

[106] **Bayram, B.**, **Duman, T. B.**, and **İnce, G.** (2021). Real time detection of acoustic anomalies in industrial processes using sequential autoencoders, *Expert Systems*, *38*(1), e12564.

[107] **Nguyen, M. H.**, **Nguyen, D. Q.**, **Nguyen, D. Q.**, **Pham, C. N.**, **Bui, D.**, and **Han, H. D.** (2021). Deep Convolutional Variational Autoencoder for Anomalous Sound Detection, *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*, IEEE, pp.313–318.

[108] **Shimada, K.**, **Koyama, Y.**, and **Inoue, A.** (2020). Metric learning with background noise class for few-shot detection of rare sound events,

ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp.616–620.

[109] **Parisi, G. I.** and **Wermter, S.** (2017). Lifelong learning of action representations with deep neural self-organization, *2017 AAAI Spring Symposium Series*.

[110] **Liu, B.** (2020). Learning on the job: Online lifelong and continual learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp.13544–13549.

[111] **Yiping, G.**, **Xinyu, L.**, and **Gao, L.** (2021). A deep lifelong learning method for digital twin-driven defect recognition with novel classes, *Journal of Computing and Information Science in Engineering*, *21*(3).

[112] **Gong, Z.**, **Cui, Q.**, **Chaccour, C.**, **Zhou, B.**, **Chen, M.**, and **Saad, W.** (2021). Lifelong Learning for Minimizing Age of Information in Internet of Things Networks, *ICC 2021-IEEE International Conference on Communications*, IEEE, pp.1–6.

[113] **Goodfellow, I. J.**, **Mirza, M.**, **Xiao, D.**, **Courville, A.**, and **Bengio, Y.** (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks, *arXiv preprint arXiv:1312.6211*.

[114] **Al-Behadili, H.**, **Grumpe, A.**, and **Wöhler, C.** (2015). Incremental learning and novelty detection of gestures in a multi-class system, *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, IEEE, pp.304–309.

[115] **Shmelkov, K.**, **Schmid, C.**, and **Alahari, K.** (2017). Incremental learning of object detectors without catastrophic forgetting, *Proceedings of the IEEE International Conference on Computer Vision*, pp.3400–3409.

[116] **Ren, M.**, **Liao, R.**, **Fetaya, E.**, and **Zemel, R.** (2019). Incremental few-shot learning with attention attractor networks, *Advances in Neural Information Processing Systems*, pp.5276–5286.

[117] **Al-Behadili, H.**, **Grumpe, A.**, **Dopp, C.**, and **Wöhler, C.** (2015). Incremental class learning and novel class detection of gestures using ensemble, *Workshop New Challenges in Neural Computation*, Citeseer, pp.122–132.

[118] **Ren, M.**, **Liao, R.**, **Fetaya, E.**, and **Zemel, R. S.** (2018). Incremental few-shot learning with attention attractor networks, *arXiv preprint arXiv:1810.07218*.

[119] **Guzmán, A.** (1968). Decomposition of a visual scene into three-dimensional bodies, *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pp.291–304.

[120] **Bregnian, A.S.** (1993). Auditory scene analysis: Hearing in complex environments.

[121] **Valin, J. M.**, **Yamamoto, S.**, **Rouat, J.**, **Michaud, F.**, **Nakadai, K.**, and **Okuno, H. G.** (2007). Robust recognition of simultaneous speech by a mobile robot, *IEEE Transactions on Robotics*, *23*(4), 742–752.

[122] **Shirkhodaie, A.**, **Elangovan, V.**, and **Rababaah, A.** (2011). Acoustic semantic labeling and fusion of human-vehicle interactions, *Signal Processing, Sensor Fusion, and Target Recognition XX*, volume8050, SPIE, pp.316–327.

[123] **Povey, D.**, **Cheng, G.**, **Wang, Y.**, **Li, K.**, **Xu, H.**, **Yarmohammadi, M.**, and **Khudanpur, S.** (2018). Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks., *Interspeech*, pp.3743–3747.

[124] **Yu, Y. Q.** and **Li, W. J.** (2020). Densely Connected Time Delay Neural Network for Speaker Verification, *Proc. Interspeech 2020*, 921–925.

[125] **Zhao, Y.** and **Hryniewicki, M. K.** (2018). XGBOD: improving supervised outlier detection with unsupervised representation learning, *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp.1–8.

[126] **Liu, F.T.**, **Ting, K.M.**, and **Zhou, Z.H.** (2008). Isolation forest, *2008 eighth ieee international conference on data mining*, IEEE, pp.413–422.

[127] **Yousef, M.**, **Najami, N.**, and **Khalifav, W.** (2010). A comparison study between one-class and two-class machine learning for MicroRNA target detection, *Journal of Biomedical Science and Engineering*, *3*(03), 247.

[128] **Liu, F. T.**, **Ting, K. M.**, and **Zhou, Z. H.** (2012). Isolation-based anomaly detection, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(1), 1–39.

[129] **Goldstein, M.** and **Uchida, S.** (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PloS one*, *11*(4), e0152173.

[130] **Principi, E.**, **Vesperini, F.**, **Squartini, S.**, and **Piazza, F.** (2017). Acoustic novelty detection with adversarial autoencoders, *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp.3324–3330.

[131] **Youden, W. J.** (1950). Index for rating diagnostic tests, *Cancer*, *3*(1), 32–35.

[132] **Abraham, W. C.** and **Robins, A.** (2005). Memory retention–the synaptic stability versus plasticity dilemma, *Trends in neurosciences*, *28*(2), 73–78.

[133] **Kemker, R.**, **McClure, M.**, **Abitino, A.**, **Hayes, T.**, and **Kanan, C.** (2018). Measuring catastrophic forgetting in neural networks, *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[134] **Phan, H.**, **Maaß, M.**, **Mazur, R.**, and **Mertins, A.** (2014). Random regression forests for acoustic event detection and classification, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 20–31.

[135] **Nakadai, K.**, **Okuno, H. G.**, **Nakajima, H.**, **Hasegawa, Y.**, and **Tsujino, H.** (2008). An open source software system for robot audition HARK and its evaluation, *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*, IEEE, pp.561–566.

[136] **Hayes, T. L.**, **Cahill, N. D.**, and **Kanan, C.** (2019). Memory efficient experience replay for streaming learning, *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, pp.9769–9776.

[137] **Stowell, D.** (2018). Computational bioacoustic scene analysis, *Computational analysis of sound scenes and events*, Springer, pp.303–333.

[138] **Ghaleb, E.**, **Popa, M.**, and **Asteriadis, S.** (2019). Multimodal and temporal perception of audio-visual cues for emotion recognition, *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, pp.552–558.

CURRICULUM VITAE

Name Surname: Barış Bayram

EDUCATION:

- **B.Sc. :** 2012, Izmir University of Economics, Faculty of Computer Engineering, Department of Computer Engineering
- **M.Sc. :** 2015, Istanbul Technical University, Graduate School of Science Engineering and Technology, Computer Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2021 - present, Senior Data Scientist, hepsiJET, R&D Department, Istanbul, Turkey
- 2018 - 2021, Machine Learning Engineer, Yapı Kredi Teknoloji, R&D Department, Istanbul, Turkey
- 2016 - 2017, Researcher, HP Scholarship, Istanbul Technical University, Istanbul, Turkey
- 2015 - 2016, Researcher, ITU and Honda Research Institute Japan, Istanbul, Turkey
- 2014 - 2015, R&D Engineer, Turkish Robotics Research, Istanbul, Turkey

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Bayram, B.**, İnce, G., 2021. An Incremental Class-Learning Approach with Acoustic Novelty Detection for Acoustic Event Recognition, Sensors 21 (19), 6622.

- **Bayram, B.**, Duman, T.B., İnce, G., 2021. Real time detection of acoustic anomalies in industrial processes using sequential autoencoders, Expert Systems 38 (1), e12564.

- **Bayram, B.**, İnce, G., 2020. Real-Time Auditory Scene Analysis using Continual Learning in Real Environments, Journal: European Journal of Science and Technology, 215-226.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- **Bayram, B.**, Ülkü, B., Aydın, G., Akhavan-Tabatabaei, R., Bozkaya, B., 2021. A Machine Learning Approach to Daily Capacity Planning in E-Commerce Logistics, In Machine Learning, Optimization, and Data Science: 7th International Conference, LOD 2021, 2021, Grasmere, UK.

- **Bayram, B.**, Köroğlu, B., Gönen, M., 2020. Improving fraud detection and concept drift adaptation in credit card transactions using incremental gradient boosting trees, 19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE.

- Yeşilkanat, A., **Bayram, B.**, Köroğlu, B., Arslan, S., 2020. An adaptive approach on credit card fraud detection using transaction aggregation and word embeddings, IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Cham, 2020.

- Khalilbayli, F., **Bayram, B.**, İnce, G., 2019. Unsupervised Visual Ego-motion Learning for Robots, 4th International Conference on Computer Science and Engineering (UBMK). IEEE, 2019.

- Duman, T.B., **Bayram, B.**, İnce, G., 2019. Acoustic anomaly detection using convolutional autoencoders in industrial processes, International Workshop on Soft Computing Models in Industrial and Environmental Applications, Springer, Cham, 2019.

- Al-Azzawi, N., **Bayram, B.**, İnce, G., 2018. Audiovisual attention for robots from a developmental perspective, 3rd International Conference on Computer Science and Engineering (UBMK). IEEE, 2018.

- **Bayram, B.**, İnce, G., 2018. Advances in Robotics in the Era of Industry 4.0, Industry 4.0: Managing The Digital Transformation.

- Aslan, E.S., **Bayram, B.**, İnce, G., 2017. Distributed face recognition system for indoor environments, 25th Signal Processing and Communications Applications Conference (SIU), IEEE, 2017.

- **Bayram, B.**, İnce, G., 2015. Audio-visual multi-person tracking for active robot perception, 2015 IEEE/SICE International Symposium on System Integration (SII).

- **Bayram, B.**, İnce, G., 2015. Audio-visual human tracking for active robot perception, 23rd Signal Processing and Communications Applications Conference (SIU).