UNIVERSITÉ PARIS XIII - SORBONNE PARIS NORD

École Doctorale Sciences, Technologies, Santé Galilée

Meta-Decomposition and Evaluation Processes in Machine Learning Applications

THÈSE DE DOCTORAT
présentée par:

Seyed Mohammad Reza Modaresi
Laboratoire d'Informatique de Paris Nord

pour l'obtention du grade de
DOCTEUR EN Informatique

soutenue le 15/12/2023 devant le jury d'examen composé de :

Nathalie Pernelle, Professor, Université Sorbonne Paris Nord	Présidente du jury
Samira Bouzefrane, Professor, CNAM Paris	Rapportrice
Karim DJouani, Professor, Université Paris-Est-Créteil	Rapporteur
Mostafa HaghirChehreghani, MCF, Amirkabir University of Technology	Examinateur
Dominique Vaufreydaz, Professor, Université Grenoble Alpes	Examinateur
Aomar Osmani, MCF HDR, Université Sorbonne Paris Nord	Directeur de thèse
Mohammadreza Razzazi, Professor, Amirkabir University of Technology	Co-Directeur de thèse
Abdelghani Chibani, MCF, Université Paris-Est-Créteil	Co-Encadrant de thèse

Acknowledgements

Words cannot express my gratitude to my supervisors for their invaluable patience, feedback, and the wealth of knowledge and expertise they generously shared. I am also grateful to my friends and colleagues for their late-night feedback sessions, and moral support. I would like to extend my deep thanks to my awesome friends for being a source of companionship and support, particularly during the challenging times of COVID-19. The moments of joy we shared together were invaluable. Their company and laughter have been a lifeline, making every moment memorable and played a significant role in keeping me comfort and happy throughout this period.

And finally, I would like to express my heartfelt gratitude to my incredible family. Dad, you've always been my biggest supporter, giving me the space to be creative and cheering me on in pursuing my dreams. Mom, your love, motivation, and patience in my life always inspired me to chase my dreams relentlessly. Your passion for science also setting me on the path of research from the beginning. A grateful appreciation to my amazing brother and sisters for their unwavering support, always being close to me even when physically apart. Your presence, encouragement, and patience have been my guiding lights. I've learned so much from each of you, shaping my worldview since childhood and influencing my future plans.

Thank you for being a fundamental part of my journey.

Abstract

Segmentation is a crucial primary step in a variety of real-world applications such as medical image analysis, activity recognition, and sound event detection. It involves partitioning input data into smaller segments, thereby inducing alterations in certain characteristics of the input data. This process introduces at least two families of uncontrollable biases. The first family of biases is introduced to the model due to the changes in problem space made by the segmentation. The second family of biases is caused by the segmentation process itself, including the fixation of the segmentation method and its parameters. This thesis presents a novel adaptive layer designed to augment existing medical image segmentation deep models, enhancing their performance. This adaptive layer dynamically adjusts the receptive field size based on pixel and neighboring information. These concepts are then extended to more intricate scenarios involving heterogeneous data types, presenting a novel meta-decomposition or learning-to-decompose approach for segmentation. This approach mitigates implicit biases while enabling adaptive segmentation for various data types, accommodating data variations and heterogeneities such as seasonal differences in activities. Recognizing the impact of segmentation on the problem space, the research scrutinizes the drawbacks of state-of-the-art evaluation methods, emphasizing the necessity for more comprehensive frameworks, focusing on point-based evaluation methods, neglects spatial or temporal relationships between instances. To validate the efficacy of the suggested evaluation techniques and the meta-decomposition approach, extensive experimentation is conducted across diverse concrete real-world datasets.

Keywords: Evaluation, Segmentation, Decomposition, Meta-Decomposition, Meta-Learning, Medical Image Segmentation, Activity Recognition, Sound Event Detection

Contents

1 General Introduction **1**

2 Context of the thesis **9**
2.1 Chapter Overview . . . 9
2.2 Targets Beyond zero-dimensional (0D) and their applications . . . 9
2.2.1 Medical Image Segmentation (MIS) . . . 10
2.2.2 Activity Recognition (AR) . . . 13
2.2.3 Sound Event Detection (SED) . . . 13
2.3 Segmentation Problem Challenges . . . 15
2.4 Evaluation Problem and Challenges . . . 19
2.5 Preliminaries on Evaluation (Point-based) . . . 22
2.6 Conclusion . . . 26

3 Dynamic Receptive Field **27**
3.1 Chapter Overview . . . 27
3.2 Related Works . . . 29
3.2.1 Related Works on Image Segmentation . . . 29
3.2.2 Related Works on Multi-scale Receptive Field . . . 31
3.3 Adaptive UCTransNet . . . 35
3.4 Experiment . . . 39
3.4.1 Environment Setup . . . 39
3.4.2 Datasets . . . 39
3.4.3 Hyperparameters and Implementation Details . . . 40
3.4.4 Model Complexity . . . 40
3.4.5 Evaluation Metrics . . . 41
3.4.6 SegPC 2021 Case Study . . . 41
3.4.7 ISIC 2018 case study . . . 43
3.4.8 Discussion . . . 44
3.5 Conclusion . . . 47

4 Segmentation **49**
4.1 Chapter Overview . . . 49
4.2 Related Works and Preliminaries . . . 50
4.3 Problem Formulation . . . 55
4.4 Decomposition . . . 56
4.5 Meta-Decomposition . . . 58
4.6 Experiment . . . 60

4.6.1 Case Studies . . . 61
4.6.2 Environment . . . 70
4.6.3 Framework Description and Baseline . . . 71
4.6.4 Performance Measurements . . . 72
4.6.5 Results and Discussion . . . 75
4.7 Conclusion . . . 78

5 Evaluation 81
5.1 Chapter Overview . . . 81
5.2 Beyond Point-based Metrics: Related Works . . . 83
5.2.1 Evaluation Methods in Activity Recognition (1D) . . . 83
5.2.2 Evaluation Methods in Sound Event Detection (SED) (1D) . . . 88
5.2.3 Evaluation Methods in Other Applications in (1D) . . . 90
5.2.4 The Relations of Intervals . . . 92
5.2.5 Evaluation Methods in Medical Image Segmentation (2D and 3D) . . . 94
5.2.6 Analysis of Current Trends . . . 100
5.2.7 Metric Learning . . . 104
5.2.8 Discussion . . . 104
5.3 Significance of Evaluation in a Real-World Application . . . 106
5.3.1 Background and a Short Review of Algorithms on Pre-symptomatic coronavirus disease 2019 (COVID-19) Detection . . . 107
5.3.2 Analysis of Evaluation Methods . . . 110
5.3.3 Analysis with a Less Biased Evaluation Method . . . 115
5.3.4 Discussion . . . 117
5.4 Proposed Evaluation Method . . . 118
5.4.1 Detection Property (D) . . . 120
5.4.2 Uniformity Property (U) . . . 121
5.4.3 Boundary Alignment Property (B) . . . 123
5.4.4 Total Volume Property (T) . . . 125
5.4.5 Relative Volume Property (R) . . . 126
5.4.6 Precision, Recall, and F-score . . . 127
5.4.7 Computational Complexity . . . 127
5.5 Experiments . . . 128
5.5.1 Activity Recognition Case Study . . . 128
5.5.2 Sound Event Detection Case Study . . . 132
5.5.3 Medical Image Segmentation Case Study . . . 137
5.6 Conclusion . . . 150

6 General conclusion and perspectives 151
6.1 Conclusions . . . 151
6.2 Recommendations and Future Work . . . 153

A Common Evaluation Metrics Formulation in Medical Image Segmentation 155

B More details of the used Datasets 157

Bibliography 163

List of Figures

1.1 Classical instances vs Durative instances . . . 3

2.1 Synapse Multi Organ Dataset . . . 11
2.2 Liver Tumor CT-scan Orthogonal View . . . 12
2.3 A smart home . . . 14
2.4 An Illustrative Day from the Aruba Dataset . . . 15
2.5 Sound Event Detection System . . . 16
2.6 Segmentation Examples . . . 18
2.7 Impact of evaluation in Activity Recognition . . . 20
2.8 Evaluation challenge in sound event detection . . . 20
2.9 The problem of evaluation in medical image segmentation . . . 21
2.10 Binary Confusion Matrix . . . 24
2.11 Multi-Class Confusion Matrix . . . 25

3.1 Illustration of Diverse Scales in Medical Images . . . 28
3.2 The Evolution of Inception Modules. . . . 31
3.3 MultiRes Block and its Dual and upgraded version . . . 33
3.4 Dilated Atrous Mechanism . . . 34
3.5 Multi Receptive Field . . . 34
3.6 Adaptive Convolution Layer added to the leading UCTransNet architecture. . . 38
3.7 The train loss and validation loss of SegPC 2021 dataset. . . . 42
3.8 The train loss and validation loss of ISIC 2018 dataset . . . 44
3.9 Visual comparisons of different methods on the Cytoplasm segmentation of SegPC 2021 dataset . . . 45
3.10 Visual analysis of various methods applied to the segmentation of the ISIC 2018 dataset . . . 46

4.1 Common steps in Activity Recognition . . . 53
4.2 Segmentation, Decomposition, and Meta-Decomposition . . . 60
4.3 Orange4Home dataset sensor configuration . . . 62
4.4 Orange4Home Activity Duration . . . 62
4.5 Daily Activity Patterns in the Orange4Home Dataset . . . 63
4.6 Sensor Event Frequency in Orange4Home Dataset (Part One) . . . 64
4.7 Sensor Event Frequency in Orange4Home Dataset (Part Two) . . . 65
4.8 CASAS Aruba, Home1 and Home2 datasets sensors configuration . . . 67
4.9 Aruba Dataset Activity Duration . . . 68
4.10 Aruba dataset Sensor Event Frequency . . . 69

4.11 Comparison of two segmentation algorithms . . . 74
4.12 An example TS-CM calculation . . . 74
4.13 The impact of two segmentation methods . . . 76

5.1 Event, Frame, Segment Analysis . . . 84
5.2 Common event analysis decision functions . . . 86
5.3 Two different predictions with similar accuracy . . . 87
5.4 Event Analysis Diagram . . . 87
5.5 Collar and PSDS decision functions . . . 90
5.6 Example activities . . . 92
5.8 Simplified interval relations . . . 93
5.7 Relations in allen's interval algebra . . . 93
5.9 Relations between multiple non-overlapping intervals . . . 94
5.10 Common Loss functions . . . 103
5.11 Generated red alarms by different algorithms in COVID-19 . . . 111
5.12 ROC of the different algorithms for pre-symptomatic COVID-19 detection . . . 113
5.13 TPR of various algorithms for pre-symptomatic COVID-19 detection . . . 114
5.14 ROC of the different algorithms using our less biased metric . . . 117
5.15 ROC of the different pre-symptomatic COVID-19 detection algorithms using our less biased metric . . . 118
5.16 Medical Image Segmentation Case Study a . . . 121
5.17 Medical Image Segmentation Case Study b . . . 122
5.18 Multiple organs and their medial axis . . . 124
5.19 Medical Image Segmentation Case Study c . . . 125
5.20 Differences between Total and Relative Volume Properties . . . 126
5.21 Example Algorithms in AR . . . 129
5.22 An artificially generated sound event data . . . 133
5.23 Detailed information on different classes in two systems . . . 136
5.24 Comparision of different SED systems . . . 136
5.25 Pancreas Dataset. Orthogonal View. . . . 138
5.26 Pancreas Dataset. 3D View. . . . 139
5.27 Liver Tumor Dataset. 3D View. . . . 140
5.28 Synapse Multi Organ Dataset, Orthogonal View. . . . 141
5.29 Medical Image Segmentation Case Study a . . . 142
5.30 Medical Image Segmentation Case Study b . . . 142
5.31 Medical Image Segmentation Case Study c . . . 143
5.32 Medical Image Segmentation Case Study d . . . 143
5.33 Medical Image Segmentation Case Study e . . . 144
5.34 Medical Image Segmentation Case Study f . . . 144
5.35 Medical Image Segmentation Case Study g . . . 145

B.1 The standard routines in Orange4Home dataset [Cum+18] . . . 157
B.2 Home1 Activity Duration . . . 158
B.3 Detailed Activities sensors hit map in Home1 . . . 159
B.4 Home1 Activities . . . 160
B.5 Home2 Activities . . . 160
B.6 Home2 Activity Duration . . . 161

B.7 Home2 dataset Hit Time 162

List of Tables

3.1 Comparision of the number of parameters in each model 41
3.2 Comparison of results for the SegPC 2021 dataset. 42
3.3 Comparison of results for the ISIC 2018 dataset 43

4.1 Performance evaluation of methods used for segmentation 75

5.1 Evaluation metrics in top conferences 101
5.2 Comparison of our metric in an artificially generated dataset 129
5.3 Comparison of our metric for a selected activity 130
5.4 Comparison of our metric with Tatbul metric 130
5.5 Ward's Metric in a given class 130
5.6 Macro average of all classes by our metric 131
5.7 Comparision of different methods in sound event detection 135
5.8 Evaluation of four methods with different metrics among the LiverTumor dataset.147
5.9 Evaluation of four methods with different metrics over Pancreas-CT dataset . 148
5.10 Evaluation of four methods with different metrics over the MultiOrgan dataset 149

Acronyms

Δ Infection Window Size. 111–114

$\circ$ Hadamard matrix product (elementwise product). 22, 23, 26, 37

$\hat{\partial}$ Border. 155, 156

$\mathcal{C}$ Correlated Function. 119, 120, 122

$\mathcal{G}$ Set of ground truth segment one hot encoded classes. 119, 120, 122, 126, 127, 130

$\mathcal{S}$ Predicted segments set in an instance. 119, 120, 126, 127, 129, 130

$\oslash$ Hadamard matrix division (elementwise division). 25

$\overline{C}$ C complement. 120, 121

∂ Border. 101, 155, 156

$\sqcap$ Intersection. 119–121, 125, 126

θ Segment Size. 115–117

$\widehat{\mathcal{S}}$ predicted segment one hot encoded classes. 120–122, 125, 126

A alarm. 107, 112, 115

$C+$ Positive (red) Alarm. 107, 112, 115

$C-$ Negative (green) Alarm. 107, 112, 115

$C?$ Unknown Alarm. 107, 115

K Reference day for $\mathrm{Participant(u)} \in \mathrm{PositiveCOVID-19patients(U+)}$. It is either the symptom onset day or COVID-19 test day (in case of asymptomatic infection). 108, 112, 115

$U+$ Positive COVID-19 patients. 107, 108, 112, 115

$U-$ Participants that do not have COVID-19. 108, 112, 115

U All Participant. 26, 107

α Needed adaption days. 107

τ . 155, 156

u Participant. 107, 108, 112, 115, 116

$\mathcal{V}$ Volume. 121, 126

CM Confusion Matrix. 22, 108

F1 F1 measure. 19, 25, 26, 41, 87, 89, 91, 94, 97, 129–131, 136, 138, 155

Fβ Fβ measure. 26, 96, 97, 101, 126, 127, 148, 155

MCC Matthews Correlation Coefficient. 111

RA Random Algorithm. 110, 112, 113, 116

TNR True Negative Rate. 109, 111–113

TS Threat Score. 111

0D zero-dimensional. v, 3, 5–7, 9–11, 13, 81, 82

1D one-dimensional. 3, 6, 10, 36, 83, 119, 121, 127

2D two-dimensional. 3, 6, 10, 83, 95, 119–121, 127

3D three-dimensional. 3, 6, 10, 11, 83, 95, 119–121, 127, 145, 146

Acc Accuracy. 24, 25, 41, 96, 97, 101, 106, 138, 155

AD Anomaly Detection. 109, 110

AHD Average Hausdorff Distance. 138

AI Artificial Intelligence. 3, 100

AJI aggregated Jaccard index. 99, 105

AR Activity Recognition. 2–4, 6, 9, 10, 13, 16, 17, 20, 22, 26, 47, 49–51, 56, 78, 81–84, 104, 105, 120, 128, 151, 153

ASD Average Surface Distance. 98, 101

ASSD Average Symmetric Surface Distance. 19, 94, 98, 101

B Boundary Alignment Property. 123, 125, 127, 142–149, 152

C C. 120–122, 125, 126

CNN convolutional neural network. 27, 31–33

COVID-19 coronavirus disease 2019. vi, 7, 106–117

CT Computed Tomography. 10, 12, 82, 96, 126, 138, 139, 141, 143–146, 148, 154

D Detection Property. 121, 142, 143, 146–149, 152, 153

DB Boundary Distance. 123, 124

DC Dice Similarity coefficient. 6, 19, 21, 26, 41, 94–97, 99–101, 105, 138, 147–149, 153, 155

DK Skeleton Distance. 123, 124

DN Normalized Distance. 124, 125

EAD Event Analysis Diagram. 87

FN False Negative. 6, 22–24, 26, 46, 82, 84–86, 89, 96, 97, 101, 108, 109, 111, 112, 115, 119–122, 124–126, 133–136, 144–146, 150, 153, 155

FP False Positive. 6, 22–26, 46, 82, 84–86, 89, 96, 97, 101, 108, 109, 111, 112, 115, 119–122, 124–126, 128, 133–137, 144, 146, 150, 153, 155

FPR False Positive Rate. 97, 116, 138

FWIoU Frequency weighted intersection over union. 26, 97

G ground truth segment one hot encoded classes. 119–126, 128, 129, 155, 156

g ground truth vector. 155

HD Hausdorff Distance. 19, 21, 94, 95, 98, 99, 101, 105, 123, 138, 147–149

IoT Internet of Things. 13, 83

IoU Intersection over Union. 6, 19, 26, 41, 94, 96, 97, 100–102, 105, 138, 147–149, 153, 155

L Labels. 22

MAD Mean Absolute Distance. 98, 101

MIoU Mean Intersection over Union. 26, 97

MIS Medical Image Segmentation. 3, 4, 6, 9–12, 20–22, 26, 28, 29, 31, 32, 36, 47, 56, 81–83, 94–96, 98–104, 106, 118, 120, 124, 127, 137, 148, 151, 153

ML Machine Learning. 1

MME Multi-modal Evaluation Metric. 5, 6, 147–149, 153

MRI Magnetic Resonance Imaging. 10

MSI medical similarity index. 99

NSD Normalized Surface Distance. 19, 94, 98, 101, 105, 147–149, 155, 156

OQM Objective Quality Metric. 105

PET Positron-Emission Tomography. 10

PQ Panoptic Quality. 105

PRC Precision. 6, 19, 24–26, 83, 91, 94, 96, 97, 101, 126, 127, 129–132, 138, 143, 144, 146–149, 153, 155

R Relative Volume Property. 26, 126, 142–144, 146–149, 152

ROC Receiver Operating Characteristic. 113, 116

S predicted segment one hot encoded classes. 119–122, 125, 128, 129, 155, 156

s predicted segment one hot encoded vector. 155

SED Sound Event Detection. 3, 4, 9, 10, 13, 15, 20, 22, 26, 56, 82, 83, 88, 89, 104, 120, 151, 153

T Total Volume Property. 125, 126, 142–144, 146–149, 152, 153

TN True Negative. 22–24, 85, 96, 97, 108, 109, 111, 112, 115, 125, 155

TP True Positive. 6, 22–26, 46, 82, 84–86, 89, 91, 96, 97, 101, 108, 109, 111, 112, 115, 119–122, 124–126, 128, 129, 133–137, 144, 145, 150, 153, 155

TPR Recall or True Positive Rate. 6, 19, 24–26, 83, 91, 94–97, 101, 108, 109, 111–114, 116–118, 126, 127, 129–132, 138, 143, 144, 146–149, 153, 155

U Uniformity Property. 122, 142–144, 146–149, 152

ViS Video segmentation. 101, 102

VS Volume Similarity. 99, 105, 147–149

Chapter 1

General Introduction

Evaluation holds significance through the time. According to Nietzsche Philosophy [Nie83], our understanding of reality is limited to our senses, subjective experiences, historical and cultural factors, and the fact that truth is subject to interpretation and perspective. However, he believed that critical self-reflection and evaluation of our assumptions and biases could broaden our perspectives and improve our understanding of truth. Similar to our understanding of truth, the way we evaluate things is also subjective and influenced by biases and assumptions.

In Machine Learning (ML), the term “bias” refers to any factor that favors one generalization over another [MM80; GD95]. Biases are often incorporated in the pre-processing, learning, post-processing, and evaluation steps of ML to reduce the complexity and learn correct concepts [GD95; DAr20]. However, it is crucial to evaluate the biases for the bias's selection. The effectiveness and quality of bias evaluation methods significantly impact the overall benefits gained from bias choices and the understanding of the underlying truth by ML models [GD95]. Moreover, evaluation of the bias evaluation can broaden the models' perspectives and reduce various types of biases.

An example of a common pre-processing step in many ML models is the segmentation process, which is our concentration in this thesis. According to the Oxford Dictionary, a segment is a part of something separate from the other parts or can be considered separately and segmentation is the act of dividing something into different parts. Segmentation extracts relevant features from multiple sources, simplifies the input data. It also overcomes the limitation of information in a single sample by providing adequate information about a concept [Che+21b; NGC15] or reducing the problem's complexity [Ber+18; LMS14]. For instance, a single door-open event is not adequate to identify whether the actor is leaving or entering the house. It is often assumed that the loss of information in the segmentation process is lower than that of the acceptable bound. However, this step can introduce two at least two families of uncontrollable biases. The first one results from changes made by the segmentation process on the original problem space, for instance, discretizing the problem into a fixed one-second frame in Sound Event Detection (SED), which impacts the understanding of machine learning models and the evaluation process about truth. The second family of bias arises from the segmentation process itself, including the selection of a particular segmentation method and its associated parameters, which are inextricably linked to the evaluation process. For example, an appropriate segmentation approach in one period of time in Activity Recognition (AR) may not be efficient for another period of time due to the changes in data over time. Usually, in these contexts, researchers have implicitly included their knowledge about the application in the segmentation and evaluation procedure, which is often associated with uncontrollable biases in their model and evaluation processes. Therefore, the model may misleadingly present convenience results in the training and testing phase while the results are not acceptable in the real-world. In addition, segmentation poses significant challenges due to its complexity and the impracticality of attaining an exact solution in a reasonable time [Asa+01]. The "no-free-lunch" theorem states that there is no particular bias that on average is the best one to be used [Ada+19; DAr20]. As a result, careful consideration and management of the segmentation process is necessary to ensure the validity and generalizability of the resulting models for the given application with respect to the biases.

The segmentation problem becomes more challenging when the target concepts deviate from the traditional point-like representations in classical machine learning instances. In

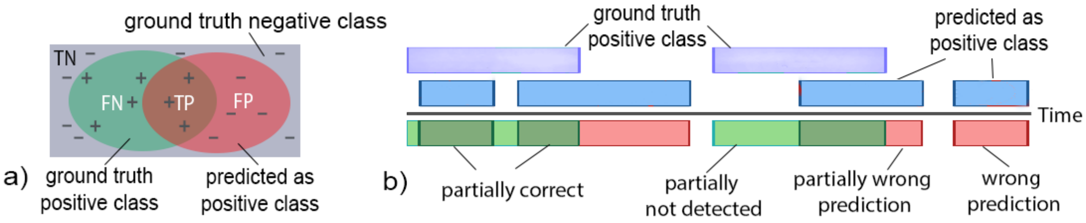

ground truth negative class
TN
FN
TP
FP
a)
ground truth positive class
predicted as positive class
ground truth positive class
predicted as positive class
Time
b)
partially correct
partially not detected
partially wrong prediction
wrong prediction

Figure 1.1: a) Classical instances (0D) b) Durative instances (1D). The horizontal line represents time, and the box shows activity duration. A durative instance may be partially correct and partially incorrect while classical instances have a binary correctness status.

the traditional scenarios, we typically work with a set of examples and one or multiple target values, for instance, predicting the category of an image or the price of a stock. However, it is not the case with several real-world applications like AR, Medical Image Segmentation (MIS), and Sound Event Detection (SED). Unlike the classical classification problems in which the predicted targets are either correct or incorrect, the targets include time intervals (one-dimensional (1D)) or shapes (two-dimensional (2D) or three-dimensional (3D) in images) and they can be correct and incorrect at the same time. For example, consider a scenario where the actual breakfast time for a person is from 8:00 to 9:00, but according to the system prediction, the person eats from 8:15 to 9:15. In this case, is this prediction correct or incorrect? This brings up two important questions: 1- How should we evaluate the methods including targets beyond 0D? 2- Do the classical evaluation methods sufficiently assess these targets with multiple dimensions? The distinction between traditional and 1D targets is highlighted in Figure 1.1.

While selecting the right dataset and ensuring that high-quality ground truth play significant roles in the evaluation process [De-+18], this thesis narrows its attention to just the metrics and criteria for evaluating competitive approaches. We assume that both the datasets and the strongly labeled ground truth are of high quality and the best prediction is the one that exactly matches the ground truth. Analysis of roughly 200 papers presented at major Artificial Intelligence (AI) conferences in 2022 (see Section 5.2.6), revealed that segmentation methods are commonly evaluated using point-based approaches that overlook spatial or temporal relationships between instances. It is particularly the case when the targets have more than zero dimensions, which is typical in segmentation problems.

The typical projection of models' output into a space with a total order relationship, such as what is done in point-based assessments, may fail to capture the different characteristics of a segment. This is particularly important because the significance of different properties changes in various applications and even different stages of an application. For example, the optimal technique for early tumor diagnosis may differ from that of used to assess treatment response. During the early diagnosis phase, it is crucial to detect even small tumoral lesions, while observing volumetric changes is essential for assessing treatment response [Li+19b]. As a result, it is necessary to review the evaluation process and project the evaluation into a multidimensional space with a partial order relationship that considers the contextual relationships between instances. This can help researchers, users, and models to make more informed decisions in selecting the appropriate technique for their specific application.

In this thesis, we address the task of segmentation and evaluation, which is a prevalent issue in numerous machine learning applications such as MIS, AR, and SED. Despite the growing interest in this research area, various challenges persist. This is particularly evident in real-world scenarios, where biases are inherent in both the segmentation and evaluation tasks. Therefore, our initial objective is first to address and reduce the potential uncontrollable biases that arise from the segmentation process. Secondly, after devising a solution, it is essential to thoroughly evaluate its performance.

As explained earlier, segmentation is a common pre-processing step in methods used in various applications. However, this step introduces at least two families of uncontrollable biases. The first one is caused by the alterations made by the segmentation process to the initial problem space, and the second one results from the segmentation process itself, including the fixation of the segmentation method and its parameters. To avoid these short comings, first, we introduce a comprehensive and unified formalization of segmentation, treating the segmentation problem as a specific case of decomposition problem. This encompasses the decomposition (segmentation), problem resolution (ML step), and composition. Incorporating the composer task in the segmentation makes it possible to evaluate the relationship between the initial problem to be solved and the problem after the segmentation, resulting in an improved evaluation and consequently selecting the appropriate segmentation method. Therefore, we can define various segmentation algorithms as hyperparameters to be optimized and automatically selected. Then, we propose a new meta-decomposition or learning-to-decompose concept, which learns to break down the original task into sub-tasks for integration with meta-learning approaches that require multiple tasks. Meta-decomposition aims to minimize segmentation biases and optimize overall system performance by learning how to generate sub-tasks instead of presuming a predetermined and fixed segmentation method. By defining segmentation as an ML hyperparameter to be adaptively learned based on the application and constraints in the outer learning algorithm, we enhance the recognition quality of the inner learning process. Our initial results demonstrate that our framework is more effective.

After developing a potential solution to the recognition problem, evaluating its performance becomes crucial, particularly when dealing with targets that extend beyond zero dimensions. Since segmentation algorithms produce heterogeneous segments (e.g., in terms of type and size), it is crucial to transfer them into a comparable space for proper evaluation; otherwise, the comparison would be invalid. In addition, selecting an appropriate algorithm is based on the objective of the application. Therefore, comparing them becomes essential. In this case, can we precisely determine one algorithm that is better than the others when dealing with more than 0D targets? These questions are crucial not only in aiding developers to optimize their systems for a given application but also in enabling different researchers to compare and contrast their approaches. Therefore, in this thesis, we first conduct a analysis of the state-of-the-art, revealing that the evaluation process needs improvement. Then, since the predictions with more than zero dimension, can be correct and incorrect at the same time, it necessitates further in-depth study. As interpreting different metrics on real data is a challenging task [Pan15], we propose a Multimodal Evaluation Metric (MME) that is adaptable for use in various applications and easily visualized and interpreted.

The MME approach refines well-known True Positive (TP), False Positive (FP), and False Negative (FN) for MIS by permitting fractional values for each target instead of binary values, accounting for partially correct predictions. It also allows a more comprehensive assessment of the segmentation method's performance. Using the updated TP, FP, and FN values, we can compute commonly employed metrics like Intersection over Union (IoU), Recall or True Positive Rate (TPR), Precision (PRC), and Dice Similarity coefficient (DC), which are easily interpretable, even for non-experts [Tat+18]. Advancing beyond prior research relaying on point-based relations, this work examines the spatial or temporal interdependencies of pixels (voxels), covering both 1D, 2D and 3D relations. To elaborate further, this metric evaluates the identification of individual segment spots by a single prediction instead of numerous fragmented ones (uniformity property), the detection of each segment (detection property), the alignment of ground truth and prediction boundaries based on their shape (boundary alignment property), and quantifies the relative and total volume of accurately predicted segments. This enables evaluating the quality of extracted targets that have more than 0D which helps experts to select a proper approach based on their requirements.

The rest of this thesis is organized as follows.

Chapter 2 presents a review of the context and foundational knowledge essential for the depth and breadth of this thesis. In addition, the applications and the research problems are introduced.

In Chapters 3 to 5, we provide our contribution followed up by a comprehensive review of existing literature on segmentation and evaluation. After identifying the gaps in the current knowledge and positioning the current research within the context of prior work, we propose our proposal to resolve the issues. In Chapter 3, we propose an adaptive layer to be added on top of the best-performing segmentation deep network. The experiments presented in this chapter show the possibility of improving the performance by dynamically changing the receptive field. In Chapter 4, we extend the finding for a more complex and heterogeneous application in AR. The proposed meta-decomposition approach is introduced in this chapter.

In chapter 5, we first deeply explore the state-of-the-art and the evaluation process in a problem in a real-world scenario, such as the well-known COVID-19 pandemic. Then, this chapter introduces a novel approach for evaluating targets that extend beyond the 0D, a common scenario in segmentation problems as explained in chapters 3 and 4. This approach projects the evaluation into high-level properties and is explored in this chapter in-depth, highlighting its framework, benefits, and potential applications.

Finally, we end with a conclusion. This concluding chapter offers a recap of the research problem, methods employed, and the primary findings. It also emphasizes the overarching significance and contributions of this study to the field. Then, suggestions for improving the current state of segmentation techniques and evaluation processes are provided. Furthermore, potential areas for future research and exploration are identified.

Reading the chapters sequentially helps to fully grasp the concepts introduced and their implications.

Chapter 2

Context of the thesis

2.1 Chapter Overview

This chapter embarks on the foundational knowledge essential for the depth and breadth of this thesis. Three elements are explored: Applications, Challenges, and Preliminaries. Beginning with applications, the focus is narrowed down to three specific domains: MIS, AR, and SED. Each of these applications, though diverse, shares the characteristic of targets spanning beyond 0D, a shared property that warrants their collective study. Transitioning to challenges, we introduce deeper the challenges in the segmentation and the evaluation. Then, in the preliminaries, we succinctly elucidate the fundamental concepts and terminologies, setting a stage for readers to delve deeper into the intricate nuances of the subject. This chapter aims to equip readers with a holistic understanding of the topic, bridging the abstract with the empirical and the foundational with the advanced.

2.2 Targets Beyond 0D and their applications

Commonly, in machine learning tasks, the target we are seeking from a given input is a single numerical value or a point devoid of any extra layers of depth [DAr20]. For instance, consider the usual classifiers used in image processing. They determine the categories an image belongs to, or when we predict a people's age based on their educational background. Diverging from this, our focus in the thesis is on targets that transcend the zero-dimension aspect. For example, in AR and SED scenarios, in addition to the category, the targets encompass dimensions like duration, which is a 1D target. Similarly, in the context of MIS, targets extend into dimensions like area or volume. The distinction between traditional targets and 1D targets is highlighted in Figure 1.1. In contrast to classical targets that are either correct or incorrect, the targets beyond 0D can be correct and incorrect at the same.

Throughout this thesis, we will approach the problem of evaluation and segmentation using concrete applications and areas namely AR, SED, and MIS. The shared property in all these applications is that all targets (concepts) have more than zero dimensions. In AR, the targets are activity events. Each activity event has an activity class and duration which is considered a 1D concept. In SED, we are trying to locate sound events in an audio track. Each sound event has also sound class and duration which is also considered a 1D concept. In MIS, the goal is to partition the entire image into a set of regions [Aza+22]. Each region has a class and area in the case of 2D images or volume in the case of 3D images. The reason for these diverse applications is their impact and their shared properties. In the following, we provide a concise description and motivation for each application.

2.2.1 Medical Image Segmentation (MIS)

As a trending subject in the field of image processing and computer vision [Asg+21], MIS involves extracting the boundaries of desired targets, such as tumors, in medical images and determining their class [Luo+22]. The accurate segmentation of medical images, such as Computed Tomography (CT), Positron-Emission Tomography (PET), and Magnetic Resonance Imaging (MRI), plays a vital role in the diagnosis and treatment of various diseases and assists physicians in patient management, including staging, assessment, and prognosis of the treatment response [Li+19b; Liu+21; Tia+21]. Nonetheless, manual segmentation of medical images, especially those with irregular geometric shapes, is a high-effort task that requires considerable expertise and attention. Therefore, automatic segmentation methods interest for years to alleviate this burden on physicians and improve the accuracy and reliability of medical diagnoses [Liu+21]. It presents a formidable challenge, as the boundaries of tumors or nodules and their surrounding tissues may not be clearly distinguishable due to the influence of adhesions, subjectivity, and other complex conditions that may obscure or confound the identification of relevant features [Tia+21].

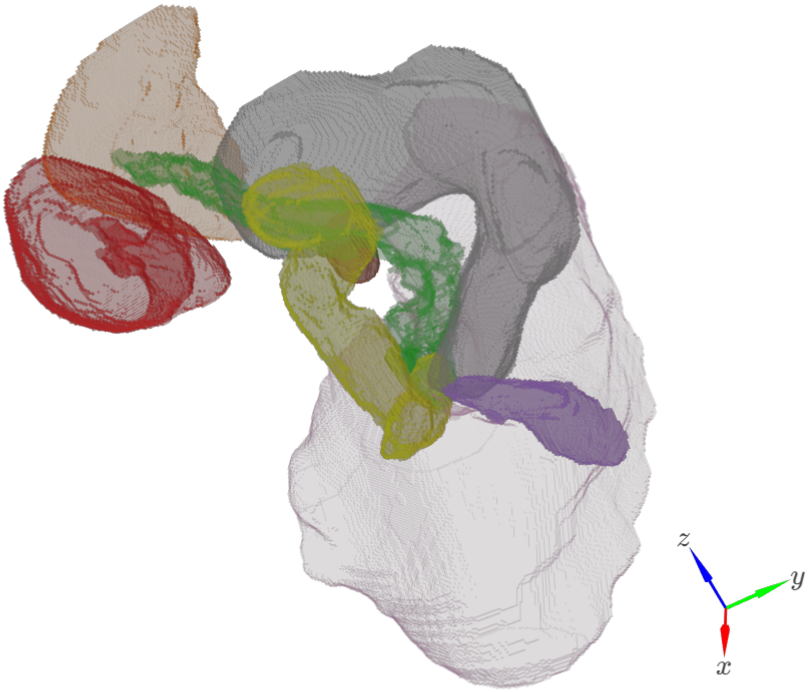
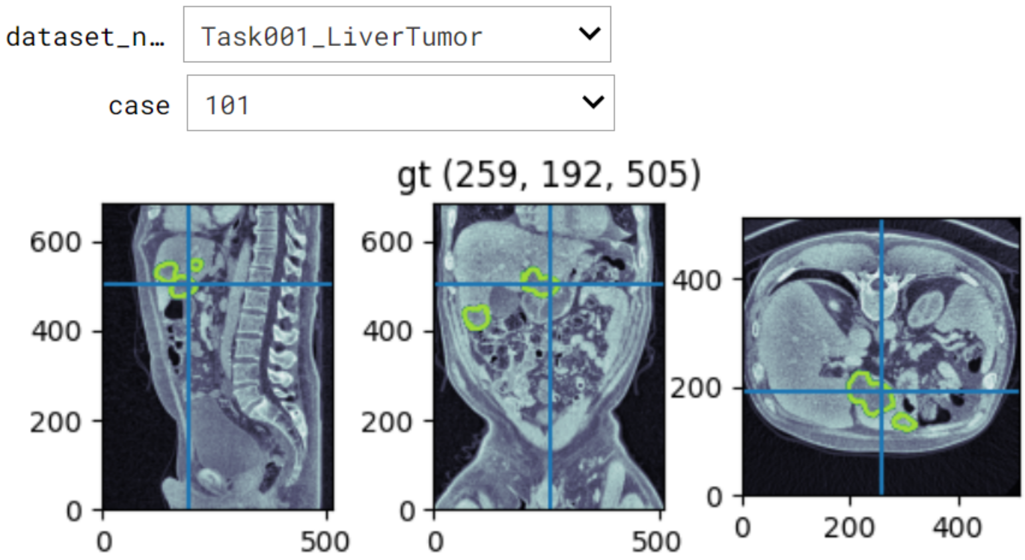


Figure 2.1: An example segmentation result in Synapse Multi Organ Dataset is represented in 3D. Different colors represent different organs.

Deep learning-based MIS has gained considerable traction in recent years [Che+23; Hou+21; Dev+21; Asg+21; Mal+22; Luo+22; KHS22]. A myriad of models has been introduced in the literature for various MIS tasks and clinical outcomes, encompassing multi-organ detection, tissue mass detection, tumor or nodule segmentation and classification, cell counting, multiple diagnoses, prognosis, and the prediction of treatment outcomes for various chronic diseases like cancers or neurodegenerative diseases [Kum+22; Dev+21; Roy+23; Sim+19; Ant+22; Rot+16; Son+22]. In the context of cancer diagnosis, for instance, deep models such as [Mel+22; Eal+22; IQ22; Ise+21; SA22] have shown improved



dataset_n… Task001_LiverTumor

case 101

Figure 2.2: Liver Tumor CT scan Orthogonal View. The tumor spot is shown in green. The vertical and horizintal lines show the slice of the CT scan.

performance in segmenting tumors or nodules. Despite the significant progress, implementing deep learning in MIS continues to pose challenges since medical images often contain noise, artifacts, adhesions, and other distortions that can negatively affect the performance of deep learning models, particularly when discerning tumoral tissue boundaries and surroundings [Kum+22; Mal+22; Tia+21]. MIS is recognized as an NP-hard problem [Asa+01] requiring heuristics for resolution. This makes the performance metrics crucial for assisting clinicians and system designers in choosing the appropriate models for the clinical problem [THT14]. While numerous studies have demonstrated that these models exhibit robust predictive capabilities, achieving results close to those of clinicians [Mal+22], recent studies highlight the existence of statistical biases in the assessment method used to evaluate these models due to the used metrics [MSK22; Rei+21].

Examples of two problems in MIS are shown in Figures 2.1 and 2.2. In the first figure, the goal is to distinguish different organs in a CT scan, while in the second figure, the goal is to recognize a tumor spot in a CT scan of the liver.

2.2.2 Activity Recognition (AR)

We are moving towards the Internet of Things (IoT), and the number of deployed sensors is rapidly increasing [Per+14]. IoT generates a long and heterogeneous series of data [Hu+19]. Recognition of human activities (AR) from these sensor data is expected to be the heart of myriad IoT applications such as healthcare, smart homes, and security [Per+14; QPM18; Che+21b]. AR is crucial to society because it allows computer devices to monitor, analyze, and improve human daily life by recognizing their behaviors [Che+21b]. However, it is challenging due to the complexity of human behaviors and their variety from person to person [BNE21]. In sensor-based human activity recognition, there are two primary sensor deployment strategies. The first embeds fixed sensors within environments, while the second uses sensors that people wear or carry. Ambient sensing may use computer vision, which has roots in public security and surveillance but raises privacy concerns in private spaces like homes. Therefore, the use of environmental sensors is preferred for lesser privacy issues. Acoustic sensing and capturing sound and speech, has also gained traction, which especially beneficial for remote health monitoring. On the mobile front, smartphones with myriad functions and sensors have been used for activity recognition, expanding their scope outdoors but with challenges due to device limitations. Additionally, the surge in wearable technology has spurred a range of health applications monitoring various physiological signs, leading to increased research in this area. Some studies even merge both technologies for more comprehensive insights [Ale15]. Without going into details, an example of a smart home and human activities is shown in Figures 2.3 and 2.4.

2.2.3 Sound Event Detection (SED)

SED has garnered significant attention in recent years due to its wide-ranging potential applications and implications in diverse fields. A SED system is designed to discern and recognize specific auditory events (category and duration) in an audio track. For example, these systems can pinpoint distinct sounds like gunshots, glass breaking, or a baby's cry

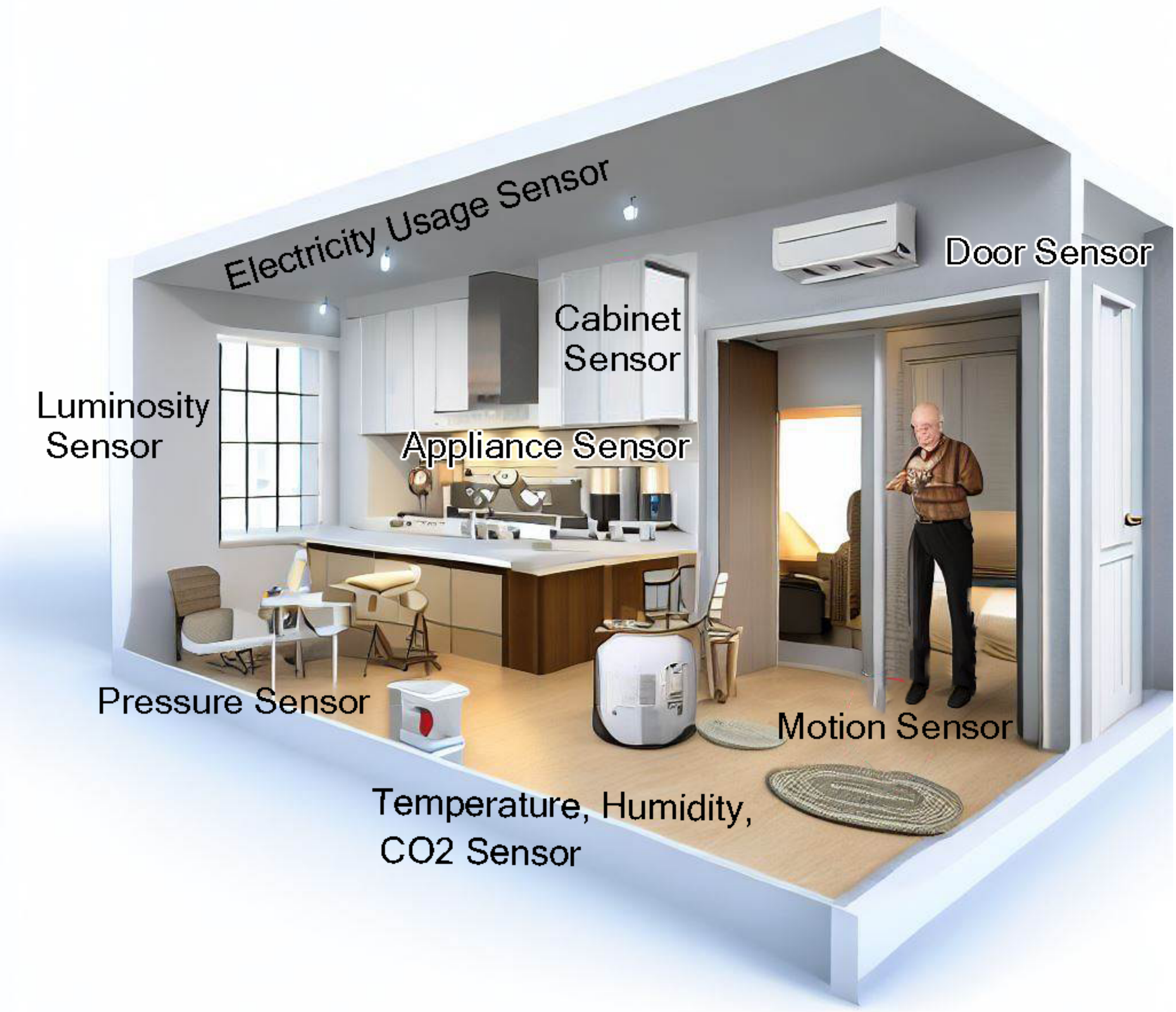


Figure 2.3: A demonstration of a smart home equipped with environmental sensors.The picture is crafted using the DallE system and Microsoft Bing.

[SHV20]. With increasing urbanization and the advent of smart technologies, such systems have found utility in a myriad of situations ranging from security to healthcare and audio content-based searching. Consider healthcare where specific sounds like a patient's coughing or breathing patterns can be monitored and analyzed remotely. Similarly, in smart homes, such systems can be used to detect events like glass breaking, signaling a potential break-in, or a baby's cry, alerting caregivers to the baby's needs [Bil+20; MHV16; CC20].

However, the inherently unpredictable nature of sound events poses a significant challenge to the accuracy and reliability of SED systems. Auditory events arise from a myriad of sources and processes. These processes often occur at nearly random intervals, making

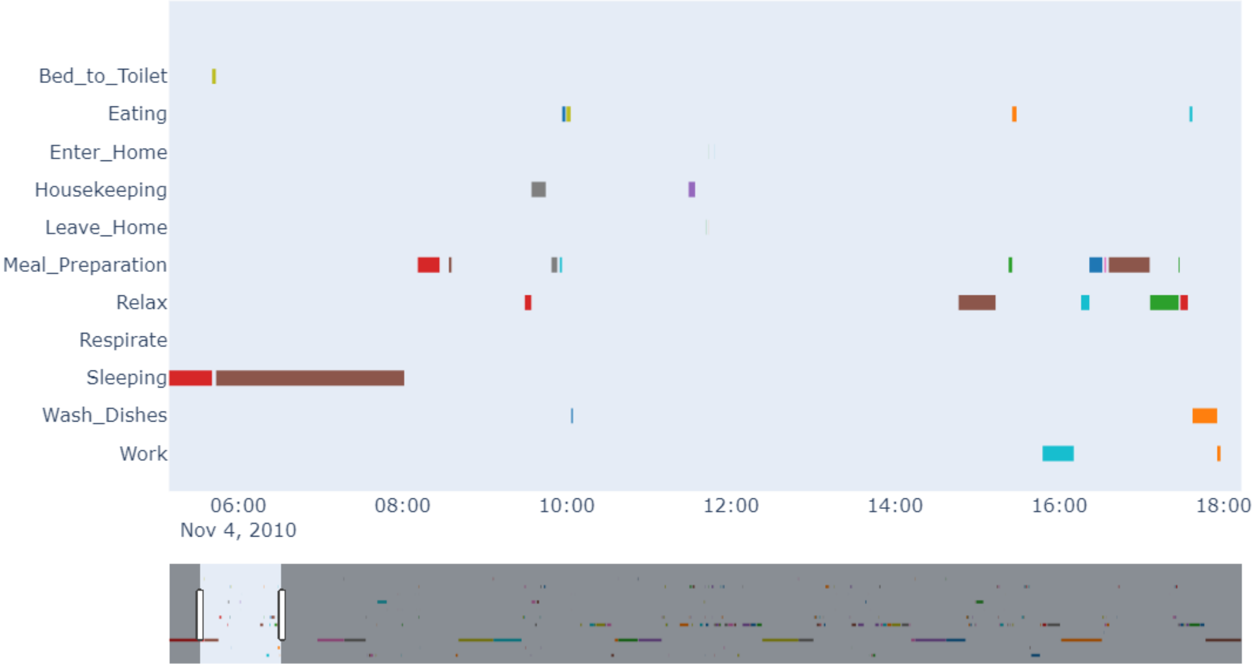


Figure 2.4: Example of Daily Activities from the Aruba Dataset. The pattern of one week is represented by a compact bar at the image's bottom, with the chosen day highlighted for clarity. The y-axis enumerates the different classes of activities, while the horizontal lines mark the length of each activity's occurrence.

the consistent identification of sound events a complex task [Bil+20]. This randomness can sometimes lead to false positives or negatives, where the system might mistakenly identify a sound or fail to recognize an event.

Furthermore, the variability in environmental acoustics, background noises, and overlapping sounds add layers of complexity to the problem. An understanding of these challenges is essential for the development of robust and efficient SED systems. To provide a visual representation of what sound events look like when represented, refer to Figure 2.8.

2.3 Segmentation Problem Challenges

Segmentation is a concept widely used across diverse domains, including marketing [Dol20; EBK21], data transmission [BBB19], audio processing [VMM22; McC19], image analysis [WWZ20], time series [Fu11], biology [Vic+19], genetics [ANJ+22], healthcare [Fer+20], text and document processing [Bar+20], and activity analysis [Li+19a]. In the Oxford Dic-

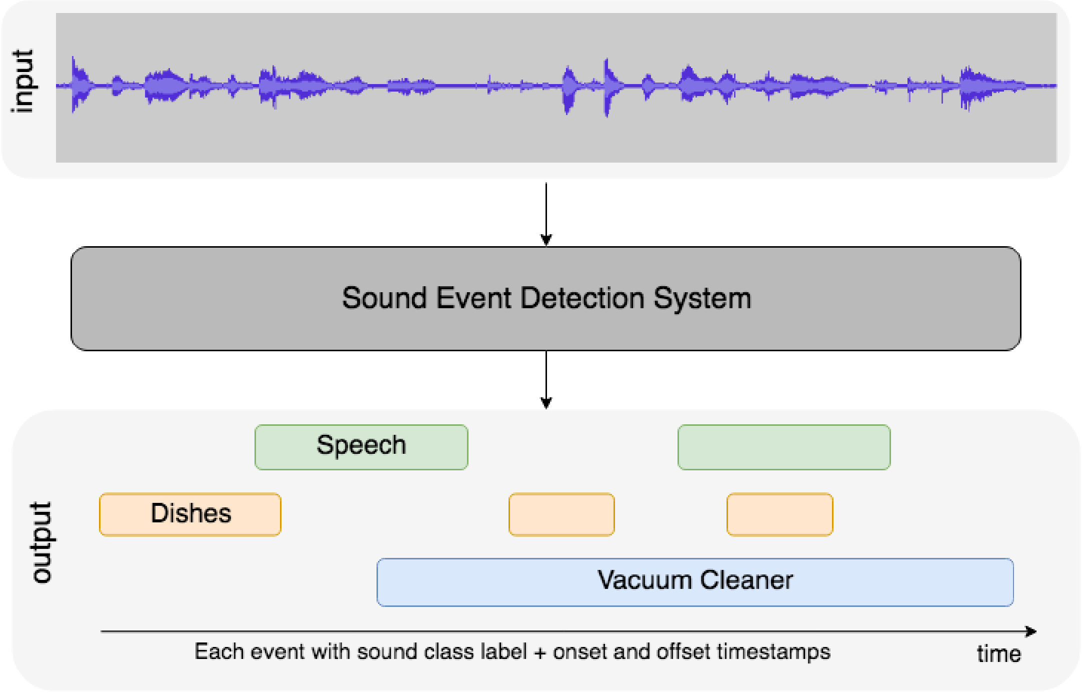

input
Sound Event Detection System
Speech
Dishes
Vacuum Cleaner
output
Each event with sound class label + onset and offset timestamps
time

Figure 2.5: Sound Event Detection System [Mes+18]. The system receives an audio track and outputs the duration and categories of sound events.

tionary, a segment is defined as "a part of something that is separate from the other parts or can be considered separately," and segmentation is "the act of dividing something into different parts." Segmentation is a method of dividing a big (maybe infinite) and complex structure into a set of smaller meaningful finite pieces that may have similar properties and help reveal underlying mechanisms and foster recognition [Ber+18; LMS14]. In the area of image processing and computer vision, segmentation refers to the process of dividing a digital image into multiple segments (sets of pixels) to simplify or change the representation of an image into something more meaningful and more straightforward to analyze [Luo+22; Asg+21; Aza+22]. In AR, segmentation is employed in order to overcome the limitation of a single sample, thereby providing adequate information regarding an activity [Che+21b; NGC15; DTP21]. Using these smaller segments helps to have more straightforward modeling and not to process unrelated information. Therefore, it helps to obtain a more accurate and appropriate model in less time. Segmentation methods are also used as a pre-processing step to reduce the number of data points in the original data [WGS16].

Segmentation helps to deal with the complexity of problems, although it directly impacts the recognition quality [NGC15]. On the one hand, inadequate information in one segment may lead to poorly detection of the targets; on the other hand, if a segment contains too much information, extra complexity may be added for future data processing [KC14; NGC15]. Consequently, a trade-off exists between the sufficiency of information in each segment, minimizing the number of segments, and reducing the processing complexity of each one to discover the expected concepts. Segmentation can also be expressed as a discretization process [Fu11]. In time series, segmentation is also defined by finding meaningful segments corresponding to the state changes in the underlying process [SEL21]. Few examples for segmentation are shown in Figure 2.6. For instance, to recognize the text content from an image, dividing the image into several similar repeated segments, such as characters, facilitates the recognition process. There are 36 different alphanumeric characters to focus on, making it less complex than that of trying to recognize the entire chunk of text [RNN99; Chr21]. Another common example in speech recognition is finding boundaries between words, syllables, and phonemes to help better recognition [GGY10]. Segmentation in images is a common approach to removing unnecessary information [Asg+21]. In Figure 2.6, two examples of segmentation in the image are shown that select the important area and estimate the pose to extract key points of a human. Sometimes, certain limitations make segmentation not just useful but essential for the case that not all data can be accessed at the required time.

Even though the segmentation process helps to deal with the complexity of problems, it alters some characteristics of the input data. It introduces at least two families of uncontrollable biases. The first family of biases is introduced to the model due to the changes in problem space made by the segmentation itself. For instance, in AR, a common approach consists of segmenting the data and feeding them to the model to identify the activity in each segment. It is often assumed that the classifier performance follows the whole system performance [Per+14; Bil+20; QPM21; CN15; BNE21]. This hard hypothesis may misleadingly present convenient results because different segmentation algorithms generate dissimilar segments. For example, a time window that generates equal-length windows cannot be

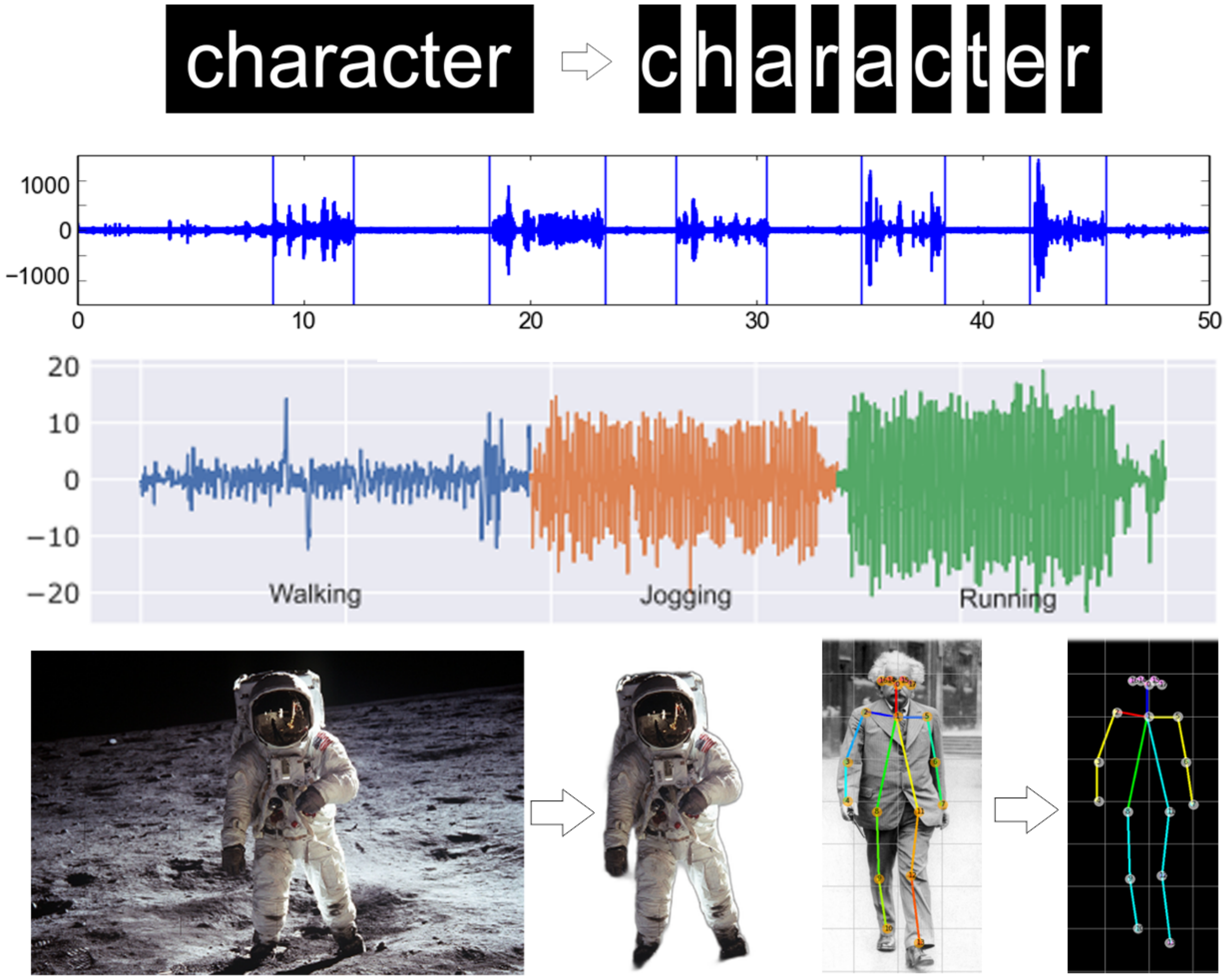


Figure 2.6: Segmentation: Examples from Image, Voice, and Activity Recognition

compared to a dynamic windowing method. In Section 4.6.4, this will be explained in detail. Additionally, activities have some properties related to their duration [Mod+22b]. For example, the fixed event window approach presents an irrelevant number of segments to their duration [Ale15; Mod+22b]. As an additional example, steady recognition of the sleeping activity is critical; otherwise, it may misleadingly present a disorder [Ale15; Mod+22b]. However, segmentation may break these properties.

The other bias of segmentation is concerned with the segmentation approach itself, including the fixation of the segmentation method and its parameters. The segmentation approach is often implicitly incorporated with the prior knowledge or assumptions originated from the developers, researchers, or experts during the selection and tuning of the segmentation approach. It is a crucial step since the accurate recognition of concepts (particularly complex ones) highly depends on the quality of the segmentation method [Che+21b; NGC15]. For instance, by including a bias to have smaller segments than the required segment size for learning a specific activity, the machine learning process may not properly identify that activity. Furthermore, it is unrealistic to assume that individual activities (concepts) remain unchanged for a long time; for example, the daily activities in winter are different from those in summer [Che+21b]. Therefore, an appropriate segmentation approach in one period may not be efficient for another period. While there has been a rise in the use of AutoML techniques to automate algorithm and hyperparameter selection [Mu+22; Tay+18], they do not dynamically change the algorithm over time. In addition, it is essential to include the segmentation process inside the machine learning pipeline in order to evaluate the global quality of the recognition system [Mod+22b].

2.4 Evaluation Problem and Challenges

Segmentation is recognized as an NP-hard problem [Asa+01] requiring heuristics for resolution. This makes the performance metrics crucial for assisting clinicians and system designers in choosing the appropriate models for the clinical problem [THT14]. While numerous studies have demonstrated that these models exhibit robust predictive capabilities and achieving results close to those of clinicians [Mal+22], recent studies highlight the existence of statistical biases in the assessment method used to evaluate these models due to the used metrics [MSK22; Rei+21].

Despite the existence of numerous evaluation metrics in the literature, there are concerns regarding these metrics. Limited understanding and interpretability of these metrics may result in significant bias when selecting a suitable segmentation method for a particular application [Nai+21]. Common evaluation metrics, such as TPR, PRC, DC or F1, IoU, also called Jaccard Index, Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD), and Normalized Surface Distance (NSD), are widely used to assess the

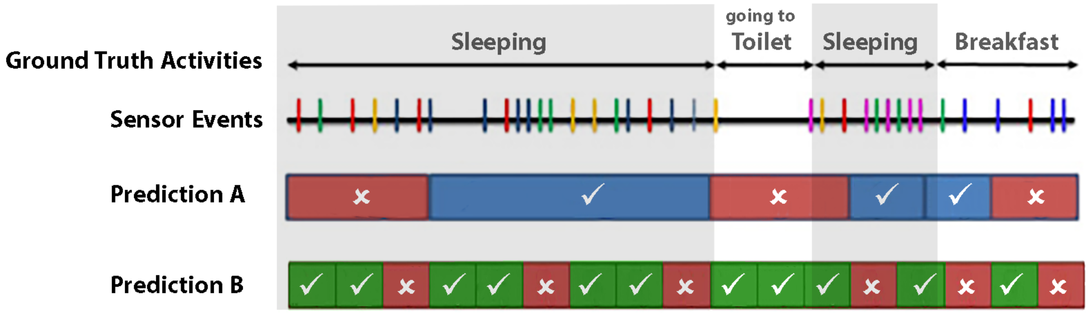


Figure 2.7: Examples of human activities. The first line shows the different activities and their duration. The second line shows the sensor events where the colors of vertical lines show the occurrence of different sensors. In this figure, two prediction systems are depicted; the checkmark shows the correct prediction, while the cross shows the incorrect one. While both systems predict correctly similar duration of sleeping activity, the second prediction presents the existence of a disorder in sleeping activity while the first prediction has an error at the beginning of sleeping activity.

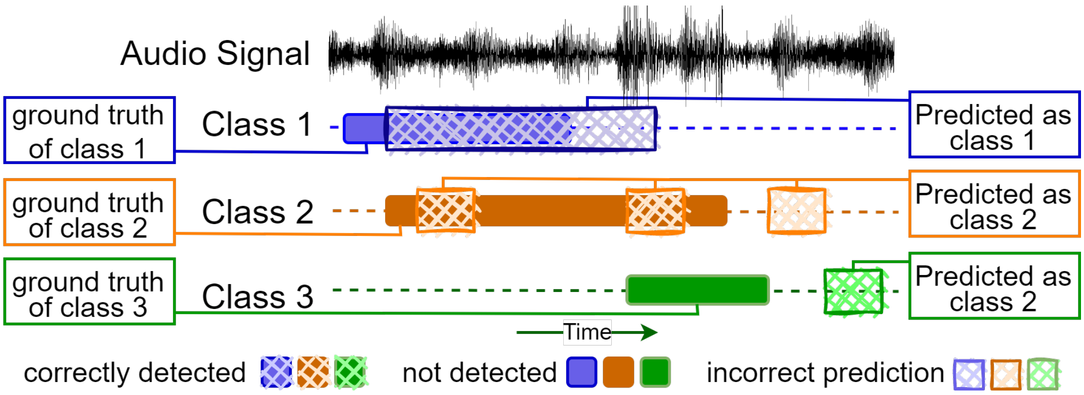


Figure 2.8: Examples of sound events. SED instances are partially correct and partially incorrect, while the classical ones are either correct or not. How should we consider different types of events in the evaluation? For instance, is the importance of the duration the same for gunshot events and dog breaking? How should we provide interpretable information for the users to help them choose an appropriate segmentation approach for a specific task or phase?

performance of MIS, AR, and SED systems [Che+23; Hou+21; Eel+20; Kum+22; MSK22; Dev+21; Che+22a; Ren+22; TC22; Tar+21; Abd+23; Ma+21; Luo+22; Ker+21; TH15; Mal+22; Son+22; WZ20; KHS22]. Evaluating these systems often involves simplifying the multidimensional concepts to point-based ones [Kum+22; MSK22; Ma+21]. However, voxels (pixels, times) are interdependent, and neglecting these dependencies can lead to incomplete assessments of proposed models and their outcomes (particularly the clinical ones) [Kim+15]. For instance, early identification of tumoral regions necessitates detecting

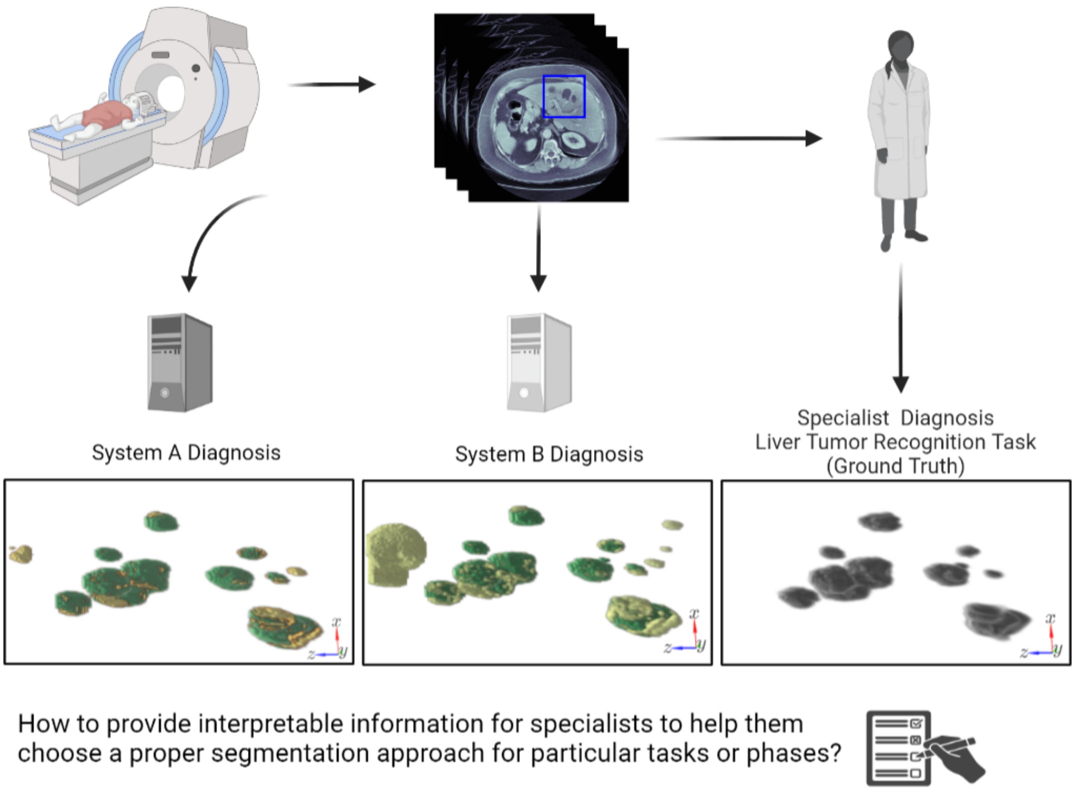

System A Diagnosis
System B Diagnosis
Specialist Diagnosis
Liver Tumor Recognition Task
(Ground Truth)
How to provide interpretable information for specialists to help them
choose a proper segmentation approach for particular tasks or phases?

Figure 2.9: Example of MIS and its evaluation problem. Providing interpretable information helps with the selection of an appropriate segmentation approach depending on the specific goal. For instance, during initial tumor detection, even marginal tumor identification is critical, while in the treatment response assessment, the changes in the volume are important.

tumor presence regardless of size, while treatment response evaluations require monitoring volume changes [Tia+21]. Additionally, the presence of a dominant spot of lesions (activities or sound events) of the same type might result in overlooking smaller lesions (activities or sound events) during assessment [TH15; Kim+15]. The quality assessment of these systems extends beyond these factors; evaluations should also consider the uniformity or fragmentation of predictions and the preservation of segment shapes, which assist experts and models in identifying tumor types [Tia+21]. Furthermore, predicted segments can be partially correct and partially incorrect simultaneously, unlike point-based predictions that are either entirely correct or incorrect. For example, medical treatment outcomes can vary significantly even if two tumor segments have similar measures of the aforementioned metrics, such as DC, and HD [Kim+15]. Very recent studies highlight the need for reliable model performance assessments, as well as the presence of statistical biases in the assessment of both binary and multi-class problems [MSK22; Rei+21; Nai+21; WWZ20; Kum+17; Kim+15; TH15; GSC22; Hoe+22; Rei+22; Koe+22; Jav+22; Lee+22; Fag+22; BJ22]. In conclusion, a more appropriate way to evaluate the performance of MIS techniques involves considering their various aspects. To provide a clearer illustration of the evaluation in SED, AR, and MIS, we have depicted these evaluations using self-descriptive images in Figures 2.7 to 2.9.

2.5 Preliminaries on Evaluation (Point-based)

The point-based evaluation considers each concept as an individual instance; therefore, a prediction is either correct or incorrect. Assuming y_i is a vector representing the i-th target in the ground truth and $\hat{y}_i$ is the prediction of the system for i-th example. They are vectors of real values in regression and vectors of the one-hot-encoded labels in the classification problem. In this thesis, our concentration is on the classification problem.

Confusion Matrix Let us consider $\{0 : Negative(Background), 1 : Positive(Foreground)\}$ as the classes (labels) for binary classification and $\mathrm{L} = \{0, 1, ..., k\}$ in multi-class cases with k interesting classes plus a background (null or 0) class. The Confusion Matrix (CM) is commonly used to evaluate the performance of a classification system. It provides an overview of how the model operates by presenting a summary of the number of TP, True Negative (TN), FP, and FN predictions made by the system [Pan15; Rei+21]. These terms are defined in the following: [True Positive] TP refers to the number of correctly predicted positive instances [Pan15]. In multi-class cases, we calculate TP for each class c as the sum of instances where the ground truth label y_i is labeled c, and the system's prediction $\hat{y}_i$ matches the class c. It is formulated in Equation (2.1) using the Hadamard matrix product (elementwise product) (∘) notations and one hot encoded representation.

$$\mathrm{TP} = \sum_i y_i \circ \hat{y}_i \tag{2.1}$$

[True Negative] TN is the number of correctly predicted negative instances in binary classification [Pan15]. In multi-class cases, for each class c, the TN is calculated as the sum of instances where both the ground truth label y_i and the system's prediction $\hat{y}_i$ are not equal to c. This can be represented mathematically in Equation (2.2).

$$\mathrm{TN} = \sum_i (1 - y_i) \circ (1 - \hat{y}_i) \tag{2.2}$$

[False Positive] FP are negative instances that have been incorrectly identified as positive by the classification system [Pan15]. In multi-class cases, we calculate FP for each class c as the sum of instances where the ground truth label y_i is not labeled c, but the system's prediction $\hat{y}_i$ is labeled as c. Using the $\circ$ notations and one hot encoded representation, we can formulate it as shown in Equation (2.3).

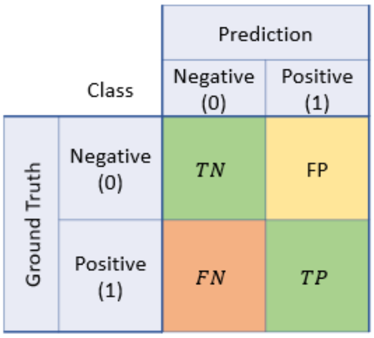
$$\mathrm{FP} = \sum_i (1 - y_i) \circ \hat{y}_i \tag{2.3}$$

[False Negative] FN refers to the number of positive instances that are incorrectly predicted as negative [Pan15]. In multi-class cases, we calculate FN for each class c as the sum of instances where the ground truth label y_i is labeled c, but the system's prediction $\hat{y}_i$ does not match the class c. It is formulated in Equation (2.4).

$$\mathrm{FN} = \sum_i y_i (1 - \circ \hat{y}_i) \tag{2.4}$$

Therefore, the confusion matrix is a 2×2 matrix in the binary case, as shown in Figure 2.10, where the top left element represents TN, the top right element represents FP, the bottom left element represents FN, and the bottom right element represents TP.

In multi-class cases, TP, FN, FP, and TN are vectors, where TP_c, FN_c, FP_c, and TN_c represent the corresponding value for class c. Therefore, we can generate a confusion matrix of size $(k+1) \times (k+1)$, to visualize the classification performance. Each element of this matrix m_{ij} represents the count of ground truth instances of label i that the system predicted as label j. Clearly, m_{ii} denotes accurate predictions, while all other elements signify inaccurate predictions. In this context, $\mathrm{TP}_c = m_{cc}$ represents the number of correct predictions for class c, $\mathrm{FP}_c = \sum_{j=1}^{k} m_{jc} - \mathrm{TP}_c$ signifies the number of predictions incorrectly labeled as class c, and $\mathrm{FN}_c = \sum_{j=1}^{k} m_{cj} - \mathrm{TP}_c$ represents the number of ground truths of class c that were predicted as other classes. Counting all predictions that do not belong to class c and were not predicted as class c, denoted as TN_c, represents the remainder of the predictions. These concepts are visually demonstrated in Figure 2.11.



		Prediction	
	Class	Negative (0)	Positive (1)
Ground Truth	Negative (0)	*TN*	FP
	Positive (1)	*FN*	*TP*

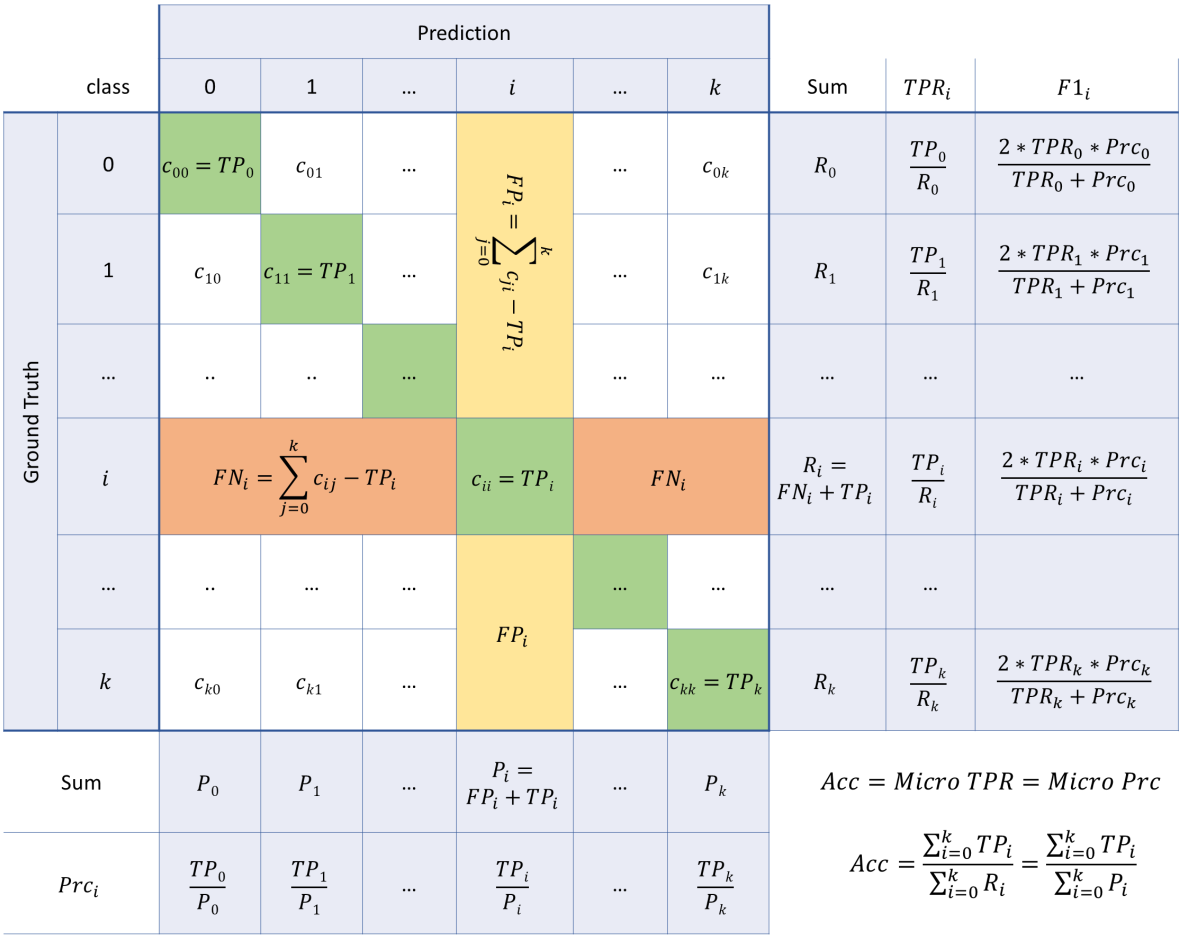
Figure 2.10: Binary Confusion Matrix

Accuracy Acc is a metric that quantifies the ratio of correct predictions over the total number of predictions [Pan15]. In the binary classification scenario, Acc is calculated using Equation (2.5). In the multi-class classification scenario, when the number of predictions equals to the number of ground truth, Acc is equivalent to the micro-average of the TPR or PRC that is represented in Equation (2.6).

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2.5}$$

$$\mathrm{Acc} = \frac{\sum_{c=1}^{k} \mathrm{TP}_c}{\sum_{c=1}^{k} \mathrm{TP}_c + \mathrm{FN}_c} = \frac{\sum_{c=1}^{k} \mathrm{TP}_c}{\sum_{c=1}^{k} \mathrm{TP}_c + \mathrm{FP}_c} \tag{2.6}$$

Precision The Precision (PRC) is another performance metric commonly used in classification tasks. It measures the rate of correctly recognized positive instances among all predicted instances as positive [Pan15]. It is particularly used in scenarios where the cost of false positives is high. In such cases, we want to minimize the number of false positives, i.e., instances that are predicted as positive but are actually negative. Precision helps us achieve this by measuring the proportion of true positives among all predicted positives. A



		Prediction								
	class	0	1	...	i	...	k	Sum	TPR_i	$F1_i$
Ground Truth	0	$c_{00} = TP_0$	c_{01}	...	$FP_i = \sum_{j=0}^{k} c_{ji} - TP_i$	...	c_{0k}	R_0	$\frac{TP_0}{R_0}$	$\frac{2 * TPR_0 * Prc_0}{TPR_0 + Prc_0}$
	1	c_{10}	$c_{11} = TP_1$	...		...	c_{1k}	R_1	$\frac{TP_1}{R_1}$	$\frac{2 * TPR_1 * Prc_1}{TPR_1 + Prc_1}$
	...	..	..	...		...	...	...	...	...
	i	$FN_i = \sum_{j=0}^{k} c_{ij} - TP_i$			$c_{ii} = TP_i$	FN_i		$R_i =$ $FN_i + TP_i$	$\frac{TP_i}{R_i}$	$\frac{2 * TPR_i * Prc_i}{TPR_i + Prc_i}$
	...	..	...	...	FP_i	...	...	...	...	
	k	c_{k0}	c_{k1}	...		...	$c_{kk} = TP_k$	R_k	$\frac{TP_k}{R_k}$	$\frac{2 * TPR_k * Prc_k}{TPR_k + Prc_k}$
Sum		P_0	P_1	...	$P_i =$ $FP_i + TP_i$	...	P_k	$Acc = Micro\ TPR = Micro\ Prc$		
Prc_i		$\frac{TP_0}{P_0}$	$\frac{TP_1}{P_1}$	...	$\frac{TP_i}{P_i}$	...	$\frac{TP_k}{P_k}$	$Acc = \frac{\sum_{i=0}^{k} TP_i}{\sum_{i=0}^{k} R_i} = \frac{\sum_{i=0}^{k} TP_i}{\sum_{i=0}^{k} P_i}$		

Figure 2.11: Multi-Class Confusion Matrix including TPR, PRC, F1 and Accuracy (Acc)

high precision value indicates that the model is correctly identifying positive instances, and there are relatively few false positives. The mathematical definition of PRC using Hadamard matrix division (elementwise division) ($\oslash$) notation is presented in Equation (2.7).

$$\mathrm{PRC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.7}$$

Recall or True Positive Rate Recall or True Positive Rate (TPR), also known as sensitivity, measures the proportion of true positive instances that are correctly identified by the model among all the positive instances in the ground truth [Pan15]. In other words, TPR quantifies the ability of a classification model to correctly identify positive instances and is an important metric for evaluating the model's performance in scenarios where the cost of false negatives is high. The mathematical definition of TPR is shown in Equation (2.8).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.8}$$

Fβ measure To gain a comprehensive understanding of the performance of a classification model, it is common practice to evaluate both TPR and PRC metrics together [Pan15]. It can show better the strengths and weaknesses of the model and make informed decisions about how to optimize its performance. Fβ measure (Fβ) is the weighted harmonic means of TPR and PRC [Pan15]. Its special case F1 when $\beta = 1$ also called DC is commonly used in situations where both precision and recall are equally important.

$$\mathrm{F\beta} = \frac{(1 + \beta^2) \cdot \mathrm{TPR} \circ \mathrm{PRC}}{\beta^2 \cdot \mathrm{PRC} + \mathrm{TPR}} \tag{2.9}$$

$$\mathrm{F1} = 2\frac{\mathrm{TPR} \circ \mathrm{PRC}}{\mathrm{PRC} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} = \frac{2\mathrm{TP}}{\mathrm{R} + \mathrm{U}} = 2\frac{\sum_i y_i \circ \hat{y}_i}{\sum_i y_i + \hat{y}_i} \tag{2.10}$$

Intersection over Union IoU or Jaccard metric is a performance metric commonly used in image segmentation where the main objective is to evaluate the similarity between the predicted and ground truth segments [TH15]. It is calculated by dividing the number of TP by the sum of TP, FN, and FP [Asg+21; Luo+22; LN22] and is formulated in Equation (A.2). Commonly, Mean Intersection over Union (MIoU) is referred to the macro average of IoU in each class, or Frequency weighted intersection over union (FWIoU) is their weighted average where the weights are their frequencies [Zhe+21; Luo+22].

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{DC}}{2 - \mathrm{DC}} \tag{2.11}$$

2.6 Conclusion

This chapter explored the fundamental concepts necessary for the comprehensive understanding of this thesis and the practical applications in AR, SED, and MIS. It also provided a short overview of the segmentation and evaluation challenges and described some fundamental and basic concepts.

Chapter 3

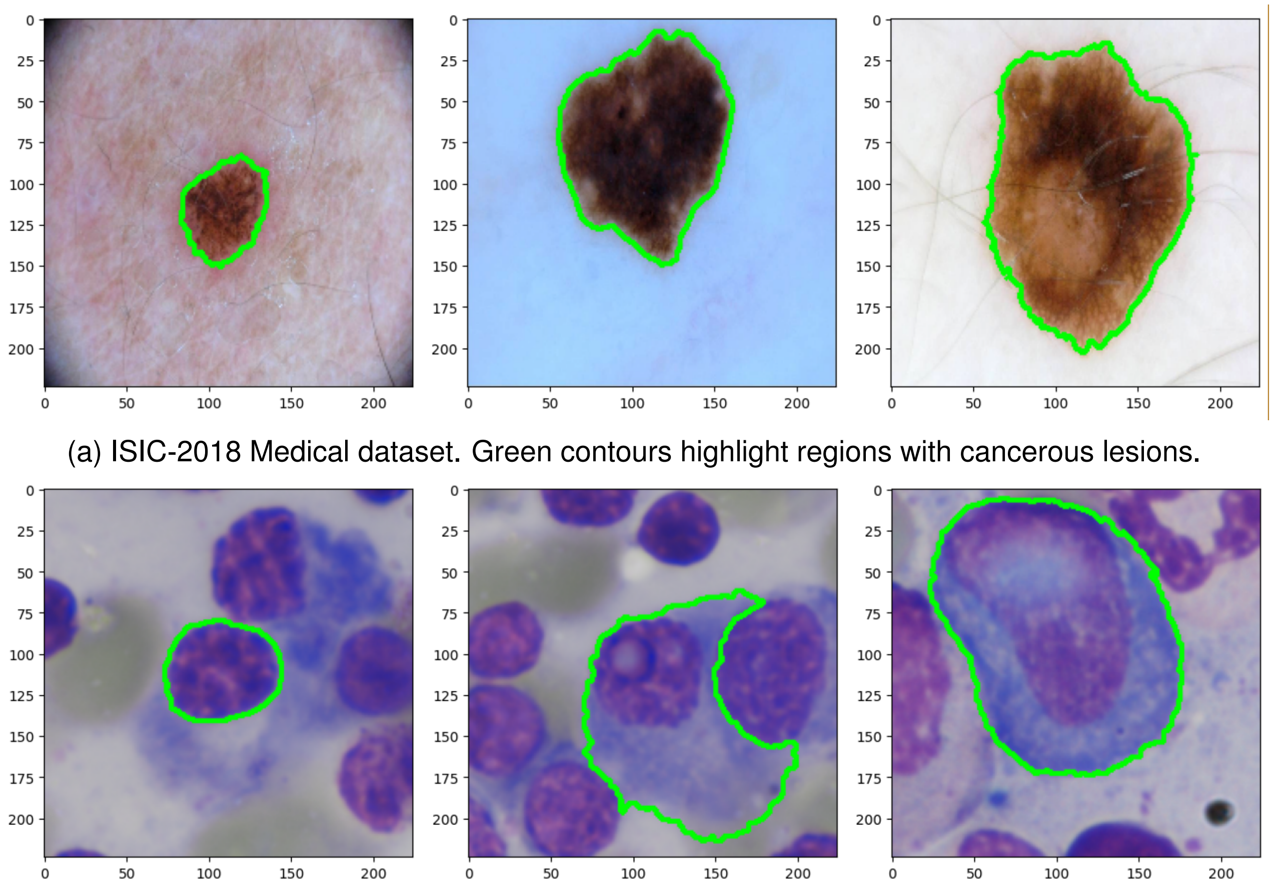
Dynamic Receptive Field

3.1 Chapter Overview

In the context of image segmentation, the goal is to partitioning an image into meaningful regions or segments. convolutional neural network (CNN) models are widely used for analyzing and processing medical images. The fundamental block of CNN is the convolution layer, which contains tiny adjustable weight grids (known as kernels) that convolved on the input image. This operation involves moving the kernel over the image, and at every stop, it performs a mathematical operation (dot product) using the weights of filters and the pixel values of the image to get a single output. This process provides Local Receptive Fields for every pixel, assisting the model in concentrating on nearby characteristics. As we go deeper, the model starts recognizing more intricate and broader patterns, enhancing its understanding of the image [Liu+21].

Typically, kernels of a predetermined size are employed, often configured as a 3x3 grid. This can be likened to a windowing technique where the kernel moves over the image in small segments, processing each segment to generate a corresponding output in a new feature map. This process involves the kernel striding across the image, analyzing one small region at a time. A number of researchers highlight that the size of these receptive fields plays a crucial role in enhancing the efficiency of the models [IR20; MNA16;

Wan+21c; Qiu+18]. Consequently, sticking to a kernel of a fixed size might not work well for every image. For instance, larger input images or targets might require bigger receptive fields. As demonstrated in Figure 3.1, having varying scales in different images suggests the usefulness of having diverse receptive fields to accommodate all of them [IR20].



(a) ISIC-2018 Medical dataset. Green contours highlight regions with cancerous lesions.

(b) SegPC-2021 Medical dataset. Green contours highlight cytoplasm and nucleus regions.

Figure 3.1: Illustration of Different Scales in Medical Images. As visible, differences in the scale of the images needs to be taken into account.

Before we delve into the topic of meta-decomposition in the upcoming chapter, we sought to understand the recent progress in the Computer Vision domain concerning the dynamic choice of receptive fields. Then, at the end of this chapter, we introduce a novel method that incorporates a single dynamic receptive field layer into the best state-of-the-art models in MIS, which enhances their performance.

3.2 Related Works

3.2.1 Related Works on Image Segmentation

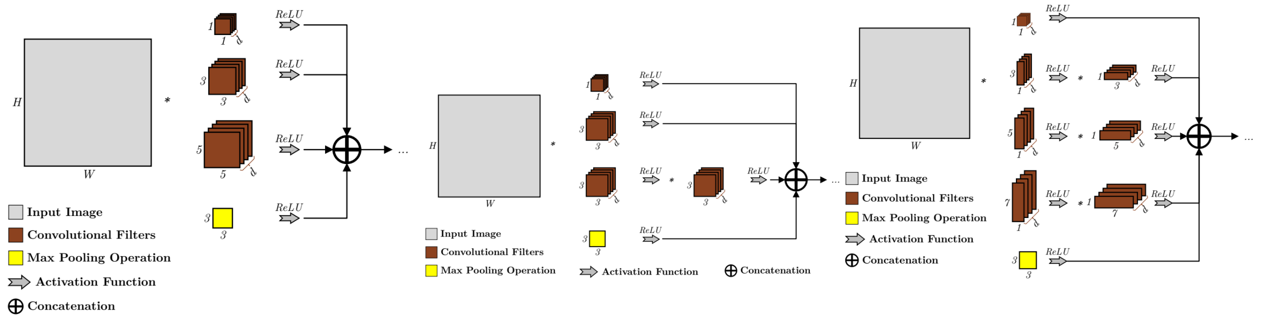
In the field of computer vision, image segmentation often employs the technique of pixel-wise dense prediction. It involves assigning a label to each pixel in an image, where the aim is to classify every pixel into predetermined classes. The group of similarly labeled pixels forms a segment [YK16; SG16]. Deep learning-based MIS has gained considerable traction in recent years [Che+23; Hou+21; Dev+21; Asg+21; Mal+22; Luo+22; KHS22]. The well-known U-Net model proposed by Ronneberger et al. [RFB15] gained significant attention and is an influential architecture in the field of deep learning. Similar to the Autoencoder models, the U-Net model contains the encoder (contracting) and the decoder (expanding) paths. The unique feature of U-Net is the incorporation of skip connections that enable the flow of information from the encoder to the decoder at different scales, facilitating the preservation of spatial details and improving the localization accuracy of the segmentation results [RFB15; Aza+22]. Numerous extensions of U-Net have been proposed to improve recognition quality in medical tasks. Azad et al. [Aza+22] provide a comprehensive survey on U-Net and categorize the U-Net extensions into *Skip Connection Enhancements*, *Backbone Design Enhancements*, *Bottleneck Enhancements*, *Transformers*, *Rich Representation Enhancements*, and *Probabilistic Design*. The MISSFormer model [Hua+23] redesigns the U-Net architecture by incorporating a position-free and hierarchical U-shaped transformer. It utilizes the Enhanced multi-scale Transformer module to bridge the gap between the encoder and decoder feature maps. It has a slightly higher performance in the Synapse dataset [Aza+22].

The recent UCTransNet [Wan+22b] proposes replacing simple skip connections in U-Net with a multi-scale channel-wise module to solve the semantic gaps for an accurate MIS. It also includes an attention mechanism and transformer sub-module. The attention mechanism implicitly learns to suppress irrelevant regions while emphasizing the regions of interest and the transformer aids in capturing long-range dependencies and addresses the limitation of local receptive fields. The transformer sub-module tokenizes feature maps in each stage within the appropriate patch sizes.

The MissFormer [Hua+23] and UCTransNet [Wan+22b] performs better comparison to other approaches such as TransUNet [Che+21a], Residual U-Net [ZLW18], MultiResUNet [IR20], U-Net++ [Zho+18], Att-UNet [Okt+18] and original U-Net [RFB15] over ISIC 2018, SegPC 2021 and Synapse datasets [Aza+22].

In addition to altering skip connections, transformers, and attention mechanisms, alternative backbones are commonly used to improve U-Net performance. ResNet [Cic+16] is also a common backbone for the U-Net architecture which addresses the issues of stacking many layers in deep neural networks that causes vanishing gradient problem. The Google inception module, widely utilized for extracting features across multiple scales, was initially introduced in InceptionV1, where kernels of different sizes were concatenated in parallel [Sze+15; Sze+16; Sze+17; Zha+21]. This architecture underwent further refinements in subsequent versions, with InceptionV2 replacing the 5x5 convolution with two stacked 3x3 convolutions and InceptionV4 breaking down square convolutional kernels into two vectors to reduce computational operations while increasing the receptive field [Sze+16; Sze+17; ALB23]. However, this led to a limitation where larger kernels could not be broken down, resulting in fewer selectable features. MultiRes blocks, which employ a series of convolutional layers with residual connections, have been utilized to provide features at different scales, although with limited efficacy for small images and fuzzy objects [IR20; Hos+23; LGL21]. To overcome these limitations, the dual-channel UNet (DC-UNet) was proposed to incorporate more different-scale features at the cost of increased network parameters and GPU memory consumption [LGL21; Ans+22]. Gridach [Gri21], Jiang et al. [Jia+19], Lou et al. [Lou+22], Yang et al. [Yan+20], Fu et al. [FLH23], Wang et al. [Wan+21b], and Zhan et al. [Zha+23] consider fixed numbers of features for each multi-scale dilated atrous receptive field in parallel to not increase the computational complexity, which is an alternative to representing information at various scales. This strategy increases the receptive field of the layer without adding additional network parameters. A convolutional filter in a CNN can be decomposed as a linear combination of pre-fixed bases particularly Fourier-Bessel bases [Qiu+18]. Wang et al. [Wan+21c] combine the concept of adaptive convolutional kernel and combination of pre-fixed bases by replacing all convolution filters with adaptive atoms. This approach are slightly shown better in image classification tasks particularly in handling intra-image variance. However, it is not yet applied for image segmentation tasks. Moreover, this method is more efficient when the number of channels is relatively high (e.g., 256), which contrasts with the scenario in the first layer where, for example, there are only 3 RGB channels. In the next subsection, we study in more detail the works considering multiple receptive fields.

3.2.2 Related Works on Multi-scale Receptive Field



(a) Inception V1 module [Sze+15; ALB23]

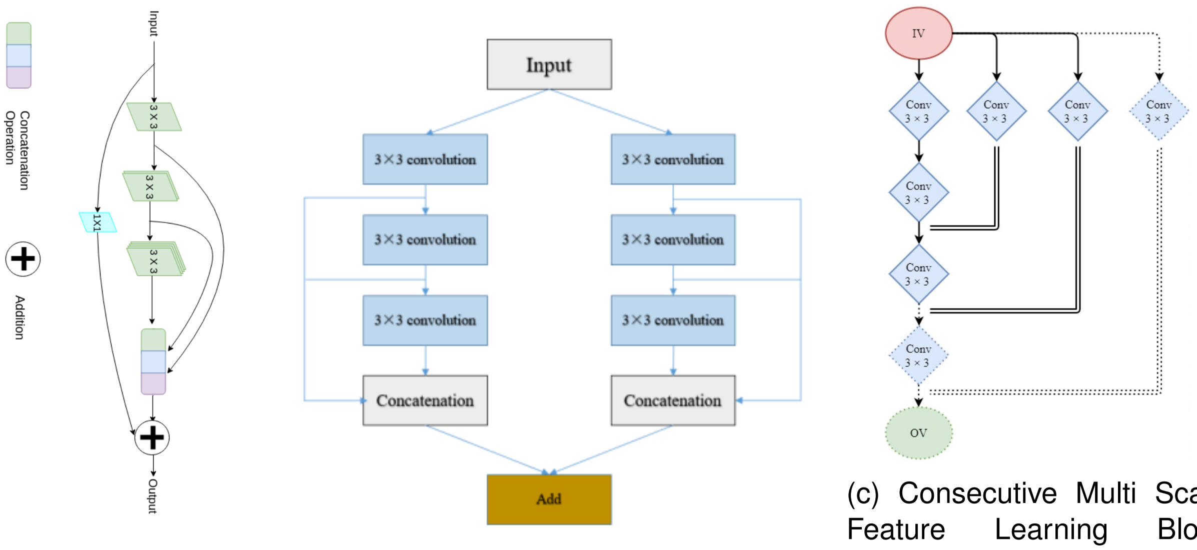
(b) Inception V2 module [Sze+16; ALB23]

(c) Inception V4 module [Sze+17; ALB23]

Figure 3.2: The Evolution of Inception Modules. (a) concatenates kernels with size 1, 3, and 5, (b) Replace intensive 5x5 convolution with two stacked 3x3 convolutions, (c) Replaces square convolutional kernels with two vectors to expand the receptive field while reducing the computational operations.

The size of receptive fields is a pivotal factor in improving the model efficiency, as highlighted in several studies [IR20; MNA16; Wan+21c; Qiu+18]. The items studied in MIS are not all the same and can be oddly shaped [IR20]. In CNN, the traditional approach utilizes shared filters across all samples and pixels, maintaining the translation equivariant property - a fundamental feature that ensures stability in recognizing patterns irrespective of their position in the input space. However, to address the challenges of sample scalability, a strategy of dynamically selecting the filter size for each image has been introduced. While this method aims to provide diverse features of different samples, it increases the computational complexity due to adding extra parameters and not respecting the translation equivariant property of CNN. Moreover, these strategies cannot be extended to efficiently manage intra-image variances and are computationally considered infeasible [Wan+21c], addressed in numerous studies in image classification tasks [Che+20; Wan+21c; Ser+07; Sze+15; ALB23]. In addition, due to specific challenges posed by irregularly shaped items in MIS, various strategies also have been devised. In [Ser+07], static Gabor filters of different sizes are used, while [Sze+15] proposes a new inception architecture that applies convolutional layers of varying kernel sizes in parallel and based on the success of even low dimensional embeddings, they also add one 1x1 convolution to reduce the computational requirements (Figure 3.2a). The Google inception module, widely utilized for extracting features across multiple scales, was initially introduced in InceptionV1 (Figure 3.2a), where kernels of different sizes were concatenated in parallel [Sze+15; Sze+16; Sze+17; Zha+21]. This architecture underwent further refinements in subsequent versions, with InceptionV2 (Figure 3.2b) replacing the 5x5 convolution with two stacked 3x3 convolutions and InceptionV4 (Figure 3.2c) breaking down square convolutional kernels into two vectors to reduce computational operations while increasing the receptive field [Sze+16; Sze+17; ALB23]. However, this led to a limitation where larger kernels could not be broken down, resulting in fewer selectable features.

MultiRes blocks (Figure 3.3a), which employ a series of convolutional layers with residual connections, have been utilized to provide features at different scales, although with limited efficacy for small images and fuzzy objects [IR20; Hos+23; LGL21]. To overcome these limitations, the dual-channel UNet (DC-UNet) was proposed to incorporate more different-scale features (Figure 3.3b) at the cost of increased network parameters and GPU memory consumption [LGL21; Ans+22]. The Adaptive Convolutions with Dynamic Atoms (ACDA) dynamically generate a kernel for each pixel based on its surroundings, which is implemented by a fast two-layer network to optimize the receptive field without escalating computational complexity [Qiu+18; Wan+21c]. Kaur et al. [KKS21] stacked several 3x3



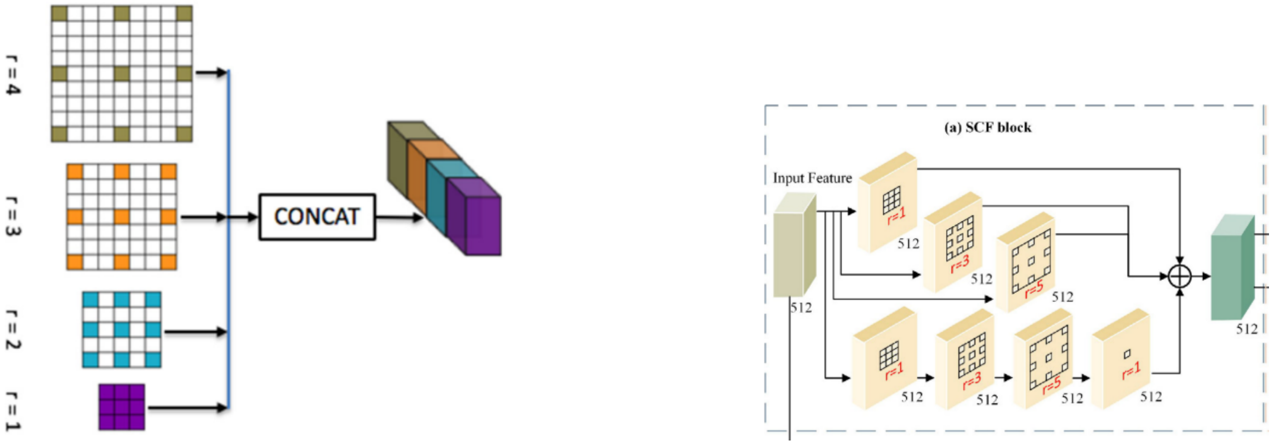
(a) MultiRes Block [IR20]

(b) Dual Channel Block [LGL21]

(c) Consecutive Multi Scale Feature Learning Block (CMSFL) [Oli+23]

Figure 3.3: MultiRes Block (right) and its dual version (center). It uses a chain of convolutional layers with residual connections. Lou et al. [LGL21] shows that dual MultiRes block provide better results. (c) presents the newer extension of MultiRes block, which uses a convolution layer instead of residual connection.

convolutions to replace 7x7 convolutions. In [Zha+21], authors use consecutive multiple 3 x 3 convolution kernels followed up with the 1 × 1 convolution layer, which replaces the 5 × 5 or 7 × 7 convolution kernels. It also adds a shortcut connection to improve the diversity of feature learning and the robustness of CNN. Olimov et al. [Oli+23] introduce consecutive multiscale feature learning blocks (Figure 3.3c) that require fewer 3x3 blocks to increase the receptive fields. Wu et al. [Wu+21] combine multi-scale convolution modules and residual connections in their Multi-scale Residual Block (MSRB) to improve the feature extraction capability. They also use chains for 3x3 kernels to avoid the computational complexity of 5x5 convolution. Multi receptive field [Liu+20] consider multiple paths of convolution, decoder, and encoder for different dilation rate in parallel and then concatenate the extracted features (Figure 3.2). The Selective Kernel module [Li+19c] consists of Split, Fuse, and Select operators. It first generates multiple paths with various kernel sizes (split). Then, it aggregates the information from multiple paths to obtain a global representation (fuse). Finally, it selects the feature maps from those paths [Byr+20]. In [Kha+22], authors proposed four convolution input layers in parallel with various kernel size.Bao et al. [BZL23] proposed using large multi-scale kernels of size 4x4, 8x8, 16x16, and 32x32 in the first layer with lower channels for larger kernels to control the total cost and to capture the information of different scales. Several works concentrate on multi-scale inputs and layers, such as [Liu+23; Yan+20].



(a) Pyramid Dilated Network Block [Gri21]

(b) Spatial Context Fusion (SCF) block [Wan+22d]

Figure 3.4: Dilated Network Blocks. Instead of using intensive kernels, theyconsider fixed numbers of features for each multi-scale dilated atrous receptive field to not increase the computational complexity.

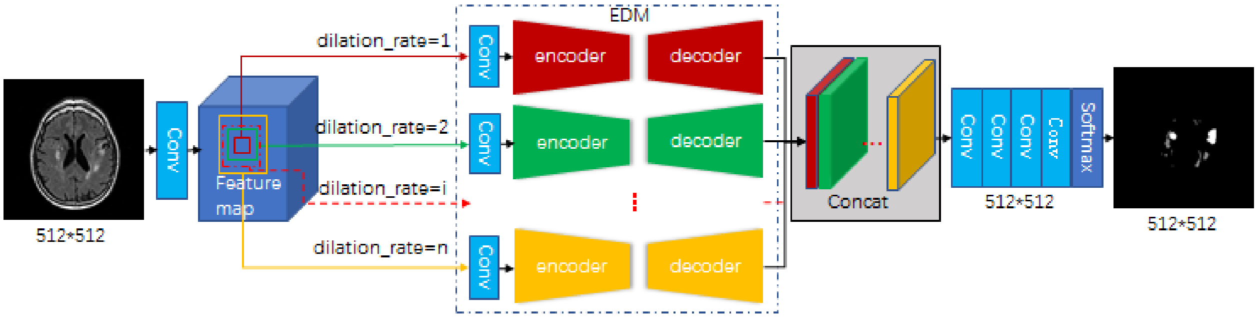


Figure 3.5: Multi Receptive Field [Liu+20]. It considers multiple independent paths of convolution, decoder, and encoder for different dilation rate in parallel and then concatenate the extracted features to obtain a multi-scale representation.

Gridach [Gri21], Jiang et al. [Jia+19], Lou et al. [Lou+22], Yang et al. [Yan+20], Fu et al. [FLH23], Wang et al. [Wan+21b], Zhan et al. [Zha+23], and Gao et al. [Gao+21] considered fixed numbers of features for each multi-scale atrous receptive field in parallel to prevent the increasing of computational complexity (Figure 3.4a). In [Gri21], authors used multiple kernel size in parallel in each layer. In [Wan+22d], authors proposed a spatial context fusion (SCF) block to address the limitation of using fixed-scale convolutional operations in the inner layers (Figure 3.4b). Yang et al. [Yan+21], Mahmud et al. [MPF21], Gao et al. [Gao+21], and Ibtehaz et al. [IR20] combine MultiRes and Res2Net block with atrous receptive fields [Gri21] to decrease the number of parameters and extracts the features of multi-scale receptive fields.

In summary, recently many approaches are proposed to improve segmentation performance that are mainly concentrated on the improving skip connections, including attention mechanism and transformer and improving the backbone such as including dynamic convolution. The UCTransNet [Wan+22b] performs better compared to other approaches such as TransUNet [Che+21a], Residual U-Net [ZLW18], MultiResUNet [IR20], U-Net++ [Zho+18], Att-UNet [Okt+18] and original U-Net [RFB15] over several datasets such as ISIC 2018, and SegPC 2021 [Aza+22].

Although employing a dynamic receptive field offers theoretical advantages, it is still challenging to implement in practice. On one side, certain approach necessitate extensive computational resources and memory to calculate the dynamic receptive field; on the other, they may not yield any enhancement in performance. After a comprehensive study of various adaptive methods, the rest of this chapter will present a novel dynamic receptive field layer that can be placed ahead of leading models. It improves the recognition performance while maintaining a similar number of parameters.

3.3 Adaptive UCTransNet

In this section, we present an adaptive convolution layer to be incorporated into leading models. Though it is not exclusive to UCTransNet, we showcase its integration at the top of this architecture, naming it AdaptUCTransNet. The AdaptUCTransNet dynamically adjusts the kernel size based on the local context of the input image, enabling the network to capture relevant features at multiple scales from the first layer, leading to improved segmentation accuracy, robustness, and better consider diverse anatomical structures, irregular shapes, and varying feature scales in medical images. Additionally, leveraging the benefits of the UCTransNet architecture, which integrates U-Net and transformer models, AdaptUCTransNet combines spatial information with self-attention mechanisms to extract meaningful features from medical images.

In MIS, the presence of diverse segments necessitates the adaption of a flexible approach. By adjusting the kernel size based on the specific context of each pixel, we can have a higher level of discernment when extracting features within the inner layers of deep networks. By liberating the network from strict reliance on the hyperparameter associated with the dataset for determining the kernel size, we empower it to adapt and optimize its performance according to the unique characteristics of the input data. Therefore, the proposed approach is to add the adaptive multi-size kernel convolution layer to the best deep learning model for MIS. We leverage the fact that convolution layers can be mathematically expressed as a linear combination of predetermined bases [Qiu+18] inspired from [Wan+21c]. By employing a limited number of Fourier-Bessel (FB) bases, we substantially reduce the number of parameters. Notably, this reduction in parameters does not compromise the accuracy of image classification tasks [Qiu+18; Wan+21a], while it has not yet been explored for MIS. Additionally, using Fourier-Bessel bases improve the recognition of the structural information of the input image and effectively mitigates the impact of high-frequency noise and addresses the computational complexities associated with employing multiple kernel sizes within the convolution layer.

Formally, the symbol $[p]$ is employed to denote the spatial coordinates on a feature map Z. These coordinates are applicable to a range of dimensional structures, extending from 1D to more complex, higher-dimensional forms. The notation $\mathcal{N}^{\delta}_{Z[p]}$ is used to represent the receptive field surrounding the feature vector $Z[p]$, where the distance is within a δ. The size of this receptive field can vary, ranging from being quite small to encompassing the entirety of the input feature. Considering this receptive field, the $\mathcal{T}$ is used to represent transformed inner receptive field. For instance, in the dilated receptive field, $\mathcal{T}$ selects a subset of input features at specified intervals. Conversely, in the case of a traditional convolution layer,

$\mathcal{T}$ encompasses all pixels, incorporating them without any alterations. Then, traditionally, this inner receptive field is convolved with a set of kernels, denoted as K. The convolution operation can be mathematically expressed as $Z'[p] = K * \mathcal{T}(\mathcal{N}^{\delta}_{Z[p]})$, where Z' denotes the resultant feature map post-convolution, and $*$ operation sums the element-wise product. i.e., this can be expressed as $K * N = \sum_q (K \circ N)_q$, where the $\circ$ signifies Hadamard element-wise multiplication and the sum is taken over all elements of the product. These kernels are learned end-to-end by the network during the training process, typically via backpropagation and gradient descent. The inclusion of these kernels results in the addition of extra parameters to the network. As expected, using larger kernels further increases the total parameter count. Therefore, it will be more challenging to dynamically select the most suitable kernel size for each spatial coordinate.

Based on the finding in [Qiu+18], a convolution kernel can be decomposed as a combination of Fourier-Bessel bases. Therefore, instead of using a learnable kernel, we can learn the weights (W) for the pre-fixed Fourier-Bessel bases with different sizes (FS). Therefore, $Z'[p] = FS \times W * \mathcal{T}(\mathcal{N}^{\delta}_{Z[p]})$. For adaptively changing the receptive field, we use another inner network to learn these weights (W) based on the receptive field $\mathcal{N}^{\delta}_{Z[p]}$. This inner network, called the coefficient generator network, is trained end-to-end with the backpropagation and gradient descent. In the process of selecting appropriate weights for kernels of varying sizes, we stack several layers of smaller kernels to control the complexity, as suggested by Simonyan et al. [SZ15]. This approach ensures that the output comprehensively covers the entire receptive field. For example, by stacking a minimum of four layers of 3x3 kernels, we can achieve 9x9 receptive field. This multi layer network convolved through the receptive field $\mathcal{N}^{\delta}_{Z[p]}$ with a smaller kernel size, and the output of this network is the local weights $(W(\mathcal{N}^{\delta}_{Z[p]}))$ for the pre-fixed Fourier-Bessel bases with different sizes (FS). This inner network remains fixed across all receptive fields. Consequently, this approach maintains the translation invariance characteristic of convolutional networks while taking into account the local context.

Building upon the specific attributes of medical images, we present our framework to en-

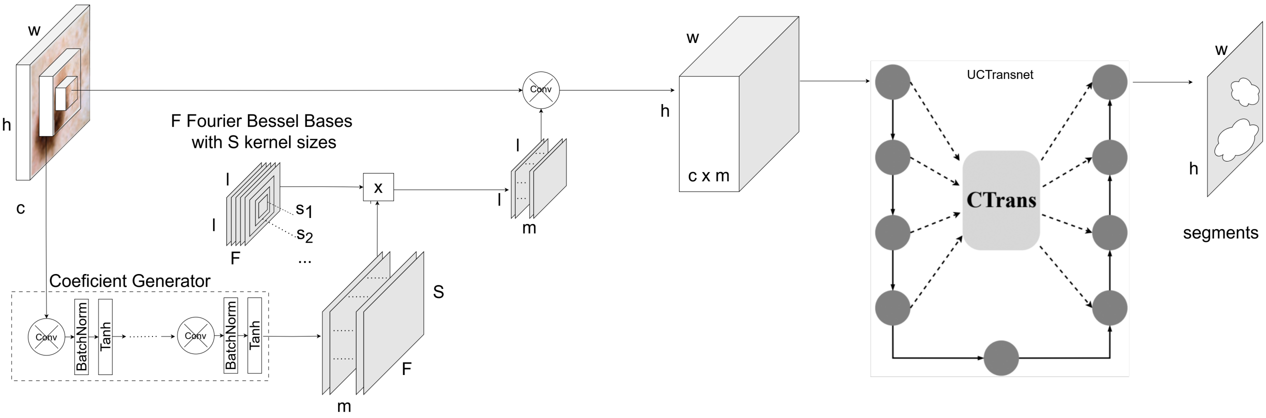


Figure 3.6: Adaptive Convolution Layer added to the leading UCTransNet architecture. The coefficient generator network generates the weights for Fourier-Bessel bases with different sizes for each pixel and channel. It results a fixed kernel to be convolved for that pixel.

hance the accuracy of the best segmentation deep network. The inclusion of our adaptive layer does lead to an increase in the number of parameters, but this increase constitutes only a minor fraction of the entire parameters of the network. A graphical representation of the network architecture is provided in Figure 3.6. Given set F of Fourier-Bessel bases with $|S|$ different sizes denotaed as FS, for every pixel and channel another receptive field (local neighbors of the corresponding pixel which is greater than S) will be convolved in the coefficient generator network using smaller kernels. This network will be trained end to end in the training phase of the whole network. It produces W with size $|F| \times |S| \times m$ weights, where m is the number of intermediate channels. Matrix multiplication of Fourier-Bessel bases (FS) and these weights (W), generates a kernel for the given pixel and channel. The adaptive layer convolves the input image with the explained kernel, resulting in m intermediate features for each pixel and channel. Then, these intermediate features will be fed to the leading segmentation technique UCTransNet [Wan+22b], which replaces the simple skip connections in U-Net with a multi-scale channel-wise module. By using this combination, we can improve the performance of the model.

3.4 Experiment

In the experiments, we conduct an extensive evaluation of the adaptive layer added ahead of UCTransNet, attunet and well-known U-Net architecture using benchmark medical image datasets. By comparing the results with traditional CNNs that employ fixed kernel sizes, we demonstrate the superior performance and generalizability of our adaptive approach. Experiments are conducted on various public testbeds, including the Multiple Myeloma Plasma Cell Segmentation (SegPC) 2021 [Gup+23; Gup+21] and the International Skin Imaging Collaboration (ISIC) 2018 datasets [Cod+19], which will be explained in details in the next sub section. Then, after explaining the details of implementation, we present the experimental results.

3.4.1 Environment Setup

In order to foster transparency and repeatability of our work, All the codes, datasets, and documentation are freely accessible on our GitHub repository: https://github.com/modaresimr/adaptive_mis. All experiments are run on an NVIDIA DGX-1 machine featuring a Tesla V100-32 GPU, Intel Xeon E5-2698v4 CPUs, and 512 GB of RAM. However, we use only a part of these resources.

3.4.2 Datasets

Experiments are conducted on various public testbeds, including the Multiple Myeloma Plasma Cell Segmentation (SegPC) 2021 [Gup+23; Gup+21] and the International Skin Imaging Collaboration (ISIC) 2018 datasets [Cod+19]. SegPC contains a collection of 775 microscopic 2D images from the bone marrow samples of Multiple Myeloma patients. It significantly helped hematologists in making more accurate diagnoses and facilitates cancer screening. The dataset from ISIC 2018 boasts a large collection of 2,594 RGB dermoscopy images. Robust segmentation of these images plays a crucial role in medical diagnosis and is challenging due to inconsistent lighting conditions, varying lesion sizes, texture disparities, and differences in color and positioning. Moreover, the presence of unrelated elements like air bubbles, hair strands, or ruler markers further adds to the complexity [Has+20; Cod+19; Gup+23; Gup+21; Aza+22]. Both datasets are illustrated in Figures 3.9 and 3.10 with the segmentation result of our approach.

Similar to the work by Azad et al. [Aza+22], we allocated 70% of images for training, 10% for validation, and the remaining 20% for testing, and our research focused on the segmentation of Cytoplasm components in SegPC 2021 and segmentation of cancer lesions in ISIS 2018 datasets.

3.4.3 Hyperparameters and Implementation Details

Our pipeline infers the segments from raw image data. All images underwent a sizing down operation to a standard size of 224 x 224 pixels. The pipeline is composed of an adaptive convolution layer with the kernel size of 3, 5, 7, and 9. We also select six Fourier Bessel bases similar to [Wan+21c]. For the coefficient generator network, we have used six intermediate features (m), which, is responsible for encoding the weights of the prefixed bases. This network will be trained end to end during the global training process. We maintain early stopping with a patience of 20 epochs during the training. For better comparison, we make the other hyperparameters similar to those of given in [Aza+22], such as the batch size of 16, epochs limit of 100, Adam optimizer with a learning rate of 0.0001, and the average of cross-entropy loss and dice loss for the loss function. The entire implementation, along with hyperparameters, is accessible and verifiable through our publicly available open-source repository.

3.4.4 Model Complexity

An essential factor in the assessment of models is the computational complexity. The number of trainable parameters of these components is listed in Table 3.1. Therefore, although the training complexity is similar (the differences are less than 2%), its performance is better as shown in Tables 3.2 and 3.3.

Table 3.1: Number of parameters in each model, including the adaptive variant. This indicates that although our method is effective in determining the ideal dynamic kernel size, it keeps the number of parameters almost the same as those of the original model.

Methods	Normal	with Adaptive Layer
U-Net	19.487	19.850
Att-UNet	34.879	35.242
UCTransNet	66.431	66.794

3.4.5 Evaluation Metrics

We used a series of performance metrics for a comprehensive analysis of our model's effectiveness. Acc, DC, and IoU served as the primary metrics. Accuracy gives a general idea of the model's overall performance, which is crucial to be interpreted alongside other metrics due to data imbalance. We utilize the IoU to measure the overlap between the predicted and actual segmentation. DC is used as an alternative to the F1 score due to its increased relevance in medical imaging. It places double emphasis on true positives, which is the harmonic mean of precision and recall. It effectively gauges the spatial overlap of the predictions, which is particularly useful in biomedical image segmentation tasks. Through these varied metrics, we ensured a robust evaluation of our model's segmentation performance.

3.4.6 SegPC 2021 Case Study

SegPC contains a collection of 2D microscopic images from the bone marrow samples of Multiple Myeloma patients. It significantly helped hematologists in making more accurate diagnoses and facilitates cancer screening. Similar to previous studies [Aza+22], we allocated 70% of 775 images for training, 10% for validation, and the remaining 20% for testing, and focused on the segmentation of Cytoplasm components in SegPC 2021.

We have showcased the visual segmentation results from the SegPC 2021 dataset in Figure 3.9. The strength of our adaptive multi-size-kernel representation effectively demonstrates its aptitude to generate accurate segmentation maps for cells of diverse scales and backgrounds. A comparison of results for the SegPC 2021 dataset is detailed in Table 3.2, further highlighting the effectiveness of our methodology. In Figure 3.7, we have plotted the training and validation loss curves for the SegPC 2021 dataset. It shows the model's robust performance, as it is neither underfitting nor overfitting.

Table 3.2: Comparison of results for the SegPC 2021 dataset. Each experiment is repeated five times. We have added our adaptive layer to two leading models (AttUNet, UCTransnet), and traditional UNet, improving not only the performance of these models but also their consistency (as shown by the standard deviation). For a more comprehensive comparison, other deep models, such as missformer, resunet, and multiresunet, are included at the second part of the table. Their models with our adaptive layer are accessible in our repository.

model	Accuracy	Dice	IoU
Adapt_UCTransnet	98.66±0.01	92.11±0.02	91.96±0.02
UCTransnet	98.61±0.04	91.85±0.23	91.71±0.22
Adapt_AttUNet	98.71±0.01	92.41±0.03	92.25±0.03
AttUNet	98.65±0.02	92.10±0.08	91.95±0.08
Adapt_UNet	98.22±0.01	89.58±0.13	89.60±0.11
UNet	98.07±0.05	88.69±0.35	88.80±0.30
missformer	98.35±0.04	90.38±0.16	90.32±0.15
resunet	97.74±0.04	86.70±0.15	87.04±0.14
multiresunet	96.15±0.41	80.29±1.70	81.46±1.41

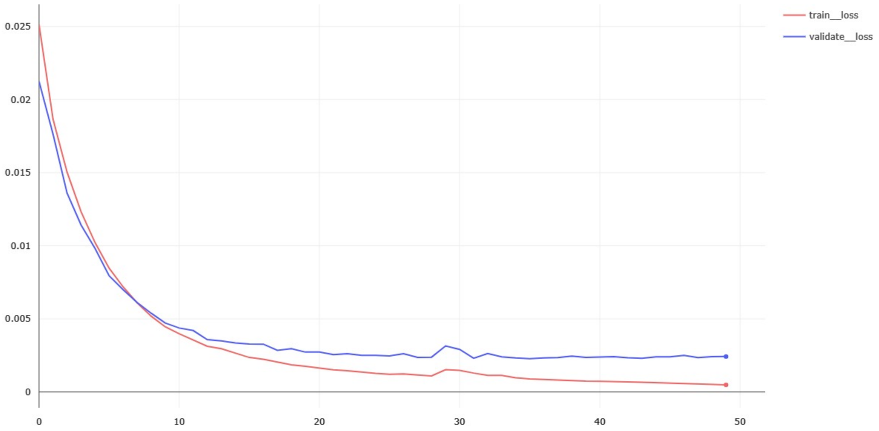


Figure 3.7: The train loss and validation loss of SegPC 2021 dataset. They indicate that the model is neither overfitting nor underfitting.

3.4.7 ISIC 2018 case study

The dataset from ISIC 2018 boasts an extensive collection of 2,594 RGB dermoscopy images. Robust segmentation of these images plays a crucial role in medical diagnosis and is challenging due to inconsistent lighting conditions, varying lesion sizes, texture disparities, and differences in color and positioning. Moreover, the presence of unrelated elements like air bubbles, hair strands, or ruler markers further add to the complexity [Has+20; Cod+19; Gup+23; Gup+21; Aza+22].

Table 3.3: Comparison of results for the ISIC 2018 dataset. It shows the effectiveness of our methodology. Similar to Table 3.2, Each experiment is repeated five times and we have added our adaptive layer to two leading models (AttUNet, UCTransnet), and traditional UNet. This approach improves the performance of these models. For a more comprehensive comparison, other deep models, such as missformer, resunet, and multiresunet, are included at the second part of the table. Their models with our adaptive layer are accessible in our repository.

model	Accuracy	Dice	IoU
Adapt_UCTransnet	95.64±0.13	89.31±0.18	87.68±0.16
UCTransnet	95.54±0.07	89.04±0.27	87.40±0.25
Adapt_AttUNet	95.57±0.16	88.96±0.28	87.36±0.32
AttUNet	95.44±0.15	88.66±0.25	87.04±0.29
Adapt_UNet	94.80±0.21	87.09±0.29	85.41±0.36
UNet	94.43±0.25	86.18±0.39	84.49±0.46
missformer	95.25±0.18	88.38±0.42	86.69±0.43
resunet	94.35±0.09	85.84±0.07	84.19±0.09
multiresunet	92.83±0.63	84.01±1.02	81.80±1.15

Similar to previous studies [Aza+22], we allocated 70% of 775 images for training, 10% for validation, and the remaining 20% for testing, and our research focused on the segmentation of cancer lesions in ISIS 2018 dataset. This dataset is illustrated in Figure 3.10 with the segmentation result of our approach. Skin lesions typically manifest within the texture and seldom adhere to a definite shape or geometric pattern. This unpredictable behavior might explain why transformer-based networks may not yield substantial benefits for texture-related patterns [Aza+22].

Yet again, the adaptive multi-size-kernel representation capability of our methodology demonstrates its proficiency. Compared to other approaches, it's remarkably effective at

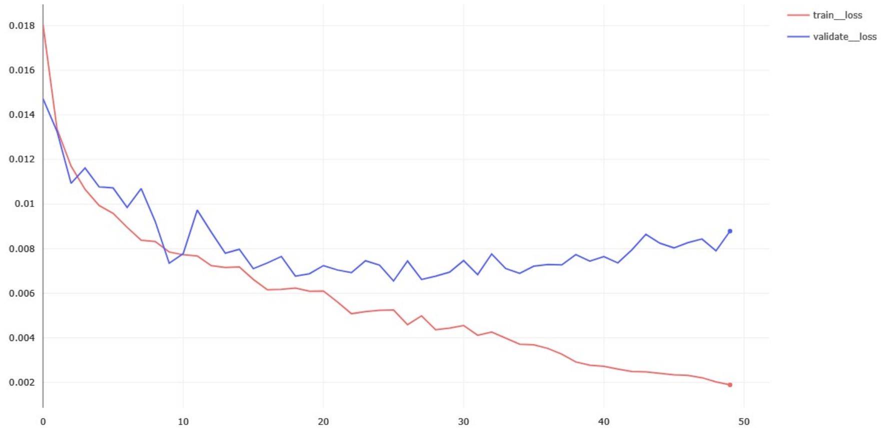


Figure 3.8: The train loss and validation loss of ISIC 2018 dataset. Similar to Figure 3.7, they indicate that the model is neither overfitting nor underfitting.

localizing abnormal regions, which is clearly illustrated in the segmentation results shown in Figure 3.10. This calls for a deeper exploration of the robustness and applicability of our approach. In Figure 3.8, we have plotted the training and validation loss curves for the ISIC 2018 dataset, which shows stability with a minimal gap between training and validation losses. This figure shows that the model is likely performing well.

3.4.8 Discussion

The conducted experiments substantiate the efficacy of integrating an adaptive layer at the initial stage of deep networks, enhancing their resilience to diverse scales. This layer augments the network's capability to discern structural information across varying sizes, while maintaining a comparable parameter count. The experiments demonstrate that the inputs including larger segments are better recognized by the proposed method. The noteworthy aspect of these experiments was the enhancement of all existing models through the integration of the adaptive layer, without necessitating any modifications to their structure. This improvement was observed even in models that inherently feature a multi-scale module (such as UCTransnet). The accuracy of our experiment and number of parameters aligns with the recent survey by Azad et al. [Aza+22], further validating the credibility of our experimental results.

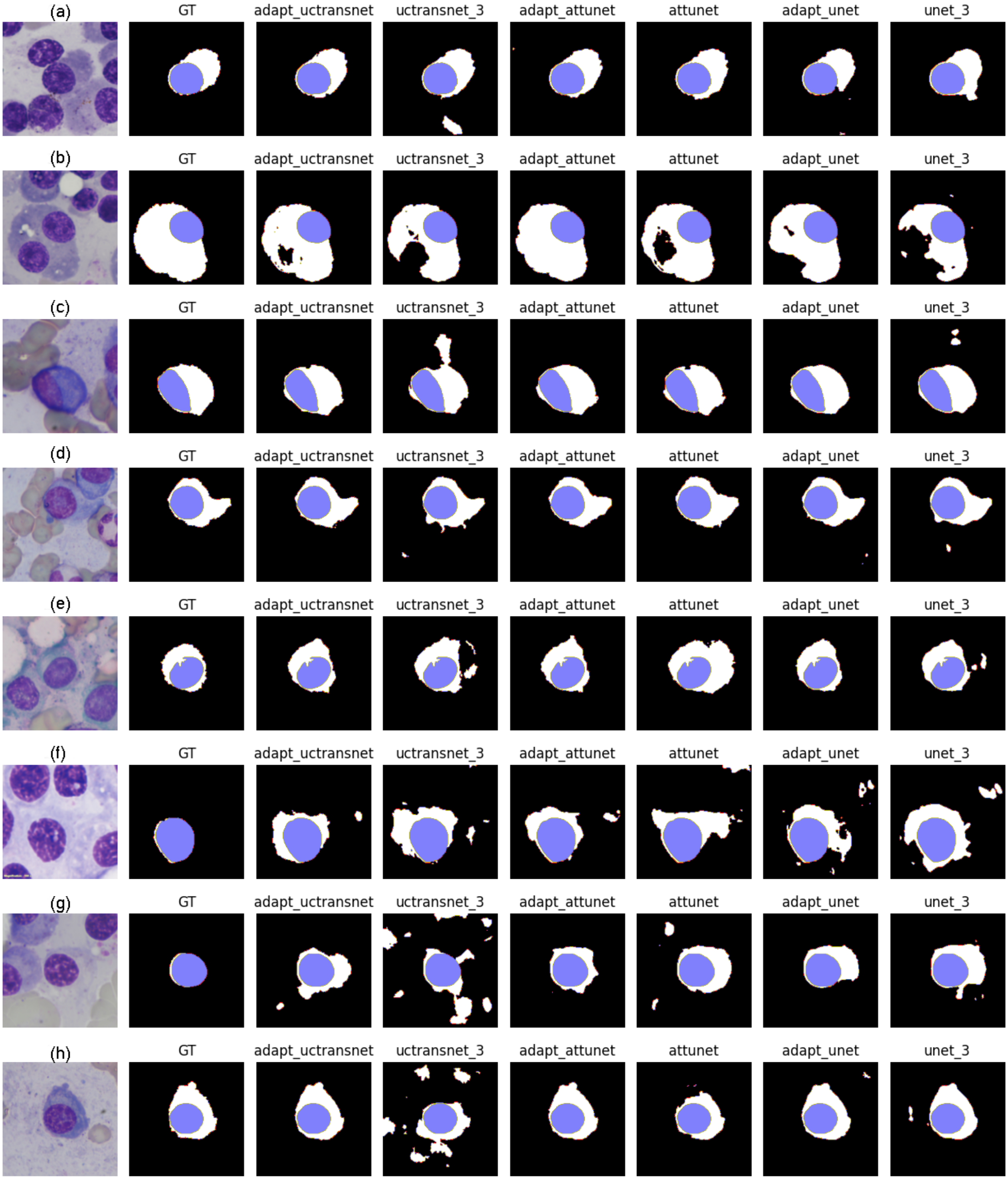


Figure 3.9: Visual comparisons of different methods for cytoplasm segmentation (depicted as the white region) on the SegPC 2021 dataset. The blue region denotes the Nucleus area of a cell. The initial column displays the input image, while the second column presents the ground truth. Following these, the subsequent columns feature the models along with their adaptive versions. As is evident, models incorporating the adaptive layer more accurately recognize the shape of the cytoplasm, and this improvement is particularly greater in larger segments.

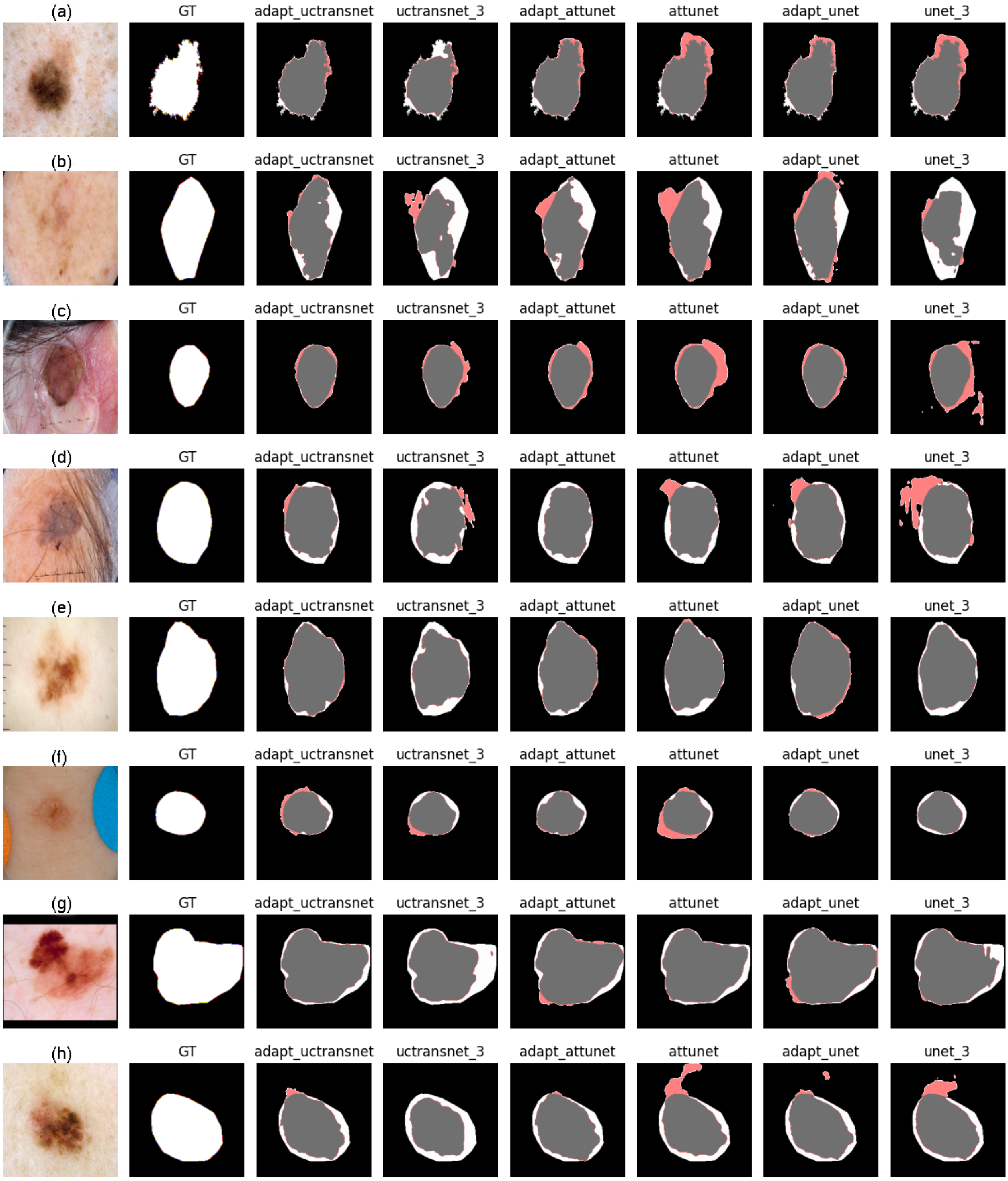


Figure 3.10: Segmentation output of various deep model in ISIC 2018 dataset. The white region represents the ground truth that remains undetected (FN), while the gray region represents the detected ground truth (TP), and red denotes the FP. The columns orders are similar to Figure 3.9. Once again, our model is more effective in identifying target regions, particularly noticeable in larger ones where traditional models with fixed kernels face difficulties in detecting intra-size features.

3.5 Conclusion

In this chapter, we delved into recent advancements in the field of computer vision, focusing on the dynamic modification of the receptive field, a strategy bearing resemblance to the previously described windowing approach. We introduced a novel adaptive layer designed for integration into MIS, aiming to enhance their overall performance. We have shown the effectiveness of our adaptive layer approach by including a dynamic layer on the top of the best segmentation deep network. This approach improves the recognition performance by dynamically changing the receptive field in the first layer, resulting better identification of structural information and various size targets, and reducing high frequency noises while keeping the number of parameters nearly unchanged.

This exploration raised pertinent questions regarding the applicability of this approach in handling tasks involving images that different parts are mostly homogeneous and whether its utility extends to more complex and heterogeneous applications. Consequently, it has prompted us to consider the potential of employing this dynamic receptive field modification strategy in segmentation tasks in AR, a topic we plan to elaborate on in the forthcoming chapter.

Chapter 4

Segmentation

4.1 Chapter Overview

The process of segmentation assumes diverse interpretations within various applications at various levels. In the scope of this thesis, we specifically delve into the realm of segmentation at the pre-processing stage before feeding the data to the machine learning models. Therefore, for instance, in the field of Activity Recognition, segmentation refers to the process of dividing the input sensor events into segments before feeding to the model such common approaches are Event Window and Time Window.

Segmentation is a common step in the processing pipeline of many Internet of Things applications, such as AR. This step introduces at least two families of uncontrollable biases. The first is caused by the changes of the segmentation process on the initial problem space, and the latter results from the segmentation process itself, including the fixation of the segmentation method and its parameters. For example, an appropriate segmentation approach in one period may not be efficient for another period due to the changes in data over time. To avoid these biases, we first redefine the segmentation problem as a special case of the decomposition problem, including a decomposer (traditional segmentation), resolutions, and a composer. In the literature, the composer task is often ignored in machine learning models. However, incorporating the composer task in the segmentation makes it possible to evaluate the relationship between the initial problem to be solved and the problem after the segmentation, resulting in an improved evaluation and consequently selecting the appropriate segmentation method. It addresses the first families of the aforementioned biases. Then, we formally present our novel meta-decomposition or learning-to-decompose concept. It learns how to decompose the original task into sub-tasks to be combined with the meta learning approaches which require multiple tasks. Meta-decomposition reduces the second family of segmentation biases by considering the segmentation as a hyperparameter to be optimized by the outer learning problem. Therefore, meta-decomposition improves the overall system performance by dynamically selecting the appropriate segmentation method without including the aforementioned biases in this process. As mentioned before, without considering the composer part, meta-decomposition introduces an additional bias in the comparison of different segmentation approaches due to the inconsistency in the segments. Extensive experiments on four public datasets demonstrate the feasibility of finding a proper segmentation method and its hyperparameter by our proposal with simple and effective data-driven approach.

Our approach consists of two major steps: a) redefining the segmentation problem as the data decomposition problem, and b) formalizing our novel meta-decomposition model. In what follows, we first introduce the related works, the definition and terminology and then elaborate on each step.

4.2 Related Works and Preliminaries

As described previously, the segmentation process assumes diverse interpretations within various applications at different levels. In the scope of this thesis, we specifically delve into the realm of low-level segmentation, while evaluating its impact at the higher level. Within the domain of activity recognition, segmentation denotes the process of dividing the input sensor events into segments that carry significant meaning. In AR using ambient sensors, data usually comes as a continuous flow of raw sensors' data. One challenge is to achieve a proper division of these long sequences of raw and continuous data flow into smaller blocks of information and reduce the needed computational resources [Via18; NGC15]. The segmentation should provide enough information for recognizing the activity and it has a direct effect on the accuracy of activity recognition [NGC15]. This segmentation enables the identification and analysis of distinct activities or events within a continuous stream of sensor data. By segmenting the input, the system can better understand and interpret different actions or behaviors. Similarly, in the context of image segmentation, the goal is to partition an image into meaningful regions or segments. One approach to achieve this is by creating a grid or dividing the image into smaller sections. In the domain of audio processing, segmentation involves breaking down an audio file into smaller frames or segments. This division facilitates subsequent analysis, such as speech recognition or audio classification. By dividing the audio into frames, specific characteristics, and patterns can be extracted and analyzed within each segment, leading to improved understanding and processing of the audio data.

In sensor-based HAR, the objective is to identify activities including both activity classes and their temporal duration based on a sequence of heterogeneous input sensor events [CN15]. In AR, activities are durative and may occur in parallel by various participants. Sensors track the actions and interactions over time. These sensors are used to capture the human actions and activities of daily life. Sensors have various types, quality, noise and collection rates. For example, several sensors transmit their current state only when there's a change due to communication and battery concerns (event-based), thus, they could be activated at different times. It causes the sampling over time to be patchy and not uniform [NLL22]. The information provided by a single sensor event is inadequate for identifying a particular activity. Accordingly, it is crucial to partition sensor events into a collection of segments that can be mapped to a specific activity [Min+20; BNE21; DTP21]. Segmentation can significantly affect the system's performance because it alters some characteristics of the underlying data. Segmentation is widely considered to be the most common task before the feature extraction and resolution step [YFF17]. In activity recognition using ambient sensors, data usually arrives as a continuous flow of raw data. One challenge is to achieve a proper division of these long sequences of raw and continuous data flow into smaller blocks of information and reduce the needed computational resources [Via18; NGC15]. The segmentation should provide enough information for recognizing the activity and it has a direct effect on the accuracy of activity recognition [NGC15]. Segmentation can also add extra complexity in further data processing when one segment covers two or more activities or events belonging to one activity are spread into several segments [NLL22]. This segmentation enables the identification and analysis of distinct activities or events within a continuous stream of sensor data. By segmenting the input, the system can better understand and interpret different actions or behaviors. The common pipeline is shown in Figure 4.1

Therefore, a trade-off exists among the amount of information in each segment (segment size), the number of segments, and the processing complexity of each segment.

Since segmentation can significantly affect the final results, several studies in AR mainly work on utilizing pre-segmented sequences [NLL22]. Nevertheless, this approach is not practical in real-world scenarios [KC14; De-+18; BNE21; XWG20; Cum+18]. Figures 4.6, 4.7 and 4.10 clearly indicate that pre-segmented data offers an easier path for recognizing activities than that of non-segmented data. Therefore, other studies rely on a segmentation approach that is based on temporal information [BNE21; Fu11; KC14; NGC15], similarity or dissimilarity between segments [NLL22; YFF17; WOO15; SB18], ontology and domain knowledge [Tri+19; SB18; WM18; NLL22], learning the segment size [KC14; SB18], sensor events [NGC15; BNE21; OOB11a], activity or explicit segments [NGC15; BNE21; Lun+18], gathering sufficient features [NGC15; KC14; CN15; YFF17; SB18], evolutionary computation [Tak+01], detecting change points [AC19; Zam+20], feasible space window [Hu+17], and hybrid approaches [NLL22]. It has been proved that the dynamic segmentation approaches perform better than static ones [Fu11]. However, the aforementioned studies are designed to be used in a particular application and dataset. e.g., some of them need continuous senses, which means all sensors' values should be presented at each time point;

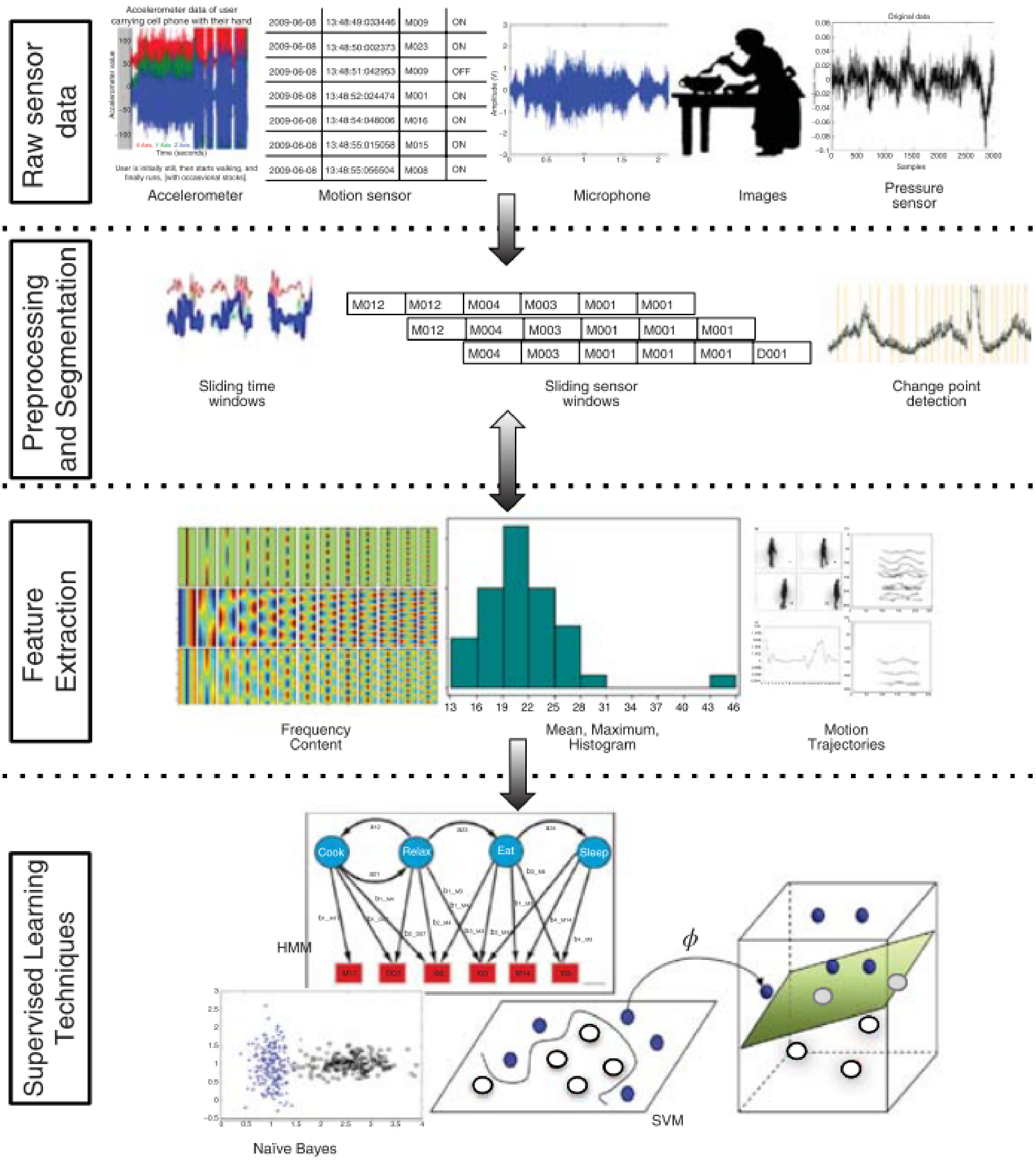


Figure 4.1: Stages of activity recognition include raw data collection, pre-processing, feature extraction, classifier training, and data classification [FC17].

others work on sparse sensor streams where sensor events are triggered only because of human activities, like motion sensor sequence [KC14].

Regarding the recent surveys of human activity recognition in smart homes [BNE21; Ari+22; Min+20; Wan+21a], the most common approaches in data segmentation are Time Windows (TW), Event Windows (EW), and Dynamic Windows (DW). The selection of the optimal parameters is the biggest challenge for these techniques [NLL22]. Kasteren [Kas11] determined that a time window of 60 seconds in TW provides a high classification performance for binary sensors, and it has been used as a reference in several recent works [BNE21; Ham+21; Med+18; Ham+20]. Moreover, for EW that has variable window duration due to the occurrence of events at various times, a window of 20 to 30 events is usually selected [BNE21; AC19]. However, these parameters are completely dependent on a given dataset. They are hard to tune and may not be efficient for complex activities [BNE21; KC14; Qui+18; SWL17]. The importance of selecting an appropriate window size is studied [Min+20]. Quigley et al. [Qui+18] demonstrate that although TW reaches a high accuracy, it fails to properly identify all classes. On the other hand, DW uses a non-fixed window size and tries to estimate the activity duration based on the sensor events. However, this approach is inefficient for complex activities [BNE21; KC14; Qui+18; SWL17]

There is a rise in deep learning approaches for HAR [Wan+19; Che+21b; Lic+20; BNE21; Bou+21b]. Guan et al. [GP17] claim that the deep learning method can be insensitive to the window size in HAR. Yet, it is not the case for HAR scenarios with the sparse data stream, since either those works are on the pre-segmented data [Bou+21b] or the parameters (e.g., length and moving step) need to be carefully tuned to achieve satisfying performance [Che+21b]. Therefore, providing a unified method usable in various applications is helpful. It prevents including implicit knowledge about the program and data in the algorithm, implementation, and evaluation.

AutoML techniques are used to automate algorithm and hyperparameter selection [Mu+22; Tay+18]. However, they do not adapt the algorithm regarding the incoming data, while an important concern in HAR is due to the fact that the proper segmentation method in one period may be inappropriate for another period [Ros+14; Ada+19]. Meta-learning, often known as learning to learn, has been successfully used for algorithm selection [Agu+19] and provides automatic and systematic guidance on algorithm selection based on the information gained through a set of algorithms on various tasks [Ros+21]. Meta-learning is not a novel concept [Sch87; SS10]; however, there has been a recent rise of interest in meta-learning [Gre+19]. Hendryx et al. [Hen+19], and Aguiar et al. [Agu+19] use a meta-learning approach across different image tasks to select the proper algorithm for generating the mask for a given image [SL21].

One of the considerations in this thesis is the fact that previous studies consider each experience as an independent instance and do not consider the decomposition of each task into sub-tasks to be used by meta-learning. Additionally, in IoT, the input sensor events in each experience are characterized by their continuous and relative nature; therefore, they are not independent [Fu11; CN15].

The previously mentioned approaches are not appropriate for the IoT data. Therefore, we initially describe the problem formally. There are a few works around the formal definition of the segmentation problem. They typically formalize their algorithms for a given application and assumptions [Oke+14; OOB11b; CN15]. In [Oke+14; OOB11b; CN15] studies, an *Activity* is assumed as a class label for each segment. This definition does not consider the continuity of activities that are occurring during a time interval. Alevizos et al. [Ale+17] formulate probabilistic complex event recognition by presenting simple event algebra for probabilistic events. Cook et al. [CN15] propose two formalizations for event segmentation approaches and sliding window approaches. To the best of our knowledge, no work has been done on dynamically selecting the appropriate segmentation method and its hyperparameter over time, despite the fact that the appropriate segmentation method and its hyperparameter in one period may be inappropriate for another period [Ros+14; Ada+19]. In addition, the previous studies never viewed the segmentation problem as a data decomposition.

4.3 Problem Formulation

Following the notations of [SS10; HRP21; Hos+22], let us consider domain A as a set of experiments. Experience $s \in A$ is a broad term used on both supervised and non-supervised learning problems that may refer to an input-target tuple, a single data point, a sequence of events (e.g., in IoT), etc. We define m_θ as a task with hyperparameter

θ (includes, e.g., the initial model parameters, choice of the optimizer, and learning rate schedule) that takes the input experiments and outputs the target concepts (e.g., activities in AR, sound events in SED, and tumor spots in MIS) based on the constraints, objectives, etc. $\phi(m_\theta, s)$ is associated with measuring the performance of task m_θ on the experiment s. We denote the expected performance of task m_θ on $S \subseteq A$ by $\mathcal{L}(m_\theta, S)$, such that:

$$\mathcal{L}(m_\theta, S) = \mathbb{E}_{s \in S}\left[\phi(m_\theta, s)\right] \tag{4.1}$$

Measurement ϕ is as diverse as the application domains. For instance, in supervised learning, ϕ might be the differences between task outputs and teacher-given values [SS10]. Without losing generality, we consider the optimum hyperparameter minimizes the expected performance ($\theta^*_S = \mathrm{argmin}_\theta \mathcal{L}(\mathrm{m}_\theta, \mathrm{S})$). Finding globally θ^* is computationally infeasible [HRP21]. Therefore, we approximate it ($\theta^*_S \approx g_\omega(S)$) guided by pre-defined meta-knowledge ω which includes, e.g., the initial model parameters (θ_0), choice of the optimizer, and learning rate schedule [Hos+22].

For instance, in sensor-based AR, each experiment contains a sequence of various sensor occurrences, and activity information (such as their label and duration); task m_θ refers to the activity recognition model and its hyperparameters; and $\phi(m_\theta, s)$ evaluates the performance of the activity recognition model on the experiment s. For instance, ϕ can evaluate the duration of correctly identified activities.

4.4 Decomposition

Even though decomposition is a well-known approach in designing algorithms [Cor+09], the data segmentation problem has never been viewed as a data decomposition problem, which consists of a decomposer that splits the input sequence into a set of smaller data (traditional segmentation), resolutions that find the concepts from these smaller segments (usually less complex than original resolutions), and a composer that combines the sub-results to generate the overall results.

To the best of our knowledge, previous studies in the literature have not considered the composer component [KC14; QJE21; YFF17; De-+18; Cum+17; Wan+19; Che+21b; HSZ20; Ber+18; Via18]. These studies have made the implicit assumption that the segmentation process preserves the integrity of the whole problem, and as such, the overall system performance is assessed based on the output of each segment. This hard hypothesis may misleadingly present convenient results without even reducing the complexity of the problem, which is explained in Section 4.6.4. Therefore, this study explicitly redefines the segmentation problem as a data decomposition that incorporates all three components: the decomposer, the resolutions, and the composer. The introduced biases, loss of information, and performance of each component can affect the overall performance of the system and should be carefully evaluated. Furthermore, neglecting the composer component would lead to inconsistencies in the comparison of different segmentation algorithms, which will be discussed in detail in Section 4.6.4. Therefore, including the composer component in the segmentation problem is crucial for accurately evaluating and comparing the performance of various segmentation algorithms.

Accordingly, the decomposer task (d_δ parameterized by δ) decomposes m_θ into M resolution sub-tasks ($\Pi_\Psi = \{\pi^i_{\psi_i}\}_{i=1:M}$ such that each sub-task is parameterized by $\psi_i \in \Psi$), and the composer task (c_σ parameterized by σ) that combines the results of sub-tasks to produce the overall system results. Task m_θ is decomposable under the measurement $\mathcal{L}$ to Π_Ψ and c_σ if and only if the composition of sub-tasks does not perform worse up to ϵ than task m_θ. Formally:

$$d_{\delta,\mathcal{L}}(m_\theta) \approx_\epsilon \Pi_\Psi, c_\sigma \iff \mathcal{L}(c_\sigma(\Pi_\Psi), A) - \mathcal{L}(m_\theta, A) \leq \epsilon \tag{4.2}$$

The task's performance after decomposition cannot surpass the original task due to the severed relationship between events, leading to information loss. When ϵ is zero, it means that m_θ is strongly decomposable to Π_Ψ and c_σ without any loss of information, otherwise, it is a weakly decomposable task, and some information is lost. δ, Ψ, and σ show the dependencies of the model on pre-defined assumptions about the decomposition, for example, the segmentation approach such as time window and event window, and their internal parameters such as window size. These assumptions can affect the global system's performance. For example, the optical character recognition (OCR) task is commonly decomposed into sub-tasks with sub-images (each contains one character), and the composition task merges the results of those tasks to produce the whole problem result (full text). Equation (4.2) shows that the decomposability of a task depends on the task target, i.e., on the objective function measured by $\mathcal{L}$. For instance, while we can decompose the face recognition task to analyze only the color frame task with an accuracy of 99.5% [LSX21]; this decomposition may be inadequate in a highly secure application, where a more detailed decomposition involving depth and color sub-tasks may be required [App22].

Obviously, all the mentioned components (decomposer, resolution, and composer) play a crucial role in measuring the performance, the introduced biases, loss of information, and complexity of the whole system and should be considered in designing and evaluating segmentation approaches. In particular, ignoring the composer component leads to inconsistencies and difficulties in comparing different segmentation algorithms, which are elaborated more in Section 4.6.4. The decomposer task itself can be decomposed into several sub-tasks and a composer task. For example, in an IoT data processing task, we can decompose the task into sub-tasks each one containing a meta-segment (for example one day, week, or month).

4.5 Meta-Decomposition

In a traditional segmentation, the probability distribution of data is supposed to be unknown but stationary. Nevertheless, the underlying distribution of data in real-world IoT systems naturally changes over time [Ros+21]. Additionally, fixing the segmentation and its hyperparameter is the second family of biases that is often implicitly incorporated with the prior knowledge or assumptions originating from the developers, researchers, or experts.

Therefore, we propose our novel Meta-Decomposition approach to resolve these concerns. Learning-to-decompose or meta-decomposition is defined as a model that can dynamically and systematically select and tune the decomposition algorithm. Formally, in Equation (4.3), we consider $d^i_{\delta_i} \in \mathcal{D}$ as the i-th decomposer task which generates sub-tasks $\Pi^i_{\Psi_i}$ and the composer task $c^i_{\sigma_i}$; then, we define the meta-decomposition task $(\widehat{d}_{\hat{\delta}})$ as the selection of sub-tasks $(\widehat{\Pi}_{\hat{\Psi}})$ and the meta-composer task $(\widehat{c}_{\hat{\sigma}})$, such that in the meta-evaluation, the meta-decomposition task outperforms those decomposition tasks individually.

$$\widehat{d}_{\hat{\delta},\mathcal{L}}(m_\theta) \approx \widehat{\Pi}_{\hat{\Psi}}, \widehat{c}_{\hat{\sigma}}, \quad s.t., \widehat{\Pi}_{\hat{\Psi}} \subseteq \bigcup_i \Pi^i_{\Psi_i} \wedge \mathcal{L}(\widehat{c}_{\hat{\sigma}}(\widehat{\Pi}_{\hat{\Psi}}), A) - \min_i \mathcal{L}(c^i_{\sigma_i}(\Pi^i_{\Psi_i}), A) \leq 0 \tag{4.3}$$

This definition is illustrated in Figure 4.2. The meta-decomposition task can be carried out in several ways to efficiently and dynamically select the proper segmentation approach and its hyperparameters depending on the incoming data, application, and constraints. Moreover, the proposed model can be easily extended to an arbitrary number of meta-levels and is not limited to a single layer of meta-decomposition. For example, meta-meta-decomposition algorithm can generate the sub-tasks for the inner meta-decomposition algorithm. For example, we can select a subset of sub-tasks (e.g., person detection and object detection) in order to recognize the human-object interaction. As mentioned before, meta-learning [SS10; Agu+19; Van19] has been successfully used for algorithm selection over various tasks (e.g., [SL21]); however, it requires multiple tasks and is not applicable to one single HAR task. In meta-decomposition, we generate the sub-tasks from a single task. These sub-tasks can be fed to the meta-learning approaches [SS10]. Therefore, we can enhance the overall system performance without including hard prior biases about the fixation of the segmentation method and its hyperparameters. The next section describes more about it with an experiment.

The current definition is open to interpretation regarding the distinction between decomposition and meta-decomposition. Specifically, combining a decomposition with a meta-decomposition, as well as any number of further meta-meta-decompositions, can always be seen as a single ”flat” decomposition algorithm. On the other hand, some decomposition methods can be seen as a type of meta-decomposition. For instance, the one that decomposes data and changes the segmentation size over time while composing the results can

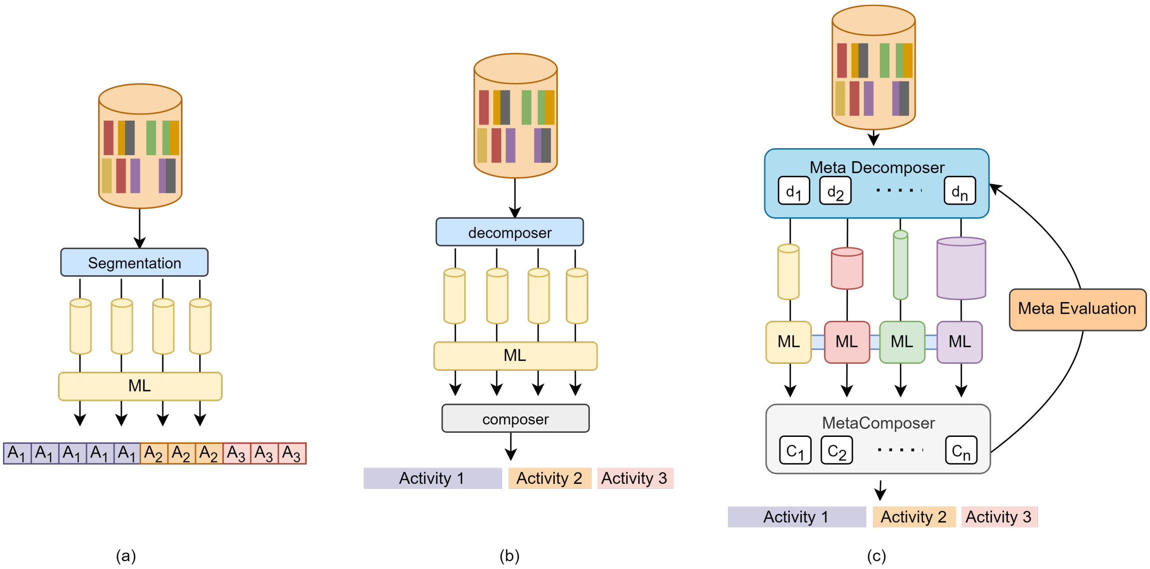


Figure 4.2: (a) shows the conventional segmentation approach, which creates a set of segments in data preparation and then treats them as individual instances to be inputted into the ML system. (b) illustrates the proposed formulation of segmentation as a data decomposition problem, including the decomposer, the ML model, and the composer. (c) provides an overview of our proposed meta-decomposition model. It can dynamically select the proper decomposition sub-tasks from multiple decomposition algorithms.

be considered a basic form of meta-decomposition.

4.6 Experiment

In this section, we proposed a meta-decomposition method to improve the overall performance of the system where the change in the environment is inevitable and a decomposer task in one period may become inappropriate in another period. In other words, meta-decomposition adaptively learns the proper decomposition for different data in dynamic environments [Ros+21].

In the experiments, we demonstrate the meta-decomposition effectiveness in IoT data in HAR. To select appropriate sub-tasks from the available decomposer tasks, we first decompose this long data into a set of meta-segments and then extract the meta-features of these meta-segments to learn how to select a suitable base decomposition task. We can also extend this step to include a higher-level meta-meta-decomposition task. However, we only consider one level in these experiments for simplicity. The details of these experiments are in the following subsections, which we explain our experiments, selected datasets, environment, framework, baselines, and evaluation method in detail. Then, we present a discussion on the results.

4.6.1 Case Studies

Experiments are conducted on various public testbeds, including the widely-used [De-+18; BNE21; Ari+22] WSU CASAS Home1, Home2 [KC14], and Aruba [Coo12] datasets that have around 32 sensors and between 250,000 to 1,700,000 events, Orange4Home (Orange4H) dataset [Cum+17] that has 207 sensors and about 700,000 events. Each testbed consists of heterogeneous sensor events and the daily activities of an individual in a smart apartment. They have imbalanced activity classes, activity durations, and sensor events. e.g., bathroom activities are frequent and last a few minutes with a few sensor events, while cooking activities may occur once a day, last about an hour, and involve numerous sensor events [Med+18]. These datasets are detailed in the following subsections.

Orange4Home Testbed

The Orange4Home dataset, as detailed in [Cum+17], provides a snapshot of events captured from a two-floor smart home that spans 87m^2, as visualized in Figure 4.3. This dataset has an extensive network of sensors, strategically placed to capture a multitude of parameters. The apartment is equipped with an impressive array of 236 sensors. This includes 83 binary sensors that typically record 'on/off' states such as door opening, presence, and switches. In addition to this, there are 55 Integer sensors such as total cold water consumption, appliance power, and humidity, 67 float sensors such as luminosity, voltage, CO2 levels, and area noise levels, and 31 categorical sensors such as weather, heater modes, AC modes, and wind direction. These sensors, in their collective capacity, monitor a vast spectrum of environmental and activity-related factors. Motion sensors track move-

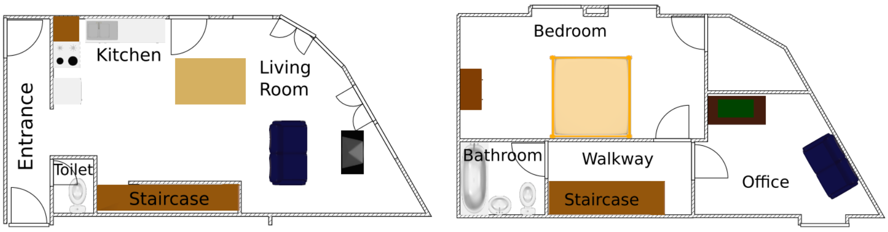


Figure 4.3: Orange4Home Dataset Sensor Configuration - This dataset represents a two-floor home with 236 sensors, including 83 binary, 55 integer, 67 float, and 31 categorical sensors. These sensors capture data related to motion, switches, humidity, water consumption, luminosity, temperature, weather conditions, and heater settings over a 5-month period.

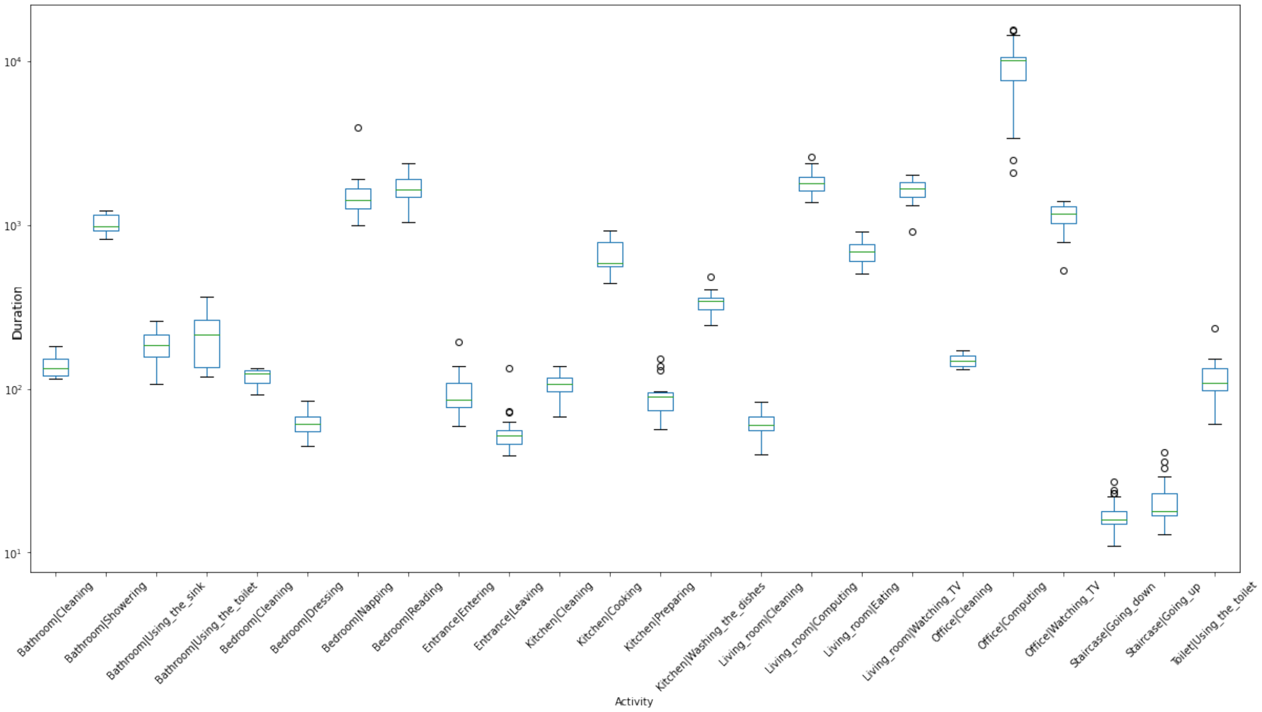


Figure 4.4: Orange4Home Activity Durations: This box-plot depicts the duration of various activities within the Orange4Home dataset. The y-axis measures time in seconds on a logarithmic scale, showcasing the diversity in average activity durations. Activities are listed on the x-axis, with durations ranging from 10 seconds (e.g., ’going down’ activity) to approximately 3 hours (e.g., ’computing’ activity).

ment within the apartment, while switches register appliance usage. Climatic conditions inside the apartment, like humidity and temperature, are diligently noted. In a more advanced setting, the dataset even captures parameters like water consumption, luminosity, and specific weather conditions. Furthermore, for those keen on understanding the energy dynamics, heater settings data provides insights into the heating preferences and practices within the residence.

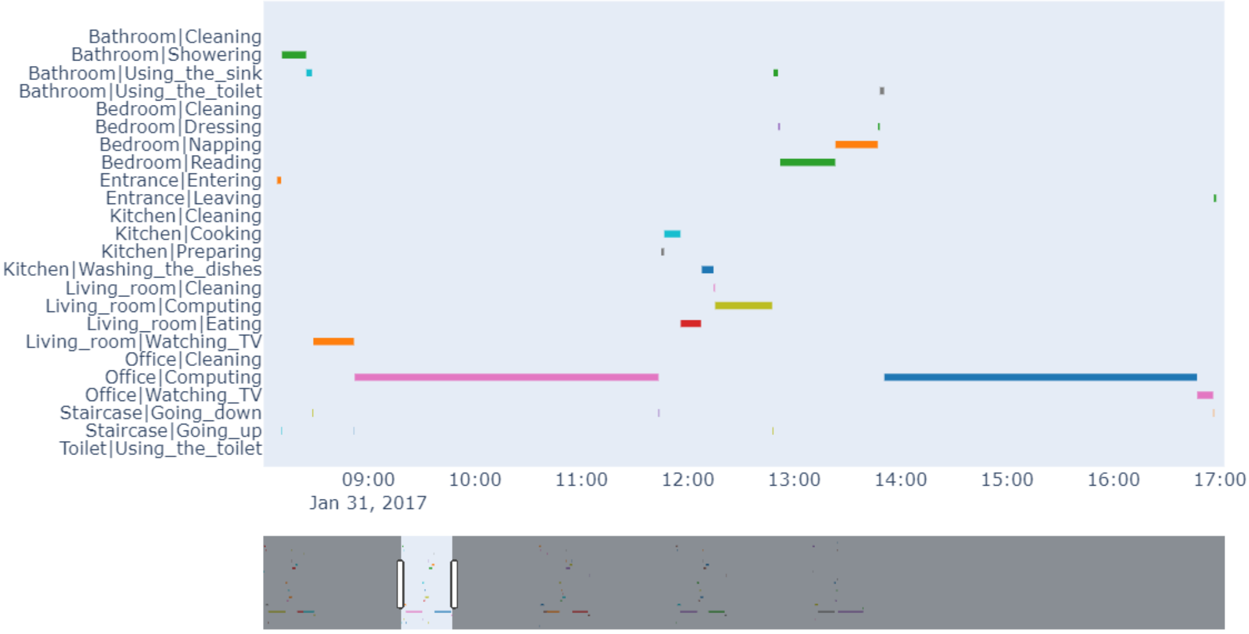


Figure 4.5: Visualization of daily activity patterns in the Orange4Home dataset, distinguished by different colors. The entirety of the week is condensed into a series of compact bars at the bottom of the image, with the selected day accentuated for emphasis. The y-axis categorizes the activities, while horizontal lines denote the length of each activity instance. Notably, the consistency in activity patterns across various days calls for careful analysis to avoid overfitting by overlooking to temporal factors.

Data collection was methodically planned. Activities from 8h00 to 17h00, aligning with standard working hours, were recorded. This was performed over four consecutive weeks, ensuring that any patterns or anomalies could be clearly discerned. As outlined in [Cum+18; Cum+17], this approach was designed to yield a comprehensive understanding of a regular day's routine within the apartment.

A snapshot of ground truth activities of this dataset can be seen in Figure 4.4. This latter visualization aids researchers and developers in correlating sensor data with actual human activities, forming a critical component in the realm of activity recognition.

To better visualize how these sensors capture and represent activities, consider the diverse response patterns across the 236 sensors for different activity types, as depicted in Figures 4.6 and 4.7. This sensor heatmap reflects the dataset's pre-segmentation by activity, highlighting the frequency of sensor events through five distinct phases of each

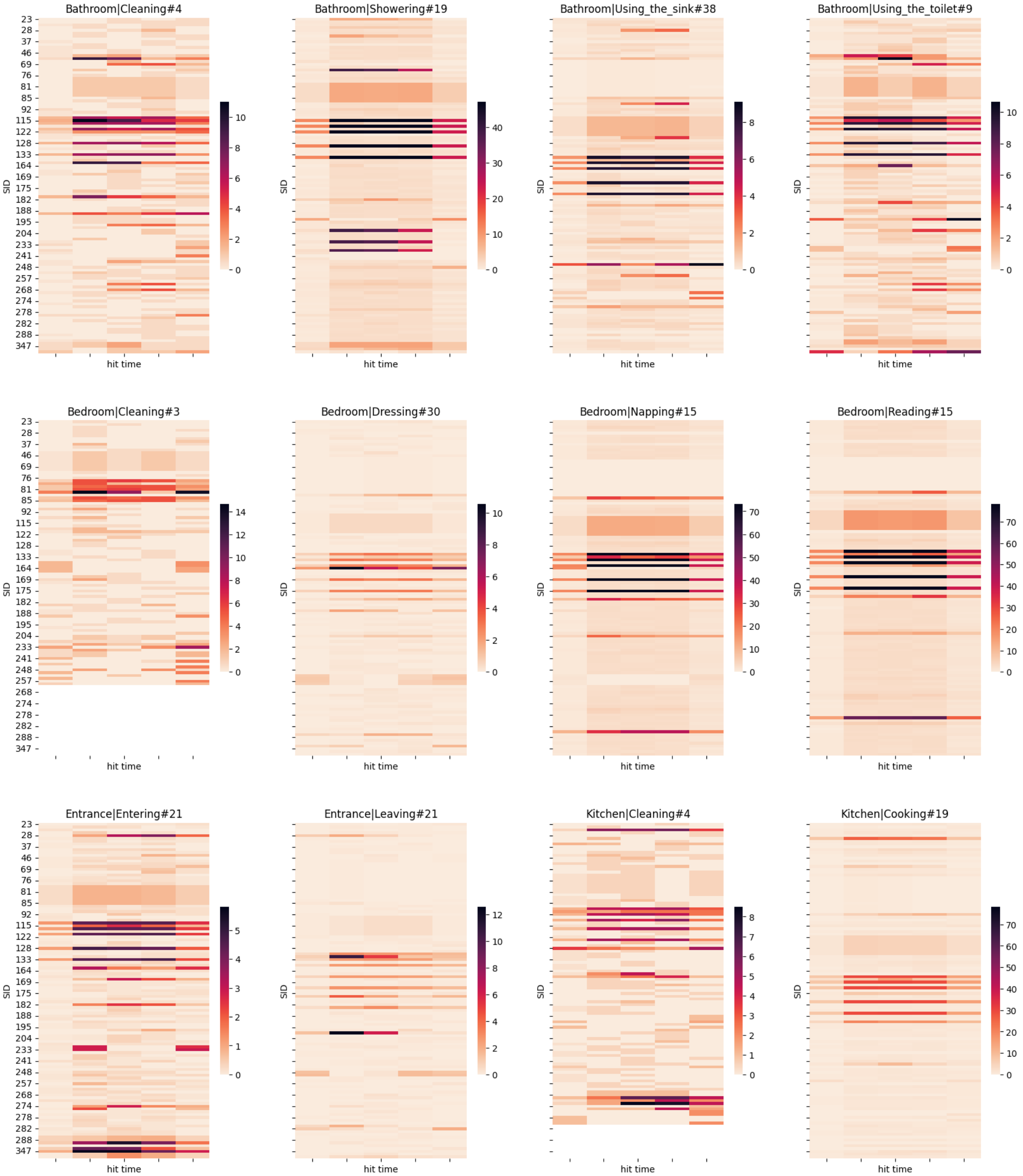


Figure 4.6: Sensor Event Frequency in the First Part (12 of 24 activities) of the Orange4Home Dataset. Each hit map indicates the frequency and distribution of sensor events for an individual activity, with the Y-axis representing specific sensors. The absence of events is shown in white, while darker tones indicate higher event frequencies. These visual patterns reveal significant differences in the occurrence and timing of sensor-triggered activities, such as between 'Kitchen Cooking' and 'Kitchen Cleaning'. Pre-segmented data showcases distinct patterns, reducing ambiguity and enhancing recognition performance of activities. Although valuable for experimental analysis, the practicality of pre-segmented data in real-world scenarios may be limited.

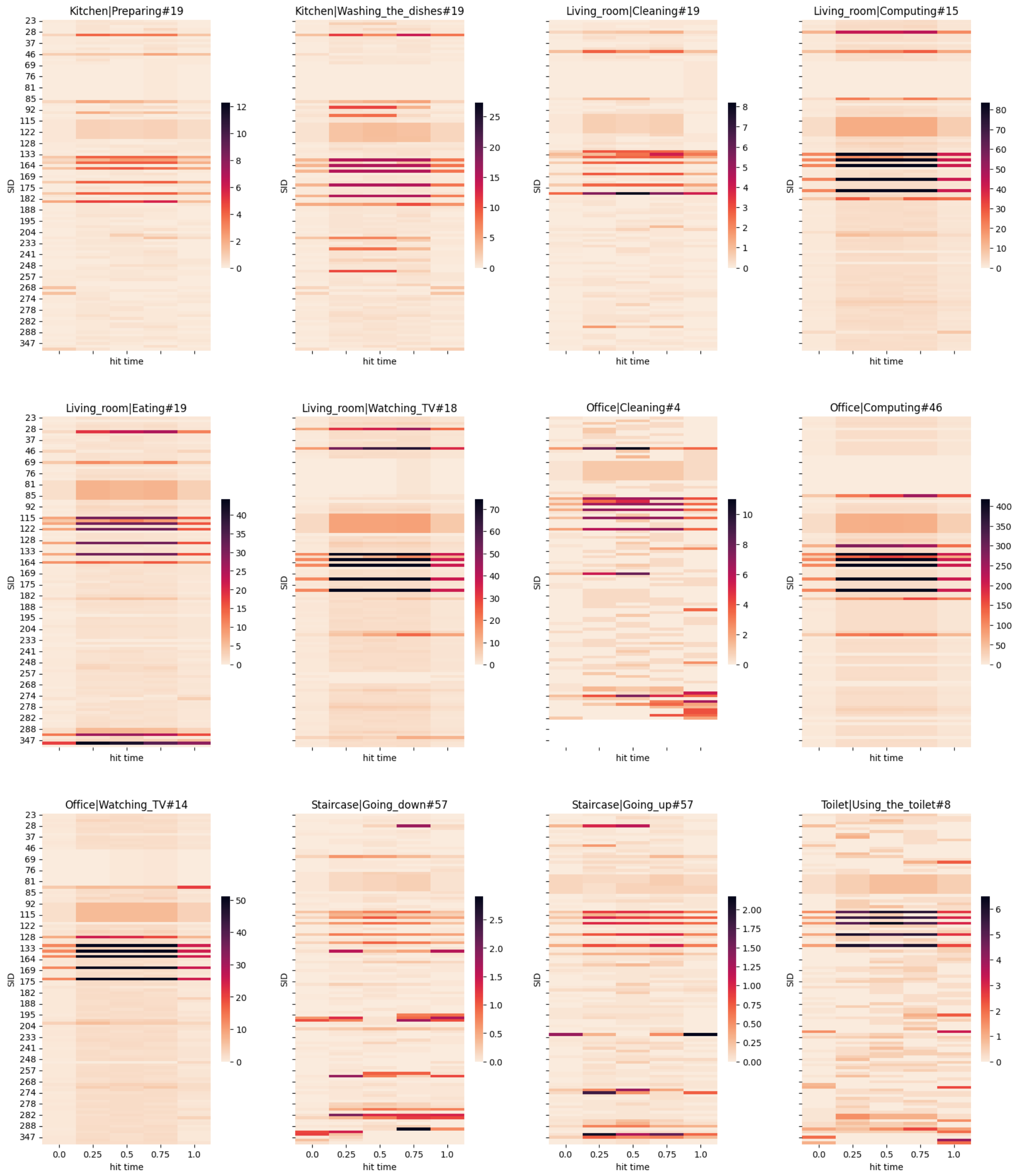


Figure 4.7: Sensor Event Frequency for the remaining activities in the Orange4Home Dataset, complementing Figure 4.6. Each hit map continues to represent the frequency and distribution of sensor events, with darker shades denoting more frequent occurrences. This part maintains the distinct activity patterns identified in the first half, supporting the more straightforward recognition of pre-segmented data across all 24 activities.

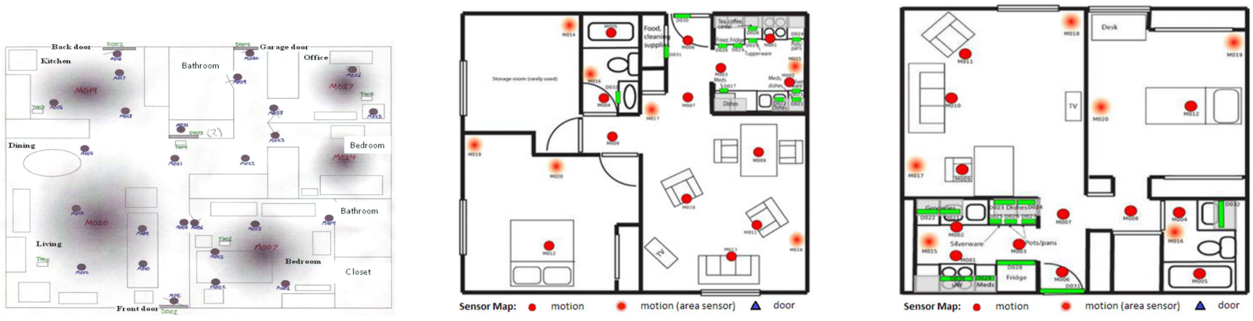
activity: the very beginning, early-mid, exact middle, late-mid, and the very end. This division provides a granular perspective on how sensor responses evolve throughout the entire duration of a given activity.

An essential aspect to bear in mind regarding this dataset is the potential pitfall of using time as a direct predictor. The primary focus should be on gleaning insights from the sensor events themselves. Given the minimal daily variations in activities, directly factoring in time can be misleading. It could result in the mistaken impression of a model's superior performance when, in reality, it might simply be echoing the predictable nature of the time-based activities.

CASAS Testbeds

Researchers at Washington State University introduced CASAS datasets that contain sensor data collected in the homes of volunteers [Coo12]. The dataset labels these activities: • Meal preparation, • Relax, • Eating, • Work, • Sleeping, • Wash dishes, • Bed to toilet, • Personal Hygiene, • Bathing, • Take Medicine, • Enter home, • Leave home, and • Housekeeping. The duration of these activities is varied. For example, on average, the duration of sleeping is 3h:35, while the duration of eating is 9 minutes [NLL22]. The CASAS team provides several datasets with the mentioned configuration such as Aruba Figure 4.8a [Coo12], Home1 Figure 4.8b, and Home2 Figure 4.8c [KC14]. Each experimental setup is performed in a one-room apartment. An elderly individual carries out its regular unscripted daily activities which are labeled later by human experts [Asg+20].

Aruba Dataset The data in the Aruba dataset was collected from a single resident living in an apartment over a period of seven months. The apartment was equipped with a variety of sensors, including motion sensors, door sensors, temperature sensors, and others, in order to monitor and log the resident's activities. Figure 4.8a illustrates these sensor configurations. When deciphering the sensor identifiers within this dataset, those begin with the letter 'M' are associated with motion sensors, while those prefixed with 'D' point towards door sensors.


Back door
Garage door
Kitchen
Bathroom
Office
Dining
Bedroom
Bathroom
Living
Bedroom
Closet
Front door
Storage room (rarely used)
Desk
TV
Silverware
Pots/pans
Fridge
Meds
Sensor Map: ● motion ● motion (area sensor) ▲ door
Sensor Map: ● motion ● motion (area sensor) ▲ door

(a) Aruba. It contains the events from 34 binary sensors for a time interval of 7 months.

(b) Home1. It contains the events from 32 binary sensors for a period of 5 months.

(c) Home2. It contains the events from 30 binary sensors for a period of 5 months.

Figure 4.8: CASAS Aruba [Coo12], Home1 and Home2 datasets [KC14] sensors configuration. Around 70% of the activities are unlabeled.

This dataset's composition includes an array of different activities. It records instances like moving from the bed to the toilet, eating, entering or leaving the home, housekeeping tasks, meal preparations, relaxation, yoga, sleeping, washing dishes, and working. The volume of events for each activity varies significantly. For instance, there's a high frequency of 'meal preparation' and 'relax' events, and fewer instances of activities like 'yoga' or moving from the 'bed to the toilet.' Figure 4.9 presents the duration of these activities in a box plot.

A big part of this dataset is the "Other events" category. These are instances where the exact activity couldn't be determined, leading to absent labels. These unlabeled events constitute a substantial 55% of the entire dataset. An example for the annotated data is presented in Figure 2.4. In Figure 4.10, we present the sensor activation of each activity for better understanding the datasets.

CASAS Home1 Dataset The CASAS Home1 dataset represents one of the two distinct datasets explored in this study, gathered from a one-bedroom apartment inhabited by a single older adult going about their routine daily activities. Spanning a timeframe of five months, data was collected from a total of 32 sensors. Among these, 20 were motion sensors and 12 were specifically designated as door or cabinet sensors which are shown in Figure 4.8b. Over the course of the data collection period, there were 371,925 sensor

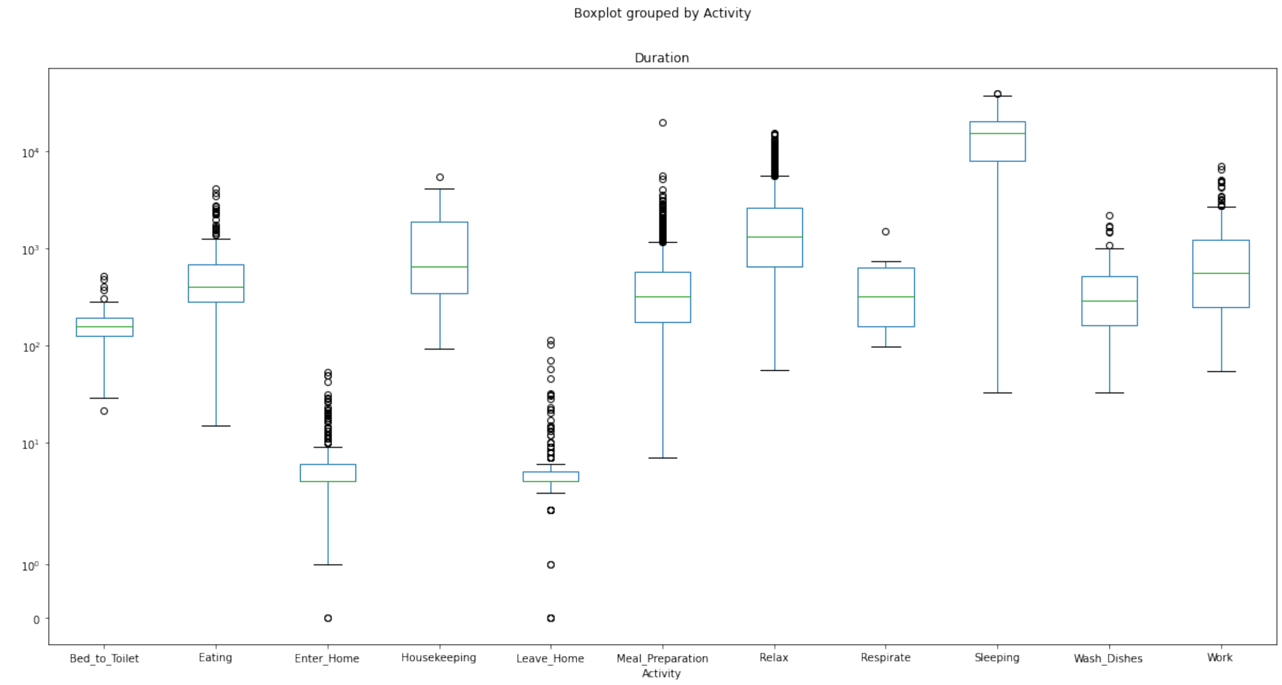


Figure 4.9: Casas Aruba Dataset Activity Duration - The y-axis represents the duration in seconds on a logarithmic scale. In comparison to Figure 4.4, this dataset contains fewer activities. Nonetheless, it still illustrates the diversity of average activity durations, ranging from 1 second (e.g., 'enter home' activity) to approximately 7 hours (e.g., 'sleeping' activity). Different activities in this dataset are displayed on the x-axis.

events registered for Home1. After the collection period, human experts took on the task of labeling the data. Eleven different classes or categories of activities were identified within this dataset as shown in Figure B.4, encompassing actions like personal hygiene, leaving or entering the home, meal preparation, sleeping, among others. Notably, there was a significant data imbalance in this dataset, particularly when considering the housekeeping activity. This category recorded a mere 13 instances, standing in stark contrast to the other, more heavily populated activity categories. This disparity in data distribution could potentially be one of the primary reasons for certain inadequacies in the model's performance when applied to the Home1 dataset [Asg+20; KC14]. The duration of each activity in this dataset is shown in the appendix Figure B.2.

In appendix Figure B.3, a visual analysis of the Home1 dataset's activities is presented. From this figure, it's evident that activities such as Bathing and Sleeping, Meal Preparation, Personal Hygiene, and Eating have been distinctly categorized, suggesting that the dataset

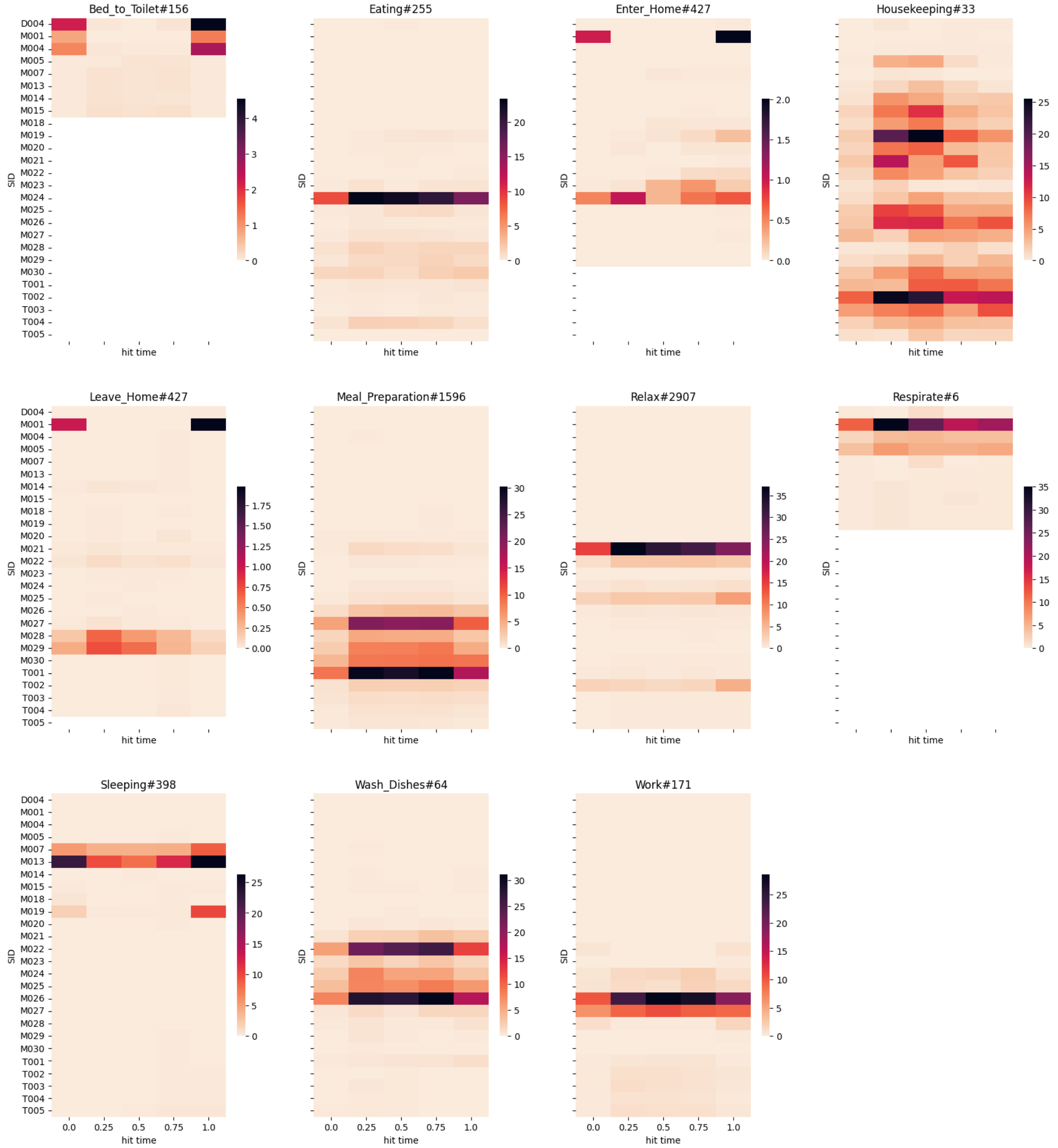


Figure 4.10: Sensor Event Frequency for the activities in the Aruba Dataset, each hit map continues to represent the frequency and distribution of sensor events, with darker shades denoting more frequent occurrences. On the legend, the average number of sensor's heats are shown. Similar observation to Figure 4.6 is shown clearer. This part maintains similar observation about the distinct activity patterns identified in Figure 4.6, supporting the more straightforward recognition of pre-segmented data across all activities.

can accurately differentiate between these two actions. Additionally, transitional activities, especially the movement from the Bed to the Toilette, have been captured with a commendable degree of accuracy. This figure, thus, offers a comprehensive view into the granularity with which the Home1 captures and distinguishes various day-to-day activities.

Home2 Dataset Complementary to Home1, the Home2 dataset also derives from a one-bedroom apartment, where a single older adult engaged in spontaneous daily activities. This dataset also stretches over a five-months period but was accumulated using 30 sensors – 18 dedicated to motion and 12 for door and cabinet monitoring (Figure 4.8c). Throughout this period, Home2 registered a total of 274,920 sensor events. The data, once collected, underwent meticulous labeling by human professionals, ensuring that each event aligned with one of the eleven predefined activity classes. These classes mirror those in the Home1 dataset and include actions such as bathing, sleeping, taking medicine, and transitions from the bed to the toilet (appendix Figure B.5). A closer examination of the Home2 dataset highlights certain disparities and challenges compared to Home1. For instance, some overlapping activities, like the resident eating a meal on a couch—a spot also associated with napping—resulted in difficulties distinguishing between the two activities. Nevertheless, a positive distinction for Home2 was observed in the housekeeping activity class. Appendix Figure B.6 presents the duration of each activity in this dataset [ASN19; KC14].

Similar to Home1 dataset, in appendix Figure B.3, a visual analysis of the Home2 dataset's activities is presented. From the figure, it's evident that some activities in the dataset can accurately differentiate when we have all the events for an activity. However, this is not the case, when the data are not pre-segmented.

4.6.2 Environment

All experiments are run on an NVIDIA DGX-1 machine featuring a Tesla V100-32 GPU, Intel Xeon E5-2698v4 CPUs, and 512 GB of RAM. However, our framework works also on a personal computer. All the codes, datasets, and documentations are freely accessible on our GitHub repository[1] .

4.6.3 Framework Description and Baseline

Our pipeline infers activities and their duration from raw sensor data. The pipeline is composed of several stages: data pre-processing, meta-decomposition, feature extraction, classification, and meta-composition. For each stage, various techniques are implemented in our repository. However, without losing generality in this experiment, we have fixed the parameters of the inner learner to focus on the meta-decomposition. For the inner learner, a fully convolutional network (FCN) with embedded layers is selected. It outperforms the long-short-term memory (LSTM) networks, while it is significantly quicker [BNE21; Bou+21a; Bou22]. It treats sensor events as words and activity sequences as text sentences. Therefore, they encode each sensor event as a word containing the sensor name and its value. For example, if a sensor with id ‘door1’ fires an ‘open’ event, it will be encoded as “door1open”. Then, based on the frequency of each word, it will be indexed from 1 (index zero is reserved for padding). Then, each sequence of the sensor events in a window is mapped to an activity. A sequential model with three layers of conv1D, batch normalization, and ReLU activation with 128, 64, and 128 filters, 1D global average pooling, and softmax layers is used. The hyperparameters of the model in the training phase include the batch size of 1024, epochs limit of 100 with early stop conditions, validation split of 0.2, Adam optimizer, and categorical cross-entropy loss. Afterward, the composition step converts the ML results to the problem space. As explained before, this step has been ignored by several studies. This step itself is a challenging problem and directly impacts the result. We demonstrate its importance using a basic combiner that combines overlapped and neighbor windows.

The idea is demonstrated through experiments using a straightforward yet effective method called SWMeta. Although there is potential for multiple higher-level meta-meta-

[1] https://Github.com/modaresimr/unified_ar

decompositions, this study only focuses on one layer of meta-decomposition, breaking down the data into one-day meta-segments. Then, we randomly select J meta segments (in this experiment, $J = 8$) and use grid search to find the inner decomposer's hyperparameters for each meta-segment. The inner model parameters are then updated with the new decomposer. This process is repeated 100 times in this experiment. Next, for each meta-segment, proper decomposer parameters will be selected by starting from the global knowledge obtained from the previous step to update the local knowledge that is proper for this meta-segment. We add the meta-features from this meta-segment and the selected decomposer parameters to the new train set. For this, we extract the meta-features containing the number of events triggered by each sensor (normalized by the mean and scaling to unit variance) and the spline-transformed day of week and month [EM96]. After training the new model using this new training set, we estimate the proper decomposer parameters for each meta-segment in the test set. Next, we generate the segments, predict the activity of each segment, and compose the predicted activities. After that, we apply the meta-composer to generate global problem solutions. Based on the recent surveys of HAR in smart homes, TW, EW, and DW (probabilistic [KC14]) are the most used segmentation approaches [BNE21; Ari+22; Min+20; Wan+21a]. Therefore, our meta-decomposer selects the decomposer's hyperparameters (segmentation algorithm and its parameters) dynamically among them. Finally, we use a multi-layer perceptron model with four hidden layers (three sequential dense layers with 16 ReLU activations and batch normalizations and one layer with softmax and linear activation) to train our model to estimate the inner segmentation hyperparameters. The general idea of this algorithm that is inspired from MAML [FAL17] is shown in the Algorithm 1.

4.6.4 Performance Measurements

Evaluating the model quality is essential to compare and optimize different approaches. As described in the decomposition definition, the processing performance depends on the decomposer, composer, segments and their size (structure), and resolution. Decom-

Algorithm 1 Simple Meta-Decomposition (SWMeta)

```
Input: Training dataset $A^{train}$, Testing dataset $A^{test}$
Input: Hyperparameters $\widehat{\delta}$                    ▷ e.g., meta-segment size, $\gamma$
Output: Predicted activities for $A^{test}$
    Initialize primary model $M$ and segment decomposer $D$ using $\widehat{\delta}$.
    Generate meta-segments for training: $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_n\}$ from $A^{train}$.
    while termination criterion not met do
        Sample a batch $B$ of $J$ tasks from $\mathcal{T}$.
        for each $(Z^{train}, Z^{val}) = \mathcal{T}_j$ in $B$ do
            Optimize $D$ for best segmentation of $M$ using $Z^{train}$.
            Decompose validation data: $S = D(Z^{val})$.
            Update and train $M$ using segmented data $S$.
        end for
    end while
    Initialize meta-feature matrix $X$ and decomposer vector $y$.
    for each task $\mathcal{T}_i$ in $\mathcal{T}$ do
        Optimize $D'$ for best segmentation starting from $D$ on $\mathcal{T}_i$.
        Extract meta-features: $F =$ MetaFeatures($\mathcal{T}_i$).
        $X$.append($F$), $y$.append($D'$).
    end for
    Train model $N$ on the $(X, y)$.
    Generate meta-segments for testing: $\mathcal{T}'$ from $A^{test}$.
    Initialize predicted activities list $C$.
    for each task $\mathcal{T}_i$ in $\mathcal{T}'$ do
        Extract meta-features: $F =$ MetaFeatures($\mathcal{T}_i$).
        Predict decomposer: $D' = N(F)$.
        Decompose task: $S = D'(\mathcal{T}_i)$.
        Initialize result list $R$.
        for each segment $s$ in $S$ do
            Predict activity $k$ for segment $s$ using $M$.
            $R$.append($k$).
        end for
        $C$.append(compose($R$)).
    end for
    return metaCompose($C$).
```

posers generate segments with various sizes and structures. Thus, it is impossible to compare their quality without transforming the results into a unified space. Figure 4.11 illustrates two examples of HAR systems that use different segmentation algorithms. Activity A is not appropriately detected in half of the segments in the first segmentation method, while it is not detected in the 40 percent for the second segmentation method. Classifier metrics are frequently used to analyze the performance of HAR systems [KC14; NGC15;

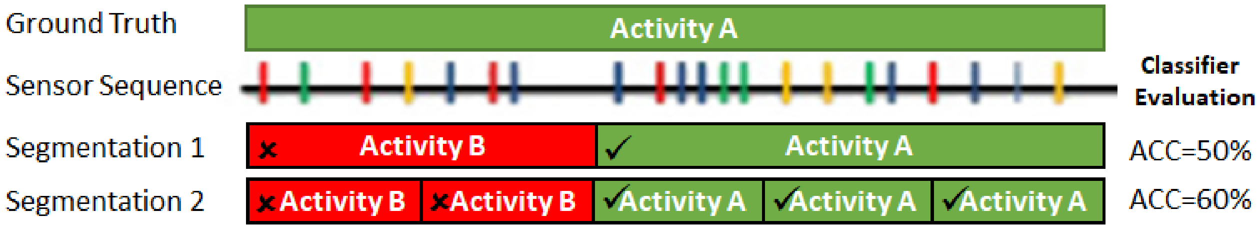


Figure 4.11: Comparison of two segmentation algorithms. One of the segments in the first method and two of those in the second method fail to detect Activity A accurately. The box shows the activity and its duration. The vertical colored lines in the sensor sequence represent the activation of various sensors at different time intervals (represented by the horizontal line). Correct predictions are denoted by '✓' while incorrect ones are denoted by '×'.

CN15; Fu11; QJE21; YFF17; De-+18; Cum+17; BNE21; Che+21b; Ber+18; Via18]. However, it may lead to biased results when comparing different segmentation approaches. For instance, in Figure 4.11, the class accuracy is 50% in the first segmentation method, while it is 60% in the second one. Obviously, their performances are similar in terms of duration. However, the aforementioned metric fails to represent the situation correctly as the various segmentation approaches can alter the problem space substantially. Moreover, activities have some properties related to their duration [Mod+22b]. For example, steady recognition of the sleeping activity is critical; otherwise, it may misleadingly present a disorder [ATE15]. However, the segmentation process may break these properties.

Therefore, after applying the composition step, we adapt the time slice (TS) based confusion matrix (CM) [KAE11] to evaluate different segmentation methods in a unified space. This TS-CM helps us to compare f-score, accuracy, recall (TPR), and other CM measures in an identical space. Figure 4.12 shows the calculation of TS-CM on an activity. To obtain

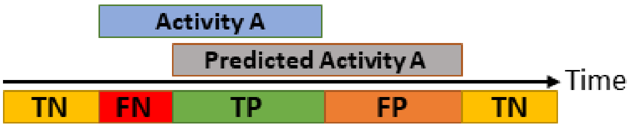


Figure 4.12: An example TS-CM calculation after composing the classifier results. (abbr. T=True, F=False, P=Positive, N=Negative)

the generalized performance, five-fold cross-validation is used for model evaluation, which is a wide approach used for model evaluation in HAR [BNE21]. It splits the dataset into five parts based on temporal occurrence. At each step, four parts are selected for training and the remaining part for testing. Then, we iterate on the parts until all the parts are used for testing. To preserve the continuous nature of events, the events of each day appear on only one part. Each configuration is repeated five times, and the average and standard deviation of its results are presented.

4.6.5 Results and Discussion

Table 4.1: Performance evaluation of methods used for segmentation. For the first three methods, the best hyperparameter is selected and shown in parentheses. Our proposed method (SWMeta) dynamically selects the appropriate segmentation method and its hyperparameter at each time. The preliminary results demonstrate that SWMeta outperforms the other methods alone. These experiments show uncontrollable biases introduced by the segmentation process can be reduced, as is expected in the formulation of meta-decomposition. In this table, w and s refer to window size and shift.

Dataset	Segmentor	TPR	F1
Home1 (11 classes)	EW (w=5, s=2)	0.65±0.04	0.42±0.03
	TW (w=30, s=20)	0.48±0.09	0.41±0.05
	DW	0.35±0.01	0.27±0.01
	SWMeta (*)	0.65±0.03	0.43±0.03
Home2 (11 classes)	EW (w=6,s=3)	0.56±0.08	0.40±0.02
	TW (w=50, s=40)	0.52±0.07	0.42±0.04
	DW	0.32±0.01	0.21±0.01
	SWMeta (*)	0.50±0.12	0.39±0.09
Aruba (11 classes)	EW (w=3, s=3)	0.59±0.05	0.34±0.04
	TW (w=60, s=50)	0.47±0.06	0.33±0.05
	DW	0.26±0.03	0.21±0.01
	SWMeta (*)	0.61±0.04	0.39±0.05
Orange4H (24 classes)	EW (w=40, s=20)	0.27±0.07	0.32±0.05
	TW (w=60, s=60)	0.32±0.08	0.35±0.06
	DW	0.30±0.01	0.34±0.01
	SWMeta (*)	0.34±0.04	0.36±0.03

The impact of the segmentation methods and their parameters are shown in Figure 4.13. It highlights the importance of the composition step in the segmentation process and shows that our segmentation reformulation as a decomposition problem improves the evaluation of the biases introduced by the segmentation step. For instance, in the initial subfigure of Fig-

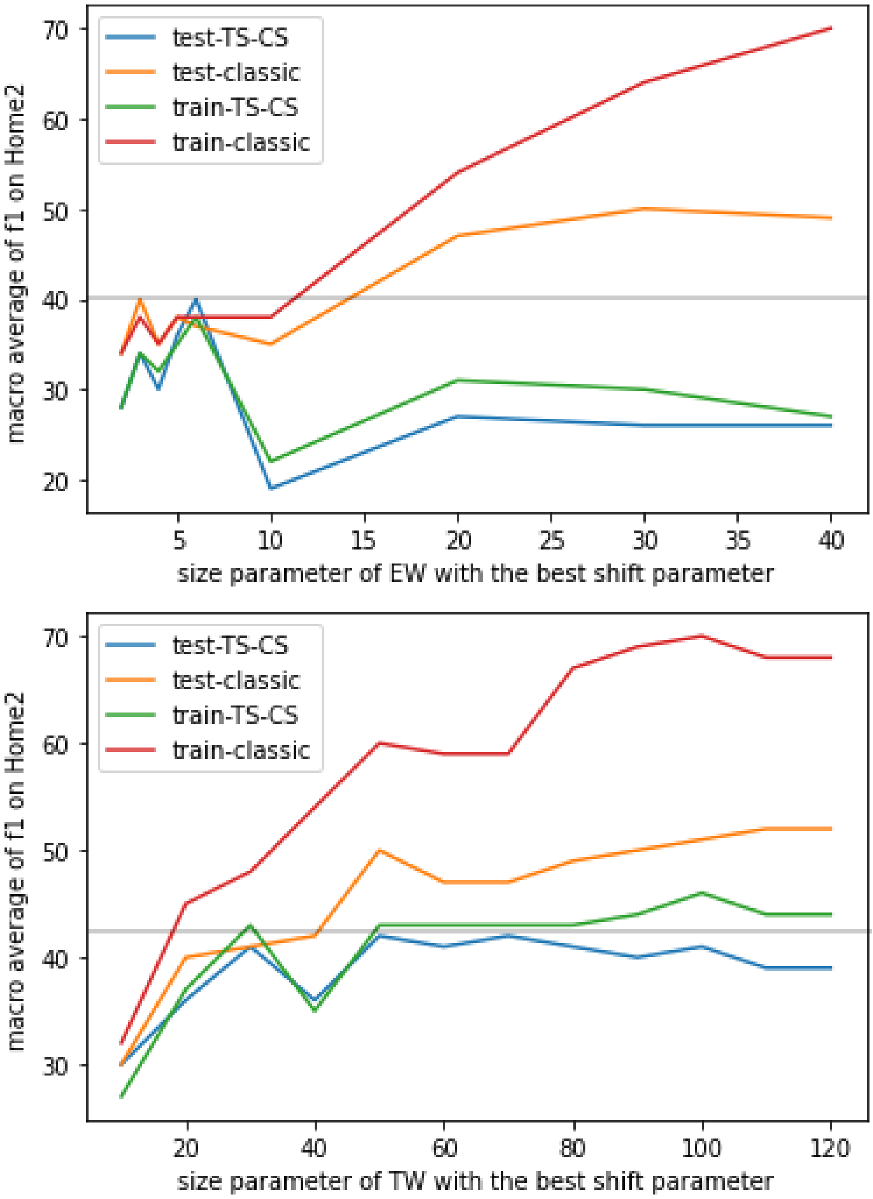


Figure 4.13: The impact of two segmentation methods (EW and TW) and their parameters (window size and the best shift parameter with that window size) on both train and a test set of the Home2 dataset with classifier metric and TS-CM metric. Interestingly, we can observe that increasing the window size increases the performance measured by the classifier metric, while when we apply the composer to the results and calculate the TS-CM measure, the performance decreases. It demonstrates once more how crucial it is to include the composition component in the segmentation process.

ure 4.13, it is noticeable that augmenting the size parameter in EW results in an enhancement in performance in the absence of the composer (classic), however, it leads to a reduction in performance after the inclusion of the composer (TS-CS). Furthermore, it is worth mentioning that employing 30 events size, as utilized in various types of research such as

[BNE21; AC19], introduces a substantial bias. The same applies to TW, thereby necessitating caution in interpreting the results. To show the usefulness of the meta-decomposition concept, in these experiments, our method learns the appropriate segmentation method each time in the training phase. Then, in the test phase, it selects the appropriate segmentation method and its hyperparameter dynamically at each time. To demonstrate the superiority of this approach, we compare it with the **best** hyperparameter of each method individually, which heavily rely on human experience or domain knowledge. The results are summarized in Table 4.1. To find this best parameter for the baseline, we conducted a grid search on each dataset. As we can observe from the table, the recommended window size of 60 seconds for TW in [BNE21; Med+18; Ham+21] introduces a bias, as the optimal hyperparameter varies between 30 and 60 seconds for different datasets. Our proposed approach dynamically selects the best segmentation method at each time among those methods and outperforms those methods individually, except for the Home 2 dataset, which contains few sensor events, thus meta-segment does not have enough data to predict the proper segmentation method and its hyperparameters.

For our inner learning model, we adapt the deep learning model proposed by Bouchabou et. al. [Bou+21a] which is described in Section 4.6.3. Assuming identical settings, our results would have been equivalent. However, we introduced three distinct differences in this experiment. First, we include the composition step, which means we rebuild the initial problem results and evaluate the results on that space instead of considering the classification performance, which is more described in Section 4.6.4. As shown in Figure 4.13, it produces noticeable disparities. Second, they assumed that the input sensor events are pre-segmented based on the activity duration; then, they applied the windowing approach to each segment before the deep-learning step, while we do not have such an assumption, which results in the inclusion of significant noise in the learning model. Third, we use the macro average while they use the weighted average, which gives more weight to the dominant activities.

These experiments show that segmentation may introduce uncontrollable biases and reduce recognition quality, while, our meta-decomposition concept lessens uncontrollable biases in segmentation by dynamically choosing the proper decomposer and its parameters based on meta-features. They also show that ML can better decompose (traditionally segment) the recognizing activities from sensor data into multiple subtasks without implicitly including knowledge about the problem domain.

4.7 Conclusion

Segmentation is often discussed without giving due attention to the composer component. Nevertheless, the composer component plays a crucial role in establishing the connection between machine learning outcomes, segments, and the overall results of the problem. This is essential for mitigating implicit bias and managing diversity in segmentation. As a result, we redefine the segmentation problem as a data decomposition problem, comprising a decomposer, resolutions, and a composer.

Furthermore, while most existing literature primarily focuses on fixed segmentation methods that heavily rely on human expertise or domain knowledge, we introduce the concept of meta-decomposition, or the process of learning how to decompose. This approach treats segmentation as a hyperparameter within the outer optimization loop, allowing for adaptive selection based on incoming data, potentially utilizing machine learning techniques. This adaptive approach helps control and thereby reduce additional biases introduced during the segmentation step.

This framework marks the initial step toward enhancing the quality of AR without implicitly incorporating human biases related to the application and dataset used in algorithms, implementations, and evaluations. Since the segmentation step alters the problem space, and different segmentation algorithms produce diverse segments, this section discusses and proposes a unified space for their evaluation. However, this alone is insufficient. It is imperative to reconsider the evaluation of such concepts, especially those extending beyond zero-dimensional (0D), which will be explored further in the next chapter by projecting the evaluation onto five high-level dimensions.

Chapter 5

Evaluation

5.1 Chapter Overview

Despite the existence of numerous evaluation metrics in the literature, there are concerns regarding these metrics. Limited understanding and interpretability of these metrics, particularly, for targets beyond 0D may result in significant bias when selecting a suitable segmentation method for a particular application [Nai+21]. Upon reviewing the literature and analyzing trends discussed in the previous chapter, it has become evident that multiple aspects are of interest to users, experts, and models. To address these concerns, in this chapter, we propose a novel formulation that replaces these extensive metrics and projects the evaluation into a high-level latent space. This formulation makes the evaluation easily comprehensible and interpretable by both machines and experts.

Considering a set of n test cases, $T = \{t_1, ..., t_n\}$, the goal is to provide a comprehensive evaluation for each test case. In this chapter, we concentrate on individually evaluating each sample, without delving into the various aggregation techniques, such as micro-average, macro-average, weighted average, and ranked average, which can be employed to obtain a global evaluation. A test case may consist of multiple targets, denoted as $t = \{c_1, c_2...\}$. For instance, in the context of AR, there could be several instances of moving to the toilet within a single episode, or in the case of MIS, multiple tumor spots may be present in a single CT scan, each treated as an individual target.

The evaluation of algorithms with more than 0D requires consideration of various properties due to the interdependencies inside the targets (such as the dependency of voxels in the context of MIS). The significance of these properties varies depending on the application or even at different stages within the same application. For instance, the detection of tumors in their early stages is more critical compared to their size, while changes in size and shape are more important in evaluating treatment response. Therefore, we propose a novel evaluation method that defines multiple properties and measurements (based on the well-known TP, FP, and FN). These measurements can be aggregated (e.g., in a weighted manner) to produce a scalar value or used collectively as multi-objective metrics. Furthermore, our evaluation method is modular, allowing for the straightforward inclusion of new measurements for additional properties. The objective is to project the evaluation onto an interpretable latent space, which can provide valuable insights for domain experts to make informed decisions regarding the application and stage in question.

Our proposed evaluation method extends point-based metrics to handle partial matches between ground truth and predicted segments. In contrast to point-based metrics, where each voxel is either correctly predicted or not (i.e., the value of TP, FP, or FN for each instance is either 0 or 1), our method generalizes these terms for more than 0D data by allowing partial value to each target. This enables a more nuanced and detailed evaluation of segmentation performance, providing insights into the situation of matching between predicted and ground truth segments.

In the following sections, we present the key properties of evaluating more than 0D targets, particularly in MIS, AR, and SED drawn from state-of-the-art studies. We also introduce the formulas for measuring these properties.

5.2 Beyond Point-based Metrics: Related Works

The evaluation of system performance has been a longstanding concern for computer science and engineering researchers, system designers, operators, and end-users [HLR00]. While it is essential to measure the performance of intelligent systems to create reliable and cost-effective solutions, using consistent settings for comparing different systems is equally important [Mad+09]. Despite the availability of several metric formulations, the interpretation of real-world data remains a significant challenge [Pan15; Nai+21].

Evaluation metrics can encompass more than just point-based measurements and may extend to range-based assessments for 1D, 2D, and 3D data. In this case, in contrast to point-based targets that are either correct or incorrect, the targets can be simultaneously correct and incorrect. In this section, we explore the evaluation metrics used in a wide range of applications including AR, SED, and MIS to analyze the trends and the methods for evaluating these targets.

5.2.1 Evaluation Methods in Activity Recognition (1D)

AR is the process of automatically identifying and categorizing human activities based on sensor data from wearable devices or other sources. Human activity can be defined as a set of actions performed by an individual over a period of time and is typically associated with a specific activity label. In this section, we consider an activity that is characterized by its duration (start and end) and its label.

AR is expected to be a core component in numerous future IoT applications such as healthcare, smart homes, and security [QPM21; Per+14; CN15]. Therefore, evaluating the effectiveness of different AR algorithms is essential. Some metrics such as accuracy, observing the TPR against PRC are common metrics that are easy to understand and interpret even by non-experts. These metrics are well-used for discrete instances and pre-segmented data sequences [CN15]; where, a predicted instance is either correct or incorrect. However, a predicted target in AR can be correct in one period and incor-

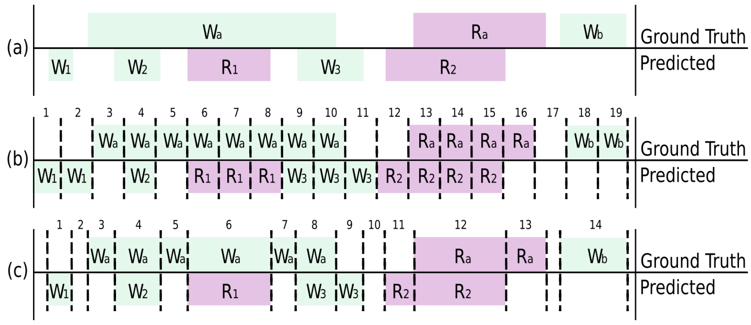


Figure 5.1: Differences between event, frame, and segment analysis. W and R represents walking and running activity [Min+06]

rect or partially correct in another [Tat+18]. Figure 1.1 illustrates the traditional evaluation metrics including TP, FP, and FN for the point instances, which are either correct or incorrect, and their unsuitability for AR where the targets have duration, which can be partially correct and partially incorrect [Tat+18]. However, traditional evaluation methods on the discretized range instances are often assumed to reflect the overall performance of the system [Per+14; Bil+20; QPM21; CN15]. This assumption neglects practical scenarios and may misleadingly present convincible results. Despite the importance of evaluating range-based targets, it is not well-developed. Still, there is no universally accepted formula for evaluating the effectiveness of these systems [Ser+20].

Evaluating the performance of these systems is usually performed by comparing predictions with the ground-truths [MHV16]. It can also be viewed as the matching of two time-series. This can be challenging due to imperfect time boundaries of ground truth labels and unclear distinctions between some targets [CN15]. For example, the transition between walking and running activity is subjective and hard to enforce in practice. Therefore, some decision functions accommodate offsets using *ambiguous range* [Hwa+19], *fuzzy event boundaries* [NIS04], time series matching techniques (such as *dynamic time warping*, *longest common sub-sequences* [Fu11]), or *categorical probability distribution* [HK11]; however, these techniques fail to distinguish sources of errors (e.g., fragmentation) [War+11]. Common approaches to evaluate AR systems include time-frame, event-based, and classifier performance [Min+06; KAE11; QPM21; MHV16], which are detailed in the following:

Time-frame based methods divide the entire range into atomic units of fixed period intervals, allowing for easy comparison between algorithms [Min+06; KAE11]. Each frame is independent of both the ground truth and prediction and can be classified as TP, TN, FP, and FN. The duration of this atomic unit follows the timescale of the domain such as one second. The method is illustrated in the second subfigure of Figure 5.1.

These techniques typically decrease the time resolution to an atomic unit, such as one second, with the aim of accommodating some level of misalignment between the reference and prediction [MHV16]. However, this approach can lead to an increase in false positives when a predicted frame is partially incorrect but the ground truth frame is negative, which is counterproductive to the intended goal of the method. On the other hand, the hypothesis in the segment method tends to decrease the time resolution, potentially leading to an erroneous influence on the final outcome. For instance, the fragmentation that occurs within a frame may not be detectable. Furthermore, it should be noted that different types of targets may have varying requirements, with one-second serving as an appropriate threshold for some classes while being inappropriate for others.

Segment Based Methods define a segment as the maximum interval in the ground truth and the predictions that remain constant. Therefore, each segment has a different duration but there are no ambiguities for the boundaries of each interval, which is the case with frame-based methods [Min+06]. Therefore, each segment can be classified as TP, TN, FP, and FN. The value of these categories is similar to frame-based methods when the frame's size tends to zero.

Event Based Methods consider individual target occurrences as the basic atomic units for comparison, irrespective of their duration. Event-based methods are essential to consider alongside time-frame methods [War+11]. In these methods, the occurrence time and the order of the events can be important factors [Min+06]. When time is important, the correctness of the prediction needs to be defined. Common decision functions employed in event-based methods include majority, mid-point, etc. [War+06; Fer+21; Bil+20; MHV16;

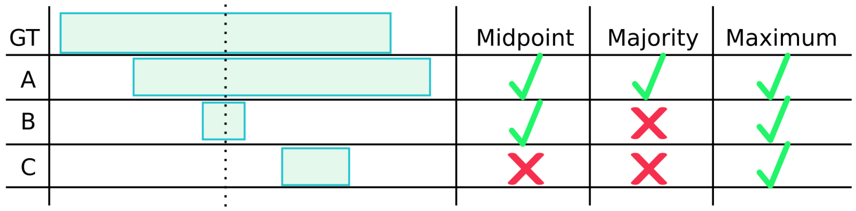


Figure 5.2: Common decision functions for the correctness of an event, such as midpoint, majority, and maximum. The horizontal rows represent the ground truth and different predictions. The vertical dashed line represents the midpoint of the ground truth.

Lea+17]. For example, when a prediction covers the majority of ground truth, it is classified as a TP. On the other hand, if it does not cover the majority, the segment is considered a FP. Some of these decision functions are shown in Figure 5.2. These decision functions help to evaluate the performance of a model using traditionally well-defined metrics for traditional machine learning evaluations.

Minnen's Metrics The interpretation of errors varies across different applications. Therefore, Minnen's metrics classify each frame/segment's errors into *insertion*, *overfill*, and *merge* as sources of FP errors, and *deletion*, *substitutions*, *underfill*, and *fragmentation* as sources of FN errors [Min+06] which are visualized the errors using the Error Division Diagram. These errors are defined as follows:

- **FP-insertion** detection of an activity when nothing actually happened.
- **FP-overfill** time before and after the occurrence of an activity that is incorrectly identified as part of the activity.
- **FP-merge** covering multiple ground truth by a single prediction
- **FN-deletion** failure to detect a target,
- **FN-substitutions** wrongly detected with another class
- **FN-underfill** not detected duration at the beginning and end of the activity,
- **FN-fragmentation** detecting a ground truth by multiple predictions in a fragmented way

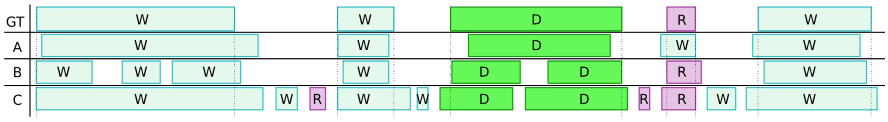


Figure 5.3: Sample ground truth with multiple classes and three different predictions. All these predictions have similar accuracy but the types of their errors are different [Min+06].

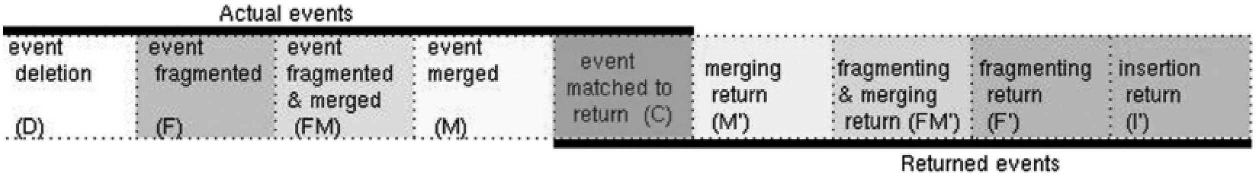


Figure 5.4: Event Analysis Diagram (EAD) showing the different event error types [War+11].

Ward's Metrics Ward et al. [War+11] argue that an evaluation metric should do more than just assign a grade to a system. It should also characterize the performance and quantify the strengths and weaknesses of the system, providing insight for designers to improve it. They emphasize the importance of considering event-based methods alongside time-frame methods. Ward's metrics specify eight distinct event error types. Four of them apply to ground truth events: deletions (D), fragmented (F), fragmented and merged (FM), and merged (M). The remaining four apply to predicted events: merging (M'), fragmenting and merging (FM'), fragmenting (F'), and insertions (I'). These errors, alongside with correct events (C), are shown in a single diagram (Figure 5.4), known as the Event Analysis Diagram (EAD). Each of these errors can be normalized depending on the size of predictions and ground truth, making it easier to compare different databases.

Compact Error Metrics Kasteren et al. [KAE11] proposed a compact error metric based on time frame analysis. This metric aims to provide a concise summary of the performance of an event detection system by offering a macro average of each event type with respect to the ground truth or predictions. The average column in the compact error metric table is calculated based on the F1 score, which considers both precision and recall for each event type.

From the behavior analysis perspective, evaluating each activity needs a different evaluation method [ATE15]. e.g., duration-sensitive activities need to be evaluated differently from frequency-sensitive ones. Recognizing the sleeping activity in a fragmented manner shows a disorder, while an imprecise alignment of the beginning and end of an activity does not [ATE15].

Timeliness is another metric used for online and real-time prediction [RK13]. It is defined as the duration of continuous correct prediction of an activity without switching to an inaccurate prediction. To compare different AR algorithms in a similar situation, a competition is held and time frame f_1-score, recognition delay, installation complexity, user acceptance, and interoperability are used as the evaluation criteria [Gjo+15].

Therefore, an expert must perform a time-consuming analysis of these massive and heterogeneous diagrams, matrices, and information. Accordingly, combining them as a scalar metric is complex. Besides, these approaches also consider the total duration of positional errors and do not provide an event-based tunable model for it.

5.2.2 Evaluation Methods in Sound Event Detection (SED) (1D)

A Sound Event Detection (SED) system recognizes sound events in audio tracks. It is a developing research field from both academia and industry due to its potential applications in healthcare, medical telemonitoring, surveillance, smart home, monitoring, security, audio content-based searching, etc. [Bil+20; MHV16; CC20].

The time interval included in sound events is one of the most important dimensions in evaluating SED systems. These systems should determine the occurrence time interval of sound events in addition to their sound classes. Moreover, sound events can simultaneously happen (e.g., opening door sound events during a speech event) [MHV16]. Evaluating the performance of SED systems is often done by comparing their predictions with the references [MHV16]. One of the first evaluation methods was defined in CLEAR[1] 2006 challenge named acoustic event error rate [Sti+07]. It marks a reference event as correctly identified when the temporal center of the predicted event is inside it [Sti+07]. It also defines

[1] CLassification of Events, Activities and Relationships

insertion, deletion, and substitution errors. The metric was ambiguous in some cases, e.g., whenever a part of a reference is well detected, and the other part has a substitution error [Tem+09]. In CLEAR 2007 challenge, recall, precision, and f-score (considering the above definition for correct prediction) were used; however, they redefined the acoustic event error rate by using frame-based methods [Sti+08; MHV16; Sto+15]. Frame-based (segment) methods take fixed-duration intervals (e.g., 10 ms) as the basic atomic unit. It facilitates comparing different algorithms since each frame is independent of both the references and predictions [KAE11]. In the DCASE[2] 2013 challenge [Sto+15], the frame-based error rate, precision, recall, f-score, and collar-based method were used. In collar-based methods, a reference is considered as correctly detected if the beginning (onset), the ending (offset) or both are within a specific tolerance (e.g., 200 ms). This tolerance is necessary because of the inexact labeling of the data [MHV16]. PSDS[3] is a recent method for SED evaluation which is proposed for more robust defining of TP, FN, and FP by considering the intersection rate based on references and predictions [Bil+20]. It overcomes the dependency of the evaluation on the sound event's duration and provides robustness to labeling subjectivity [Fer+21]. Researchers in [Ton+20] explore the inequality of missing events in different scenarios. They break down FP into related and unrelated by considering the scene and sound event relation; then, they give a double penalty to unrelated FPs in calculating F1. The IEEE AASP[4] challenges on detection and classification of acoustic scenes and events [Sto+15] highlights the need for an appropriate metric. Still, no universally accepted metric is defined [Ser+20]. The previously mentioned methods consider few situations of errors and have some certain deficiencies. e.g., they are highly biased by their assumptions [Fer+21] and may misleadingly present convincible results.

Still, researchers mainly used collar, segment (time-frame based), and PSDS (polyphonic sound detection score) methods in SED [MHV16; Bil+20]. However, they cannot show the different sources of errors.

[2] Detection and Classification of Acoustic Scenes and Events
[3] Polyphonic Sound Detection Score
[4] Audio and Acoustic Signal Processing

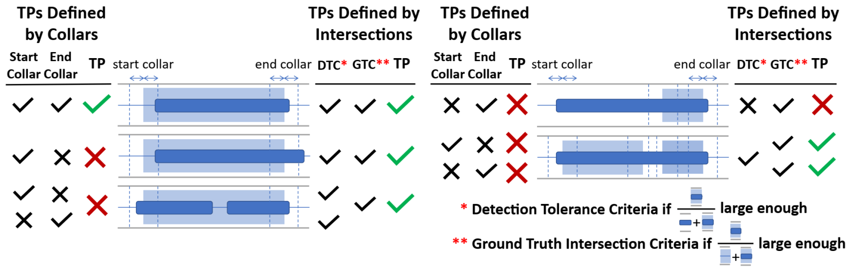


Figure 5.5: Collar and PSDS decision functions.

5.2.3 Evaluation Methods in Other Applications in (1D)

In video action detection [Awa+21], anomaly detection [Tat+18], and video abnormal event detection [Ion+19], etc., targets are also durative. The National Institute of Standards and Technology (NIST) developed a challenge for detecting activities in video (ActEV) [Awa+21]. It first used false alarm rate (instance-based) and missed detection probability (instance-based) as evaluation metrics. However, in 2019, it used the time-frame method for calculating false alarm rates [Awa+21]. Other metrics in abnormal event detection in the video are false rejection rate, equal error rate, decidability index, receiver operating characteristic curves, and area under the curve [DB15; Ion+19]. However, the equal error rate can be misleading in the anomaly detection setting [Lu+19]. Numenta anomaly benchmark [LA15] is designed to evaluate different anomaly detection algorithms. It uses a scaled sigmoidal scoring function for the relative position of each detection; however, it ignores fragmented predictions. To resolve previously mentioned issues, researchers in [Tat+18] redefine precision and recall for time series (particularly in anomaly detection). They need some functions to be explicitly defined for a given application. Those functions are: γ (to consider fragmented events), δ (to consider the positional relation between PE and GTE), $\mathrm{overlap}$ (the rate of the correctly detected events (e.g., $\mathrm{overlap}(x, y, \delta()) = \mathcal{T}(x \cap y)/\mathcal{T}(x)$), and α which is a coefficient. They are formulated in Equation (5.1).

$$\mathrm{exist}(e, X) = [e \cap X \neq \emptyset], \qquad \mathrm{score}(e, X) = \gamma(e, X) \times \sum_{x \in X} \mathrm{overlap}(e, e \cap x, \delta()), \tag{5.1}$$

$$\mathrm{Recall} = \frac{1}{|R|} \sum_{r \in R} \alpha \times \mathrm{exist}(r, P) + (1 - \alpha) \times \mathrm{score}(r, P), \qquad \mathrm{Precision} = \frac{1}{|P|} \sum_{p \in P} \mathrm{score}(p, R)$$

Analysis of [Tat+18] Despite its potential to enhance the evaluation process, the proposed method by [Tat+18] exhibits some limitations, which are explored in the following discussion:

1. It disregards the coefficient α in the precision calculation. Consequently, the $overlap$ function is given inconsistent weights in the calculation of precision and recall. This inconsistency can lead to a misinterpretation of the results, rendering the method unsuitable for complementary use, such as in the calculation of F1.

2. Fragmented predictions can result in a significantly positive PRC score. For instance, in Figure 5.6, the PRC of (a) is considerably higher than that of (b) due to the presence of fragmented predictions. This situation can also occur for TPR.

3. To account for the impact of duration, the proposed method normalizes the duration of events. Specifically, precision is calculated as $\underset{p\in P}{avg}(\frac{tp}{T(p)})$, and recall is calculated as $\underset{r\in R}{avg}(\frac{tp}{T(r)})$. Although this normalization appears to work well for a single prediction and ground truth, it yields different values for TP in TPR and PRC when applied to the entire dataset. Thus, they are not calculated in a similar mathematical model and cannot be used as complementary measures, such as in the calculation of F1. Equation (5.2) presents these calculations for Figure 5.6 (d).

$$\begin{aligned}\text{Precision} &= \frac{\frac{\mathrm{TP}_1}{\mathrm{P}_1}+\frac{\mathrm{TP}_2}{\mathrm{P}_2}}{1+1} = \frac{\Sigma\text{normalized TPs based on PEs}}{\Sigma\text{normalized PEs}} \\ \text{Recall} &= \frac{\frac{\mathrm{TP}_1}{\mathrm{R}_1}+\frac{\mathrm{TP}_2}{\mathrm{R}_2}+\frac{0}{\mathrm{R}_3}}{1+1+1} = \frac{\Sigma\text{normalized TPs based on GTEs}}{\Sigma\text{normalized GTEs}}\end{aligned} \tag{5.2}$$

4. Defining an appropriate cardinality function for the proposed method is complex. Additionally, it is challenging to adjust and fine-tune the formula because the dependencies between cardinality, position, and overlap are not clearly defined [Hwa+19]. For instance, in Figure 5.6 (c), the first and second ground truths have the same TPR of 0.33 when using $\gamma(e, X) = |e \cap X|^{-1}$ as suggested by the authors. A similar situation can occur when calculating PRC for merged predictions.

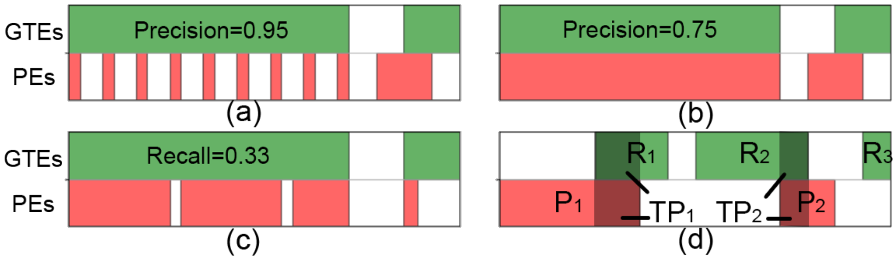


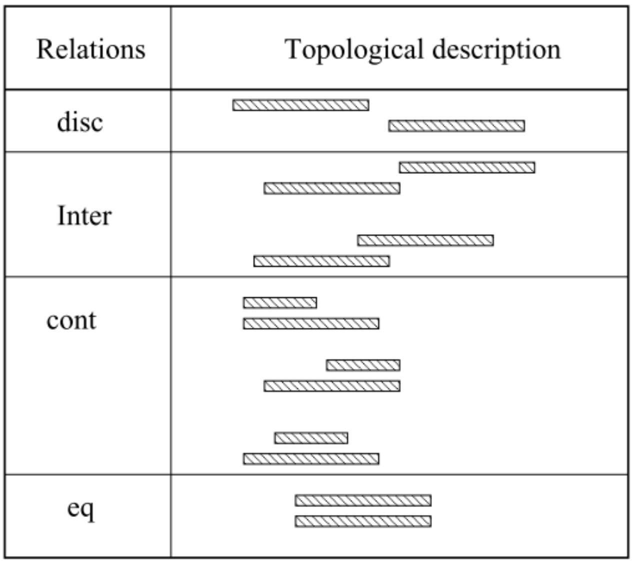
Figure 5.6: Example activities that help to explain the drawbacks in [Tat+18].

5. The [Tat+18] approach cannot be applied to duration-sensitive activities [ATE15].

6. Incorporating a new property, such as total duration, into this metric is not straightforward.

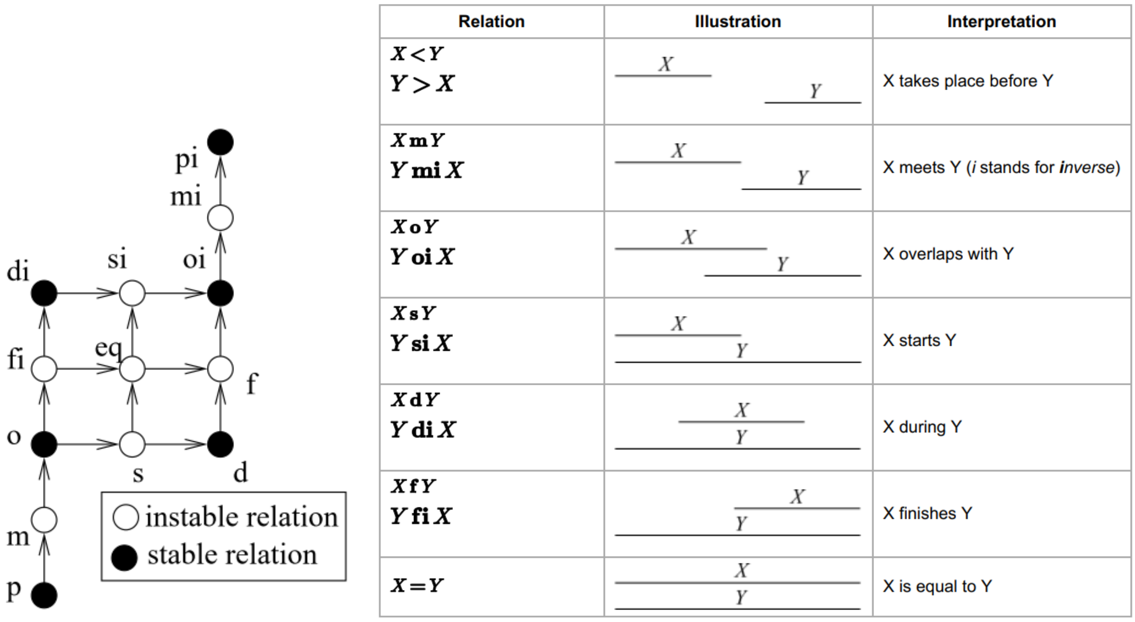
5.2.4 The Relations of Intervals

For the evaluation purpose, our goal is to calculate the error between prediction and ground truth. Therefore, when the targets are points, a prediction is either correct or incorrect. However, for interval targets, it is more complicated. Based on Allen's interval algebra, the number of different relations between two chains of intervals is at least exponential in the number of intervals, e.g., it is 13 for two intervals, and 8989 for two sequences of three intervals [Sch07]. Figure 5.7 shows the relations between two intervals. Osmani [Osm02] groups Allen's relations into four categories for a simpler representation of their learning problem. Those relations are $disconnect(disc)$ which corresponds to Allen's $before$ relation; $intersect(inter)$ which corresponds to Allen's $meets$ and $overlaps$ relations, $contain(cont)$ which groups together Allen's $finish$, $start$, and $during$ relations and $equality(eq)$ which shows the equality of two intervals. The opposite of the relations $disc$, $inter$, and $cont$ are $disc^{-1}$, $inter^{-1}$, and $cont^{-1}$. They are displayed in Figure 5.8.



Relations	Topological description
disc	
Inter	
cont	
eq	

Figure 5.8: Simplified relations between two intervals [Osm02]



Relation	Illustration	Interpretation
$X < Y$ $Y > X$	X Y	X takes place before Y
$X \mathbf{m} Y$ $Y \mathbf{mi} X$	X Y	X meets Y (*i* stands for *inverse*)
$X \mathbf{o} Y$ $Y \mathbf{oi} X$	X Y	X overlaps with Y
$X \mathbf{s} Y$ $Y \mathbf{si} X$	X Y	X starts Y
$X \mathbf{d} Y$ $Y \mathbf{di} X$	X Y	X during Y
$X \mathbf{f} Y$ $Y \mathbf{fi} X$	X Y	X finishes Y
$X = Y$	X Y	X is equal to Y

Figure 5.7: Relations in Allen's interval algebra. The abbreviations "p", "m", "o", "s", "d", "f", "eq", and "i" refer to the relations "preceding", "meet", "overlap", "start", "during", "finish", "equal", and "inverse", respectively[Osm02].

When only one target of the same class can happen at each time, we can enumerate relations between more than two intervals as $disc$, $inter$, $cont$, $merge$, eq, and their inverse relations. As shown in Figure 5.9, the possible relations between one interval and multiple non-overlapping intervals include at most two "disc" and "inter", and several "cont". Considering multiple non-overlapping intervals on both sides that are not disconnected, will only

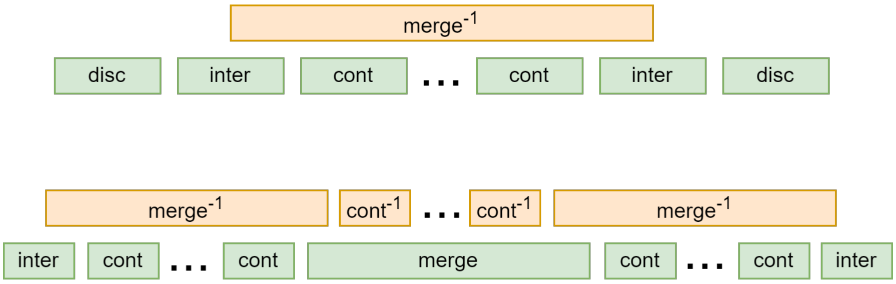


Figure 5.9: Relations between multiple non-overlapping intervals

increase the number of "cont" relations.

5.2.5 Evaluation Methods in Medical Image Segmentation (2D and 3D)

Deep learning-based MIS has gained considerable traction in recent years [Dev+21; Asg+21; Mal+22; Luo+22]. MIS is recognized as an NP-hard problem [Asa+01] requiring heuristics for resolution. This makes the performance metrics crucial for assisting clinicians and system designers in selecting the appropriate models for the clinical problem [THT14]. While numerous studies have demonstrated that these models exhibit robust predictive capabilities, achieving results close to those of clinicians [Mal+22], recent studies highlight the existence of statistical biases in the assessment method used to evaluate these models due to the used metrics [MSK22; Rei+21].

Common evaluation metrics, such as TPR, PRC, DC or F1, IoU, also called Jaccard Index, HD, ASSD, and NSD, are widely used to assess the performance of MIS systems [Che+23; Hou+21; Eel+20; Kum+22; MSK22; Dev+21; Che+22a; Ren+22; TC22; Tar+21; Abd+23; Ma+21; Luo+22; Ker+21; TH15; Mal+22; Son+22; WZ20; KHS22]. Evaluating MIS often involves simplifying the multidimensional targets to point-based ones [Kum+22; MSK22; Ma+21]. However, voxels (pixels) in MIS are interdependent, and neglecting these dependencies can lead to incomplete assessments of proposed models and their clinical outcomes [Kim+15]. For instance, early identification of tumoral regions necessitates detecting tumor presence regardless of size, while treatment response evaluations require monitoring volume changes [Tia+21]. Additionally, the presence of dominant lesions might result in overlooking smaller lesions during assessment [TH15; Kim+15]. The quality assessment of an MIS system extends beyond these factors; evaluations should also consider the uniformity or fragmentation of predictions and the preservation of segment shapes, which assist experts and models in identifying tumor types [Tia+21]. Furthermore, predicted segments in MIS can be partially correct and incorrect simultaneously, unlike point-based predictions that are either entirely correct or incorrect. For example, medical treatment outcomes can vary significantly even if two tumor segments have similar measures of the mentioned metrics such as DC, and HD [Kim+15]. Very recent studies highlight the need for reliable model performance assessments in MIS, as well as the presence of statistical biases in the assessment of both binary and multi-class problems [MSK22; Rei+21; Nai+21; WWZ20; Kum+17; Kim+15; TH15; GSC22; Hoe+22; Rei+22; Koe+22; Jav+22; Lee+22; Fag+22; BJ22]. In conclusion, a more appropriate method to evaluate the performance of MIS techniques involves considering their various aspects.

Assessment of Medical Image Segmentation is a crucial task in medical image analysis, where it should evaluate the correctness of the predicted labels as well as the boundaries of the targets [Nai+21]. In this context, the input image can be represented as a 2D or 3D matrix, where each element, also known as a pixel or voxel, may have one or more characteristics, such as radio density.

The performance of medical image segmentation techniques can be evaluated using a variety of metrics and typically compared against expert-extracted ground truth [Ker+21]. The selection of the best evaluation metric for image segmentation depends on the specific application and the characteristics of the images being segmented.

According to [TH15], the selection of a suitable algorithm for MIS requires the evaluation of various properties of these systems. These properties include accounting for outliers, evaluating small and large segments, handling different segment shapes and complex boundaries, considering TPR for low-density segments, and assessing the volume and alignment of segments. For example, over-segmenting a tissue can often lead to unfortunate consequences, whereas missing a tumor can easily have catastrophic consequences [Kim+15].

Moreover, four types of basic errors in segmentation based on human visual tolerance are inside hole, border hole, added background, and added region [WWZ20; SNL13].

In addition to computer science studies, radiomics analysis has shown that accurately segmenting tumors requires the detection of approximate tumor locations as a critical first step [Li+19b; Tia+21]. Therefore, an effective segmentation metric should consider this property. Furthermore, tumor morphology (that characterizes tumor margin) is often the most challenging aspect of detection [Tia+21]. Other important characteristics for clinical treatments include tumor size (dimensions), shape (3-D geometry), and uniformity (irregular or ellipsoid shape, sphericity, lobulation, speculation, roughness, the longest and shortest diameters, margin sharpness, surface area, volume, and surface-area-to-volume ratio) [Li+19b; Tha+18; Tia+21; SJ19]. For small tumors in which even slight changes can significantly impact the radiomics measures, robust segmentation is critical [Li+19b]. As a result, in addition to dominant tumors, small tumors should also be considered in the metric.

Moreover, in MIS, voxel size can vary greatly between scans, and ignoring voxel size can lead to inaccurate segmentation evaluation [TH15]. For example, the slice thickness of CT scans can range from 0.5 mm to 5 mm [Li+19b]. Therefore, it is important to consider voxel size in the metric used for evaluating segmentation algorithms, while it is missing in some state-of-the-art analyses [TH15].

The following subsections present a concise yet comprehensive overview of prevalent evaluation techniques, their limitations, and contemporary evaluation methods by scholars.

Common Evaluation Methods

In general, TP (resp. TN) is used to indicate the correct positive (resp. negative) predictions. Similarly, for incorrect predictions, there are two situations. FN (resp. FP) refers to the positive (resp. negative) instances predicted wrongly as negative (resp. positive) [Pop+07]. Voxel-wise metrics such as Acc, PRC, TPR, Fβ, IoU, and DC [SR22; Tar+21; Eel+20; Ma+21; Luo+22; Ker+21; MS19; Aer10; Sch+20; Tag+19] are commonly used to evaluate MIS systems. These methods treat each voxel as an independent instance, enabling the straightforward classification of predicted voxels as either correct or incorrect. The primary components of these metrics, including TP, FP, FN, and TN, are formulated in appendix Equation (A.1). The IoU, also known as the Jaccard metric [Asg+21; Luo+22] measures the intersection between the predicted and the ground truth over their union and is formulated in Equation (A.2). The MIoU provides an overall measure of the segmentation accuracy across all classes, by macro averaging the IoU values for each class, while FWIoU gives more weight to the IoU values of the more frequent classes, making it useful for imbalanced datasets [Zhe+21; Luo+22]. The Acc metric is calculated as the ratio of correctly predicted instances to the total number of instances, regardless of their class. On the other hand, the PRC calculates the ratio of correctly predicted positive instances to the total number of positive predictions. In other words, it measures the accuracy of positive predictions. The TPR, also known as sensitivity or recall, is the ratio of correctly predicted positive instances to the total number of actual positive instances. It measures the completeness of positive predictions [Pop+07; MSK22; Kum+22; TH15]. The Acc, PRC, TPR, and their weighted harmonic means (Fβ) for binary classes are formulated in Equation (A.3). When $\beta = 1$ in Fβ (F1), the resulting metric is known as the Dice Similarity coefficient (DC), which is shown in Equation (A.4). The DC measures the overlap between the predicted and the ground truth segmentation masks by considering both the FP and FN rates [Dic45]. It ranges between 0 (no overlap) and 1 (perfect overlap). Receiver Operating Characteristic (ROC) curve is generated by plotting the TPR against the False Positive Rate (FPR) for different classification thresholds, while the area under the ROC curve (AUC) measures the area beneath the ROC curve [TH15].

As explained, point-based evaluation methods may disrupt the relationships between voxels. To address this problem, some studies have expanded the point-based evaluation method by introducing a threshold to determine correct or incorrect predictions. By adjusting this threshold, we can gain insights into the performance of the prediction system. For example, P@X, and R@X metrics indicate PRC and TPR of a segmented image, with correct predictions being defined as those with IoU scores above X. The average of those values (AP and AR) is also commonly used in image segmentation [BZB22; Lin+14; Lea+17; FVS22]. However, this thresholding approach has limitations. It only considers whether a predicted instance overlaps with the ground truth above or below the threshold and does not consider the actual situation of overlap. This means that the evaluation does not differentiate between instances that are over-segmented, under-segmented, or whether they are fragmented into multiple segments, or merged with other segments.

While metrics based on voxels provide a straightforward approach to evaluating the segmentation performance, they fail to capture the spatial dependency and consistency of the segmented regions [Li+19b]. To address this limitation, HD is used in MIS evaluation [Ayd+22]. It measures the distance between the boundaries of the ground truth and predicted segmentation [Ker+21; TH15; MSK22]. Three variations of HD, namely HD_{max}, HD_{95}, and HD_{mean}, are typically used to represent the maximum, 95th percentile, and average distances between the boundary points of the ground truth and the prediction. While HD_{max} is sensitive to outliers and can be influenced by a few points with very large distances, HD_{95} and HD_{mean} are less influenced by noise and outliers. HD calculates the distance without considering the segment size; therefore, the balanced HD metric is proposed in [Ayd+22] to normalize the distance. Other metrics such as Average Surface Distance (ASD), ASSD, Mean Absolute Distance (MAD), Mahalanobis Distance, Manifold Distance, and Chamfer Distance are also commonly used [Pan+22; Qiu+22; Sch+19; TH15]. While these distance-based metrics can provide useful information about the segmentation accuracy, they have limitations in providing interpretable information on the source of errors and may not be sensitive to the topology of the segmented object.

The points in the boundary of the ground truth and prediction should not be considered equally [Nik+21]. For instance, misidentifying a group of points in close proximity is more severe than misidentifying the same number of points that are distributed sparsely along the boundary. To address this issue, authors of [Nik+21] proposed NSD that allows a certain tolerance on the boundaries of the ground truth and the prediction. It is defined in the appendix Equation (A.5). However, this metric also has limitations mentioned for point-based methods.

The incompatibility between the metrics and the expected clinical outcome is another important issue. As it is demonstrated in [Kim+15], clinical outcomes can be highly different for two predictions even when common metrics like DC and HD provide similar results. The authors of [Kim+15] propose a medical similarity index (MSI) that is compliant with the clinical requirements. They use local distance [Kim+12] between prediction and ground truth aligned by their center to determine the dissimilarity. Then, they employ an asymmetric Gaussian function to impose more penalties on over- or under-segmentation. However, this metric needs statistically significant data for all clinical cases [Ahn+19] and has limited correlation with visual assessment [Nai+21]. Furthermore, this metric imposed that each segment has one single center which is not applicable for all image segmentation cases such as segmentation of irregularly shaped tumors or lesions [Tia+21; Bur+04; Rei+21].

A good evaluation criterion should penalize both object-level and pixel-level errors, including missed detection, false detection, under-segmentation, and over-segmentation [Kum+17]. To this end, a combined metric called aggregated Jaccard index (AJI) for nuclear segmentation, which takes into account both object and pixel level errors, is introduced [Kum+17]. However, this metric has limitations in providing an overview of different sources of errors and interpretable information on different segmentation approaches. For instance, Kromp et. al. [Kro+21] were unable to analyze their findings using this metric.

The volume of segments is another useful property. However, comparing the total volume of the prediction and ground truth without considering their alignment (for instance, Volume Similarity (VS) and Relative Volume Difference) may yield a perfect similarity score even if there is no actual overlap between the prediction and ground truth, which limits its usefulness in evaluating segmentation accuracy, particularly when precise overlap information is required [TH15; Rei+21; Nai+21].

Without being exhaustive, there are numerous other metrics have been exploited to evaluate the Medical Image Segmentation (MIS) performance in the state-of-the-art including mutual information, interclass correlation, variation of information, probabilistic distance, global consistency error, Cohen kappa coefficient, rand index (and its adjusted and probabilistic variation), Segmentation covering, C-Factor, bookmaker informedness, relative volume difference, bidirectional local distance, objective quality metric, Root Mean Square Symmetric Contour Distance (RMSD), Number Of generated Proposals, and deformation vector field [MSK22; Rei+21; Nai+21; WWZ20; Kum+17; Kim+15; TH15; GSC22; TC22; Rue+14]. Notably, some of these metrics, such as the DC, IoU, rand index, Cohen kappa coefficient, interclass correlation, probabilistic distance, and adjusted rand index, have been found to closely approximate each other both relatively and absolutely [Eel+20; GSC22; Kum+17; Nai+21; Bou+22]. Despite the wide range of metrics available, assessing the performance of segmentation algorithms in MIS remains a challenging task [WWZ20; Nai+21]. The limited understanding and interpretability of these metrics can lead to significant bias in selecting an appropriate segmentation method for a particular application [Nai+21].

5.2.6 Analysis of Current Trends

In addition to the aforementioned review of evaluation methods in the state-of-the-art literature, we conduct a comprehensive analysis of the evaluation metrics employed by researchers to obtain an overview of the current trends in segmentation evaluation methods. For instance, the papers about image segmentation in the year 2022 from two randomly highly-ranked conferences in image processing and general AI are selected. A total of 200 papers were analyzed: 40 papers from AAAI Conference on Artificial Intelligence and 160 papers from Computer Vision and Pattern Recognition (CVPR). Notably, around 60% of the analyzed papers (28 at AAAI and 90 at CVPR) employed IoU either alone or in conjunction with other evaluation methods, indicating its widespread use. This information is summarized in Table 5.1. Our analysis provides insight into the prevalence of voxel-based methods (groups I, II, III) in evaluating MIS. However, the limitations of these methods have driven researchers to explore alternative methods, such as boundary-based methods (IV). Moreover, distance-based methods (V) rely on actual distances, which may make them less effective in evaluating small segments when large segments are present. Some works

Table 5.1: Evaluation metrics for segmentation evaluation used by the authors in two randomly selected top A* conferences as an example. This table demonstrates that voxel-based approaches are commonly used. Some references have been omitted for brevity. “Conf.,” stands for “conference”, and the symbol “%” represents the proportion of papers utilizing the evaluation method mentioned in the row among all segmentation papers published in that conference.

Evaluation metric	Conf.	%	Selected References
(I) IoU	AAAI	70%	[Qin+22; Kun+22; Hua+22], ...
	CVPR	56%	[Du+22; Xie+22; Mei+22], ...
(II) PRC, TPR, Acc, Fβ, their IoU based average	AAAI	25%	[HXC22; Ahm+22; SRY22], ...
	CVPR	29%	[Ke+22; Kim+22; BZB22], ...
(III) DC	AAAI	15%	[YHL22; Liu+22; Pan+22], ...
	CVPR	4%	[Cip+22; ZZ22; Qiu+22], ...
(IV) Border (∂)IoU, ∂PRC, ∂Fβ	AAAI	10%	[Lan+22; Wan+22a; Xu+22], ...
	CVPR	8%	[Tan+22; Zhu+22; Din+22], ...
(V) HD, ASD, MAD, ASSD, NSD, Chamfer	AAAI	8%	[Liu+22; Wan+22c; Pan+22], ...
	CVPR	4%	[Zho+22; Che+22b; Pen+22], ...
(VI) Execution Time	AAAI	10%	[Lan+22; Ahm+22; LZW22], ...
	CVPR	3%	[Han+22; Zha+22; Che+22c], ...

(e.g., [BZB22; Ahm+22]) employ a threshold based on the IoU to classify correct and incorrect predictions (TP, FN, and FP). Specifically, some studies in group (II) use multiple thresholds or the average of different thresholds to evaluate the system performance.

This review has revealed a significant gap in effectively capturing the diverse characteristics of two-dimensional or three-dimensional segments in MIS systems. This gap coupled with the highlights from several recent studies [MSK22; Rei+21; Nai+21; WWZ20; Kum+17; Kim+15; TH15; GSC22] hinders the ability to accurately compare different systems and select the appropriate system for different conditions or applications. To address these limitations, our study explores alternative evaluation metrics that account for different types of segmentation errors and provide a more comprehensive assessment of algorithm performance in different clinical scenarios which will be presented in the following sections.

In addition to MIS, in Video segmentation (ViS) which is simultaneous detection, segmentation, and tracking of object instances in videos [YFX19]. It has three dimensions: two are related to each frame (image) and one in the time. Similar to image segmentation field, a common metric is IoU [YFX19; Per+16]. As explained before, IoU does not take into account the relationships between the pixels. In [Per+16], authors proposed evaluating video object segmentation based on region similarity (IoU), contour accuracy (boundary detection), and temporal stability [Yao+20]. Temporal stability considers the segmentation smoothness between frames. To avoid misinterpreting occlusions and deformations as instability, Perazzi et al. [Per+16] measure it on sequences without these effects. In boundary detection formalization by Canny [Can86], three objectives are single detection, high detection rate, and accurate localization.

Galasso et al. [Gal+13] present two metrics for evaluating video segmentation based on boundary and volume. In the matter of boundary, they use each frame separately and calculate its boundary metric; therefore, they do not consider time relations. In the matter of volume, they convert the frame and time data into a cube and calculate the total volume by equalizing the temporal and spatial data. They also resolve the inconsistency of the data annotated by different people. They consider the union of the boundaries (resp. average of borders intersections) in all annotations for calculating precision (resp. recall) and the average of intersection between the best-matched frame for volume calculation. They urge that their metric is non-degeneracy, does not have assumptions about data generation, and supports multiple human annotations and adaptive spatial and temporal refinement.

Bounding Box-based metrics are also used for ViS in which the overlap ratio and distance to the center are common metrics [Yao+20; MM16]; however, we do not take them into account.

Loss Functions

Loss functions are one of the most important ingredients in learning-based models to measure the dissimilarity between the ground truth and the predicted segments [Ma+21]. They are designed to help the network learn meaningful targets closely aligned with the ground truth. In MIS, loss functions can be broadly categorized into distribution-based, compound

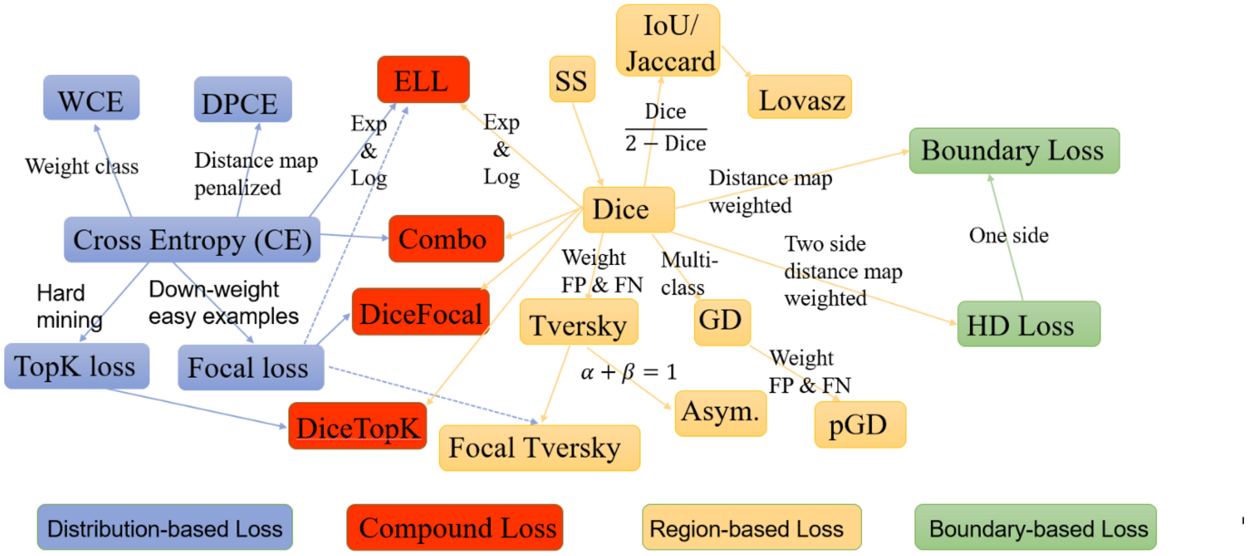


Figure 5.10: Common Loss functions used in MIS [Ma+21]

loss, region-based, and boundary-based types. Some common examples include the Mean Squared Error (MSE) and Mean Absolute Error (MAE), which quantify the average squared and absolute differences between predictions and ground truth, respectively. The SVM loss ensures the score of the correct category surpasses that of the incorrect ones by a safety margin. Cross-Entropy Loss (CEL) and its weighted version (WCEL) are prevalent in classification issues, dealing with predicted probabilities and actual labels, and can address class imbalances by incorporating class weights. Focal Loss further mitigates the class imbalance by focusing on challenging instances, while Dice Loss (DL) and its generalized form (GDL) measure set similarity, crucial in segmentation tasks, with the latter adjusting for class imbalances. Tversky loss introduces a trade-off between false positives and false negatives, and Boundary Loss, a recent innovation in image segmentation, emphasizes the importance of boundary pixels in segments. Each loss function is specifically tailored and applied based on the unique challenges and requirements presented by the model and data in various MIS tasks, thereby facilitating more accurate and insightful predictions and segmentations in medical images. Without going into detail, the commonly used loss functions in MIS are summarized in Figure 5.10.

5.2.7 Metric Learning

Metric learning is the process of learning a distance metric tuned to a particular task that accurately reflects the similarity or dissimilarity between data points [Cak+19; Kul12]. In metric learning, the objective is to transform data into an embedded space, wherein data points with high similarity (depending on the given application) are positioned in close proximity, while those exhibiting dissimilar characteristics maintain a considerable distance from one another [MBL20]. For instance, employing the K-nearest neighbor algorithm, which is dependent on the distance between various data points, learning an effective distance metric can significantly enhance recognition performance [YJ06].

5.2.8 Discussion

In the previous sub-sections, we reviewed various metrics in several applications. Significant achievements and compelling studies on point-based metrics have prompted researchers to reduce the evaluation problem with higher dimensions into a point-based problem by introducing various forms of bias. An instance of this approach is seen in SED and AR, where the underlying targets are continuous over time. One common approach is to split the time range into atomic units and treat each as an independent instance in point-based methods. This often reduces time resolution to one second to address alignment issues to address some degree of misalignment between the reference and prediction [MHV16]. However, this may lead to unexpected false positives when a predicted frame is partly incorrect but the ground truth is negative. Additionally, the fragmentation that occurs within a frame may be undetectable. Moreover, different targets may have varying requirements; one second may work for some but not others. Reducing resolution in many cases is unwise as it introduces unnecessary bias, increasing errors without saving computation time. Instead of resorting to resolution reduction, it is more prudent to utilize segment-based methodologies or to select the smallest viable atomic unit (e.g., preserving the input data's original resolution) to avoid the inclusion of such biases. In MIS, the situation is different because of the constraint of acquiring devices. Despite their large size, they have a low resolution (e.g., 512x512x512 = 128 Mega voxels), and the smallest possible unit used is usually a pixel or a voxel. However, the acquiring devices vary, and thus, the resolutions vary on a case-by-case basis. It should be taken into account during the evaluation since these differences in resolution make each case distinct. Another issue is the utilization of classifier metrics in the segmentation pre-processing step. For instance, in AR, the input sensor events are typically partitioned into segments, and a label is assigned to each segment. This label is then used as a reference, and classification metrics are applied accordingly [KC14; NGC15; CN15; Fu11; QPM21]. The segmentation pre-processing step alters the problem space. For example, in AR, the commonly used sliding event window and sliding time window segmentation approach change the problem space. As this reduction, is done by breaking the inter-correlation in the targets, they cannot capture several important properties of the models. For example, they cannot capture the uniformity of predictions. To address the aforementioned issues, researchers in [Tat+18] redefine precision and recall for time series (see Equation (5.1)). Despite its potential to enhance the evaluation process, it exhibits some limitations, which are explored in Section 5.2.3. Several attempts are also made for 2D and 3D targets which are explored in [Eel+20; GSC22; Kum+17; Nai+21; Bou+22]. It is evident that many of the metrics commonly used to evaluate segmentation are either relatively or absolutely approximated by one another [Eel+20; GSC22; Kum+17; Nai+21; Bou+22]. These metrics can be broadly categorized as follows: Discretize-wise metrics based on voxel-wise, instance-wise or boundary-wise, such as IoU, DC, and NSD; distance-based metrics, such as HD, and Mahanabolis distance; volume-based metrics, such as VS and relative volume similarity; and hybrid metrics that combine several metrics such as AJI, Objective Quality Metric (OQM) and Panoptic Quality (PQ).

Metric learning aims to learn a distance function or similarity measure tailored to a specific task, enabling the effective comparison of data points [MH19]. Nonetheless, the focus of this thesis is not on learning an appropriate distance metric; instead, it aims to evaluate machine learning models that may even incorporate metric learning within their internal processes. For instance, numerous metric learning algorithms are assessed using accuracy (Acc) as an evaluation metric [MBL20].

Despite the wide range of metrics available, assessing the performance of segmentation algorithms in MIS remains a challenging task [WWZ20; Nai+21]. Nonetheless, the limited understanding and interpretability of these metrics can lead to significant bias when selecting an appropriate segmentation method for a specific application [Nai+21].

These issues can lead to more serious problems since the variety of metrics with similar properties can create ambiguities for users, models, and experts when selecting an appropriate approach for evaluation. In conclusion, a more appropriate method to evaluate the performance of these techniques is needed, which involves considering various aspects of the technique while also taking into account the high-level and easily understandable properties required for different applications.

5.3 Significance of Evaluation in a Real-World Application

In this section, we delve into the critical significance of the evaluation function in a real-world application. Our goal is to illustrate how the choice of evaluation methodologies can profoundly impact the outcomes and misleadingly present convincing results.

During the period of this thesis, the COVID-19 has been stated as a global pandemic [22]. One illustrative example of the practical implications of this issue can be found in the realm of COVID-19 research. We embarked on an in-depth examination of the algorithms for early diagnosis of this infection, specifically focusing on Nature Medicine journal, a reputable source of research in this domain. Our analysis uncovered a notable flaw in most models related to the choice of the evaluation function, such that, all the tested algorithms perform worse (from the evaluation function perspective) than an algorithm that generates alarms randomly from a binomial distribution.

For more detail, viral shedding of SARSCoV2 begins 5-6 days earlier than the symptom onset and decreases 14.6 days after it. The peak period is within two days before and one day after the symptom onset [Xin+21; BS21; He+20]. It is shown that COVID-19 patients have anomalous heart rates based on their daily steps [Sha+21; Sma+20]. This data can be easily retrieved from common smartwatches. Therefore, several studies such as [Mis+20; Ski+21; BS21; Ala+22] are done to develop algorithms for identifying COVID-19 infection during its incubation period (the period from the start of the infection to the first clinical sign or symptom [Eli+21]).

Our study on the state-of-the-art for predicting pre-symptomatic COVID-19 discovered a notable flaw related to the choice of evaluation approaches. One reason is that the evaluation method components (true and false positive and negative rates) are not computed in the unified space. As a showcase, we explain it with the results presented in the recent article published in the famous Nature Medicine journal, which uses wearable gadgets to predict COVID-19 infection. In this section, we explain how the used evaluation metric in the literature misleadingly provides better performance for a random algorithm in comparison to that of all other state-of-the-art algorithms. We also analyze the results provided by the latest pre-symptomatic COVID-19 detection systems in the literature which highlights the need for a new metric to evaluate these systems.

5.3.1 Background and a Short Review of Algorithms on Pre-symptomatic COVID-19 Detection

The pre-symptomatic COVID-19 detection systems need α days to be adapted to the user-specific patterns. Then, for each day (d) and each participant (u) they provide an alarm $\mathrm{A}_d^{\mathrm{u}}$ that is either positive ($C+$), negative ($C-$), or unknown ($C?$). Unknown alarms are triggered during the adaptation phase ($d < \alpha$) or when there are many missing sensor occurrences. The daily calculation is usually considered in the literature due to the inability to extract the exact starting date of COVID-19 infection (because of its incubation period) and the fact that the behaviors of each person change over time in a day [Ala+22; Mis+20]. In this thesis, all participated users (U) are divided into positive COVID-19 patients ($U+$), and non-COVID-

19 participants ($U-$). We also define $\mathrm{K^u}$ as the symptom appearance day (or the positive test day in the asymptomatic cases) for each $\mathrm{u} \in \mathrm{U}+$.

The most commonly used metrics in the literature are based on the CM which is built by calculating TP, TN, FP, and FN [VCV22]. They assumed that instances are independently and identically distributed. The sequential nature of sensors and the specific properties of COVID-19 infection means that instances are not independent. Additionally, the targets are durative based on the incubation period of COVID-19 infection. Thus, classical metrics are not appropriate [Mod+22b; Mod+22a; Mod+22c].

Regarding the algorithms for pre-symptomatic COVID-19 detection, Mishra et al. [Mis+20] propose online and offline methods for COVID-19 detection before its symptom onset. They evaluate their model by calculating the number of false alarms per month, correct detection rate (TPR), relative detection date to symptom onset date, and the number of alarms before and after the onset of symptoms. In [Ski+21], authors use various machine learning approaches to identify the incubation period of COVID-19 patients by comparing this pre-defined period with previous healthy data (with a 7-day gap). However, one of the limitations of their study is the limited number of participants (only 27 COVID-19 and 27 non-COVID-19 participants). In addition, they assumed that heart rate and step count data are pre-segmented into two specific fixed windows prior to the onset of symptoms, which is not applicable to real-world scenarios.

Bogu et al. [BS21] developed a deep-learning approach to predict COVID-19 infection. They use the symptom onset date in COVID-19 patients and random dates for non-COVID-19 participants as the reference date. Then, they consider the period between 7 days before and 21 days after the reference date as the infectious period, the period between 10 days and 20 days before the infectious period as the non-infectious period, and 21 days after the symptom onset as the recovery time. Afterward, they use data earlier than the non-infectious period to train the model and test it with the rest of the data. They use a sliding time window approach with a duration of 8 hours and an offset of 1 hour. Then, they evaluate their method with the number of correctly predicted windows as TP or TN and the number of incorrect windows as FP or FN. Pre-symptomatic COVID-19 detection requires a high level of inference (whether or not the user has COVID-19), regardless of the window information. The failure to detect COVID-19 in one window does not mean that the algorithm's performance is degraded because the heart rate pattern changes over time during the COVID-19 infection period. Also, comparing the results of algorithms with different windowing techniques is challenging [Mod+22b].

Abir et al. [Abi+22] extend the study in [BS21] by using Long Short-Term Memory (LSTM) Variational Autoencoder (VAE). They use a similar configuration as [BS21]. However, this very recent paper uses a small dataset containing only 25 health and 25 COVID-19 patients, which was published at the beginning of this pandemic. Since they do not publish their framework, we cannot reproduce their results in our experiments, which are done on a recent dataset containing 2048 users. Therefore, we cannot include that work in the comparison.

In the cutting-edge paper published in Nature Medicine, Alavi et al. [Ala+22] propose a new method to identify COVID-19 patients using heart rate and step data from various wearables. They define the infection window as the period from 21 days prior to the onset of COVID-19 symptoms (for symptomatic cases) or the date of diagnosis (for asymptomatic cases) and the non-infection window as the non-COVID-19 periods (preceding 21 days to a negative test result, the entire time frame for untested participants and the period before infection window for positive COVID-19 cases). Then, they calculate TP, TN, FP, and FN and show the effectiveness of their method by TPR and True Negative Rate (TNR).

There are several other works around using machine learning methods to control COVID-19 pandemic such as [Abd+21; Als+21; Mot+21; Cha+21; NSH20; Mos+22]. The effectiveness of their approach is evaluated using traditional approaches. However, it does not involve intervals in the targets; instead, it deals with discrete elements, which is apart from the approach adapted in this study.

This can be considered as a special case of Anomaly Detection (AD) in time series. Several approaches such as the works in [KSH19; Car+21; Hwa+19; Tat+18; LA15; Emm+15] are used in AD, however, they usually do not consider interval targets, which is the case with the incubation period, or they consider strict constraints (such as the studies in [Hwa+19; Tat+18]). Therefore, there is a need to define a proper evaluation method for the early detection of COVID-19.

5.3.2 Analysis of Evaluation Methods

In this section, we provide an analysis (in Section 5.3.2) of different pre-symptomatic COVID-19 detection systems (described in Section 5.3.2) on the largest public dataset for pre-symptomatic COVID-19 detection (described in Section 5.3.2) along with evaluating the evaluation method used in the literature.

Dataset

We select the latest and largest public dataset[5] that is used in the cutting-edge Nature paper [Ala+22]. It provides heart rate, and step count data retrieved from smart watches for 2048 participants, which contains 18 asymptomatic and 66 symptomatic COVID-19 patients (84 total). In this dataset, the participants with Fitbit wearables have high-resolution heart rate data.

Algorithms

We select the latest algorithms in pre-symptomatic COVID-19 infection detection, such as nightsignal [Ala+22], CuSum [Mis+20], isolationforest (offline) [Mis+20], laad [BS21], rhrad [Ala+22; Mis+20], and random algorithm (Random Algorithm (RA)) which is described in Section 5.3.2. It's worth noting that the rhrad and CuSum methods (in contrast to nightsignal and laad ones) are susceptible to high-resolution data (Fitbit wearables), and isolation forest also works better by using this data. Therefore, we use only high-resolution data for rhrad,

[5] The anonymous step count and raw heart rate data used in this study is downloadable from: https://storage.googleapis.com/gbsc-gcp-project-ipop_public/covid-19-Phase2/covid-19-Phase2-Wearables.zip

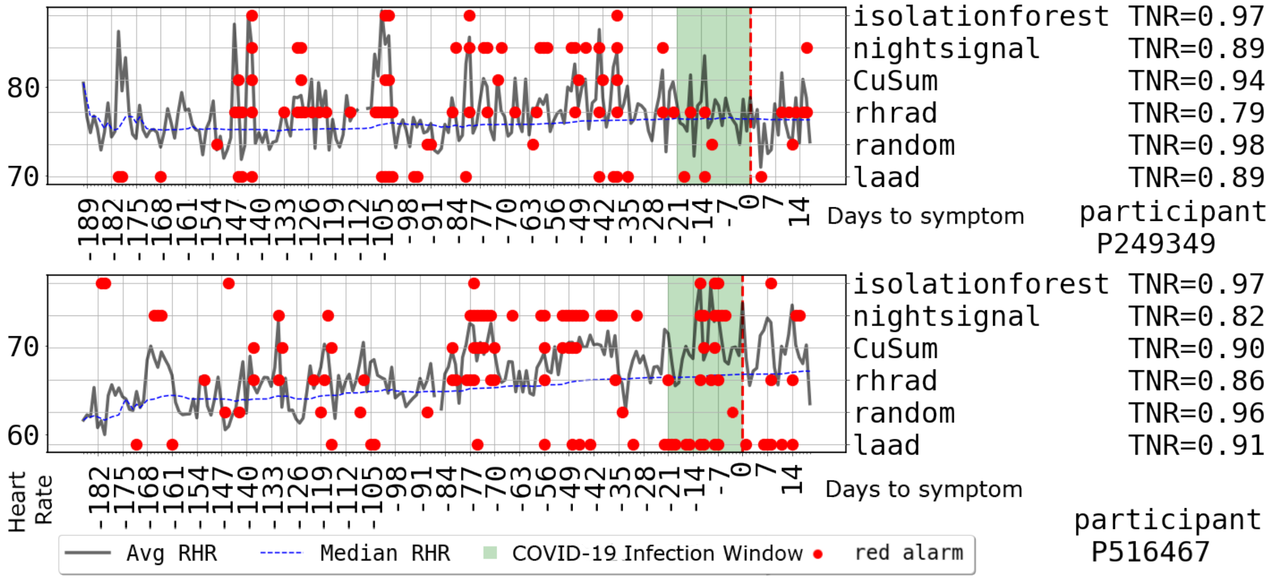


Figure 5.11: Generated red alarms by different algorithms for two participants. The green part shows the infection window (Infection Window Size (Δ)). It shows that current COVID-19 detection approaches generate many false alarms and a basic algorithm that generates an alarm at random has better TNR.

CuSum, and isolation forest algorithms.

Analysis

The most used metrics in evaluating these systems are TPR, TNR, and the median of the distance between the alarms and the symptom [Ala+22]. In this context, TP (resp. FN) is defined as the existence (resp. inexistence) of red alarms during the infection detection window (which has a duration of 21 days in [Ala+22]) and TN (resp. FP) is the number of raised green alarms (resp. red alarms) in the non-COVID-19 period [Ala+22]. However, there are some biases in the calculation of these metrics.

Since COVID-19 occurs a limited number of times (once in the test dataset) for each user, the units of TP and FN are the number of patients, however, the units of FP and TN are the number of days. Therefore, accuracy, precision, Threat Score (TS), and Matthews Correlation Coefficient (MCC) cannot be calculated because their components are in different units: the unit of TP and FN is patients, and the unit of FP and TN is days. This difference in the units causes the inability to use TPR and TNR as complementary.

To analyze this definition better, we formally describe the aforementioned TP, FP, FN, and TN in Equation (5.3). In this equation, Δ is the infection window size (which is pre-defined as 21 days before the symptom onset in [Ala+22]), and [.] shows the Iverson bracket, that is either 1 (when the enclosed condition is satisfied) or 0.

$$
\begin{aligned}
&\forall \mathrm{u} \in \mathrm{U+} : \qquad \text{//For positive COVID-19 users}\\
&\quad \mathrm{TP}^{\mathrm{u}} = [\exists d, (\mathrm{K}^{\mathrm{u}} - \Delta) \leq d \leq \mathrm{K}^{\mathrm{u}} \wedge \mathrm{A}_d^{\mathrm{u}} = \mathrm{C+}]\\
&\quad \mathrm{FN}^{\mathrm{u}} = [\nexists d, (\mathrm{K}^{\mathrm{u}} - \Delta) \leq d \leq \mathrm{K}^{\mathrm{u}} \wedge \mathrm{A}_d^{\mathrm{u}} = \mathrm{C+}]\\
&\quad \mathrm{FP}^{\mathrm{u}} = \sum_{d<\mathrm{K}^{\mathrm{u}}-\Delta} [\mathrm{A}_d^{\mathrm{u}} = \mathrm{C+}]\\
&\quad \mathrm{TN}^{\mathrm{u}} = \sum_{d<\mathrm{K}^{\mathrm{u}}-\Delta} [\mathrm{A}_d^{\mathrm{u}} = \mathrm{C-}] \qquad (5.3)\\
&\forall p \in \mathrm{U-} : \qquad \text{//For healthy users}\\
&\quad \mathrm{TP}^{\mathrm{u}} = \mathrm{FN}^{\mathrm{u}} = 0\\
&\quad \mathrm{FP}^{\mathrm{u}} = \sum_d [\mathrm{A}_d^{\mathrm{u}} = \mathrm{C+}]\\
&\quad \mathrm{TN}^{\mathrm{u}} = \sum_d [\mathrm{A}_d^{\mathrm{u}} = \mathrm{C-}]
\end{aligned}
$$

Then, we define a Random algorithm (RA^p), that generates a red alarm ($C+$) randomly with a probability p and a green alarm with a probability $1-p$, and make predictions drawn from a binomial distribution of all participant data for each day. We use RA^p as a baseline to compare the performance of state-of-the-art algorithms. Obviously, an algorithm is unacceptable when it performs worse than a random algorithm.

Based on Equation (5.3), the TPR of RA^p is equal to the probability of having at least one red alarm in the Δ days before the symptom onset. Therefore, its TPR is $p \times \Delta$ and TNR is the probability of correctly negative prediction. For that, the prediction should be outside the infection window period (Δ). Since, in general, the Δ is small (a few days) compared to the non-infection period (e.g., six months), we can approximate TNR of RA^p with $1-p$ for long-duration experiments. These calculations can be observed in Figure 5.13; however, because the number of positive COVID-19 patients is limited (around 78) and the missing days (the days without any sensor events) are ignored, the TPR and TNR in this figure are

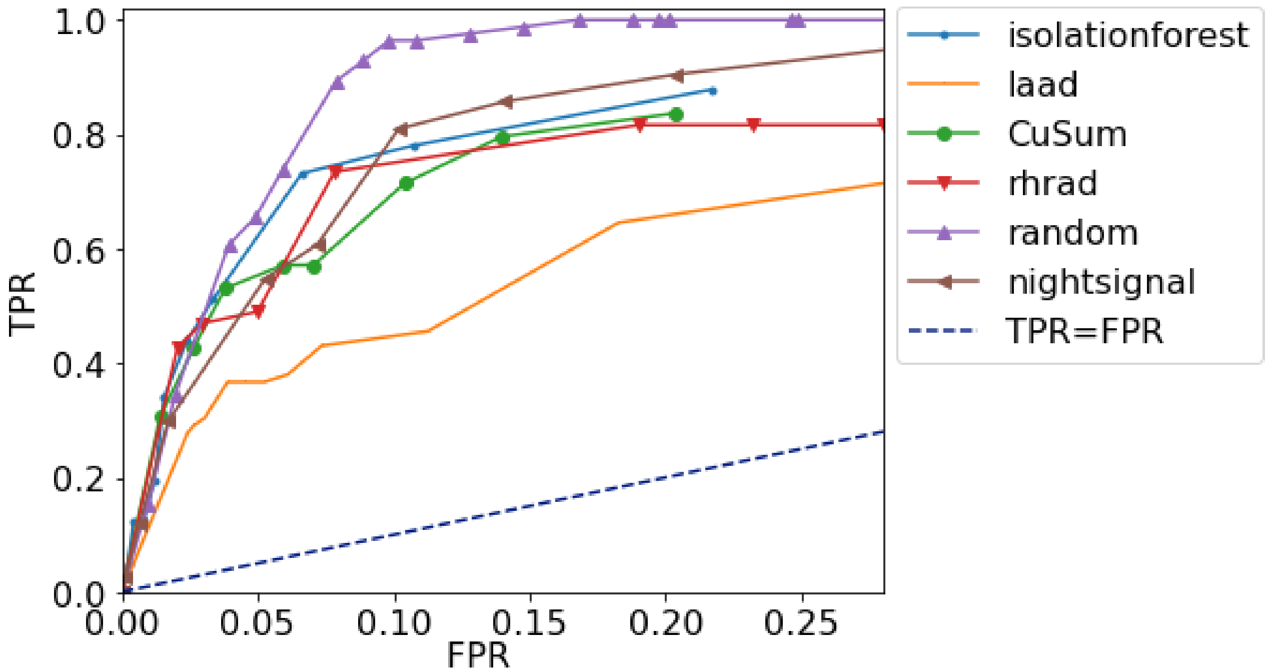


Figure 5.12: ROC of the different algorithms for pre-symptomatic COVID-19 detection. An algorithm that generates an alarm at random provides better TPR for all the parameters. It means that either none of the state-of-the-art algorithms work properly or the evaluation approach is flawed.

not matched exactly to the above calculation.

We demonstrate the red alarms generated by various algorithms in the literature along with RA for two selected participants in Figure 5.11. This figure illustrates that these algorithms generate many false alarms; however, to see the effectiveness of different algorithms, we need a global view of all participants. Figure 5.12 displays the Receiver Operating Characteristic (ROC) diagram of those algorithms. As is expected by our previous calculation, the random algorithms provide better ROC.

The evaluation method has a parameter for specifying the infection window period (Δ). All alarms triggered during this period are considered correct alarms. The used period in [Ala+22] (21 days before symptom onset) is too long, according to the current studies [Xin+21; BS21; He+20]. Therefore, in Figure 5.13, we show the changes in TPR by changing Δ. Changing this has a small impact on TNR. This is reasonable because the total experiment duration (e.g., six months) is usually much greater than the infected window (e.g., 21 days). Therefore, we have displayed TNR as a fixed value in the legend of each

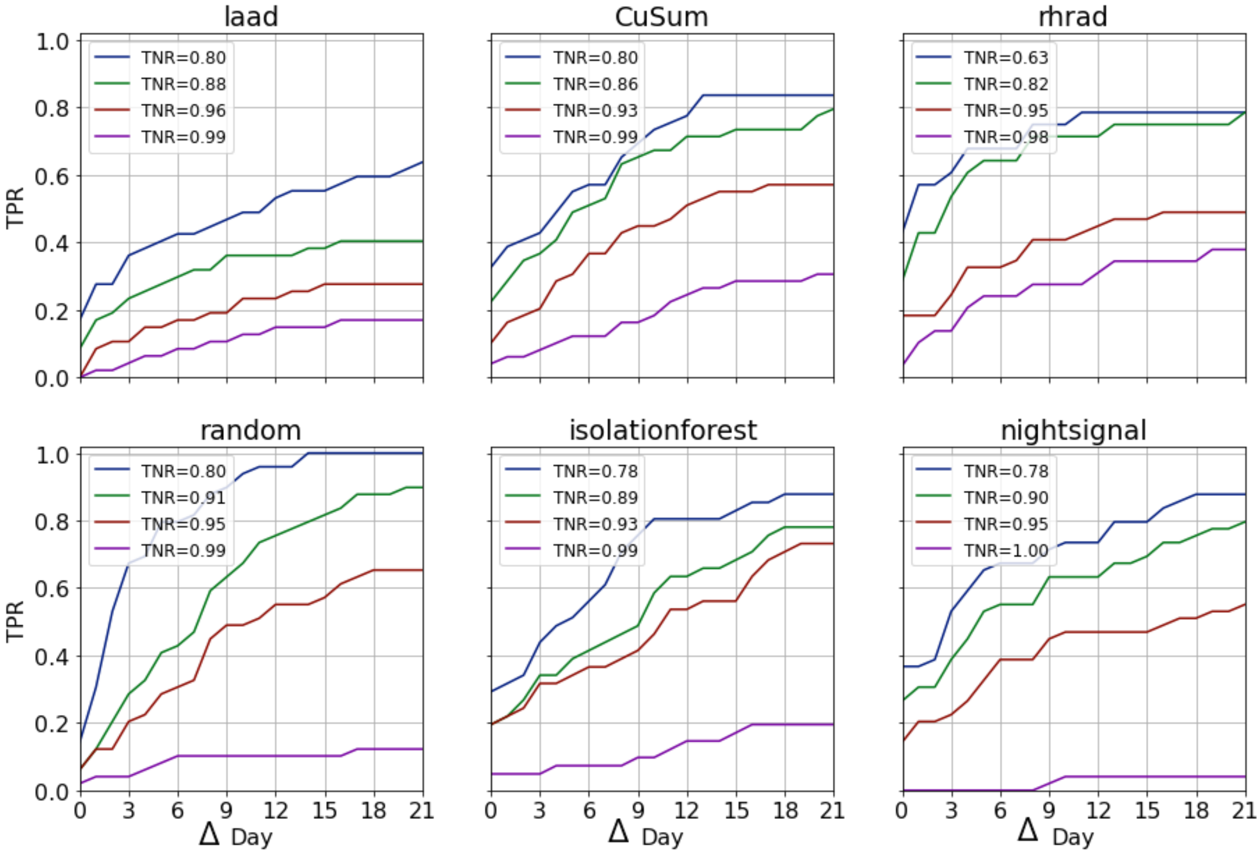


Figure 5.13: TPR of various algorithms for pre-symptomatic COVID-19 detection and parameters based on the infection window size (Δ) (TNR is in the legend). It shows even by changing the Δ, a random algorithm performs better (except for tiny window sizes).

plot to display the different parameters of each algorithm. This figure indicates that the random algorithm outperforms current methods even when changing the infection window size.

The other metric is the median duration between the red alarm and the symptom onset. For instance, in [Ala+22], the authors indicate that approximately 50% of alarms are generated within three days before the symptom onset (i.e., TPR=40% because 20% of infected individuals are not detected at all). Figure 5.13 represents the number of detected COVID-19 patients cumulatively per day relative to the symptom's day. This demonstrates that the random algorithm is still superior.

These analyses demonstrate that either none of the state-of-the-art algorithms are effective or these evaluation methods are flawed.

5.3.3 Analysis with a Less Biased Evaluation Method

In the previous section, we demonstrated that either the used evaluation method or current COVID-19 detection algorithms are flawed. Therefore, we define a simple but less biased evaluation method to show the feasibility of calculating all the TP, FN, FP, and TN in a unified space (in contrast to previous ones). Briefly, it divides the alarms into multiple segments, each containing θ days; one of them begins with the onset of symptoms (if symptomatic) or on the COVID-19 test date (if asymptomatic). Accordingly, a segment is considered a COVID-19 segment ($C+$) when it contains at least one red alarm or a non-COVID-19 segment ($C-$) if it contains no red alarms but has enough green alarms (θ_0). For patients with COVID-19 infection, the preceding segment to symptom onset is TP (resp. FN) if it is a COVID-19 (resp. non-COVID-19) segment. Other segments are considered TN (resp. FP) if they are non-COVID-19 (resp. COVID-19) segments.

Since the goal is pre-symptomatic COVID-19 detection, the alarms after the symptom onset until 21 days after that are ignored [Shr+20]. Moreover, we ignore the alarms in the second segment before symptom onset because it is ambiguous between correct and incorrect detection [Xin+21; BS21; He+20]. In Equation (5.4), we present the formula for calculating TP, FP, FN, and TN.

$$
\begin{aligned}
w_i^{\mathrm{u}} &= \{\mathrm{A}_d^{\mathrm{u}} | i \times \theta \leq d - \mathrm{K}^{\mathrm{u}} < (i+1) \times \theta\} \\
S(w) &= \begin{cases} \mathrm{C+} & \textbf{if } \exists_{a \in w}, a = \mathrm{C+} \\ \mathrm{C-} & \textbf{else if } len(\{a \in w | a = \mathrm{C+}\}) \geq \theta_0 \\ \mathrm{C?} & otherwise \end{cases} \\
TP^{\mathrm{u}} &= [\mathrm{u} \in \mathrm{U+} \wedge S(w_{-1}) = \mathrm{C+}], \\
FN^{\mathrm{u}} &= [\mathrm{u} \in \mathrm{U+} \wedge S(w_{-1}) \neq \mathrm{C+}] \\
\Gamma^{\mathrm{u}} &= \{i \in \mathbb{Z} | \mathrm{u} \in \mathrm{U-} \ \vee \ i < -2 \ \vee \ i > \frac{21}{\theta}\} \\
FP^{\mathrm{u}} &= [\Sigma_{i \in \Gamma} \ S(w_i) = C+], \\
TN^{\mathrm{u}} &= [\Sigma_{i \in \Gamma} \ S(w_i) = C-]
\end{aligned}
\tag{5.4}
$$

In this equation, w_i^{u} is the i-th segment for user u, $S(w)$ is a function that shows the segment status, and Γ^{u} is the index of segments in the non-COVID-19 period.

This section describes the experiment settings using the cutting-edge dataset and state-of-the-art algorithms to support our claims. These algorithms are modified to raise a red alarm when they detect the COVID-19 pattern. We also publish used methods, datasets, and manuals in our open-source repository[6].

The observation in Section 5.3.2 indicates that either the state-of-the-art algorithms are inappropriate or the used evaluation method is incorrect because our presented random alarm generator (RA) performs better than these algorithms. As we explained before, this section aims to compare the state-of-the-art algorithms, which shows the importance of a proper evaluation method. Then, we proposed a simple but less biased evaluation approach to analyze the situation.

In the first experiment, we explore different values for the θ parameter (Figure 5.14). We also use $\theta_0 = \theta/2$ to filter segments that do not have enough information in at least half of the segment size. It shows that $\theta = 10$ provides better performance for rhrad, nightsignal, isolationforest, and CuSum algorithms; however, as it is expected, it does not affect too much on the random algorithm. Recent studies show that the mean and median incubation period of SARS-CoV-2 is 6.38 days and 5.41 days, ranging from 2.33 days to 17.60 days [Eli+21]. Additionally, the World Health Organization (WHO) defines an incubation period is up to 14 days [22]. Besides, SARS-CoV-2 viral shedding begins 5-6 days before the symptom onset [He+20]. Therefore, we select $\theta = 7$ for the next experiment.

Then, in the last experiment, we compare various algorithms using our new metric. Figure 5.15 displays the ROC of these algorithms using this new metric. It shows the random algorithm (RA) is close to the line of TPR = FPR, which means this new metric behaves normally. Interestingly, the algorithm rhrad which was one of the worst algorithms using the other metrics, outperforms other algorithms in the new metric; and RA, which was the best algorithm in that metric, is the worst one in our metric. It indicates that this new metric

[6] https://github.com/modaresimr/covid

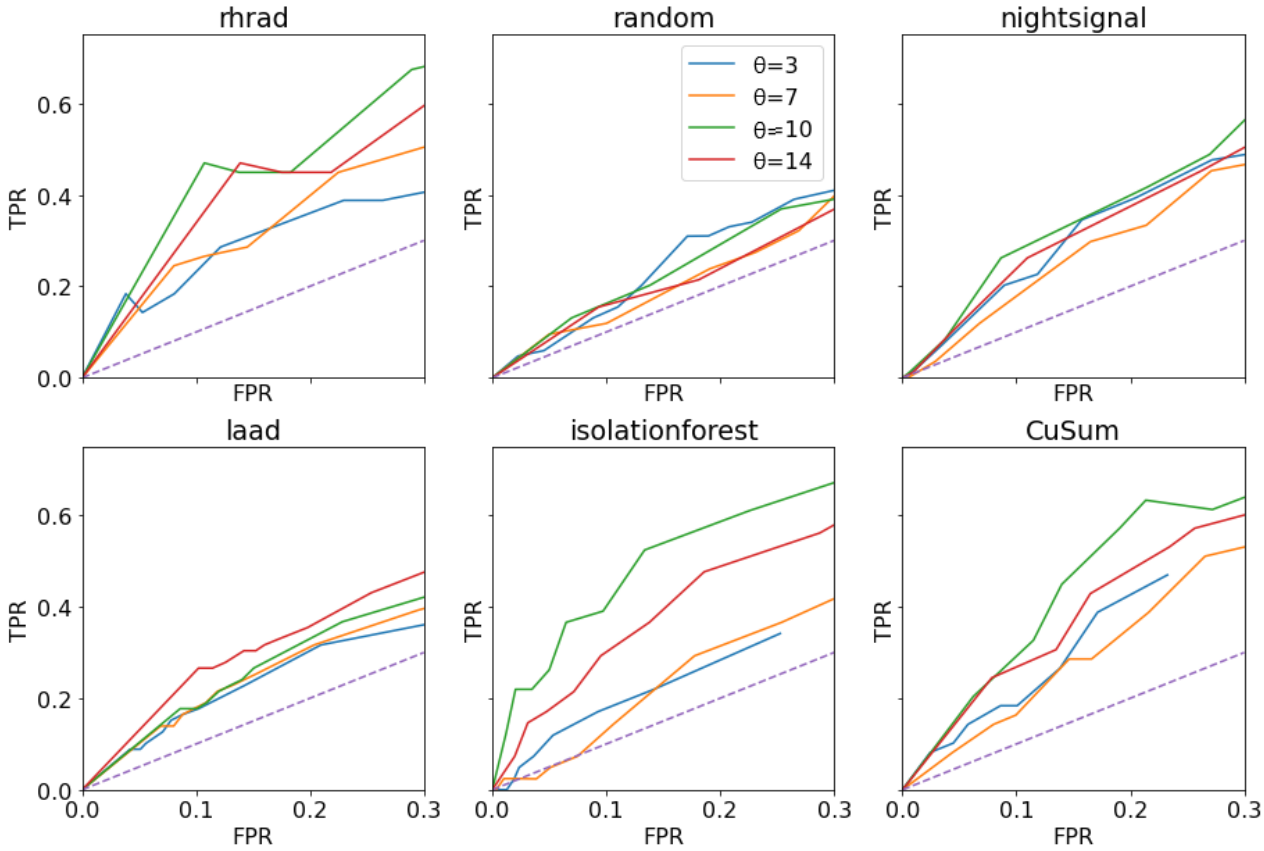


Figure 5.14: ROC of the different algorithms using our less biased metric with different θ. As it is visible, the random algorithm provides the worst TPR for all θ, and it is near the TPR=FPR line. The selection of θ depends on the incubation period of COVID-19 and should be selected by an expert. We suggest $\theta = 7$ based on the recent studies on COVID-19 infection.

eliminated the impacts of randomness in prediction.

Our metric shows that state-of-the-art algorithms perform slightly better than the random ones, which was not visible by the other evaluation methods. However, they need more improvement to be used in a real-world setting, since they generate many false alarms.

5.3.4 Discussion

In this section, we first compare different algorithms for pre-symptomatic COVID-19 infection detection using different evaluation metrics. Then, we highlight that the evaluation functions are as important as the learning algorithms, and validation of the evaluation function itself plays a critical role. We have experimented with an article published in the famous Na-

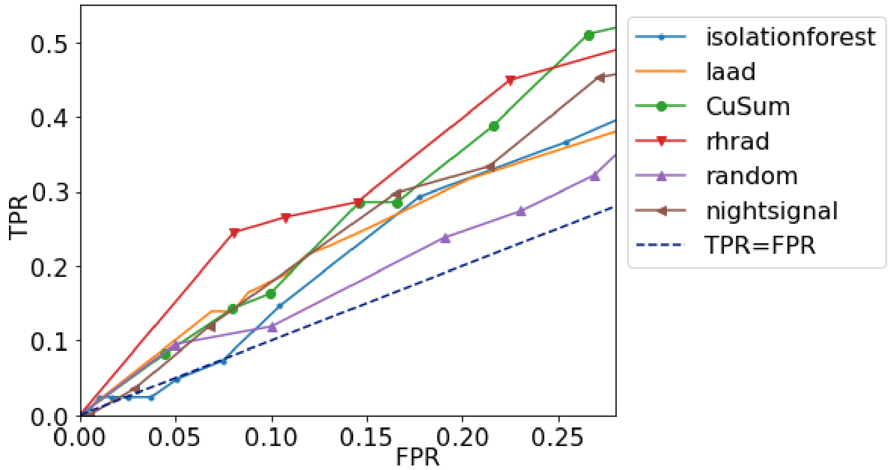


Figure 5.15: ROC of the different algorithms using our less biased metric. As it is visible, an algorithm that generates an alarm at random provides the worst TPR for all the parameters. As we expected, it is near the TPR=FPR line. It also shows the rhrad algorithm works better while it was one of the worst algorithms in [Ala+22] evaluation method. However, it shows all state-of-the-art methods need to be improved, and the current results are not good enough.

ture Medicine journal to demonstrate how the improper evaluation method that is currently used for this application leads to misleadingly convincing results for a random algorithm. We reveal that these algorithms perform slightly better than a random algorithm, and rhrad algorithm provides a better result than the others, while by using the traditional metrics, it was one of the worst algorithms. Moreover, as demonstrated in this section, the presented algorithms in the state-of-the-art are all ineffective since they generate many false alarms.

5.4 Proposed Evaluation Method

The proposed metric aims to address the limitation in understanding and interpreting the numerous evaluation metrics in the literature [Nai+21]. The evaluation of MIS algorithms requires consideration of various properties due to the spatial dependencies between voxels (pixels). Indeed, the significance of these properties varies depending on the application or even at different stages of the medical diagnosis and treatment. For instance, detecting tumors in their early stages is more critical than their size, while changes in size and shape are more important in evaluating treatment response. The proposed evaluation method presents five high-level properties and measurements (based on well-known TP, FP, and FN). These measurements can be combined in a weighted manner to produce a scalar value or used collectively as multi-objective metrics. Furthermore, the modularity of the proposed method enables the straightforward inclusion of new measurements for additional properties. Our proposal leads to a significant advantage in making informed decisions concerning the expected clinical outcome.

The proposed method considers the presence of multiple segments, such as tumor spots, in medical images and treats each segment as an individual instance. As these instances have 2D or 3D shapes, the predicted segments can simultaneously be partially correct and partially incorrect at the same time. Our formulation assumes that each $\mathrm{G} \in \mathcal{G}$ represents one individual ground truth segment, and each $\mathrm{S} \in \mathcal{S}$ represents one individual predicted segment. For instance, in a medical image containing three tumor spots, $\mathcal{S}$ would have three members, and each S would refer to the label and the corresponding voxels. Our proposed evaluation method is based on the following assumptions:

1. The set of individual segments in the ground truth ($\mathcal{G}$) and prediction ($\mathcal{S}$) are given as input.

2. The inputs are given as 3D matrices of size $(w \times h \times d)$ where w, h, and d denote the width, height, and depth of the image. In the case of 2D images or 1D intervals, the input is reshaped to a 3D image with a depth (and height) of one.

3. A perfect prediction is one that exactly matches the ground truth.

4. The G and S are correlated when they overlap, i.e., $\mathrm{G} \sqcap \mathrm{S} \neq \emptyset$ where $X \sqcap Y$ returns all the overlaps between all elements of X and Y. For simplicity, we define $\mathcal{C}(x, Y) = \{y \in Y | x \sqcap y \neq \emptyset\}$. Therefore, $\mathcal{C}(\mathrm{S}, \mathcal{G})$ returns the correlated ground truths segments ($\mathcal{G}$) with respect to S, and $\mathcal{C}(\mathrm{G}, \mathcal{S})$ returns the correlated predicted segments ($\mathcal{S}$) with respect to G.

5. Each voxel may belong to multiple classes (e.g., both liver and tumor). Therefore, we evaluate each class separately as positive and the rest as negative. This way allows

us to use individual settings for each class.

6. To account for the effects of varying machine resolutions and slice thickness on shape properties, we incorporate a vector (dx, dy, dz) that denotes the voxel size for each dimension.

Our method normalizes the targets based on the ground truths, which are independent of the predictions. We achieve this by clustering all G and S into a set called C. More specifically, $\mathrm{C} = \{(\mathrm{G}, \widehat{\mathcal{S}}) | \mathrm{G} \in \mathcal{G} \wedge \widehat{\mathcal{S}} = \{\mathrm{S} \in \mathcal{S} | \mathcal{C}(\mathrm{S}, \mathcal{G}) \neq \emptyset\}\}$, where $\widehat{\mathcal{S}}$ contains all predicted segments that are correlated with at least one ground truth segment. Orphan predictions, which are denoted by $\overline{\mathrm{C}} = \{\mathrm{S} \in \mathcal{S} | \mathrm{S} \sqcap \mathcal{G} = \emptyset\}$, contain predicted segments that are unrelated to any ground truths. Occasionally, a predicted segment can be associated with multiple clusters, resulting in division among those clusters.

Our proposed evaluation method extends point-based metrics to handle partial matches between ground truth and predicted segments. In contrast to point-based metrics, where each voxel is either correctly predicted or not (i.e., the value of TP, FP, or FN for each instance is either 0 or 1), our method generalizes these terms for 2D and 3D data by allowing partial value to each segment. This enables a more nuanced and detailed evaluation of segmentation performance, providing insights into the situation of matching between predicted and ground truth segments. In the following sections, we present the key properties of MIS, SED, and AR drawn from state-of-the-art studies. We also introduce the formulas for measuring these properties.

5.4.1 Detection Property (D)

Detection Property determines the detection of a ground truth target even with a small prediction (at least θ portion to the ground truth). In other words, it checks for the existence of overlaps between a single ground truth (G) and a single predicted (S). This property is crucial in applications such as alarm systems. For example, early detection of all tumor spots is the most critical component, and then other properties are taken into account.

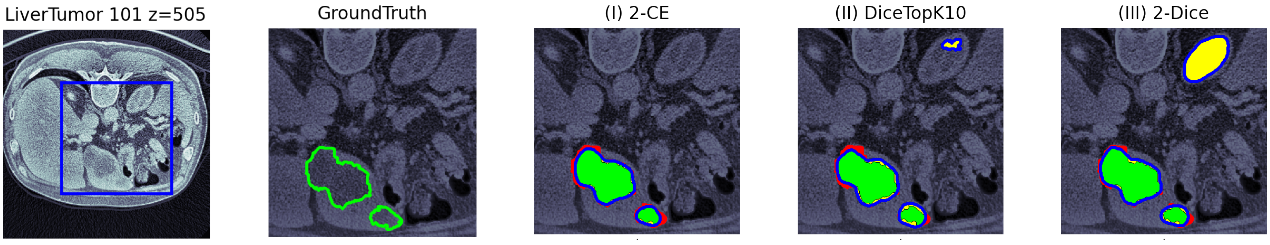


Figure 5.16: The Detection Property (D) shows that system (I) identifies all tumor spots, while systems (II) and (III) incorrectly identify an additional tumor spot (the top right yellow spot is falsely predicted as a tumor while it is not actually a tumor).

In another example, detecting gunshot sound events may be more critical in urban area surveillance. G is TP when at least one S recognizes a part of it and is FN otherwise. Each S with no intersection with any G is considered FP. Measuring this property is formulated in Equation (5.5), where Volume ($\mathcal{V}$) computes the volume of the target in 3D, its area in 2D, or its length in 1D. $[.]$ is the Iverson bracket, which equals 1 when the enclosed condition is true and 0 otherwise.

$$\mathrm{TP}^{\mathrm{D}} = \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \left[\sum_{\mathrm{S}:\widehat{\mathcal{S}}} \frac{\mathcal{V}(\mathrm{G} \sqcap \mathrm{S})}{\mathcal{V}(\mathrm{G})} > \theta_{tp} \right], \qquad \mathrm{FN}^{\mathrm{D}} = |\mathrm{C}| - \mathrm{TP}^{\mathrm{D}},$$

$$\mathrm{FP}^{\mathrm{D}} = \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \left[\sum_{\mathrm{S}:\widehat{\mathcal{S}}} \frac{\mathcal{V}(\mathrm{S}) - \mathcal{V}(\mathrm{G} \sqcap \mathrm{S})}{\mathcal{V}(\mathrm{G})} > \theta_{fp} \right] + |\overline{\mathrm{C}}| \tag{5.5}$$

Therefore, one G is considered TP when at least θ_{tp} fraction of it is correctly identified, and FN otherwise. Predicted segments that are not detected (i.e., $|\overline{\mathrm{C}}|$) and those in which the rate of their incorrectly predicted parts is higher than θ_{fp} are counted as FP.

5.4.2 Uniformity Property (U)

Uniformity Property evaluates the detection of a single ground truth segment (G) by a single prediction (S) instead of multiple fragmented ones or, conversely, evaluates if a prediction covers only one ground truth segment instead of multiple ones. For instance, in the case of tumor detection, false detections may arise due to merging multiple segments into a single prediction, leading to mischaracterizations of the tumor by including surrounding tissues in the tumor or splitting the tumor into multiple segments. These errors can result in changes

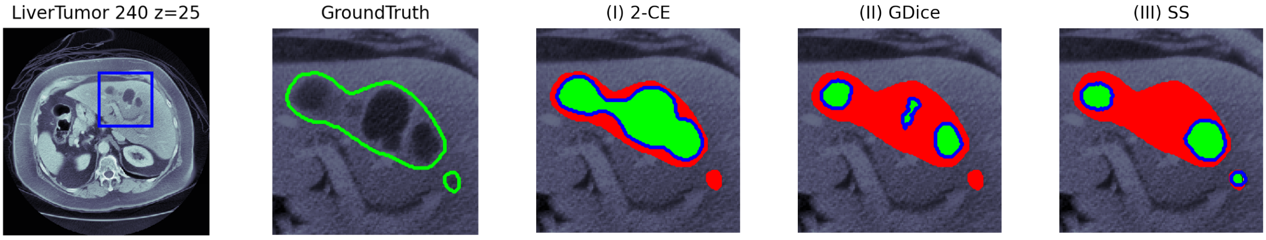


Figure 5.17: In (II), a tumor spot is predicted in a fragmented form, resulting in a lower Uniformity Property (U) value. On the other hand, since one of the two spots is detected without fragmentation and the other with fragmentation, system (III) has a better U than that of (II).

to the texture, statistics, and shape of the tumor, which may impact the nature and features of the tumor [Tia+21]. In another example, the fragmented detection of sleeping activity may incorrectly suggest a sleep disorder [AE17].

One ground truth segment (G) is considered TP if it is recognized by a prediction (all elements in C). However, when a ground truth segment is identified by multiple predictions, the prediction with the highest overlap will be considered as TP, and the rest will be considered as FN. Conversely, when a prediction identifies more than one ground truth segment, it will be considered FP. The formulas for measuring Uniformity Property (U) are given in Equation (5.6).

$$\mathrm{TP}^{\mathrm{U}} = |\mathrm{C}|, \qquad \mathrm{FN}^{\mathrm{U}} = \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} |\widehat{\mathcal{S}}| - 1, \qquad \mathrm{FP}^{\mathrm{U}} = \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} |\bigcup_{\mathrm{S}:\widehat{\mathcal{S}}} \mathcal{C}(\mathrm{S},\mathcal{G})| - 1 \tag{5.6}$$

Consequently, all detected ground truths that are correlated with any predictions are considered TP. When the number of predictions is one, there are no FN; otherwise, it means that the ground truth is detected in a fragmented manner, and all (except one) of the predictions are counted as FN. Conversely, a prediction covering multiple ground truths is considered a FP.

The uniformity property evaluates the detection of a ground-truth segment in a non-fragmented manner, whereas the detection property focuses on correctly identifying a ground-truth segment regardless of whether it is detected in a fragmented or non-fragmented manner.

5.4.3 Boundary Alignment Property (B)

The Boundary Alignment Property (B) is designed to reward when the ground truth boundaries are precisely detected while penalizing for any inaccuracies in the boundaries. In addition, this property takes into account the boundaries based on the shape of the segment since the misclassified voxels are not the same, and they should be locally targeted [Bur+04; Nik+21; WWZ20]. They can significantly change the segment shape and features (see Figure 5.35).

The shape of a segment, including factors such as compactness, roundness, sphericity, lobulation, speculation, and roughness, is crucial in clinical treatments [Li+19b]. This property is related to the work of [Ker+21; Ma+21]. We examine the ability of segmentation algorithms to recognize the boundaries of the shape normalized by their local key points. Using this approach, the normalized boundaries of a spiculated lesion, for example, are not affected by the large volume in its center.

To measure this property, first, we utilize the medial axis of the ground truth segment, which provides a thin representation of the segment [LKC94; Wal+14]. This representation provides the base points for normalizing the distance between the prediction and the ground truth. The normalized distance is used to determine the reward or penalty for the prediction based on how accurately it identifies the boundary of the ground truth segment. For an interval, the center point of the interval is considered as the thin representation.

The segmentation of multiple organs and their medial axis is displayed in Figure 5.18. Unlike HD and voxel-based metrics, which treat all parts of a segment (or its border) equally, we use a normalized distance based on the segment's key points. This is important, particularly for small tissues, as even minor misdetection can significantly impact the segment's radiomic characteristics [Li+19b].

The Boundary Alignment Property (B) is formulated in Equation (5.8). In this equation, the functions $\mathrm{DK}(v, \mathrm{G})$ and $\mathrm{DB}(v, \mathrm{G})$ estimate the distance of voxel v from the medial axis and boundary of the ground truth segment (G). These functions can be computed once during the pre-processing step for each ground truth segment and need not be recalculated for

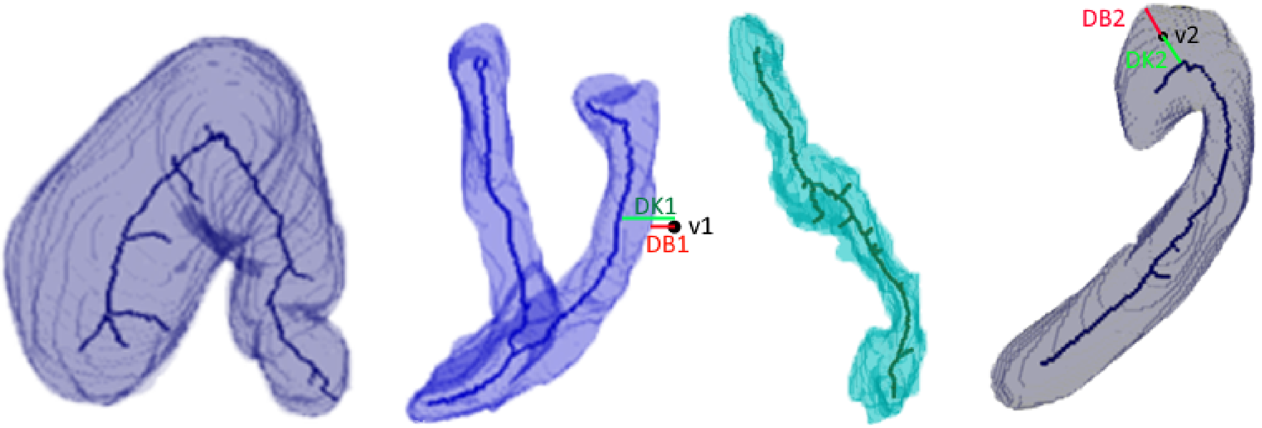


Figure 5.18: The segmentation of multiple organs and their medial axis (shown by dark lines in 3D space). Normalizing the error based on the distance to the medial axis allows us to evaluate the normalized performance of each predicted segment in MIS systems.

analyzing each prediction. The normalized distances of each voxel v are then calculated in Equation (5.7). $\mathrm{DN}^{in}(v, \mathrm{G})$ and $\mathrm{DN}^{out}(v, \mathrm{G})$ represent the normalized distance for voxels inside and outside the ground truth in this equation. The denominators in both formulas are the distance from the boundary to the key point. For instance, the normalized distance for $v1$ (resp. $v2$) in Figure 5.18 is calculated by $\mathrm{DB1}/\mathrm{DK1} - \mathrm{DB1}$ (resp. $\mathrm{DB2}/\mathrm{DK2} + \mathrm{DB2}$). Furthermore, considering the variable voxel sizes commonly present in medical images, we have incorporated the voxel size into calculating the distance functions (DN, DB, and DK).

$$
\begin{aligned}
\mathrm{DN}^{out}(v, \mathrm{G}) &= \frac{\mathrm{DB}(v, \mathrm{G})}{\mathrm{DK}(v, \mathrm{G}) - \mathrm{DB}(v, \mathrm{G})} \qquad \text{if } v \notin \mathrm{G} \\
\mathrm{DN}^{in}(v, \mathrm{G}) &= \frac{\mathrm{DB}(v, \mathrm{G})}{\mathrm{DK}(v, \mathrm{G}) + \mathrm{DB}(v, \mathrm{G})} \qquad \text{if } v \in \mathrm{G}
\end{aligned}
\tag{5.7}
$$

After calculating the normalized distance for each voxel in both ground truth and prediction segments, we classify them into TP, FN, and FP. Finally, the normalized distance is computed based on the total distances of ground truth voxels to obtain a single score for the boundary alignment property.

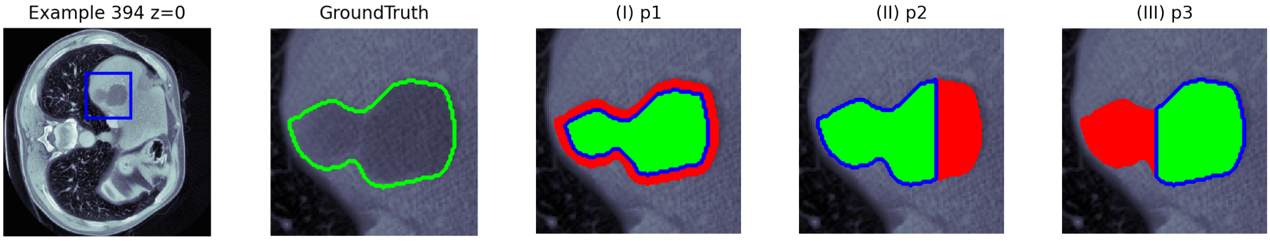


Figure 5.19: Three predictions with identical properties except for Boundary Alignment Property (B). B weights each voxel relative to the ground truth boundary and its thin representation; determining how close the prediction is to the boundaries based on the ground truth shape. Method (I) predicts the shape of the ground truth well, with only small misdetection near the boundary. However, in method (II), a group of misdetected voxels alters some parts of the shape. In (III), the misdetected parts completely change their shape; resulting in the lowest B among these predictions.

$$
\begin{aligned}
\mathrm{TP}^{\mathrm{B}} &= \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \; \sum_{v\in \mathrm{G}\sqcap\widehat{\mathcal{S}}} \mathrm{DN}^{in}(v,\mathrm{G}) \div \sum_{v\in \mathrm{G}} \mathrm{DN}^{in}(v,G) \\
\mathrm{FN}^{\mathrm{B}} &= \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \; \sum_{v\in \mathrm{G}\wedge v\notin\widehat{\mathcal{S}}} \mathrm{DN}^{in}(v,\mathrm{G}) \div \sum_{v\in \mathrm{G}} \mathrm{DN}^{in}(v,G) \qquad (5.8) \\
\mathrm{FP}^{\mathrm{B}} &= \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \; \sum_{v\notin \mathrm{G}\wedge v\in\widehat{\mathcal{S}}} \mathrm{DN}^{out}(v,\mathrm{G}) \div \sum_{v\in \mathrm{G}} \mathrm{DN}^{in}(v,G)
\end{aligned}
$$

Accordingly, based on the alignment error between predictions and ground truths, the TP for each cluster is justified, with larger gaps between the ground truth and prediction boundaries resulting in higher errors. This property focuses solely on the detected segments as it is formulated in Equation (5.8). Therefore, TP (resp. FN) for B is the sum of the normalized distance of detected (resp. not detected) voxels in ground truth in each cluster C. Then we calculate FP, which is the total normalized distances of the false predictions.

5.4.4 Total Volume Property (T)

The Total Volume Property is a widely used property for evaluating segmentation performance based on the overlap property introduced by [TH15]. This property compares the ground truth segment (G) with the predicted segment (S) voxel by voxel or time sample by time sample, where each is classified as TP, FP, FN, or TN. Total Volume Property (T) is particularly useful in assessing global treatment response, as highlighted in studies such

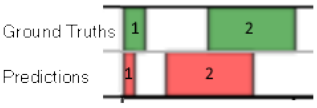


Figure 5.20: Total and Relative Volume properties. The total volume property does not present the impact of uneven segments, while it is necessary for durative sensitive activities. The detection of the first interval is less important than the second one in the total duration property.

as [Tha+18]. The mathematical formulation for T is given in Equation (5.9).

$$\mathrm{TP}^{\mathrm{T}} = \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \mathcal{V}(\mathrm{G} \sqcap \widehat{\mathcal{S}}), \qquad \mathrm{FN}^{\mathrm{T}} = \mathcal{V}(\mathcal{G}) - \mathrm{TP}^{\mathrm{T}}, \qquad \mathrm{FP}^{\mathrm{T}} = \mathcal{V}(\mathcal{S}) - \mathrm{TP}^{\mathrm{T}} \tag{5.9}$$

In this formula, $\mathcal{V}$ calculates the total volume of the given segments, taking into account the varying voxel size in each dimension commonly found in medical images for different CT scans.

5.4.5 Relative Volume Property (R)

In order to reduce the impact of uneven segments, Relative Volume Property normalizes the volume of each segment individually. This is important because even small segments can significantly impact radiomic characteristics and should not be ignored [Li+19b; TH15].

$$\begin{aligned}
\mathrm{TP}^{\mathrm{R}} &= \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \frac{\mathcal{V}(\mathrm{G} \sqcap \widehat{\mathcal{S}})}{\mathcal{V}(\mathrm{G})}, \\
\mathrm{FN}^{\mathrm{R}} &= \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} \frac{\mathcal{V}(\mathrm{G}) - \mathcal{V}(\mathrm{G} \sqcap \widehat{\mathcal{S}})}{\mathcal{V}(\mathrm{G})} = |\mathrm{C}| - \mathrm{TP}^{\mathrm{R}} \\
\mathrm{FP}^{\mathrm{R}} &= \sum_{(\mathrm{G},\widehat{\mathcal{S}}):\mathrm{C}} min(1, \frac{\mathcal{V}(\widehat{\mathcal{S}}) - \mathcal{V}(\mathrm{G} \sqcap \widehat{\mathcal{S}})}{\mathcal{V}(\mathrm{G})}),
\end{aligned} \tag{5.10}$$

Consequently, TP (FN) is calculated as the sum of the normalized volumes of correctly detected (incorrectly undetected) parts of G per ground truth segment, while the FP calculation is similar, with the constraint that the FP of each cluster cannot exceed 1. Notably, since the voxel size is included in both the numerator and the denominator, it has no impact on the final result.

5.4.6 Precision, Recall, and F-score

The evaluation metrics, including Precision (PRC), Recall or True Positive Rate (TPR), and Fβ measure (Fβ), are computed based on the standard formula given in Equation (2.9), which uses the TPs, FPs, and FNs that were defined earlier for each property. To compute the average performance across multiple images, we use the image-wise average, which independently computes the metric for each image and then averages the TPR, PRC, and Fβ across all images. Particularly in MIS, the importance of TPR can outweigh that of PRC [TH15]. This means missing certain regions can be more detrimental than having incorrect predictions in other regions. For instance, missing a tumor spot can be more harmful than incorrectly predicting a small tumor. To address this, the Fβ measure can be adjusted by increasing the value of β.

5.4.7 Computational Complexity

Computational Complexity of the presented formulas is $O(|\mathcal{G}| \times |\mathcal{S}|)$ because elements of both sets of $\mathcal{G}$ and $\mathcal{S}$ are iterated. Since each element of $\mathcal{G}$ needs only related $\mathcal{S}$; the kd-tree helps us to optimize it to $O(|\mathcal{G}|log|\mathcal{S}| + |\mathcal{S}|log|\mathcal{G}|)$. In the case that $\mathcal{G}$ and $\mathcal{S}$ are ordered and are 1D, this complexity can be reduced to $O(|\mathcal{G}| + |\mathcal{S}|)$ by considering the relationships of $\mathcal{G}$ and $\mathcal{S}$. In addition to these, we need to do some computations for each target in 2D and 3D, such as computing the distance and the key points in B. The complexity of distance transform methods can be done using 3D distance transform in $O(n)$ time complexity, where n is the number of voxels/pixels in the discretized space [Gre04]. The key points can also be computed based on the 3D distance transform in $O(n)$ time complexity [Wal+14]. Therefore, the best case complexity can be achieved in $O(|\mathcal{G}|+|\mathcal{S}|)$ for 1D and $O((|\mathcal{G}|log|\mathcal{S}|+|\mathcal{S}|log|\mathcal{G}|)\times n)$ for 2D and 3D.

5.5 Experiments

5.5.1 Activity Recognition Case Study

The main additional problem in AR systems, in contrast to traditional ones, is the importance of duration: a predicted target in AR is durative and can be correct in a period and incorrect in another. However, it is often assumed that time-frame, event-based, or classifier performance follows the whole system performance [Per+14; Bil+20; QPM21; CN15]. This assumption neglects practical scenarios and may misleadingly present convincible results. Despite the importance of evaluating durative targets, even in similar areas, few empirical attempts are proposed which are confronted with the problems of correctness and completeness. Still, there is no universally accepted formula for evaluating the effectiveness of systems with durative targets. Therefore, it is fundamental to extend the correctness vocabulary and to formalize a new evaluation system including these extensions.

This section presents an experimental study of our metric. The first experiment is done on small visualizable data. The second one compares two algorithms in a real-world dataset. The parameters of each property of our metric are as follows. The $\theta_{\mathrm{TP}}, \theta_{\mathrm{FP}}$ are needed to have an appropriate detection property. In this experiment, if a S has any overlap with G ($\theta_{\mathrm{TP}} = 0$), we consider it as TP; additionally, if an incorrect part of a S is longer than its related G's duration ($\theta_{\mathrm{FP}} = 1$), we consider it as FP. We also use $(\beta_t = 2)$ to consider near linear boundary error. The codes and datasets are existed in our repository at https://github.com/modaresimr/evalseg .

Analysis of the proposal on small data

Small data is explored in this experiment for simplicity in visualization. This data contains a subset of 13 relations between two intervals in Allen's interval algebra [Osm03]. This data and our metrics' outputs are illustrated in Figure 5.21 and Table 5.2. Clearly, more S of Alg.a are incorrectly predicted than that of Alg.b in Figure 5.21, while the number of undetected G is the same. The precision and recall in *detection* measurement confirm this observation.

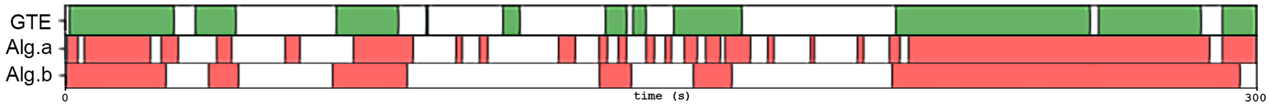
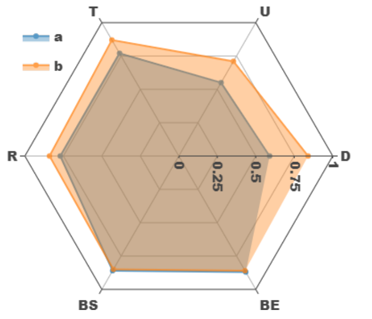


Figure 5.21: Ground truths and output of two algorithms used in [War+11].

Table 5.2: Details of our metric for algorithms of Figure 5.21. The spider chart (right image) shows the F1 for each property of those algorithms.

Algorithm	Alg.a			Alg.b		
Property	TPR	PRC	F1	TPR	PRC	F1
detection	0.73	0.50	0.59	0.73	1.00	0.84
uniformity	0.75	0.43	0.55	0.62	0.83	0.71
total duration	0.78	0.77	0.77	0.84	0.90	0.87
relative duration	0.73	0.81	0.77	0.83	0.85	0.84
boundary start	0.81	0.93	0.86	0.87	0.84	0.85
boundary end	0.99	0.78	0.87	0.85	0.87	0.86



The *uniformity* of Alg.b is higher than that of Alg.a since most of the G detected with a single S in Alg.b instead of multiple fragmented $\mathcal{S}$. For the *total duration* measurement, we can see that the correctly predicted time frames (TP) in Alg.b are more than that of Alg.a, while it is inverse for the incorrect ones. The *relative duration* independently normalizes events and applies the total duration measurement. It shows that Alg.b predicts more part of each recognized target than that of Alg.a. Since the targets' durations are similar, the *total duration* shows similar result. In the *boundary* measurement, we can observe that almost all predictions of Alg.a cover the end boundary of G. Therefore, the end parts of all G are well-detected (TPR=0.99); however, there is some part of predictions after the end of G's boundary that are incorrectly predicted (prediction=0.78).

Analysis of the proposal on a public dataset

In this experiment, we compare the non-overlapping sliding time window of 30s (SW)[7] with the Hierarchical Hidden Markov model (H-HMM) [Asg+20] to show how our metric works. WSU CASAS Home1 dataset [KC14], which contains 32 sensors, 400000 events, and about 3000 durative targets (activities) is used in this experiment. We use its first 20% for

[7] We use feature extraction in [KC14] and three layers perceptron for classifier step.

Table 5.3: Our metric and the spider chart of F1 over two algorithms for one class.

Algorithm	HHMM			SW		
Property	TPR	PRC	F1	TPR	PRC	F1
detection	0.53	0.51	0.52	0.97	0.49	0.65
uniformity	0.95	0.97	0.96	0.86	0.89	0.88
total duration	0.19	0.32	0.24	0.80	0.41	0.55
relative duration	0.39	0.58	0.47	0.92	0.54	0.68
boundary start	0.70	0.63	0.66	1.00	0.48	0.65
boundary end	0.86	0.54	0.66	0.92	0.34	0.49

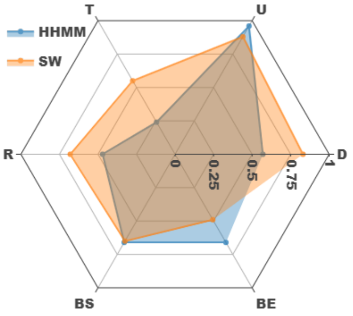


Table 5.4: Tatbul metric [Tat+18] with several parameters and its f1 chart for one class.

Algorithm	HHMM			SW		
Parameter	TPR	PRC	F1	TPR	PRC	F1
α=0, γ=1, δ=back	0.42	0.29	0.34	0.93	0.27	0.42
α=0, γ=1, δ=mid	0.39	0.37	0.38	0.92	0.36	0.52
α=0, γ=1, δ=front	0.37	0.37	0.37	0.92	0.34	0.50
α=0, γ=1, δ=flat	0.39	0.33	0.36	0.92	0.31	0.46
α=1, γ=1, δ=flat	0.53	0.33	0.41	0.97	0.31	0.47
α=0, γ=rec, δ=flat	0.39	0.33	0.36	0.92	0.30	0.45

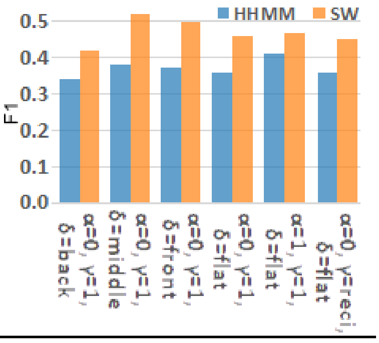


Table 5.5: Ward's proposed metrics for evaluating two algorithms for one class

(a) Event Metrics	HHMM	SW
deletions / $|\mathcal{G}|$	0.47	0.03
merged / $|\mathcal{G}|$	0.03	0.13
fragmented / $|\mathcal{G}|$	0	0.04
frag. and merged / $|\mathcal{G}|$	0	0
correct / $|\mathcal{G}|$	0.51	0.80
insertions / $|\mathcal{S}|$	0.40	0.28
merging / $|\mathcal{S}|$	0.02	0.05
fragmenting / $|\mathcal{S}|$	0	0.06
frag. and merging / $|\mathcal{S}|$	0	0
correct / $|\mathcal{S}|$	0.58	0.61

(b) Time Metrics	HHMM	SW
true positive rate	0.19	0.80
deletion rate	0.50	0
fragmenting rate	0	0.16
start underfill rate	0.29	0
end underfill rate	0.02	0.03
1-false positive rate	1.00	1.00
insertion rate	0	0
merge rate	0	0
start overfill rate	0	0
end overfill rate	0	0

testing phase and the remaining for training[8]. Then, we evaluate the effectiveness of the *take medicine* activity and the macro average of all classes[9]. We compare [Tat+18] and [War+11] metrics with ours.

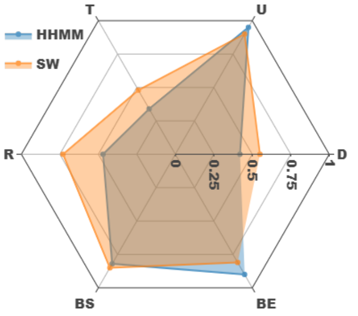
Table 5.5 (b) shows that 50% of time, the HHMM algorithm does not detect the targets,

[8] The internal steps are not important since the concentration is on the metrics.

[9] The full analysis of other classes exists in our repository.

Table 5.6: Macro average of all classes by our metric over two algorithms.

Algorithm	HHMM (macro avg)			SW (macro avg)		
Property	TPR	PRC	F1(m)	TPR	PRC	F1(m)
detection	0.44	0.42	0.41	0.86	0.34	0.51
uniformity	0.98	0.92	0.95	0.97	0.85	0.9
total duration	0.31	0.46	0.34	0.58	0.4	0.48
relative duration	0.37	0.87	0.47	0.67	0.78	0.73
boundary start	0.8	0.92	0.82	0.92	0.83	0.85
boundary end	0.94	0.89	0.9	0.89	0.79	0.81



and 29% of time, it cannot detect the start boundary, while almost none of its predictions is incorrect. For SW algorithm, it shows great performance, except around 16% of the time, the prediction is fragmented. Our metric (Table 5.3) shows this observation is incomplete. Analyzing the data shows that the duration of 5% of targets is equal to the others. Therefore, they dominate the system's quality when using the time frame metrics (e.g., Ward's time metrics) and classifier metrics[10]. Table 5.5(a) helps to understand more about the predictions with the event analysis perspective. It displays that 28% and 40% of predictions in SW and HHMM algorithms are incorrectly predicted (in contrast to the observation from Table 5.5(b)). However, almost all of the targets are recognized by SW algorithm, and nearly half of them are not recognized at all in the HHMM algorithm. It also shows that the predicted targets in both HHMM and SW algorithms are mostly uniform (have few fragmented or merged predictions). This observation is clearly shown in our detection and uniformity property in Table 5.3. Our proposed metric also correctly shows the quality of detecting the boundaries of targets, while Table 5.5 (b) displays this information totally. Since the duration of this class is much less than the total duration of this dataset, while this class constitutes 13% of targets in this dataset, the last four errors in Table 5.5 (b) are close to zero. The relative duration property in Table 5.3 shows SW either recognizes a whole ground truth target (recall=0.92) or does not recognize the target at all; however, its prediction exceeds the boundaries (precision$<$0.6).

[10] If the used segmentation algorithm generates more segments for longer events, which is the case with the well-used sliding window method.

Table 5.4 shows the metric proposed in [Tat+18] with the different parameters. We can observe that γ function, which considers fragmented and merged predictions, has a small impact on the TPR and PRC. As it is observable from our uniformity property in Table 5.3, we can see the predictions of both algorithms are uniform, while HHMM works better. This observation cannot be captured from Tatbul's metric. As analyzed at the end of Section 5.2.3, the main issue of Tatbul's metric is that recall and precision are not calculated in a similar model and cannot be used as complementary (e.g., changing α parameter affects only TPR). Lastly, δ parameter in Table 5.4 is proposed by them to consider the boundary alignment errors; however, changing that does not provide significant changes in recall and precision, while our boundary properties in (Table 5.3) clearly provide the situation of predictions. This experiment ends with Table 5.6, which compares the macro average of our metric across all classes of this dataset.

5.5.2 Sound Event Detection Case Study

Time is an important dimension in sound event detection (SED) systems. However, evaluating the performance of SED systems is directly taken from the classical machine learning domain, and they are not well adapted to the needs of these systems, such as recognizing the time, duration, detection, and uniformity of sound events. Despite its importance, it is not well-developed yet. Current methods are highly biased by their assumptions and may misleadingly present convincible results. The state-of-the-art methods consider few situations of errors and have certain deficiencies. e.g., they are highly biased by their assumptions [Fer+21] and may misleadingly present convincible results.

In classical problems, an instance is either correctly detected (TP) or not (FP or FN). However, instances in SED are durative (events start and end at a specific time). Therefore, predictions may identify parts of references (Figure 2.8). As a result, the TP, FP, and FN should have a partial value between zero and one. Additionally, the situations of predicted events (e.g., predicting a reference event by multiple fragmented predictions) should be considered in the evaluation method.

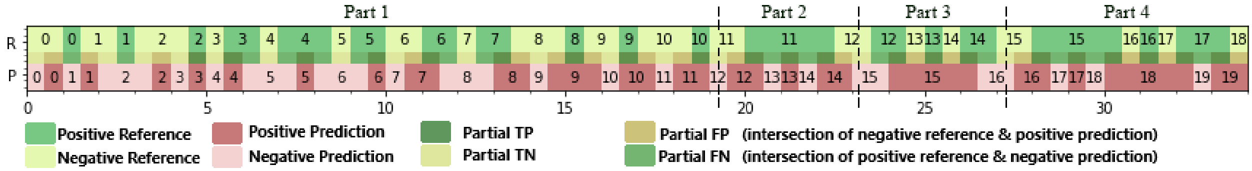


Figure 5.22: An example scenario that contains all major possible situations between references and predictions. The numbers correspond to the i-th reference (prediction) and are indicated by r_i (p_i).

In addition, the dependency on pre-defined strict parameters such as ρ_{DTC} and ρ_{GTC} in PSDS, length in collar, and time resolution in frame-based (segment) and event-based methods (that are widely used in SED evaluation) should be resolved [Sto+15; MHV16; Tur+19].

In order to analyze our method, we first demonstrate its soundness by considering all major possible situations between references and predictions. Then, we compare the sound classes of two SED systems in detail. Lastly, we re-evaluate the best ten systems presented in DCASE 2020 challenge. For the sake of reproducibility, the data, source code, and the details of the experiments are available in our repository at https://github.com/modaresimr/SED-MME-eval

Analysis on Artificially Generated Data

In this experiment, we consider all major possible situations between references and predictions using artificially generated data. This data contains four parts and is visualized in Figure 5.22. The first part is about simple situations (one reference is related to only one prediction). It includes all 13 relations described in Allen's interval algebra [Osm03]. The second part shows fragmented prediction. The third part considers a single prediction that covers multiple references. Lastly, fragmented and merged predictions are considered simultaneously. The evaluation outputs on each part are available in Table 5.7.

Verifying detection property is straightforward. We consider all reference events that have at least one common part with predicted events as TP ($r_{2...17}$), other reference events as FN ($r_{0,1}$), and falsely predicted positive predictions as FP ($p_{0...2}$) in this property.

The uniformity property captures detection of reference events by multiple predictions in a fragmented manner (e.g., r_{11} is recognized by three predictions ($p_{12...14}$); thus, each one is partial FP ($^2/_3$)), or one prediction covers multiple references (e.g., $\mathrm{r}_{12...14}$ are recognized by p_{15}; thus each one is partial FN (FN$_{\mathrm{r}_{12...14}} = {}^2/_3$) and partial TP (TP$_{\mathrm{r}_{12...14}} = {}^1/_3$)); otherwise, the predictions are complete TP (e.g., $\mathrm{TP}_{\mathrm{r}_{2...7}} = 1$). In the fourth part, similar to the second one, each reference is identified by multiple predictions, and also, similar to the third part, one prediction (p_{18}) covers three references ($\mathrm{FP}_{\mathrm{p}_{16,17}} = {}^2/_3, \mathrm{FP}_{\mathrm{p}_{18}} = {}^3/_4, \mathrm{FP}_{\mathrm{p}_{19}} = {}^1/_2$).

Total Volume Property (T) divides the predictions into independent parts and marks them as TP, FP, and FN; then it sums their time intervals (e.g., $\mathrm{TP}_{\mathrm{r}_{2...10}} = \mathrm{FP}_{\mathrm{p}_{2,7}} = \mathrm{FN}_{\mathrm{r}_3} = {}^1/_2$). The segment-based method is similar, but they produce different results since it reduces the time resolution to one second [Hei+13]).

Evaluation of long events is the dominant output in the total duration property. Therefore, the relative duration property calculates the normalized correctly recognized part of each event. Therefore, each partial TP, FP, and FN is normalized depending on its correlated reference events (e.g., $\mathrm{TP}_{\mathrm{r}_2} = 1, \mathrm{TP}_{\mathrm{r}_3} = \mathrm{FN}_{\mathrm{r}_3} = \mathrm{FP}_{\mathrm{p}_7} = {}^1/_2, \mathrm{TP}_{\mathrm{r}_4} = {}^1/_3$).

The state-of-the-art methods also exist in Table 5.7. In their definition, each of TP, FP, and FN is either zero or one; while they can have partial values in our definitions. The collar method provides only one TP (r_2) because the collar range is 200 ms [Fer+21], while the timing errors are 500ms in this data. The $psd\ d/gtc$=0.8 has a similar situation because its acceptable time shift is 200ms for each second of events. The $psd\ d/gtc$=0.1 and our detection property provide similar results because its parameter is small enough in this artificial data. However, an inconsistency exists in the third part. The FP calculated by $psd\ d/gtc$=0.8 is 1 while the TP calculated by $psd\ d/gtc$=0.5 is 3. The result produced by $psd\ d/gtc$=0.8 for the third part is similar to the p_2, which means it ignores the existence of two references in the third part. This shows that PSD method needs some improvements. However, our method does not show this inconsistency.

	Part 1			Part 2			Part 3			Part 4		
	TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP
*detection	9	2	3	1	-	-	3	-	-	3	-	-
*uniformity	9	-	-	1	-	2	1	2	-	1	2	2.6
*total durati.	4.5	4	4.5	1.5	1	1	1.5	1	1	3	1.5	2
*relative dur.	6.3	2.7	3.5	0.6	0.4	0.7	2	1	-	2.3	0.7	0.7
collar	1	10	11	-	1	3	-	3	1	-	3	4
segment	10	4	5	3	-	1	3	1	-	6	-	1
psd d/gtc=0.1	9	2	3	1	-	-	3	-	-	3	-	-
psd d/gtc=0.5	7	4	4	1	-	-	3	-	-	3	-	-
psd d/gtc=0.8	1	10	8	-	-	2	-	-	1	-	-	3

Table 5.7: Different methods for defining TP, FN, and FP on sample data. Our methods are identified by *.

Detailed comparison of two SED systems

The second analysis is made over the best systems in DCASE challenge[11] 2020 Task 4 (Miyazaki and Hao-CQU) on the public dataset [Ser+20]. We provide detailed information for sound classes by showing the TP, FN, and FP provided by different evaluation methods in Figure 5.23. By decreasing the parameter of PSD, it will be closer to our detection property, and by increasing that parameter, it will be closer to the collar method. When the duration of all references and predictions is one second, 100ms collar and $psd\ d/gtc = 0.8$ provide close results. The segment method's objective is to allow some misalignment between the reference and prediction [MHV16]; however, when a prediction is completely incorrect, this method provides more FP than usual (e.g., Blender class in Figure 5.23.b). This is opposite to its goal. The hypothesis in the segment method decreases the time resolution; thus, it may have a wrong impact on the final result. Therefore, we choose exact timing in calculating the total duration property.

Figure 5.23 shows that system (b) recognizes fewer events (detection) than those of system (a); however, its detections are less fragmented (better uniformity) and more precise in detecting the event's time interval (relative duration), while neither the collar method nor the PSD method can capture uniformity and relative duration.

[11] website: https://dcase.community/challenge2020/

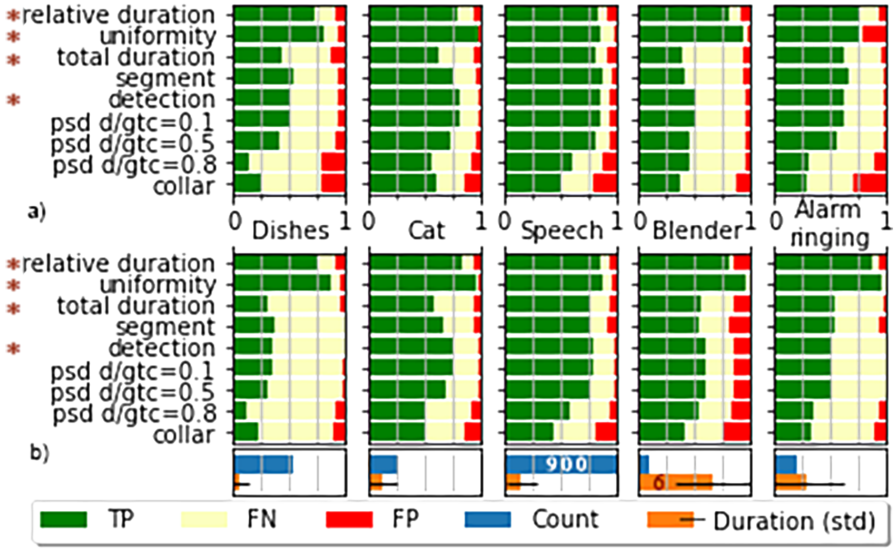


Figure 5.23: Detailed information on different classes in two systems. a) S1:Miyazaki-NU-1, b) S3:Hao-CQU-4. * specifies our methods.

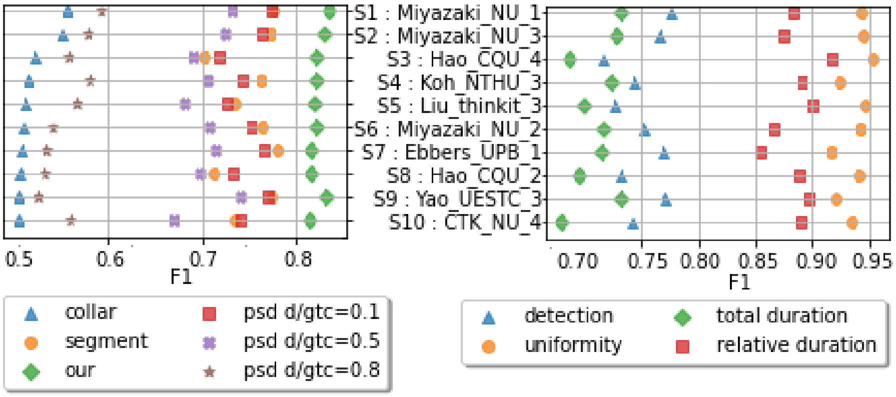


Figure 5.24: (left) F1 of different systems using event, segment, PSD, and our metric, (right) the details of our metric.

Global Comparison of Top SED Systems:

To more globally demonstrate the advantages of the proposed metric, the top ten SED systems (based on collar evaluation) in DCASE challenge 2020 Task 4 are evaluated by F_1-score over the public dataset [Ser+20] under collar, segment, PSD (with three different configurations d/gtc=0.8, 0.5, and 0.1) and ours in Figure 5.24.

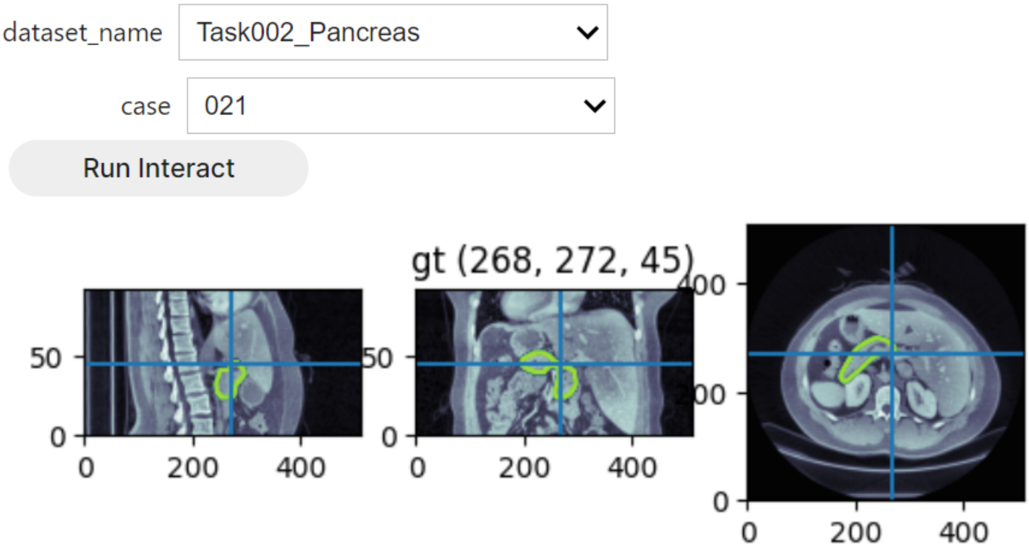
Re-interpretation of TP, FN, and FP causes changes in the rankings of those systems. The collar and PSD methods contain a strict hypothesis that a prediction is acceptable when the timing difference between reference and prediction is within fixed values. Hence, all events that satisfy these conditions are similar; besides, they provide significant differences between two events that are close to the boundaries of their preset conditions. For these reasons, the system ranked ninth by collar method takes third place using our method.

Unlike other approaches with predefined parameters, our metric is independent of strict parameters in calculating properties (detection, uniformity, total duration, and relative duration). The only optional parameter (weights) in our metric is used to prioritize the properties, which can be easily changed without recalculating the properties. Therefore, an appropriate algorithm can be easier selected for a new application with different constraints by differently prioritizing those properties. e.g., an algorithm that performs better in uniformity property; is expected to be more suitable for an application where the uniformity property is essential.

5.5.3 Medical Image Segmentation Case Study

Manual segmentation of medical images (e.g., segmenting tumors in CT scans) is a time-consuming task that can be accelerated with machine learning techniques. Evaluation is a crucial step in fine-tuning and selecting the appropriate one. However, an inappropriate evaluation method that does not entirely meet the requirements may misleadingly present convincible results. In MIS, the spatial dependencies between voxels in each segment make evaluating MIS systems challenging. For instance, in contrast to point-based targets that are either correct or incorrect, in MIS, predicted segments may be partially correct and partially incorrect at the same time.

This section presents an experimental study of our metric. The first experiment is performed on small visualizable data. The second one compares two algorithms in three real-world datasets. The parameters of each property of our metrics are as follows. The θ_{tp}, θ_{fp} are needed to have an appropriate detection property. In this experiment, if a predicted segment has any overlap with the ground truth ($\theta_{\mathrm{tp}} = 0$), we consider it as TP; additionally, if an incorrectly predicted part of a segment is greater than the volume of related ground truths ($\theta_{\mathrm{fp}} = 1$), we consider it as FP. The codes are published in our repository at



dataset_name: Task002_Pancreas

case: 021

Figure 5.25: Pancreas Dataset. Orthogonal View.

https://github.com/modaresimr/evalseg. To compare our methods, we used the recent implementation for the state-of-the-art metrics [Mül+22] by adapting it for uneven voxel size. The selected approaches are: DC (F1), IoU (Jaccard Index), TPR, PRC, FPR, Acc, HD, and Average Hausdorff Distance (AHD).

Datasets

Our experiments are performed on the datasets used in the recent survey [Ma+21] containing Pancreas-CT, LiverTumor, and MultiOrgan datasets. Pancreas-CT includes 363 CT scans, LiverTumor contains 434 liver tumor cases, and MultiOrgan includes 90 multi-organ abdominal CT cases containing eight organs (spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum).

For ease of reproducibility, we have published all the used datasets in this paper at our repository at https://www.kaggle.com/datasets/modaresimr/medical-image-segmentation with complete instructions to visualize and analyze them.

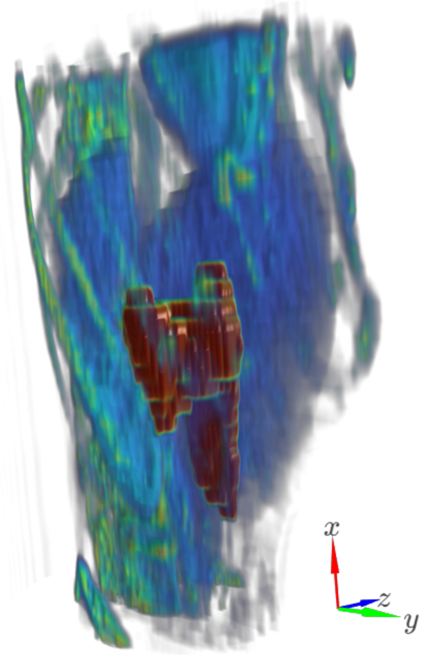


Figure 5.26: Pancreas Dataset. 3D View.

Pancreas Dataset The Pancreas-CT dataset consists of 363 CT scans collected by the National Institutes of Health Clinical Center[12] [Rot+16; Rot+15; Cla+13] and the Decathlon Task07 Pancreas dataset [Sim+19].

The National Institutes of Health Clinical Center executed 82 abdominal contrast-enhanced 3D CT scans, specifically timed to approximately 70 seconds after the administration of intravenous contrast in the portal-venous phase from 53 male and 27 female subjects. The age range of the participants encompassed a span from 18 to 76 years, with a calculated mean age of 46.8 ± 16.7 years. In terms of imaging specifics, the CT scans exhibited resolutions of 512x512 pixels, characterized by varying pixel dimensions, and the slice thickness ranged between 1.5 to 2.5 mm. An essential component of this endeavor involved the meticulous manual segmentation of the pancreas on a slice-by-slice basis. This pivotal task was carried out by a medical student. Importantly, the quality and accuracy of the initial segmentations were further enhanced through verification and modification by an experienced radiologist.

The Decathlon Task07 Pancreas dataset [Sim+19] encompasses 420 portal venous phase

[12] https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

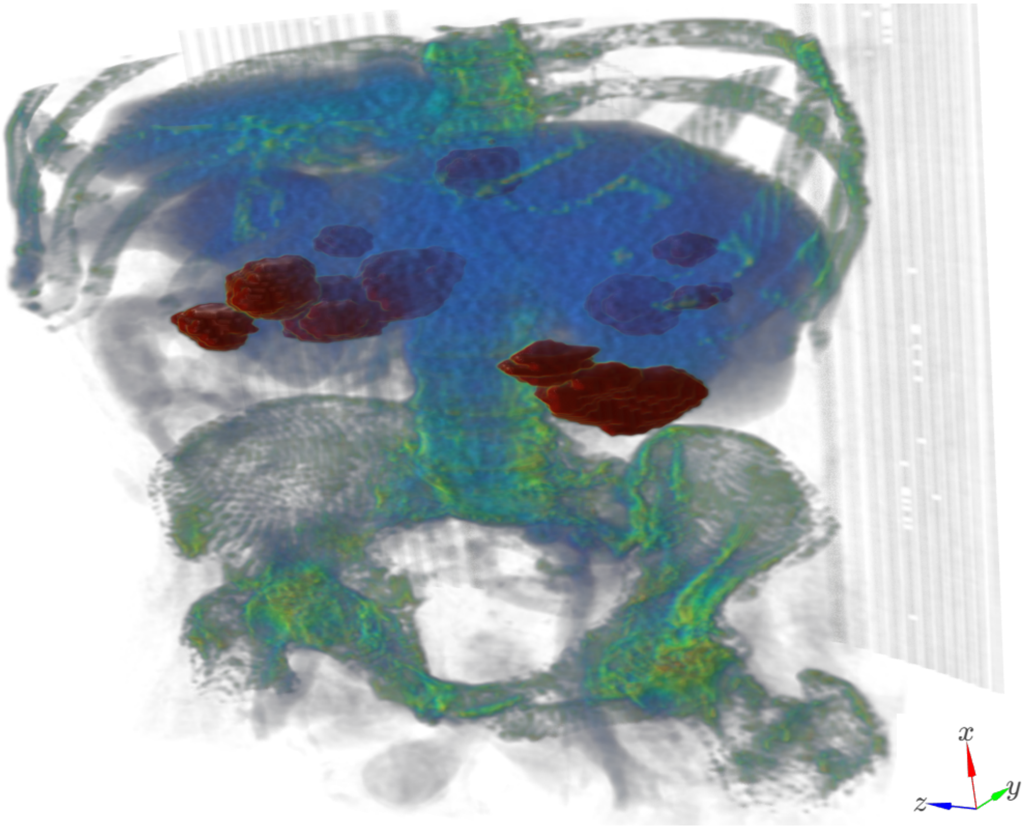
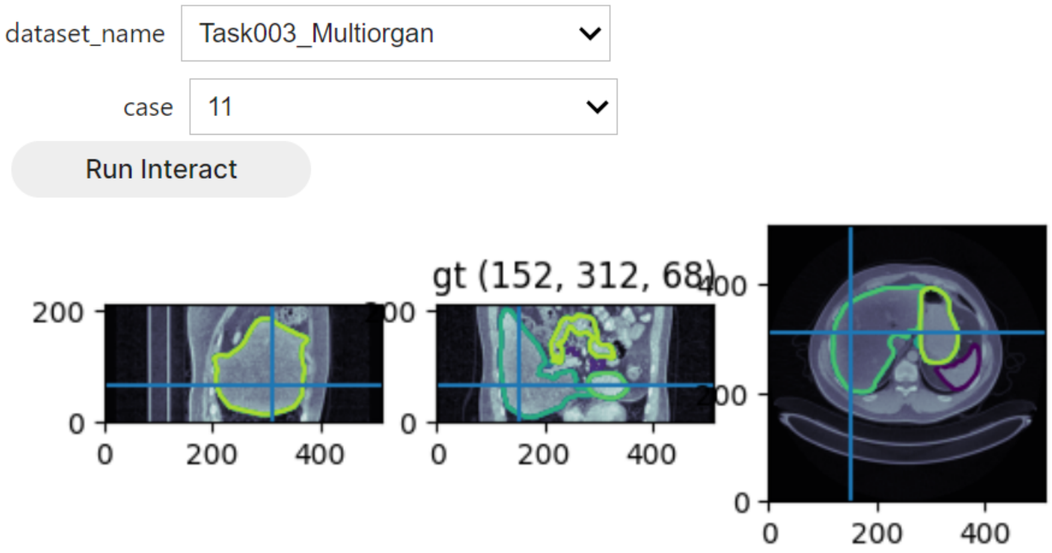


Figure 5.27: Liver Tumor Dataset. 3D View.

CT scans sourced from patients who underwent pancreatic mass resection, including various conditions. These scans were provided by the Memorial Sloan Kettering Cancer Center and are characterized by consistent acquisition parameters. An expert radiologist manually segmented the pancreatic parenchyma and masses in each scan slice using the Scout application. The dataset offers a valuable resource for radiomic research, enabling quantitative feature extraction for disease analysis and treatment response assessment. The axial slices reconstructed at 2.5 mm intervals. Figures 5.25 and 5.26 provide more details about this dataset.

Liver Tumor The LiverTumor dataset contains 434 liver tumor cases by [Sim+19; Ant+22; Bil+23]. It comprises a collection of 201 contrast-enhanced CT images sourced from diverse clinical institutions across different countries. Originating from the Liver Tumor Segmentation (LiTS) challenge, these images were obtained from patients afflicted with primary cancers, including hepatocellular carcinoma, as well as metastatic liver diseases originat-



dataset_name Task003_Multiorgan

case 11

Figure 5.28: Synapse Multi Organ Dataset, Orthogonal View.

ing from colorectal, breast, and lung primaries. These CT scans encompass a range of pre- and post-therapy images, exhibiting authentic clinical scenarios, including metal artifacts. With a resolution varying between 0.5 to 1.0 mm and a slice thickness of 0.45 to 6.0 mm, this dataset presents an intricate interplay of complexities, where expert radiologists meticulously annotated both liver and tumor regions. The challenging nature of this dataset arises from the inherent label imbalance between the relatively larger liver regions and the smaller tumor regions, necessitating the development of innovative segmentation techniques. An example image of this dataset is shown in Figures 2.2 and 5.27.

Synapse Dataset The Synapse Multi-organ Dataset stands as a significant contribution to the medical imaging, aiming to foster advancements in multi-organ segmentation. The dataset boasts with high-resolution images that provide intricate details of various organs. Each image within the dataset is carefully annotated, possessing dimensions that enable deep and comprehensive analyses.

The MultiOrgan dataset includes 90 multi-organ abdominal CT cases from [Lan+15; Gib+18], containing eight organs (spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum). These datasets are also used in the recent review [Ma+21].

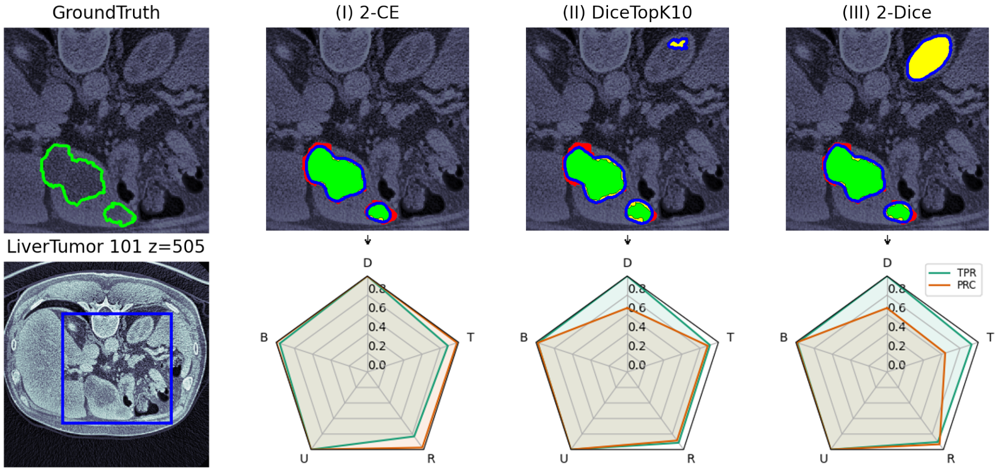


Figure 5.29: System (I) recognizes all tumor spots, however, systems (II) and (III) incorrectly identify an additional tumor spot (the top right yellow spot falsely predicted as a tumor while it is not actually a tumor). The Detection Property (D) can successfully indicate this information. Additionally, Relative Volume Property (R) presents complementary information about the accuracy of prediction in each detected segment. By comparing Total Volume Property (T), which displays the correctly identified voxels in total with D and R, we cannot identify the incorrectly detected spot in (III) is bigger than (II). These measures together can help us to correctly guess the situation of different systems without looking at their predictions.

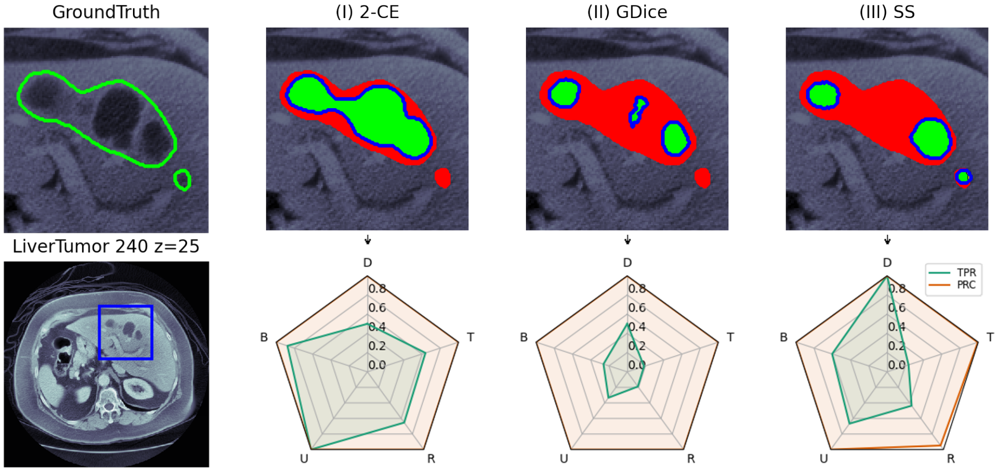


Figure 5.30: In (I) and (II), in comparison to (III), a tumor spot is not detected; however, due to the majority of the detected spots, it does not have a noticeable effect on the Total Volume Property (T). The Detection Property (D) clearly shows that one of the spots (among two spots shown in the ground truth) is not detected. Other properties show relative performance based on the detected spots. In (II), a tumor spot is predicted in a fragmented form, and its effect is visible on the Uniformity Property (U). In (III), all tumor spots are identified, however, the larger spot is not accurately identified in comparison to the smaller one. With Relative Volume Property (R) and Boundary Alignment Property (B), we can see this situation. On the other hand, since one of the two spots is detected without fragmentation and the other with fragmentation, it has a better U in comparison to (II).

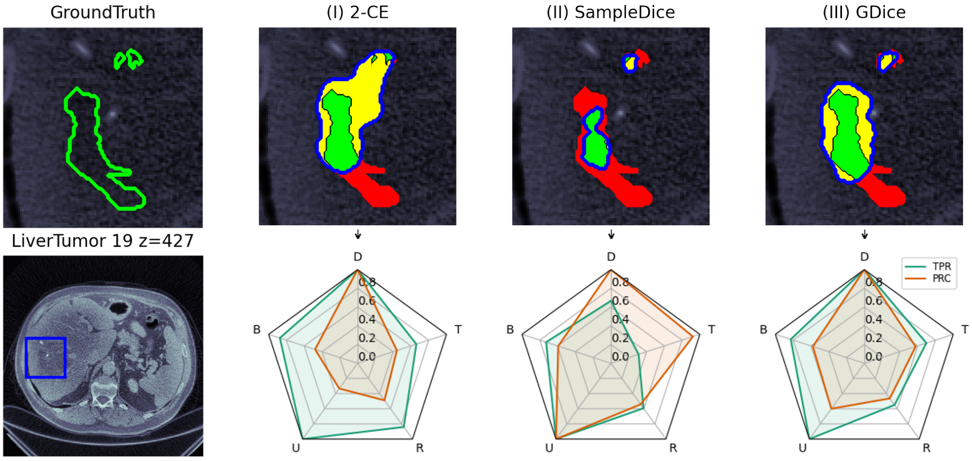


Figure 5.31: A CT scan that contains three tumor spots- two small spots and a larger area- and the output of three prediction systems. Approach (I) recognized all tumor spots as a single spot (the union of yellow and green color). On the one hand, all three tumor spots are detected ($\mathrm{TPR}^{\mathrm{D}} = \mathrm{PRC}^{\mathrm{D}} = 1$). On the other hand, all tumor spots are detected as a single spot, therefore, they are not uniform, and they do not precisely detect spots ($\mathrm{PRC}^{\mathrm{U}} \approx 0.33$). In approach (II), two tumor spots out of three are diagnosed ($\mathrm{TPR}^{\mathrm{D}} \approx 0.66$) and all the predicted spots are related to only one tumor spot ($\mathrm{PRC}^{\mathrm{D}} = \mathrm{PRC}^{\mathrm{U}} = \mathrm{TPR}^{\mathrm{U}} = 1$). The detected area of the greater spot is around 0.3 of the spot diagnosed by the expert. As it is visible, Relative Volume Property (R) shows the performance of detected spots regardless of their volume, while Total Volume Property (T) shows the general performance.

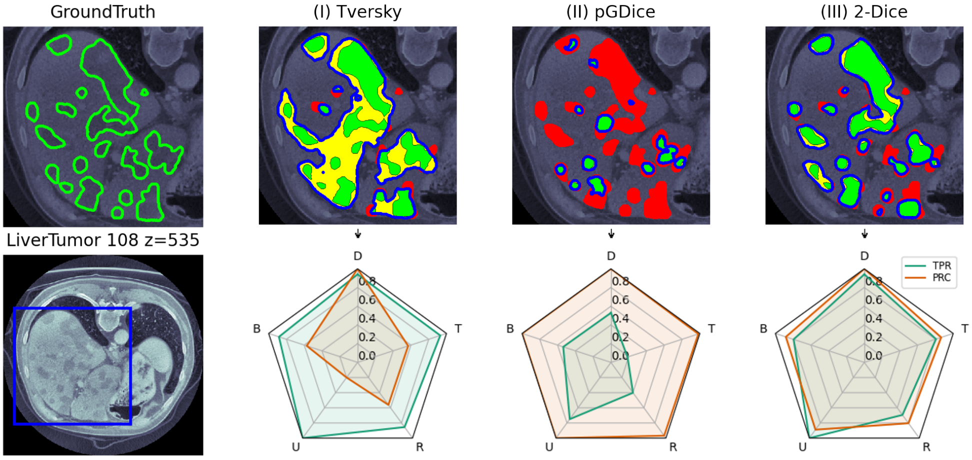


Figure 5.32: A complex ground truth. Method (I), could detect almost all ground truth segments, however, some predictions cover multiple ground truth segments ($\mathrm{TPR}^{\mathrm{D}} = \mathrm{PRC}^{\mathrm{D}} = 1$, while $\mathrm{PRC}^{\mathrm{U}} = 0.5$). Method (II) detects the ground truths by multiple predictions in a fragmented manner and also, it cannot recognize all the tumor spots, but all the predictions correspond to a segment; therefore, $\mathrm{TPR}^{\mathrm{U}} = 0.8$, $\mathrm{TPR}^{\mathrm{D}} = 0.5$ and $\mathrm{TPR}^{\mathrm{D}} = 1$. In method (III), each prediction covers only one ground truth ($\mathrm{TPR}^{\mathrm{U}} \approx \mathrm{PRC}^{\mathrm{U}} \approx 1$), which means uniformity is better than the others. Additionally, we can observe that the bigger segments are identified more precisely $\mathrm{TPR}^{\mathrm{T}} > \mathrm{TPR}^{\mathrm{R}}$. Method (I) predicts more segments, while method (II) is inverse. It can be seen in the Boundary Alignment Property (B).

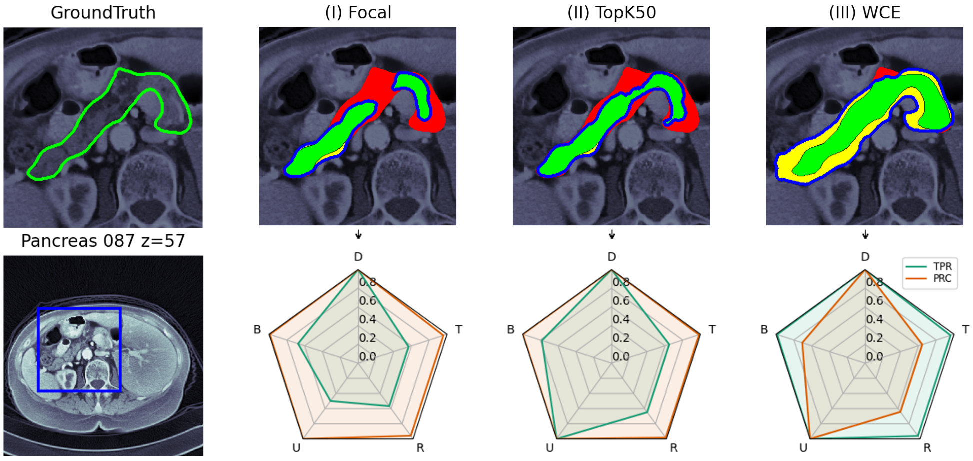


Figure 5.33: A pancreas. In method (I), the pancreas is diagnosed in two parts ($\mathrm{TPR}^{\mathrm{U}} = 0.5$), while approaches (II) and (III) correctly recognize it as a single area ($\mathrm{TPR}^{\mathrm{U}} = 1$). In both approaches (I) and (II), some parts of the pancreas are not detected ($\mathrm{TPR}^{\mathrm{R}} \approx 0.6$). While in approach (III), the system wrongly predicts some parts as the pancreas ($\mathrm{PRC}^{\mathrm{R}} \approx 0.6$). Additionally, Boundary Alignment Property (B) shows that approach (II) better recognizes the pancreas margin based on its shape. Since we have only one segment in this CT, T and R have the same value.

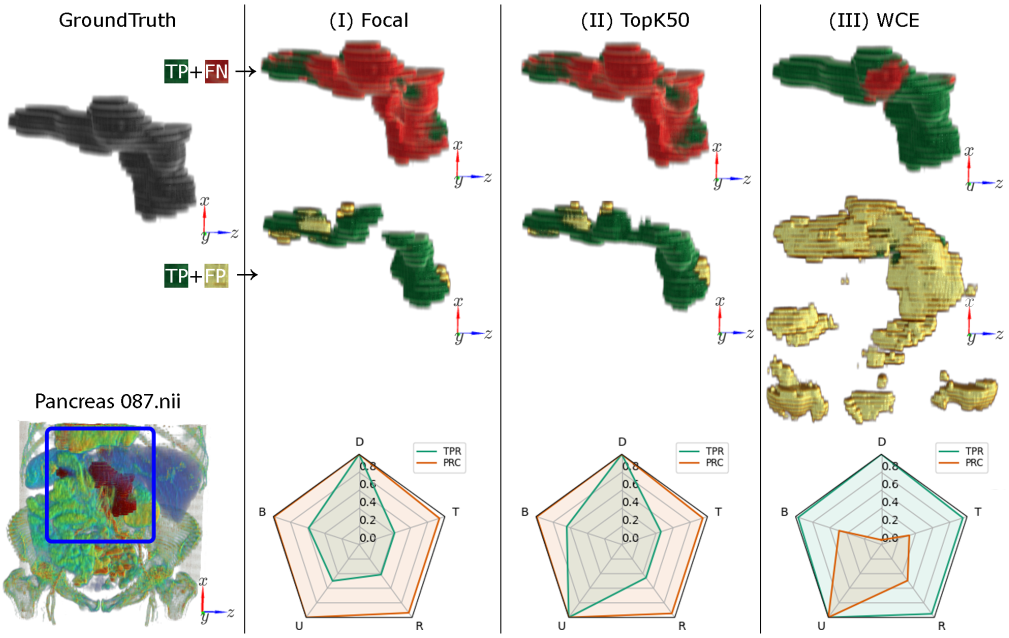


Figure 5.34: The colorized 3D CT scan (left bottom sub-figure) and the pancreas (in the dark). It is the same CT scan of Figure 5.33 without slicing. Similar to previous figures, the first sub-figure shows the ground truth. For a better understanding of the situation of predictions in each system, the missing parts of ground truth (FN) are shown in red on the first row, the prediction (TP in green and FP in yellow) is shown in the second row, the performance spider chart is shown in the bottom row.

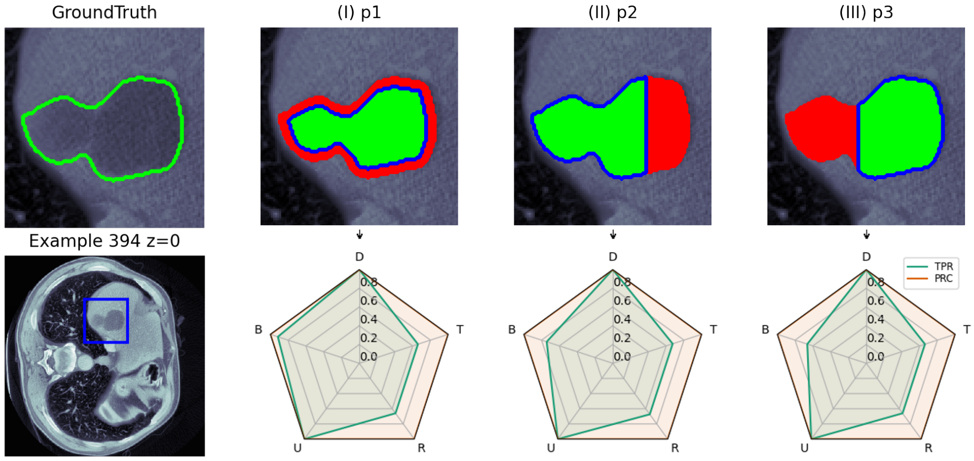


Figure 5.35: Three example predictions in which all their properties are identical except for Boundary Alignment Property (B), which weights each voxel relative to the ground truth boundary and its thin representation (explained in Section 5.4.3); therefore, it determines how close the prediction is to the boundaries based on the ground truth shape. As it is visible, method (I) predicts the shape of the ground truth well, while it has some small misdetection near the boundary. However, in method (II), a group of misdetected voxels changes the shape. In (III), the misdetected parts totally change their shape; therefore, it has the lowest B among all these predictions.

Multi-Organ segmentation is important in medical imaging, as it facilitates the precise delineation of different organs and tissues, enabling accurate diagnoses and targeted treatments. The Synapse Multi-Organ Dataset, with its vast and detailed image collection, serves this very purpose by aiding researchers and medical professionals in enhancing segmentation algorithms. An example of this dataset is shown in Figures 2.1 and 5.28.

Experiment on selected six CT scans

In this experiment, we explore our metric on six selected CT scans with binary labels (e.g., tumor or normal) from Pancreas-CT and LiverTumor datasets. For better visualization, in each 3D CT scan, the cut containing the largest diameter of lesions is selected.

These CT scans are illustrated in Figures 5.29 to 5.34. In these figures, the first row represents the exact margins of the tumors determined by a radiologist (ground truth) and the predictions made by three different systems (I, II, and III). In each prediction, the green color shows the parts that overlap with the ground truth (TP), and the red color determines the parts of the ground truth that are not identified (FN). The yellow color displays the falsely predicted parts as the positive class (FP). To enhance visibility, a blue frame is placed around each predicted segment to show the border of the whole prediction. The second row shows the original 3D CT scan (dataset and slice information are shown in the title) and the measurement of properties for each prediction (which is in the upper row). These measurements (explained in Section 5.4) compare the performance of each algorithm with the diagnosis of a radiologist expert (ground truth) and provide interpretable information with a spider chart. This chart contains five vertices: Detection Property (D), Uniformity Property (U), Total Volume Property (T), Relative Volume Property (R), and Boundary Alignment Property (B). For each vertex, TPR and PRC are shown, which show the performance of a system in recognizing the ground truth accurately and the precision of the prediction made by that system.

In Figures 5.29 and 5.30, we can observe how our system can provide meaningful information about the performance of a system on detecting tumor spots correctly (D), and the differences between R and T. Figures 5.30 and 5.31 show better how the uniformity property considers fragmented and combined predictions. In Figure 5.32, the analysis is made on a more complex CT scan, where we can observe all the metrics. In Figures 5.33 and 5.34, we show the segmentation of the pancreas and how one algorithm recognizes it in two parts while the R and T are similar. Our metric can easily show the prediction situation. In addition, it shows useful information about preserving the segment's shape with B. We have shown the CT scan of Figure 5.33 without slicing in Figure 5.34 in 3D. As it is visible, method (I) recognizes the pancreas in a fragmented manner, which is measured by the Uniformity Property (U) in the spider chart, while method (II) does not have such an error. Method (III) recognizes all the parts in the ground truth ($\mathrm{FN} \approx 0$); however, it has a lot of false predictions (FP). As it is measured by Detection Property (D) and T, many spots are wrongly predicted, and their total volume is also huge. Even for correctly detected spots, the R shows that its error is around 50% of the ground truth volume.

In Figure 5.35, we show, with an example, how boundary alignment property can provide useful information about the situation of misclassified voxels.

These experiments show that our metric can provide meaningful information which cannot be measured by other metrics, such as DC and IoU since they do not consider the spatial dependency between voxels.

Comparing four systems on real-world datasets

Table 5.8: Evaluation of four methods with different metrics among the LiverTumor dataset. We can observe that even the 23rd method (Asym) based on previous reports, works better than the others in some properties. However, the information about their actual behavior is not visible by others[*].

			SS	Asym	DiceTopK10	DiceHD
MME	D	PRC	0.23±0.23	0.64±0.33	0.62±0.35	**0.67**±0.35
		TPR	**0.89**±0.24	0.81±0.28	0.75±0.34	0.77±0.32
	B	PRC	0.59±0.26	0.73±0.28	0.74±0.31	**0.76**±0.30
		TPR	**0.86**±0.30	0.82±0.32	0.79±0.34	0.79±0.32
	U	PRC	0.84±0.28	**0.88**±0.27	0.84±0.32	0.88±0.30
		TPR	**0.93**±0.22	0.93±0.24	0.86±0.31	0.90±0.28
	R	PRC	0.55±0.19	0.67±0.23	0.67±0.28	**0.69**±0.26
		TPR	**0.83**±0.28	0.71±0.29	0.66±0.30	0.66±0.29
	T	PRC	0.36±0.29	0.61±0.29	0.64±0.32	**0.65**±0.30
		TPR	**0.84**±0.29	0.72±0.31	0.66±0.32	0.66±0.31
Other	HD[†]	avg	22.8±20.3	13.4±24.4	13.6±24.0	**11.5**±20.0
		95th	60.5±40.4	30.4±34.0	29.5±33.0	**27.9**±32.5
		max	102.5±41.2	53.6±39.0	49.4±36.8	**48.6**±39.1
	Voxel	DC	0.45±0.30	0.61±0.27	0.61±0.30	**0.62**±0.29
		IoU	0.35±0.27	0.49±0.26	0.50±0.28	**0.50**±0.26
		VS	0.51±0.31	0.71±0.27	0.73±0.30	**0.75**±0.27
	NSD[‡]	$\tau = 1$	0.06±0.05	0.14±0.09	**0**.15±0.10	**0.15**±0.10
		$\tau = 5$	0.42±0.26	**0.66**±0.28	0.64±0.30	0.65±0.30

[*] Bold and underlined values highlight the best and the second-best results.
[†] The unit of HD is in millimeters, and the lower value is better. HD is "inf" when the segmentation result is empty. Therefore, it does not represent the average of all cases.
[†, ‡] The voxel size is included in NSD and HD. Therefore, they may provide different values than other studies.

Based on the comprehensive study in [Ma+21], we have used DiceTopK, DiceHD, Asym, SS methods in that study and evaluated them on similar datasets. Based on the DC metric over all the datasets, the study in [Ma+21] ranks them first, second, twenty-third, and twenty-

Table 5.9: Evaluation of four methods with different metrics over Pancreas-CT dataset, in which all CT scans contain one pancreas[*].

			SS	Asym	DiceTopK10	DiceHD
MME	D	PRC	0.84±0.25	$\underline{0}$.92±0.19	**0.96**±0.15	0.91±0.22
		TPR	**1.00**±0.00	0.99±0.06	0.99±0.06	0.99±0.06
	B	PRC	0.80±0.16	0.90±0.12	**0.93**±0.11	$\underline{0}$.92±0.11
		TPR	$\underline{0}$.95±0.14	0.94±0.15	0.94±0.13	**0.95**±0.13
	U	PRC	0.99±0.06	1.00±0.00	1.00±0.00	1.00±0.00
		TPR	$\underline{0}$.97±0.11	0.94±0.18	**0.99**±0.08	0.94±0.18
	R	PRC	0.71±0.13	0.81±0.11	**0.85**±0.09	$\underline{0}$.84±0.10
		TPR	**0.91**±0.14	0.87±0.16	0.86±0.15	$\underline{0}$.87±0.14
	T	PRC	0.70±0.13	0.81±0.10	**0.85**±0.09	$\underline{0}$.84±0.10
		TPR	**0.92**±0.14	0.87±0.16	0.86±0.15	$\underline{0}$.87±0.14
Other	HD[†]	avg	3.76±3.51	3.01±3.51	$\underline{2}$.85±3.46	**2.79**±3.20
		95th	10.4±9.55	7.60±8.23	**6.95**±7.65	$\underline{6}$.98±7.59
		max	21.4±20.4	14.6±13.8	$\underline{13}$.7±9.86	**13.5**±9.43
	Voxel	DC	0.78±0.11	0.82±0.11	$\underline{0}$.84±0.10	**0.84**±0.10
		IoU	0.65±0.13	0.71±0.13	$\underline{0}$.73±0.12	**0.74**±0.12
		VS	0.83±0.12	0.88±0.12	$\underline{0}$.90±0.11	**0.90**±0.11
	NSD[‡]	$\tau = 1$	0.16±0.08	0.22±0.08	$\underline{0}$.24±0.08	**0.24**±0.08
		$\tau = 5$	0.84±0.13	0.89±0.12	$\underline{0}$.90±0.12	**0.90**±0.12

fifth. Moreover, the difference between DiceHD and DiceTopK is less than one percent. Tables 5.8 to 5.10 highlight the results of our metric and others over three datasets. Notably, those methods have huge tolerance (based on all the metrics) over different images in these datasets.

We can observe in those tables that some properties of Asym and SS are better than DiceTopK. Particularly in MIS, sometimes TPR is more important than PRC [TH15]. We can observe SS approach (the 23rd rank in [Ma+21]) has better TPR in most of the properties. In this situation, we can use Fβ with a higher β value to give more weight to the TPR and only use this value. Our method can provide simple and interpretable information based on five properties that show the situation of those methods. Therefore, based on the application requirement, the expert availability, and even in different stages of treatment, we can select different methods. However, other metrics do not provide this information. In

Table 5.10: Evaluation of four methods with different metrics over the MultiOrgan dataset. We have shown only the macro-average of all classes. The results for all eight classes are available in our repository*.

			SS	Asym	DiceTopK10	DiceHD
MME	D	PRC	0.29±0.17	0.70±0.18	**0.87**±0.15	$\underline{0}$.83±0.17
		TPR	$\underline{0}$.99±0.04	0.98±0.05	0.98±0.05	**0.99**±0.04
	B	PRC	0.77±0.07	0.84±0.05	**0.90**±0.06	$\underline{0}$.88±0.06
		TPR	**0.94**±0.07	$\underline{0}$.94±0.07	0.93±0.08	0.93±0.06
	U	PRC	0.98±0.04	0.98±0.05	0.98±0.04	0.98±0.04
		TPR	0.90±0.11	$\underline{0}$.92±0.10	0.91±0.10	**0.92**±0.08
	R	PRC	0.73±0.05	0.79±0.05	**0.86**±0.05	$\underline{0}$.85±0.05
		TPR	**0.91**±0.07	$\underline{0}$.89±0.07	0.85±0.08	0.86±0.07
	T	PRC	0.67±0.08	0.72±0.07	$\underline{0}$.84±0.07	**0.84**±0.05
		TPR	**0.91**±0.07	$\underline{0}$.89±0.07	0.85±0.08	0.86±0.07
Other	HD†	avg	9.34±7.22	17.5±3.94	**3.91**±4.15	$\underline{4}$.02±3.04
		95th	35.5±25.9	34.0±11.1	**11.5**±11.8	$\underline{12}$.8±10.8
		max	88.5±46.0	47.5±15.1	**20.5**±15.2	$\underline{23}$.9±16.0
	Voxel	DC	0.75±0.08	0.74±0.07	$\underline{0}$.84±0.08	**0.84**±0.06
		IoU	0.64±0.09	0.66±0.08	$\underline{0}$.75±0.09	**0.75**±0.07
		VS	0.80±0.06	0.81±0.05	$\underline{0}$.92±0.07	**0.92**±0.05
	NSD‡	$\tau = 1$	0.19±0.04	0.24±0.07	**0.30**±0.09	$\underline{0}$.30±0.09
		$\tau = 5$	0.77±0.13	0.75±0.11	**0.88**±0.11	$\underline{0}$.87±0.10

addition, our method does not contain certain issues in other metrics.

HD is "inf" when no prediction is made. In addition, it considers all the segments (e.g., big or small) similarly; therefore, the average of HD can be affected when an image contains small segments because the distances between prediction and ground truth for small segments are often smaller than that of larger segments. A similar situation exists for NSD since the border tolerance (τ) is fixed to one millimeter for all segments. In Boundary Alignment Property (B), the distance is normalized based on the shape of the segment. Therefore, it is robust to this situation. As explained before, voxel-based metrics such as DC, IoU, and VS do not take into consideration the spatial relation between voxels, and they cannot provide any information about the situation of the prediction. In addition, greater segments in contrast to smaller segments have more impact on the final results.

5.6 Conclusion

Evaluation and selection of the right metric to compare these systems are crucial. Many researchers, in the absence of a fitting metric, often resort to pixel-wise, time-frame, event-based, or classifier performance evaluations. Such measures can misleadingly indicate a system's convincing performance.

Our work introduced a new mathematical model to evaluate algorithms that produce targets exceeding zero dimensions. This model stands out for its expressiveness, capturing various properties like detection, boundary alignment, relative volume, total volume, and uniformity. It's tailored to be customizable, allowing for adjustable parameters to accommodate a broad spectrum of applications and even to emphasize certain properties over others. Another notable feature of our metric is its extensibility: introducing a new property is seamless and doesn't interfere with existing ones.

Not only have we tested our metric across multiple datasets, showcasing its robust ability to measure different algorithm properties, but it also factors in nuances such as voxel size, considering how acquisition parameters could influence shape properties. This metric enhances the expressiveness of diverse approaches, potentially influencing MIS training methodologies and paving the way for novel machine learning techniques in the upcoming years.

Lastly, our proposed metric doesn't just offer clarity for experts; it's designed to be understandable even for non-experts. By ensuring values in calculating TP, FN, and FP take into account diverse properties, we provide a measurement tool that's interpretable, adjustable, and available as open-source.

Chapter 6

General conclusion and perspectives

6.1 Conclusions

Segmentation is a common pre-processing step in many applications, including MIS, AR, and SED. However, this step introduces at least two families of uncontrollable biases. The first one is caused by the changes made by the segmentation process on the initial problem space, for instance, dividing the input into one-second frames, and the latter results from the segmentation process itself, including the fixation of the segmentation method and its parameters. To address these biases in the segmentation pre-processing step, we first reformulate the segmentation as a decomposition problem and then introduce our novel *meta-decomposition* approach to address these biases. Therefore, the segmentation problem is redefined as a particular case of data decomposition one that includes the decomposer (traditional segmentation), the resolutions (ML), and the composer steps. The composer step transforms the ML results to the global problem results to better describe and evaluate the impact of the introduced biases in the segmentation process. It addresses the first family of biases.

To overcome the second family of biases, we propose a novel approach called *meta-decomposition* or *learning-to-decompose* that learns how to decompose the original task (e.g., recognizing activities from long data) into smaller sub-tasks. Therefore, it can be integrated with meta-learning techniques that require multiple tasks to improve recognition performance. In addition, while the majority of the work in the literature focuses on fixed segmentation approaches that heavily rely on human experience or domain knowledge, the meta-decomposition seeks to reduce the segmentation biases and optimize the overall system performance by learning how to generate sub-tasks rather than assuming the segmentation method as pre-specified and fixed. In the proposed model, the segmentation is an ML hyperparameter that is learned adaptively based on the application and constraints in the outer learning algorithm to improve the recognition quality of the inner learning process. As explained before, without considering the meta-composer part, meta-decomposition introduces an additional bias in the comparison of different segmentation approaches due to the inconsistency in the segments. In the experiments, we demonstrate with a simple and effective data-driven approach, the feasibility of finding a proper segmentation method and its hyperparameter in our proposal and show the superiority of our approach compared to the other approaches with their best hyperparameters on four public datasets. As another example, we have shown its effectiveness by including a dynamic layer on the top of the best segmentation deep network. This dynamic layer improves the recognition performance by dynamically changing the receptive field while keeping the number of parameters nearly unchanged.

Evaluation also introduces several biases and is crucial process in machine learning application. e.g., a common segmentation step changes the problem space and different segmentation algorithms generate heterogeneous segments. Therefore, in this thesis, we review the evaluation process and propose to project the evaluation into a multidimensional space with a partial order relationship that considers the contextual relationships between instances. It projects the evaluation onto five high-dimensions (properties) called Detection (D), Boundary Alignment (B), Uniformity (U), Relative Volume (R), and Total Volume (T). This evaluation latent space is easily interpretable and provides a high degree of flexibility for experts to adopt it for each stage of their considered application. For example, recognizing distinct tumor spots is more crucial than their sizes in the initial scanning stage (property D is vital), while in assessing treatment response, their sizes are more relevant (property T is important). Our novel MME method evaluates segmentation techniques, emphasizing the measurement of essential properties driven by analyzes of relevant studies in MIS, SED, and AR. The MME method refines well-known TP, FP, and FN by permitting fractional values for each concept instead of binary values, accounting for partially correct predictions and enabling a more comprehensive assessment of the segmentation method's performance. Using the updated TP, FP, and FN values, we can compute commonly employed metrics like IoU, TPR, PRC, and DC, which are easily interpretable, even for non-experts. Advancing beyond prior research restricted to zero-dimensional relationships (point-based), this work examines the spatial interdependencies of pixels (voxels, times), covering one-dimensional, two-dimensional, and three-dimensional relations. To elaborate further, this metric evaluates the identification of individual segment spots by a single prediction instead of numerous fragmented ones (uniformity), the accurate detection of each segment (detection), the alignment of ground truth and prediction boundaries based on their shape (boundary alignment), and quantifies the relative and total volume of accurately predicted segments. Our approach is extensible, interpretable, adaptable, and open-source. Moreover, it considers the voxel size since some acquisition parameters, such as slice thickness and resolution, may affect the shape properties. Our evaluation method significantly improves the expressiveness of various segmentation approaches, which may have a noticeable impact on segmentation training strategies and lead to the development of new machine-learning techniques in the future.

6.2 Recommendations and Future Work

Throughout this thesis, we delved into the exploration of a dynamic layer, posited atop the best-performing segmentation deep network. Future research endeavors could broaden this exploration by comparing various other dynamic approaches. Although certain state-of-the-art dynamic approaches did not enhance the performance in our experiments, our findings substantiate that it is feasible to improve segmentation performance through the incorporation of our proposed dynamic layer.

Moreover, an avenue for further exploration arises from the potential of appending this layer to the last layer of the deep network. Preliminary results hint that there is room for additional enhancements in this direction. We also posit that the integration of an adaptive layer atop a lower-parameter inner network might achieve, or even surpass, the performance achieved with a higher-parameter inner network. This hypothesis presents another interesting aspect that needs deeper investigation in future studies.

The other suggestion is to consider the integration of the proposed meta-decomposition concept in the meta-learning approaches. Empirical evidence from our experiments substantiates that this concept enhances overall machine learning performance, thereby warranting its consideration in future studies. We hope this work will open the way for using the meta-decomposition in the meta-learning approaches. The integration of the use of these methods in the internal approaches of base learners will be studied and presented in our future work. This explicit bias description will improve the segmentation process by selecting the appropriate data decomposition according to the current tasks and, consequently, enhance the quality of the machine learning results.

Finally, for the evaluation part, a promising avenue of exploration from this model is the potential to generate a profile for each algorithm. This could serve as a heuristic for more rapid algorithm selection—a facet we aim to delve deeper into in future research. We hope also to extend to the integration of these properties in deep learning algorithm loss functions, considering iterative, weighted combinations of varied loss functions. We also foresee an expansion of these metrics to 4D-CT scans, highlighting the spatiotemporal aspects of patient lesions and organ movements.

Appendix A

Common Evaluation Metrics Formulation in Medical Image Segmentation

In order to provide a uniform formulation for analyzing such systems, we denote S and G as the predicted and ground truth segments that shows the desired concept class and its boundary. For binary classification problems, $\mathrm{g}_i \in \mathrm{G}$ (resp. $\mathrm{s}_i \in \mathrm{S}$) can be either true or false, where false represent negative or background class and true indicate positive or foreground class.
TP corresponds to the correctly predicted instances in the foreground (positive), while TN represents those in the background (negative). Similarly, for incorrect predictions, we have two situations. FN refers to the foreground instances predicted wrongly as background (negative) while FP counts the number of background instances that are wrongly classified as foreground (positive). They are formulated in Equation (A.1)

$$\mathrm{TP} = |\mathrm{G} \cap \mathrm{S}| \qquad \mathrm{TN} = |\neg\mathrm{G} \cap \neg\mathrm{S}| \tag{A.1}$$
$$\mathrm{FP} = |\neg\mathrm{G} \cap \mathrm{S}| \qquad \mathrm{FN} = |\mathrm{G} \cap \neg\mathrm{S}|$$

In this equation, $\mathrm{G} \cap \mathrm{S}$ returns the voxels with the same class in both predicitons and ground truth.
The well known IoU, Acc, PRC, TPR, Fβ, and DC are defined in the following equations.

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{A.2}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{ALL} \quad \mathrm{PRC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad \mathrm{F}\beta = \frac{(1+\beta^2)\mathrm{TPR} \times \mathrm{PRC}}{(\beta^2\mathrm{PRC}) + \mathrm{TPR}} \tag{A.3}$$

$$\mathrm{F1} = \mathrm{DC} = 2\frac{\mathrm{TPR} \times \mathrm{PRC}}{\mathrm{PRC} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} = \frac{2|\mathrm{G} \cap \mathrm{S}|}{|\mathrm{G}| + |\mathrm{S}|} \tag{A.4}$$

In [Nik+21], they proposed NSD by allowing a certain tolerance on the boundaries of the ground truth ($\partial\mathrm{G}$) and the prediction ($\partial\mathrm{S}$). This tolerance is denoted as τ and is used to define a new boundary set denoted as $\hat{\partial}\mathrm{G}$ for the ground truth and $\hat{\partial}\mathrm{S}$ for the prediction that includes all points within a distance of τ from the true boundary points. Using these sets, the Normalized Surface Distance (NSD) with tolerance τ can be defined as shown in Equation (A.5) [Ma+21].

$$\hat{\partial}\mathrm{G} = \{x | \exists \hat{x} \in \partial\mathrm{G}, \|x - \hat{x}\|_2 \leq \tau\} \tag{A.5}$$

$$\hat{\partial}\mathrm{S} = \{x | \exists \hat{x} \in \partial\mathrm{S}, \|x - \hat{x}\|_2 \leq \tau\} \tag{A.6}$$

$$\mathrm{NSD}(\mathrm{G}, \mathrm{S}) = \frac{|\partial\mathrm{G} \cap \hat{\partial}\mathrm{S}| + |\partial\mathrm{S} \cap \hat{\partial}\mathrm{G}|}{|\partial\mathrm{G}| + |\partial\mathrm{S}|} \tag{A.7}$$

Appendix B

More details of the used Datasets

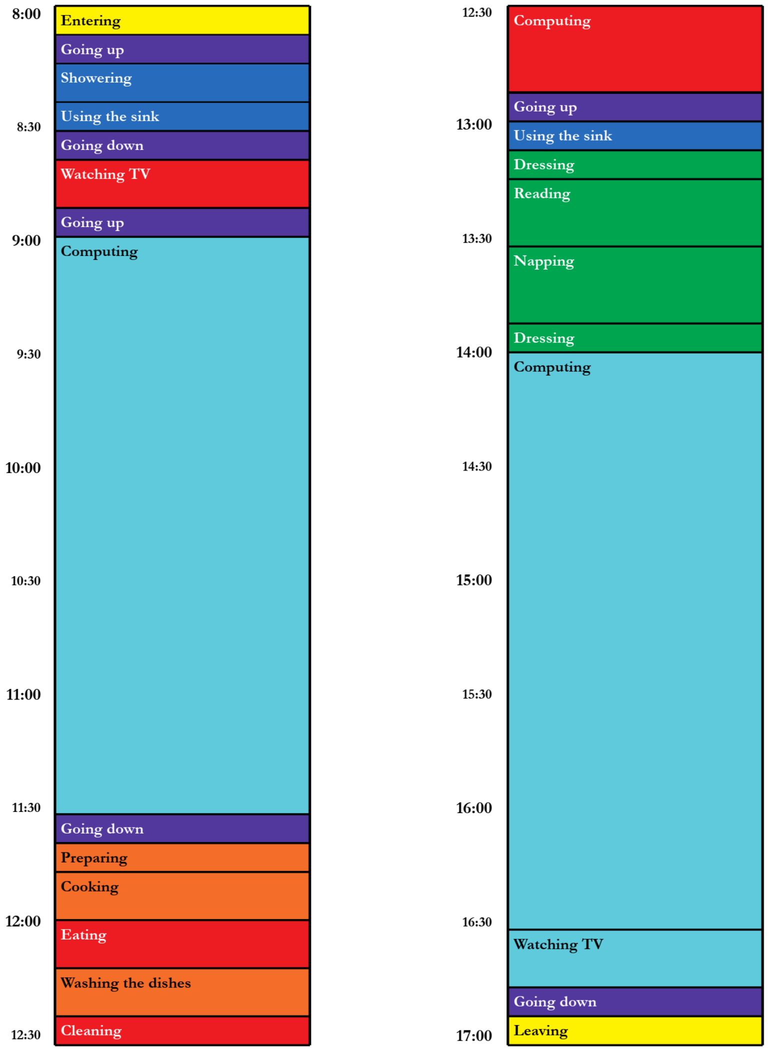


Figure B.1: The standard routines in Orange4Home dataset [Cum+18]

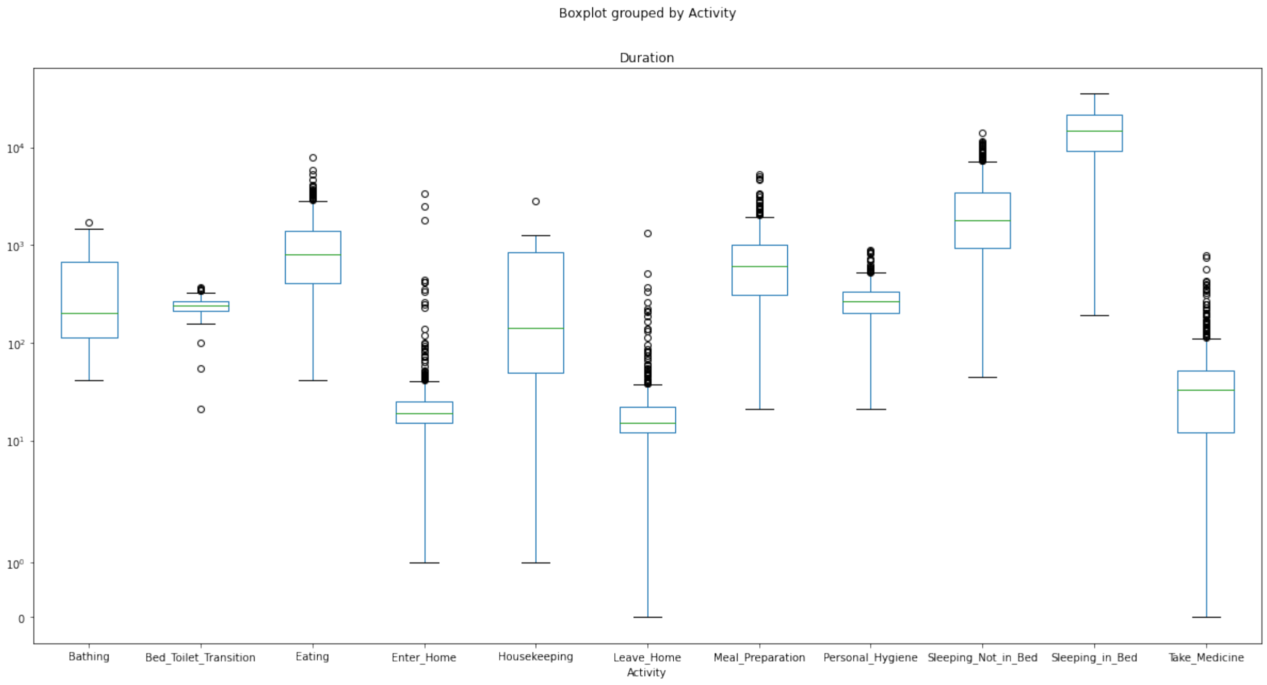
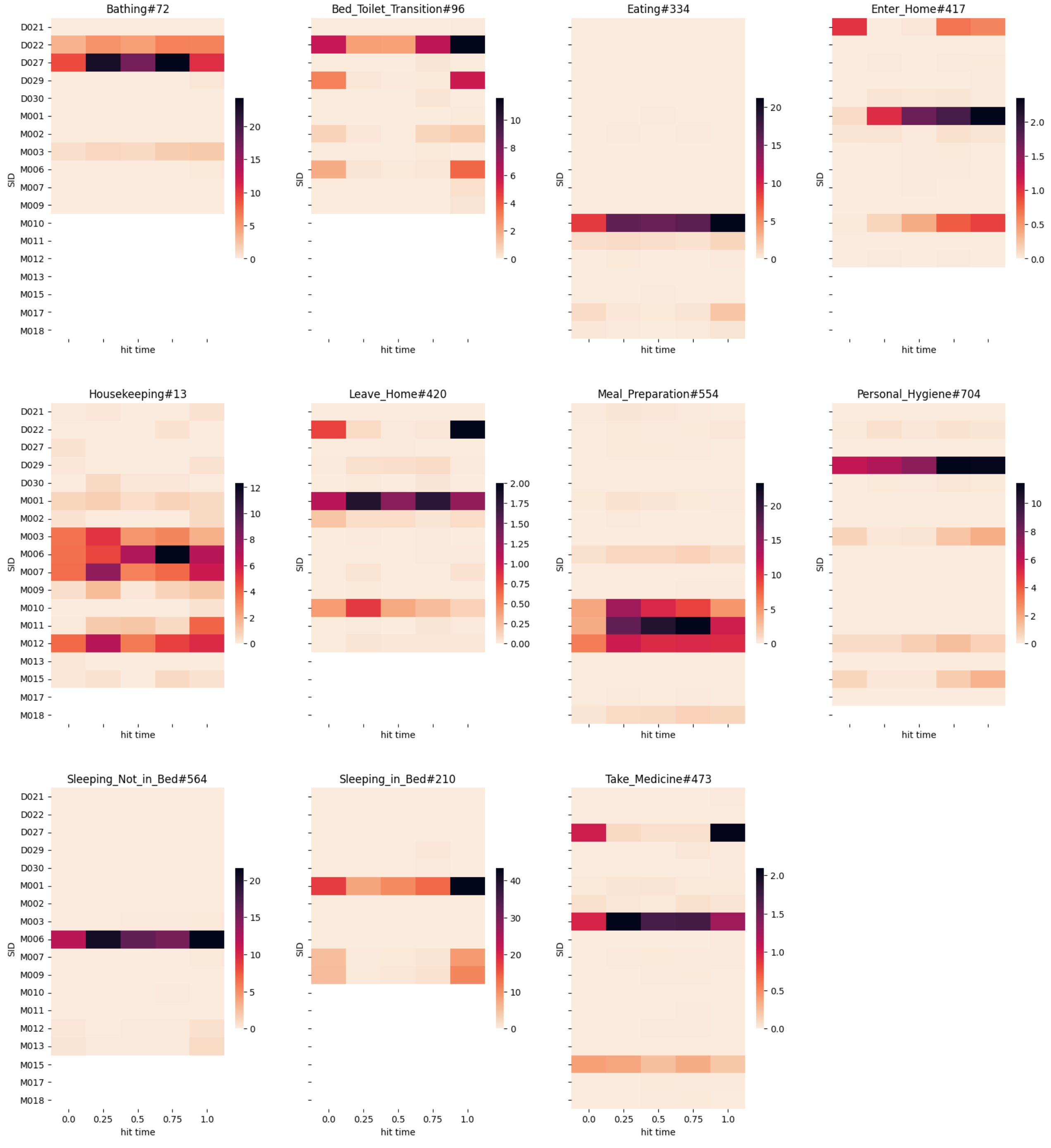


Figure B.2: Home1 Activity Duration



Bathing#72

Bed_Toilet_Transition#96

Eating#334

Enter_Home#417

Housekeeping#13

Leave_Home#420

Meal_Preparation#554

Personal_Hygiene#704

Sleeping_Not_in_Bed#564

Sleeping_in_Bed#210

Take_Medicine#473

Figure B.3: Detailed Activities sensors hit map in the Home1 Dataset. It shows the number of sensor events occurred at each time for each activity. On the legend, the average number of sensor's heats are shown.

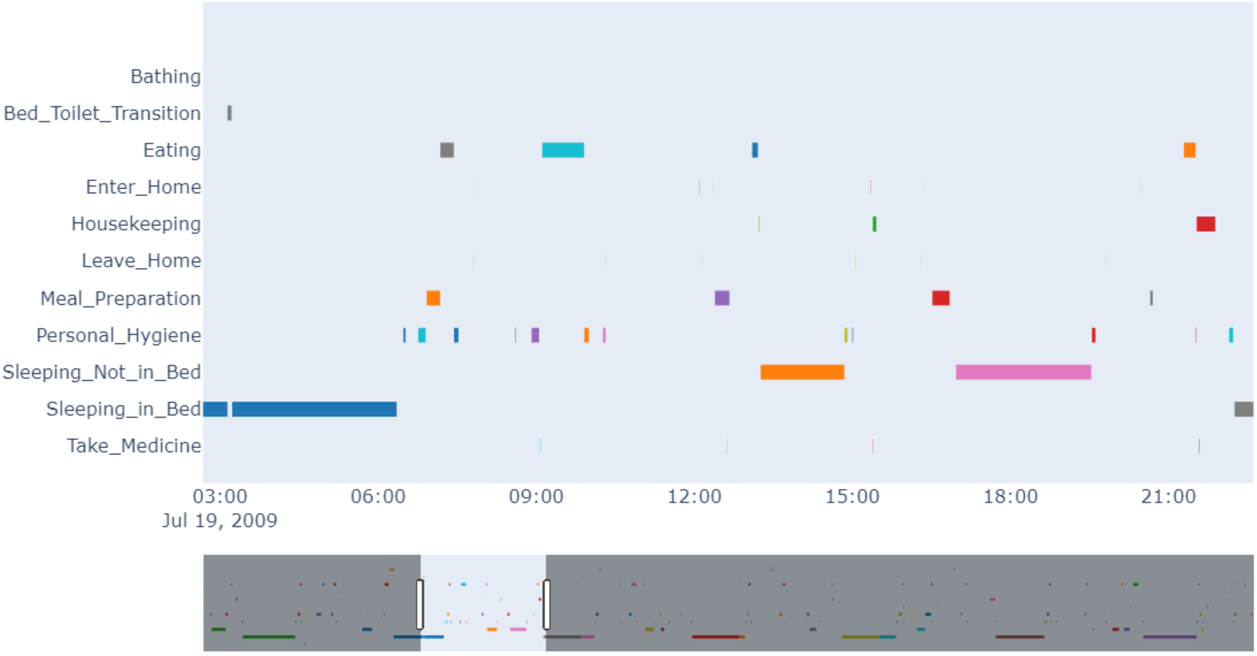


Figure B.4: Home1 Activities

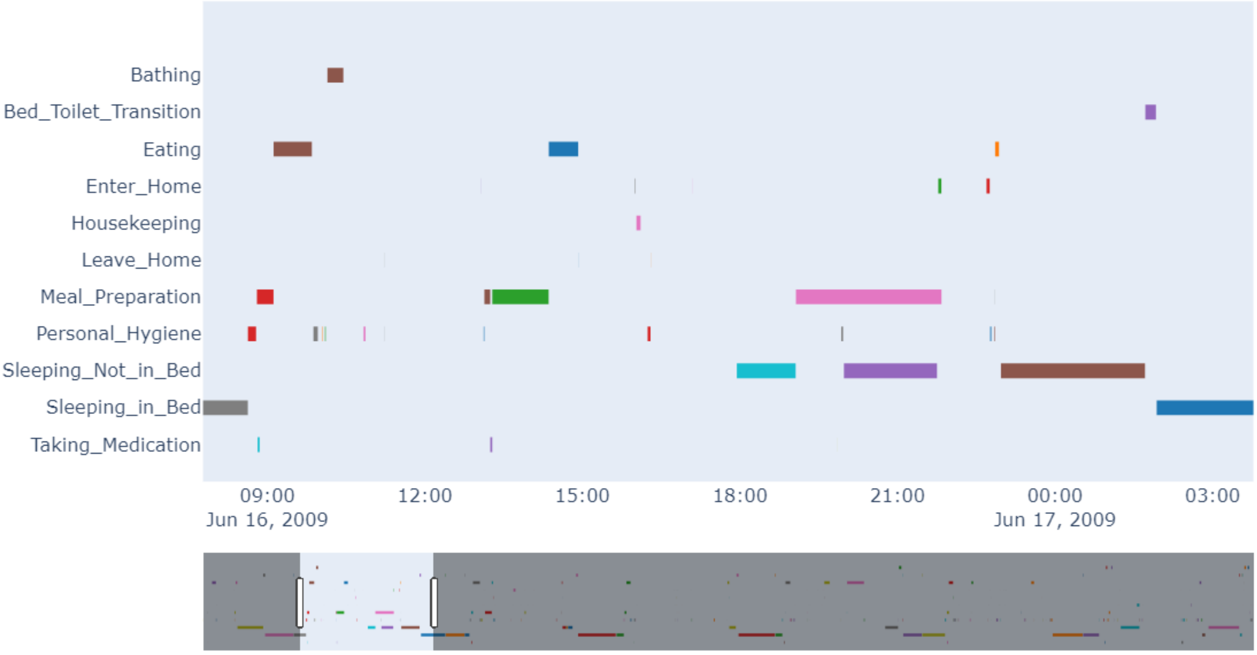


Figure B.5: Home2 Activities

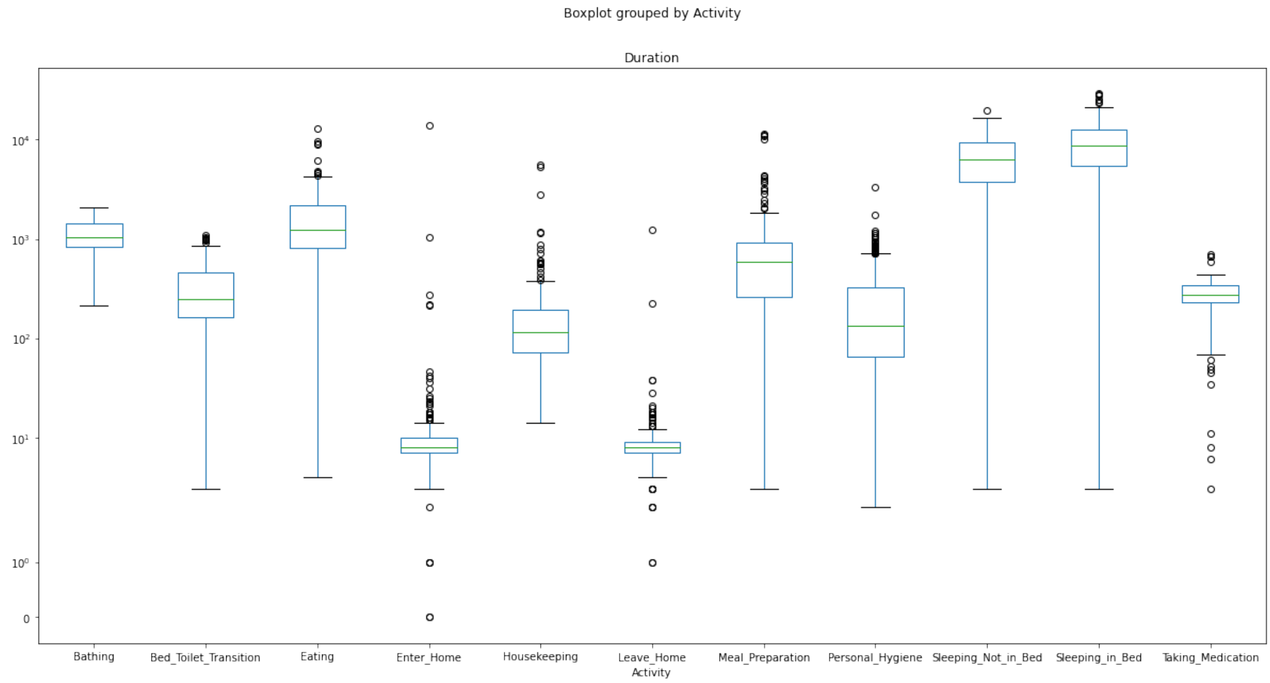


Figure B.6: Home2 Activity Duration

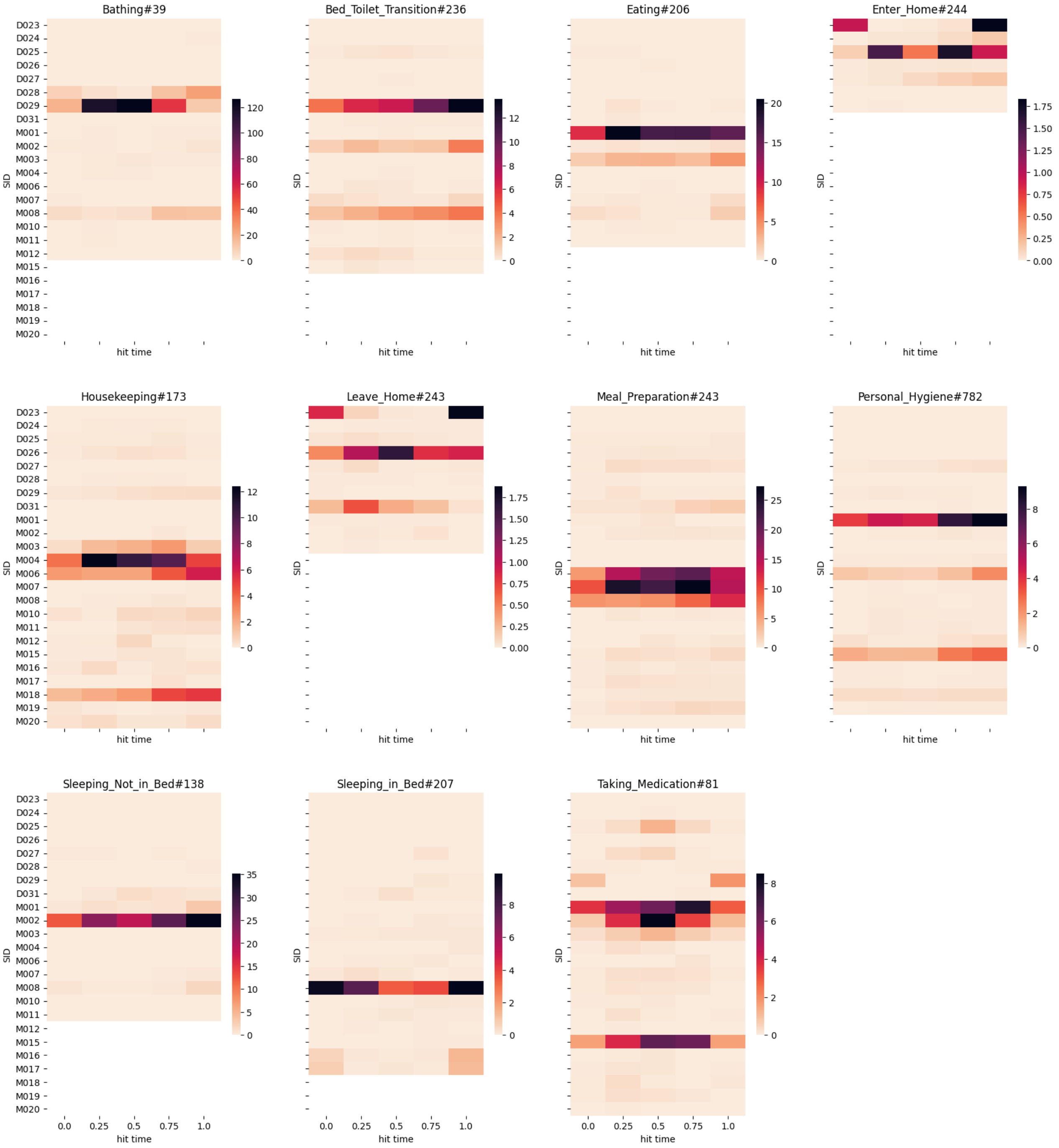


Figure B.7: Home2 dataset Hit Time. It shows the number of sensor events occurred at each time for each activity. On the legend, the average number of sensor's heats are shown.

Bibliography

[22] *Coronavirus (COVID-19) Dashboard*. 2022. URL: https://covid19.who.int/.

[Abd+21] Karrar Hameed Abdulkareem, Mazin Abed Mohammed, Ahmad Salim, Muhammad Arif, Oana Geman, Deepak Gupta, and Ashish Khanna. “Realizing an Effective COVID-19 Diagnosis System Based on Machine Learning and IoT in Smart Hospital Environment”. In: *IEEE Internet of Things Journal* 8.21 (2021), pp. 15919–15928. ISSN: 23274662. DOI: 10.1109/JIOT.2021.3050775.

[Abd+23] Mahmoud Khaled Abd-Ellah, Ashraf A.M. Khalaf, Reda R. Gharieb, and Dina A. Hassanin. “Automatic diagnosis of common carotid artery disease using different machine learning techniques”. In: *Journal of Ambient Intelligence and Humanized Computing* 14.1 (2023), pp. 113–129. ISSN: 18685145. DOI: 10.1007/s12652-021-03295-6.

[Abi+22] Farhan Fuad Abir, Khalid Alyafei, Muhammad E.H. Chowdhury, Amith Khandakar, Rashid Ahmed, Muhammad Maqsud Hossain, Sakib Mahmud, Ashiqur Rahman, Tareq O. Abbas, Susu M. Zughaier, and Khalid Kamal Naji. “PCovNet: A presymptomatic COVID-19 detection framework using deep learning model using wearables data”. In: *Computers in Biology and Medicine* 147.June (2022), p. 105682. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2022.105682.

[AC19] Samaneh Aminikhanghahi and Diane J. Cook. “Enhancing activity recognition using CPD-based activity segmentation”. In: *Pervasive and Mobile Computing* 53 (2019), pp. 75–89. ISSN: 15741192. DOI: 10.1016/j.pmcj.2019.01.004. URL: https://doi.org/10.1016/j.pmcj.2019.01.004.

[Ada+19] Stavros P. Adam, Stamatios Aggelos N. Alexandropoulos, Panos M. Pardalos, and Michael N. Vrahatis. “No free lunch theorem: A review”. In: *Springer Optimization and Its Applications* 145 (2019), pp. 57–82. ISSN: 19316836. DOI: 10.1007/978-3-030-12767-1{_}5.

[AE17] Hande Alemdar and Cem Ersoy. “Multi-resident activity tracking and recognition in smart environments”. In: *Journal of Ambient Intelligence and Humanized Computing* 8.4 (Aug. 2017), pp. 513–529. ISSN: 1868-5137. DOI: 10.1007/s12652-016-0440-x.

[Aer10] H.J.W.L. Aerts. “Molecular imaging of biologic characteristics and drug uptake : towards personalized medicine using dose painting”. PhD thesis. Maastricht University, 2010. ISBN: 9789461590220. DOI: 10.26481/dis.20101216ha.

URL: https://cris.maastrichtuniversity.nl/en/publications/48c95433-4508-4139-a5ca-64a9c0d97a2f.

[Agu+19] Gabriel Jonas Aguiar, Rafael Gomes Mantovani, Saulo M. Mastelini, André C.P.F.L. de Carvalho, Gabriel F.C. Campos, and Sylvio Barbon Junior. "A meta-learning approach for selecting image segmentation algorithm". In: *Pattern Recognition Letters* 128 (2019), pp. 480–487. ISSN: 01678655. DOI: 10.1016/j.patrec.2019.10.018.

[Ahm+22] Niaz Ahmad, Jawad Khan, Jeremy Yuhyun Kim, and Youngmoon Lee. "Joint Human Pose Estimation and Instance Segmentation with PosePlusSeg". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.1 (2022), pp. 69–76. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i1.19880. URL: https://www.aaai.org/AAAI22Papers/AAAI-6681.AhmadN.pdf.

[Ahn+19] Sang Hee Ahn, Adam Unjin Yeo, Kwang Hyeon Kim, Chankyu Kim, Youngmoon Goh, Shinhaeng Cho, Se Byeong Lee, Young Kyung Lim, Haksoo Kim, Dongho Shin, Taeyoon Kim, Tae Hyun Kim, Sang Hee Youn, Eun Sang Oh, and Jong Hwi Jeong. "Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer". In: *Radiation Oncology* 14.1 (2019), pp. 1–13. ISSN: 1748717X. DOI: 10.1186/s13014-019-1392-z.

[Ala+22] Arash Alavi, Gireesh K Bogu, Meng Wang, Ekanath Srihari Rangan, Andrew W Brooks, Qiwen Wang, Emily Higgs, Alessandra Celli, Tejaswini Mishra, Ahmed A Metwally, Kexin Cha, Peter Knowles, Amir A Alavi, Rajat Bhasin, Shrinivas Panchamukhi, Diego Celis, Tagore Aditya, Alexander Honkala, Benjamin Rolnik, Erika Hunting, Orit Dagan-Rosenfeld, Arshdeep Chauhan, Jessi W. Li, Caroline Bejikian, Vandhana Krishnan, Lettie McGuire, Xiao Li, Amir Bahmani, and Michael P Snyder. "Real-time alerting system for COVID-19 and other stress events using wearable data". In: *Nature Medicine* 57022 (Nov. 2022). ISSN: 1078-8956. DOI: 10.1038/s41591-021-01593-2. URL: https://www.nature.com/articles/s41591-021-01593-2.

[ALB23] Tariq Al Shoura, Henry Leung, and Bhashyam Balaji. "An Adaptive Kernels Layer for Deep Neural Networks Based on Spectral Analysis for Image Applications". In: *Sensors* 23.3 (2023). ISSN: 14248220. DOI: 10.3390/s23031527.

[Ale+17] Elias Alevizos, Anastasios Skarlatidis, Alexander Artikis, and George Paliouras. "Probabilistic Complex Event Recognition: A Survey". In: *ACM Computing Surveys* 50.5 (Feb. 2017), pp. 1–31. ISSN: 03600300. DOI: 10.1145/3117809. URL: http://arxiv.org/abs/1702.06379%20http://dl.acm.org/citation.cfm?doid=3145473.3117809.

[Ale15] Hande Alemdar. "Human Activity Recognition With Wireless Sensor Networks USING MACHINE LEARNING". PhD thesis. Bogazici University, 2015.

[Als+21] Shikah J. Alsunaidi, Abdullah M. Almuhaideb, Nehad M. Ibrahim, Fatema S. Shaikh, Kawther S. Alqudaihi, Fahd A. Alhaidari, Irfan Ullah Khan, Nida Aslam, and Mohammed S. Alshahrani. "Applications of big data analytics to control covid-19 pandemic". In: *Sensors* 21.7 (2021). ISSN: 14248220. DOI: 10.3390/s21072282.

[ANJ+22] ARFA ANJUM, SEEMA JAGGI, SHWETANK LALL, ELDHO VARGHESE, ANIL RAI, ARPAN BHOWMIK, and DWIJESH CHANDRA MISHRA. “Segmentation of genomic data through multivariate statistical approaches: comparative analysis”. In: *The Indian Journal of Agricultural Sciences* 92.7 (Mar. 2022), pp. 892–896. ISSN: 0019-5022. DOI: 10.56093/ijas.v92i7.118040. URL: https://epubs.icar.org.in/index.php/IJAgS/article/view/118040.

[Ans+22] Mohammed Yusuf Ansari, Yin Yang, Shidin Balakrishnan, Julien Abinahed, Abdulla Al-Ansari, Mohamed Warfa, Omran Almokdad, Ali Barah, Ahmed Omer, Ajay Vikram Singh, Pramod Kumar Meher, Jolly Bhadra, Osama Halabi, Mohammad Farid Azampour, Nassir Navab, Thomas Wendler, and Sarada Prasad Dakua. “A lightweight neural network with multiscale feature enhancement for liver CT segmentation”. In: *Scientific Reports* 12.1 (2022), pp. 1–12. ISSN: 20452322. DOI: 10.1038/s41598-022-16828-6. URL: https://doi.org/10.1038/s41598-022-16828-6.

[Ant+22] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. “The Medical Segmentation Decathlon”. In: *Nature Communications* 13.1 (July 2022), p. 4128. ISSN: 2041-1723. DOI: 10.1038/s41467-022-30695-9. URL: https://www.nature.com/articles/s41467-022-30695-9.

[App22] Apple Inc. *About Face ID advanced technology*. 2022. URL: https://support.apple.com/en-us/HT208108.

[Ari+22] Paola Patricia Ariza-Colpas, Enrico Vicario, Ana Isabel Oviedo-Carrascal, Shariq Butt Aziz, Marlon Alberto Piñeres-Melo, Alejandra Quintero-Linero, and Fulvio Patara. *Human Activity Recognition Data Analysis: History, Evolutions, and New Trends*. Apr. 2022. DOI: 10.3390/s22093401. URL: https://www.mdpi.com/1424-8220/22/9/3401/pdf?version=1651203966%20https://www.mdpi.com/1424-8220/22/9/3401.

[Asa+01] Tetsuo Asano, Danny Z. Chen, Naoki Katoh, and Takeshi Tokuyama. “Efficient algorithms for optimization-based image segmentation”. In: *International Journal of Computational Geometry and Applications* 11.2 (Apr. 2001), pp. 145–166. ISSN: 02181959. DOI: 10.1142/S0218195901000420. URL: https://www.worldscientific.com/doi/abs/10.1142/S0218195901000420.

[Asg+20] Parviz Asghari, Elnaz Soleimani, Ehsan Nazerfard, and . “Online human activity recognition employing hierarchical hidden Markov models”. In: *Journal of Ambient Intelligence and Humanized Computing* 11.3 (Mar. 2020), pp. 1141–1152. ISSN: 1868-5137. DOI: 10.1007/s12652-019-01380-5. URL: http://link.springer.com/10.1007/s12652-019-01380-5.

[Asg+21] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. “Deep semantic segmentation of natural and medical images: a review”. In: *Artificial Intelligence Review* 54.1 (Jan. 2021), pp. 137–178. ISSN: 0269-2821. DOI: 10.1007/s10462-020-09854-1. URL: https://link.springer.com/10.1007/s10462-020-09854-1.

[ASN19] Parviz Asghari, Elnaz Soleimani, and Ehsan Nazerfard. “Online human activity recognition employing hierarchical hidden Markov models”. In: *Journal of Ambient Intelligence and Humanized Computing* (July 2019). ISSN: 1868-5137. DOI: 10.1007/s12652-019-01380-5. URL: http://arxiv.org/abs/1903.04820%20http://link.springer.com/10.1007/s12652-019-01380-5.

[ATE15] Hande Alemdar, Can Tunca, and Cem Ersoy. “Daily life behaviour monitoring for health assessment using machine learning: bridging the gap between domains”. In: *Personal and Ubiquitous Computing* 19.2 (2015), pp. 303–315. ISSN: 16174909. DOI: 10.1007/s00779-014-0823-y.

[Awa+21] George Awad, Asad A Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Yvette Graham, and Georges Qu. “TRECVID 2020 : A comprehensive campaign for evaluating video retrieval tasks across multiple application domains”. In: *Proceedings of TRECVID*. USA: NIST, 2021, pp. 1–55.

[Ayd+22] Orhun Utku Aydin, Abdel Aziz Taha, Adam Hilbert, Ahmed A. Khalil, Ivana Galinovic, Jochen B. Fiebach, Dietmar Frey, and Vince Istvan Madai. “Correction: On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking”. In: *European Radiology Experimental* 6.1 (Oct. 2022), p. 56. ISSN: 2509-9280. DOI: 10.1186/s41747-022-00309-6. URL: https://eurradiolexp.springeropen.com/articles/10.1186/s41747-022-00309-6.

[Aza+22] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. “Medical Image Segmentation Review: The success of U-Net”. In: *arXiv* (2022), pp. 1–38. URL: http://arxiv.org/abs/2211.14830.

[Bar+20] Joe Barrow, Rajiv Jain, Vlad I. Morariu, Varun Manjunatha, Douglas W. Oard, and Philip Resnik. “A joint model for document segmentation and segment labeling”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2020), pp. 313–322. ISSN: 0736587X. DOI: 10.18653/v1/2020.acl-main.29. URL: https://aclanthology.org/2020.acl-main.29.pdf.

[BBB19] A. V. Bogatyrev, V. A. Bogatyrev, and S. V. Bogatyrev. "Multipath Redundant Transmission with Packet Segmentation". In: *2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*. IEEE, June 2019, pp. 1–4. ISBN: 978-1-7281-2288-5. DOI: 10.1109/WECONF.2019.8840643. URL: https://ieeexplore.ieee.org/document/8840643/.

[Ber+18] Jürgen Bernard, Christian Bors, Markus Bögl, Christian Eichner, and Et.all. "Combining the Automated Segmentation and Visual Analysis of Multivariate Time Series". In: *EuroVisWorkshop on Visual Analytics* (2018). DOI: 10.2312/eurova.20181112.

[Bil+20] Cagdas Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulovic. "A Framework for the Robust Evaluation of Sound Event Detection". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. 2020, pp. 61–65. ISBN: 9781509066315. DOI: 10.1109/ICASSP40776.2020.9052995.

[Bil+23] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C. Paetzold, Suprosanna Shit, Xiaobin Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B. Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró-i-Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. "The Liver Tumor Segmentation Benchmark (LiTS)". In: *Medical Image Analysis* 84 (Feb. 2023), p. 102680. ISSN: 13618415. DOI: 10.1016/j.media.2022.102680. URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841522003085.

[BJ22] John S.H. Baxter and Pierre Jannin. "Bias in machine learning for computer-assisted surgery and medical image processing". In: *Computer Assisted Surgery* 27.1 (2022), pp. 1–3. ISSN: 24699322. DOI: 10.1080/24699322.2021.2013619. URL: https://doi.org/10.1080/24699322.2021.2013619.

[BNE21] Damien Bouchabou, Sao Mai Nguyen, and Et.al. “A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning”. In: *Sensors* 21.18 (2021). ISSN: 14248220. DOI: 10.3390/s21186037. URL: https://www.mdpi.com/1424-8220/21/18/6037/pdf?version=1631186192.

[Bou+21a] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit LeDuc, and Ioannis Kanellos. “Fully Convolutional Network Bootstrapped by Word Encoding and Embedding for Activity Recognition in Smart Homes”. In: *Communications in Computer and Information Science*. Vol. 1370. Dec. 2021, pp. 111–125. ISBN: 9789811605741. DOI: 10.1007/978-981-16-0575-8{_}9. URL: https://hal.archives-ouvertes.fr/hal-03032449/document%20https://link.springer.com/10.1007/978-981-16-0575-8_9%20http://arxiv.org/abs/2012.02300.

[Bou+21b] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit Leduc, and Ioannis Kanellos. “Using language model to bootstrap human activity recognition ambient sensors based in smart homes”. In: *Electronics (Switzerland)* 10.20 (2021), pp. 1–25. ISSN: 20799292. DOI: 10.3390/electronics10202498. URL: http://nguyensmai.free.fr/publication/Bouchabou2021E.pdf.

[Bou+22] David Bouget, André Pedersen, Asgeir S. Jakola, Vasileios Kavouridis, Kyrre E. Emblem, Roelant S. Eijgelaar, Ivar Kommers, Hilko Ardon, Frederik Barkhof, Lorenzo Bello, Mitchel S. Berger, Marco Conti Nibali, Julia Furtner, Shawn Hervey-Jumper, Albert J. S. Idema, Barbara Kiesel, Alfred Kloet, Emmanuel Mandonnet, Domenique M. J. Müller, Pierre A. Robe, Marco Rossi, Tommaso Sciortino, Wimar A. Van den Brink, Michiel Wagemakers, Georg Widhalm, Marnix G. Witte, Aeilko H. Zwinderman, Philip C. De Witt Hamer, Ole Solheim, and Ingerid Reinertsen. “Preoperative Brain Tumor Imaging: Models and Software for Segmentation and Standardized Reporting”. In: *Frontiers in Neurology* 13.July (2022). DOI: 10.3389/fneur.2022.932219.

[Bou22] Damien Bouchabou. “Human activity recognition in smart homes : tackling data variability using context-dependent deep learning, transfer learning and data synthesis”. PhD thesis. Ecole nationale superieure Mines-Telecom Atlantique, 2022. URL: https://theses.hal.science/tel-03728064%0Ahttps://theses.hal.science/tel-03728064/document.

[BS21] Gireesh K. Bogu and Michael P. Snyder. “Deep learning-based detection of COVID-19 using wearables data”. In: *medRxiv* (2021), p. 2021.01.08.21249474. URL: https://www.medrxiv.org/content/10.1101/2021.01.08.21249474v1.

[Bur+04] Neil G. Burnet, Simon J. Thomas, Kate E. Burton, and Sarah J. Jefferies. “Defining the tumour and target volumes for radiotherapy”. In: *Cancer Imaging* 4.2 (2004), pp. 153–161. ISSN: 14707330. DOI: 10.1102/1470-7330.2004.0054. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1434601/pdf/ci040153.pdf.

[Byr+20] Michal Byra, Piotr Jarosik, Aleksandra Szubert, Michael Galperin, Haydee Ojeda-Fournier, Linda Olson, Mary O'Boyle, Christopher Comstock, and Michael Andre. "Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network". In: *Biomedical Signal Processing and Control* 61 (Aug. 2020), p. 102027. ISSN: 17468094. DOI: 10.1016/j.bspc.2020.102027. URL: https://linkinghub.elsevier.com/retrieve/pii/S174680942030183X.

[BZB22] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. "End-to-End Referring Video Object Segmentation with Multimodal Transformers". In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 4975–4985. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00493. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Botach_End-to-End_Referring_Video_Object_Segmentation_With_Multimodal_Transformers_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9880167/.

[BZL23] Hua Bao, Yuqing Zhu, and Qing Li. "Hybrid-scale contextual fusion network for medical image segmentation". In: *Computers in Biology and Medicine* 152.December 2022 (2023), p. 106439. ISSN: 18790534. DOI: 10.1016/j.compbiomed.2022.106439. URL: https://doi.org/10.1016/j.compbiomed.2022.106439.

[Cak+19] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. "Deep metric learning to rank". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2019-June. 2019, pp. 1861–1870. ISBN: 9781728132938. DOI: 10.1109/CVPR.2019.00196.

[Can86] John Canny. "A Computational Approach to Edge Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (1986), pp. 679–698. ISSN: 01628828. DOI: 10.1109/TPAMI.1986.4767851.

[Car+21] Jacinto Carrasco, David López, Ignacio Aguilera-Martos, Diego García-Gil, Irina Markova, Marta García-Barzana, Manuel Arias-Rodil, Julián Luengo, Francisco Herrera, David López, Ignacio Aguilera, Diego García, Marta García-Barzana, Manuel Arias-Rodil, Julián Luengo, and Francisco Herrera. "Anomaly detection in predictive maintenance: A new evaluation framework for temporal unsupervised anomaly detection algorithms". In: *Neurocomputing* 462 (2021), pp. 440–452. ISSN: 18728286. DOI: 10.1016/j.neucom.2021.07.095. URL: http://arxiv.org/abs/2105.12818%20https://doi.org/10.1016/j.neucom.2021.07.095.

[CC20] T K Chan and Cheng Siong Chin. "A Comprehensive Review of Polyphonic Sound Event Detection". In: *IEEE Access* 8 (2020), pp. 103339–103373. DOI: 10.1109/ACCESS.2020.2999388.

[Cha+21] Asma Channa, Nirvana Popescu, Justyna Skibinska, and Radim Burget. "The rise of wearable devices during the COVID-19 pandemic: A systematic review". In: *Sensors* 21.17 (2021), pp. 1–22. ISSN: 14248220. DOI: 10.3390/s21175787.

[Che+20] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. "Dynamic convolution: Attention over convolution kernels". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020), pp. 11027–11036. ISSN: 10636919. DOI: 10.1109/CVPR42600.2020.01104. URL: https://openaccess.thecvf.com/content_CVPR_2020/papers/Chen_Dynamic_Convolution_Attention_Over_Convolution_Kernels_CVPR_2020_paper.pdf.

[Che+21a] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation". In: *arXiv* (Feb. 2021), pp. 1–13. URL: http://arxiv.org/abs/2102.04306.

[Che+21b] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. "Deep Learning for Sensor-based Human Activity Recognition". In: *ACM Computing Surveys* 54.4 (July 2021), pp. 1–40. ISSN: 0360-0300. DOI: 10.1145/3447744. URL: http://arxiv.org/abs/2001.07416%20https://dl.acm.org/doi/10.1145/3447744.

[Che+22a] Yuantao Chen, Jiajun Tao, Linwu Liu, Jie Xiong, Runlong Xia, Jingbo Xie, Qian Zhang, and Kai Yang. "Research of improving semantic image segmentation based on a feature fusion model". In: *Journal of Ambient Intelligence and Humanized Computing* 13.11 (2022), pp. 5033–5045. ISSN: 18685145. DOI: 10.1007/s12652-020-02066-z. URL: https://doi.org/10.1007/s12652-020-02066-z.

[Che+22b] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. "C-CAM: Causal CAM for Weakly Supervised Semantic Segmentation on Medical Image". In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 11666–11675. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.01138. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Chen_C-CAM_Causal_CAM_for_Weakly_Supervised_Semantic_Segmentation_on_Medical_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879271/.

[Che+22c] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. "Sparse Instance Activation for Real-Time Instance Segmentation". In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 4423–4432. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00439. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Cheng_Sparse_Instance_Activation_for_Real-Time_Instance_Segmentation_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9880463/.

[Che+23] Gongping Chen, Yuming Liu, Jiang Qian, Jianxun Zhang, Xiaotao Yin, Liang Cui, and Yu Dai. "DSEU-net: A novel deep supervision SEU-net for medical ultrasound image segmentation". In: *Expert Systems with Applications* 223.March (2023), p. 119939. ISSN: 09574174. DOI: 10.1016/j.eswa.2023.119939. URL: https://doi.org/10.1016/j.eswa.2023.119939.

[Chr21] Christopher D.Manning. “Speech and Language Processing: An introduction to natural language processing”. In: *SPEECH and LANGUAGE PROCESSING An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition* (2021), pp. 1–18. URL: http://www.cs.colorado.edu/~martin/slp.html.

[Cic+16] Ozgun Cicek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. “3D U-net: Learning dense volumetric segmentation from sparse annotation”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9901 LNCS (2016), pp. 424–432. ISSN: 16113349. DOI: 10.1007/978-3-319-46723-8{_}49. URL: https://arxiv.org/pdf/1606.06650.pdf.

[Cip+22] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, and Costantino Grana. “Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 21105–21114. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.02046. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Cipriano_Improving_Segmentation_of_the_Inferior_Alveolar_Nerve_Through_Deep_Label_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879649/.

[Cla+13] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository”. In: *Journal of Digital Imaging* 26.6 (Dec. 2013), pp. 1045–1057. ISSN: 0897-1889. DOI: 10.1007/s10278-013-9622-7. URL: http://link.springer.com/10.1007/s10278-013-9622-7.

[CN15] Diane J. Cook and C Krishnan Narayanan. *Activity Learning: Discovering, Recognizing, and Predicting Human Behavior from Sensor Data*. 1st ed. Wiley Series on Parallel and Distributed Computing. Wiley, 2015. ISBN: 111889376X,9781118893760.

[Cod+19] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)”. In: (2019), pp. 1–12. URL: http://arxiv.org/abs/1902.03368.

[Coo12] Diane J. Cook. “Learning Setting-Generalized Activity Models for Smart Spaces”. In: *IEEE Intelligent Systems* 27.1 (Jan. 2012), pp. 32–38. ISSN: 1541-1672. DOI: 10.1109/MIS.2010.112. URL: http://ieeexplore.ieee.org/document/5567086/.

[Cor+09] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, and CLRS. *Introduction to Algorithms, Third Edition*. 3rd. The MIT Press, 2009. ISBN: 9780262033848. DOI: 10.5555/1614191.

[Cum+17] Julien Cumin, Grégoire Lefebvre, Fano Ramparany, and James L. Crowley. "A Dataset of Routine Daily Activities in an Instrumented Home". In: *International Conference on Ubiquitous Computing and Ambient Intelligence*. 2017, pp. 413–425. DOI: 10.1007/978-3-319-67585-5{_}43. URL: http://link.springer.com/10.1007/978-3-319-67585-5_43.

[Cum+18] Julien Cumin, Julien Cumin Recognizing, Human-computer Interaction, and Julien Cumin. "Recognizing and predicting activities in smart homes". PhD thesis. Université Grenoble Alpes, Dec. 2018. URL: https://tel.archives-ouvertes.fr/tel-02057332.

[DAr20] John D. Kelleher; Brian Mac Namee; Aoife DArcy. *Fundamentals of Machine Learning for Predictive Data Analytics : Algorithms, Worked Examples, and Case Studies*. MIT Press, 2020. ISBN: 9780262044691.

[DB15] Jayanta K. Dutta and Bonny Banerjee. "Online Detection of Abnormal Events Using Incremental Coding Length". In: *AAAI Conference on Artificial Intelligence*. Vol. 5. 2015. ISBN: 9781577357032. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9799.

[De-+18] Emiro De-La-Hoz-Franco, Paola Ariza-Colpas, Javier Medina Quero, and Macarena Espinilla. "Sensor-Based Datasets for Human Activity Recognition – A Systematic Review of Literature". In: *IEEE Access* 6.c (2018), pp. 59192–59210. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2873502. URL: https://ieeexplore.ieee.org/document/8478653/.

[Dev+21] Sindhu Devunooru, Abeer Alsadoon, P. W. C. Chandana, and Azam Beg. "Deep learning neural networks for medical image segmentation of brain tumours for diagnosis: a recent review and taxonomy". In: *Journal of Ambient Intelligence and Humanized Computing* 12.1 (Jan. 2021), pp. 455–483. ISSN: 1868-5137. DOI: 10.1007/s12652-020-01998-w. URL: https://doi.org/10.1007/s12652-020-01998-w%20https://link.springer.com/10.1007/s12652-020-01998-w.

[Dic45] Lee R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (July 1945), pp. 297–302. ISSN: 00129658. DOI: 10.2307/1932409. URL: http://doi.wiley.com/10.2307/1932409.

[Din+22] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. "Language-Bridged Spatial-Temporal Interaction for Referring Video Object Segmentation". In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 4954–4963. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00491. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Ding_Language-Bridged_Spatial-Temporal_Interaction_for_Referring_Video_Object_Segmentation_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9880159/.

[Dol20] Sara Dolnicar. "Market segmentation analysis in tourism: a perspective paper". In: *Tourism Review* 75.1 (Feb. 2020), pp. 45–48. ISSN: 1660-5373. DOI: 10.1108/TR-02-2019-0041. URL: http://dx.doi.org/10.1108/00251749210013050%20https://www.emerald.com/insight/content/doi/10.1108/TR-02-2019-0041/full/html.

[DTP21] Florenc Demrozi, Cristian Turetta, and Graziano Pravadelli. “B-HAR: an open-source baseline framework for in depth study of human activity recognition datasets and workflows”. In: *arXiv* (Jan. 2021), pp. 1–9. URL: http://arxiv.org/abs/2101.10870.

[Du+22] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. “Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 4310–4319. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00428. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Du_Weakly_Supervised_Semantic_Segmentation_by_Pixel-to-Prototype_Contrast_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879153/.

[Eal+22] Stephen Neal Joshua Eali, Debnath Bhattacharyya, Thirupathi Rao Nakka, and Seng Phil Hong. “A Novel Approach in Bio-Medical Image Segmentation for Analyzing Brain Cancer Images with U-NET Semantic Segmentation and TPLD Models Using SVM”. In: *Traitement du Signal* 39.2 (2022), pp. 419–430. ISSN: 19585608. DOI: 10.18280/ts.390203.

[EBK21] E. Ernawati, S. S.K. Baharin, and F. Kasmin. “A review of data mining methods in RFM-based customer segmentation”. In: *Journal of Physics: Conference Series* 1869.1 (2021). ISSN: 17426596. DOI: 10.1088/1742-6596/1869/1/012085. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1869/1/012085/pdf.

[Eel+20] Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B. Blaschko. “Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index”. In: *IEEE Transactions on Medical Imaging* 39.11 (Nov. 2020), pp. 3679–3690. ISSN: 0278-0062. DOI: 10.1109/TMI.2020.3002417. URL: https://ieeexplore.ieee.org/document/9116807/.

[Eli+21] Christelle Elias, Abel Sekri, Pierre Leblanc, Michel Cucherat, and Philippe Vanhems. “The incubation period of COVID-19: A meta-analysis”. In: *International Journal of Infectious Diseases* 104 (2021), pp. 708–710. ISSN: 18783511. DOI: 10.1016/j.ijid.2021.01.069.

[EM96] Paul H. C. Eilers and Brian D. Marx. “Flexible smoothing with B-splines and penalties”. In: *Statistical Science* 11.2 (May 1996), pp. 89–102. ISSN: 0883-4237. DOI: 10.1214/ss/1038425655. URL: https://projecteuclid.org/journals/statistical-science/volume-11/issue-2/Flexible-smoothing-with-B-splines-and-penalties/10.1214/ss/1038425655.full.

[Emm+15] Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. “A Meta-Analysis of the Anomaly Detection Problem”. In: 2015 (2015). URL: http://arxiv.org/abs/1503.01158.

[Fag+22] Shahriar Faghani, Bardia Khosravi, Kuan Zhang, Mana Moassefi, Jaidip Manikrao Jagtap, Fred Nugen, Sanaz Vahdati, Shiba P. Kuanar, Seyed Moein Rassoulinejad-Mousavi, Yashbir Singh, Diana V. Vera Garcia, Pouria Rouzrokh, and Bradley J. Erickson. “Mitigating Bias in Radiology Machine Learning: 3. Performance Metrics”. In: *Radiology: Artificial Intelligence* 4.5 (Sept. 2022). ISSN: 2638-6100. DOI: 10.1148/ryai.220061. URL: http://pubs.rsna.org/doi/10.1148/ryai.220061.

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. “Model-agnostic meta-learning for fast adaptation of deep networks”. In: *34th International Conference on Machine Learning, ICML 2017* 3 (2017), pp. 1856–1868.

[FC17] Kyle D. Feuz and Diane J. Cook. “Collegial activity learning between heterogeneous sensors”. In: *Knowledge and Information Systems* 53.2 (2017), pp. 337–364. ISSN: 02193116. DOI: 10.1007/s10115-017-1043-3.

[Fer+20] Tharindu Fernando, Houman Ghaemmaghami, Simon Denman, Sridha Sridharan, Nayyar Hussain, and Clinton Fookes. “Heart Sound Segmentation Using Bidirectional LSTMs With Attention”. In: *IEEE Journal of Biomedical and Health Informatics* 24.6 (June 2020), pp. 1601–1609. ISSN: 2168-2194. DOI: 10.1109/JBHI.2019.2949516. URL: https://ieeexplore.ieee.org/document/8883031/.

[Fer+21] Giacomo Ferroni, Nicolas Turpault, Juan Azcarreta, Francesco Tuveri, Romain Serizel, Cağdaş Cagdaş Cagdas Bilen, Sacha Krstulovic, and Sacha Krstulović. “Improving Sound Event Detection Metrics: Insights from DCASE 2020”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2021-June (2021), pp. 631–635. ISSN: 15206149. DOI: 10.1109/icassp39728.2021.9414711. URL: http://arxiv.org/abs/2010.13648.

[FLH23] Zhaojin Fu, Jinjiang Li, and Zhen Hua. “MSA-Net: Multiscale spatial attention network for medical image segmentation”. In: *Alexandria Engineering Journal* 70 (2023), pp. 453–473. ISSN: 11100168. DOI: 10.1016/j.aej.2023.02.039. URL: https://doi.org/10.1016/j.aej.2023.02.039.

[Fu11] Tak Chung Fu. “A review on time series data mining”. In: *Engineering Applications of Artificial Intelligence* 24.1 (2011), pp. 164–181. DOI: 10.1016/j.engappai.2010.09.007.

[FVS22] Benjamin Filtjens, Bart Vanrumste, and Peter Slaets. “Skeleton-Based Action Segmentation with Multi-Stage Spatial-Temporal Graph Convolutional Neural Networks”. In: *IEEE Transactions on Emerging Topics in Computing* (Feb. 2022), pp. 1–11. ISSN: 2168-6750. DOI: 10.1109/TETC.2022.3230912. URL: http://arxiv.org/abs/2202.01727%20http://dx.doi.org/10.1109/TETC.2022.3230912%20https://ieeexplore.ieee.org/document/9998567/.

[Gal+13] Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jimenez Cardenas, Thomas Brox, and Bernt Schiele. “A unified video segmentation benchmark: Annotation, metrics and analysis”. In: *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 3527–3534. DOI: 10.1109/ICCV.

2013.438. URL: https://openaccess.thecvf.com/content_iccv_2013/papers/Galasso_A_Unified_Video_2013_ICCV_paper.pdf.

[Gao+21] Shang Hua Gao, Ming Ming Cheng, Kai Zhao, Xin Yu Zhang, Ming Hsuan Yang, and Philip Torr. “Res2Net: A New Multi-Scale Backbone Architecture”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.2 (2021), pp. 652–662. ISSN: 19393539. DOI: 10.1109/TPAMI.2019.2938758. URL: https://sci.bban.top/pdf/10.1109/TPAMI.2019.2938758.pdf#view=FitH.

[GD95] Diana F. Gordon and Marie Desjardins. “Evaluation and Selection of Biases in Machine Learning”. In: *Machine Learning* 20.1 (1995), pp. 5–22. ISSN: 15730565. DOI: 10.1023/A:1022630017346.

[GGY10] Santosh K Gaikwad, Bharti W Gawali, and Pravin Yannawar. “A review on speech recognition technique”. In: *International Journal of Computer Applications* 10.3 (2010), pp. 16–24.

[Gib+18] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, and Dean C. Barratt. “Automatic Multi-Organ Segmentation on Abdominal CT with Dense V-Networks”. In: *IEEE Transactions on Medical Imaging* 37.8 (2018), pp. 1822–1834. ISSN: 1558254X. DOI: 10.1109/TMI.2018.2806309.

[Gjo+15] Hristijan Gjoreski, Simon Kozina, Matjaz Gams, Mitja Lustrek, Juan Antonio Alvarez-Garcia, Jin-Hyuk Hong, Anind K. Dey, Maurizio Bocca, and Neal Patwari. “Competitive Live Evaluations of Activity-Recognition Systems”. In: *IEEE Pervasive Computing* 14.1 (Jan. 2015), pp. 70–77. DOI: 10.1109/MPRV.2015.3.

[GP17] Yu Guan and Thomas Plötz. “Ensembles of Deep LSTM Learners for Activity Recognition using Wearables”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.2 (2017), pp. 1–28. ISSN: 2474-9567. DOI: 10.1145/3090076.

[Gre+19] Edward Grefenstette, Brandon Amos, Denis Yarats, Phu Mon Htut, Artem Molchanov, Franziska Meier, Douwe Kiela, Kyunghyun Cho, and Soumith Chintala. “Generalized Inner Loop Meta-Learning”. In: *arXiv* (Oct. 2019), pp. 1–17. URL: http://arxiv.org/abs/1910.01727.

[Gre04] George J. Grevera. “The ”dead reckoning” signed distance transform”. In: *Computer Vision and Image Understanding* 95.3 (2004), pp. 317–333. ISSN: 10773142. DOI: 10.1016/j.cviu.2004.05.002.

[Gri21] Mourad Gridach. “PyDiNet: Pyramid Dilated Network for medical image segmentation”. In: *Neural Networks* 140 (2021), pp. 274–281. ISSN: 18792782. DOI: 10.1016/j.neunet.2021.03.023. URL: https://doi.org/10.1016/j.neunet.2021.03.023.

[GSC22] Shuyue Guan, Ravi K. Samala, and Weijie Chen. “Informing selection of performance metrics for medical image segmentation evaluation using configurable synthetic errors”. In: *arXiv* (Dec. 2022). URL: http://arxiv.org/abs/2212.14828.

[Gup+21] Anubha Gupta, Ritu Gupta, Shiv Gehlot, and Shubham Goswami. "Segpc-2021: Segmentation of multiple myeloma plasma cells in microscopic images". In: *IEEE Dataport* 1.1 (2021), p. 1. DOI: 10.21227/7np1-2q42.

[Gup+23] Anubha Gupta, Shiv Gehlot, Shubham Goswami, Sachin Motwani, Ritu Gupta, Avaro Garca Faura, Dejan Stepec, Tomaz Martincic, Reza Azad, Dorit Merhof, Afshin Bozorgpour, Babak Azad, Alaa Sulaiman, Deepanshu Pandey, Pradyumna Gupta, Sumit Bhattacharya, Aman Sinha, Rohit Agarwal, Xinyun Qiu, Yucheng Zhang, Ming Fan, Yoonbeom Park, Daehong Lee, Joon Sik Park, Kwangyeol Lee, and Jaehyung Ye. "SegPC-2021: A challenge & dataset on segmentation of Multiple Myeloma plasma cells from microscopic images". In: *Medical Image Analysis* 83.October 2022 (2023), p. 102677. ISSN: 13618423. DOI: 10.1016/j.media.2022.102677. URL: https://doi.org/10.1016/j.media.2022.102677.

[Ham+20] Rebeen Ali Hamad, Alberto Salguero Hidalgo, Mohamed-Rafik Bouguelia, Macarena Espinilla Estevez, and Javier Medina Quero. "Efficient Activity Recognition in Smart Homes Using Delayed Fuzzy Temporal Windows on Binary Sensors". In: *IEEE Journal of Biomedical and Health Informatics* 24.2 (2020), pp. 387–395. DOI: 10.1109/JBHI.2019.2918412.

[Ham+21] Rebeen Ali Hamad, Masashi Kimura, Longzhi Yang, Wai Lok Woo, and Bo Wei. "Dilated causal convolution with multi-head self attention for sensor human activity recognition". In: *Neural Computing and Applications* 33.20 (2021), pp. 13705–13722. ISSN: 14333058. DOI: 10.1007/s00521-021-06007-5. URL: https://doi.org/10.1007/s00521-021-06007-5%20https://link.springer.com/content/pdf/10.1007/s00521-021-06007-5.pdf.

[Han+22] Su Ho Han, Sukjun Hwang, Seoung Wug Oh, Yeonchool Park, Hyunwoo Kim, Min-Jung Kim, and Seon Joo Kim. "VISOLO: Grid-Based Space-Time Aggregation for Efficient Online Video Instance Segmentation". In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 2886–2895. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00291. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Han_VISOLO_Grid-Based_Space-Time_Aggregation_for_Efficient_Online_Video_Instance_Segmentation_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879517/.

[Has+20] Md Kamrul Hasan, Lavsen Dahal, Prasad N. Samarakoon, Fakrul Islam Tushar, and Robert Martí. "DSNet: Automatic dermoscopic skin lesion segmentation". In: *Computers in Biology and Medicine* 120.March (2020), p. 103738. ISSN: 18790534. DOI: 10.1016/j.compbiomed.2020.103738. URL: https://doi.org/10.1016/j.compbiomed.2020.103738.

[He+20] Xi He, Eric H.Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung. "Temporal dynamics in viral shedding and transmissibility of COVID-19". In: *Nature Medicine* 26.5 (2020), pp. 672–675. ISSN: 1546170X. DOI: 10.1038/s41591-020-0869-5.

[Hei+13] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. "Context-dependent sound event detection". In: *Eurasip Journal on Audio, Speech, and Music Processing* 2013.1 (2013), pp. 1–13. ISSN: 16874714. DOI: 10.1186/1687-4722-2013-1.

[Hen+19] Sean M. Hendryx, Andrew B. Leach, Paul D. Hein, and Clayton T. Morrison. "Meta-Learning Initializations for Image Segmentation". In: *arXiv* (Dec. 2019). URL: http://arxiv.org/abs/1912.06290.

[HK11] Albert Hein and Thomas Kirste. "Generic Performance Metrics for Continuous Activity Recognition". In: *KI 2011: Advances in Artificial Intelligence*. 2011. DOI: 10.1007/978-3-642-24455-1{_}13.

[HLR00] Günter Haring, Christoph Lindemann, and Martin Reiser, eds. *Performance Evaluation: Origins and Directions*. Vol. 1769. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 17–31. ISBN: 978-3-540-67193-0. DOI: 10.1007/3-540-46506-5. URL: http://link.springer.com/10.1007/3-540-46506-5.

[Hoe+22] Katharina V. Hoebel, Christopher Bridge, Sara Ahmed, Oluwatosin Akintola, Caroline Chung, Raymond Huang, Jason Johnson, Albert Kim, K. Ina Ly, Ken Chang, Jay Patel, Marco Pinho, Tracy T. Batchelor, Bruce Rosen, Elizabeth Gerstner, and Jayashree Kalpathy-Cramer. "Is this good enough? On expert perception of brain tumor segmentation quality". In: *Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment*. Ed. by Claudia R. Mello-Thoms and Sian Taylor-Phillips. SPIE, Apr. 2022, p. 29. ISBN: 9781510649453. DOI: 10.1117/12.2611810. URL: https://doi.org/10.1117/12.2611810%20https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12035/2611810/Is-this-good-enough-On-expert-perception-of-brain-tumor/10.1117/12.2611810.full.

[Hos+22] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. "Meta-Learning in Neural Networks: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.9 (2022), pp. 5149–5169. ISSN: 19393539. DOI: 10.1109/TPAMI.2021.3079209.

[Hos+23] Md Shafayet Hossain, Sakib Mahmud, Amith Khandakar, Nasser Al-Emadi, Farhana Ahmed Chowdhury, Zaid Bin Mahbub, Mamun Bin Ibne Reaz, and Muhammad E.H. Chowdhury. "MultiResUNet3+: A Full-Scale Connected Multi-Residual UNet Model to Denoise Electrooculogram and Electromyogram Artifacts from Corrupted Electroencephalogram Signals". In: *Bioengineering* 10.5 (2023). ISSN: 23065354. DOI: 10.3390/bioengineering10050579.

[Hou+21] Essam H. Houssein, Marwa M. Emam, Abdelmgeid A. Ali, and Ponnuthurai Nagaratnam Suganthan. "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review". In: *Expert Systems with Applications* 167.October 2020 (2021), p. 114161. ISSN: 09574174. DOI: 10.1016/j.eswa.2020.114161. URL: https://doi.org/10.1016/j.eswa.2020.114161.

[HRP21] Mike Huisman, Jan N. van Rijn, and Aske Plaat. "A survey of deep meta-learning". In: *Artificial Intelligence Review* 54.6 (Aug. 2021), pp. 4483–4541. ISSN: 0269-2821. DOI: 10.1007/s10462-021-10004-4. URL: https://doi.org/10.1007/s10462-021-10004-4%20https://link.springer.com/10.1007/s10462-021-10004-4.

[HSZ20] Zawar Hussain, Quan Z. Sheng, and Wei Emma Zhang. "A review and categorization of techniques on device-free human activity recognition". In: *Journal of Network and Computer Applications* 167 (Oct. 2020), p. 102738. ISSN: 10848045. DOI: 10.1016/j.jnca.2020.102738. URL: http://arxiv.org/abs/1906.05074%0Ahttp://dx.doi.org/10.1016/j.jnca.2020.102738%20https://linkinghub.elsevier.com/retrieve/pii/S1084804520302125.

[Hu+17] Yupeng Hu, Cun Ji, Ming Jing, and Et.al. "A Continuous Segmentation Algorithm for Streaming Time Series". In: *EAI CollaborateCom*. 2017, pp. 140–151. ISBN: 978-3-319-59288-6. DOI: 10.1007/978-3-319-59288-6{_}13.

[Hu+19] Yupeng Hu, Pengjie Ren, Wei Luo, Peng Zhan, and Xueqing Li. "Multi-resolution representation with recurrent neural networks application for streaming time series in IoT". In: *Computer Networks* 152 (Apr. 2019), pp. 114–132. ISSN: 13891286. DOI: 10.1016/j.comnet.2019.01.035. URL: https://doi.org/10.1016/j.comnet.2019.01.035%20https://linkinghub.elsevier.com/retrieve/pii/S1389128619300660.

[Hua+22] Ye Huang, Di Kang, Wenjing Jia, Liu Liu, and Xiangjian He. "Channelized Axial Attention – considering Channel Relation within Spatial Attention for Semantic Segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.1 (2022), pp. 1016–1025. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i1.19985. URL: https://www.aaai.org/AAAI22Papers/AAAI-486.HuangY.pdf.

[Hua+23] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. "MISSFormer: An Effective Transformer for 2D Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 42.5 (May 2023), pp. 1484–1494. ISSN: 0278-0062. DOI: 10.1109/TMI.2022.3230943. URL: https://arxiv.org/pdf/2109.07162.pdf%20https://ieeexplore.ieee.org/document/9994763/.

[Hwa+19] Won-Seok Hwang, Jeong-Han Yun, Jonguk Kim, and Hyoung Chun Kim. "Time-Series Aware Precision and Recall for Anomaly Detection". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, Nov. 2019, pp. 2241–2244. DOI: 10.1145/3357384.3358118.

[HXC22] Patrick Huber, Linzi Xing, and Giuseppe Carenini. "Predicting Above-Sentence Discourse Structure Using Distant Supervision from Topic Segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10 (2022), pp. 10794–10802. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i10.21325. URL: https://www.aaai.org/AAAI22Papers/AAAI-10547.HuberP.pdf.

[Ion+19] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. “Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2019-June. IEEE, June 2019, pp. 7834–7843. ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.00803. URL: https://ieeexplore.ieee.org/document/8954309/.

[IQ22] Saeed Iqbal and Adnan N. Qureshi. “A Heteromorphous Deep CNN Framework for Medical Image Segmentation Using Local Binary Pattern”. In: *IEEE Access* 10 (2022), pp. 63466–63480. ISSN: 21693536. DOI: 10.1109/ACCESS.2022.3183331.

[IR20] Nabil Ibtehaz and M. Sohel Rahman. “MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation”. In: *Neural Networks* 121 (2020), pp. 74–87. ISSN: 18792782. DOI: 10.1016/j.neunet.2019.08.025. URL: https://arxiv.org/pdf/1902.04049.pdf.

[Ise+21] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18.2 (2021), pp. 203–211. ISSN: 15487105. DOI: 10.1038/s41592-020-01008-z. URL: http://dx.doi.org/10.1038/s41592-020-01008-z.

[Jav+22] Syed Ashar Javed, Dinkar Juyal, Zahil Shanis, Shreya Chakraborty, Harsha Pokkalla, and Aaditya Prakash. “Rethinking Machine Learning Model Evaluation in Pathology”. In: *arXiv* (Apr. 2022), pp. 1–9. URL: http://arxiv.org/abs/2204.05205.

[Jia+19] Yun Jiang, Ning Tan, Tingting Peng, and Hai Zhang. “Retinal Vessels Segmentation Based on Dilated Multi-Scale Convolutional Neural Network”. In: *IEEE Access* 7 (2019), pp. 76342–76352. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2922365.

[KAE11] Tim Van Kasteren, Hande Alemdar, and Cem Ersoy. “Effective performance metrics for evaluating activity recognition methods”. In: *ARCS*. 2011.

[Kas11] T van Kasteren. “Activity recognition for health monitoring elderly using temporal probabilistic models”. PhD thesis. University of Amsterdam, 2011, p. 158. ISBN: 9781267776891. URL: http://dare.uva.nl/record/374890.

[KC14] Narayanan C. Krishnan and Diane J. Cook. “Activity recognition on streaming sensor data”. In: *Pervasive and Mobile Computing* 10.PART B (Feb. 2014), pp. 138–154. DOI: 10.1016/j.pmcj.2012.07.003.

[Ke+22] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. “Mask Transfiner for High-Quality Instance Segmentation”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 4402–4411. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00437. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Ke_Mask_Transfiner_for_High-Quality_Instance_Segmentation_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879791/.

[Ker+21] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. “Boundary loss for highly unbalanced segmentation”. In: *Medical Image Analysis* 67 (2021), pp. 285–296. ISSN: 13618423. DOI: 10.1016/j.media.2020.101851.

[Kha+22] Tariq M. Khan, Muhammad Arsalan, Antonio Robles-Kelly, and Erik Meijering. “MKIS-Net: A Light-Weight Multi-Kernel Network for Medical Image Segmentation”. In: *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, Nov. 2022, pp. 1–8. ISBN: 978-1-6654-5642-5. DOI: 10.1109/DICTA56598.2022.10034573. URL: http://arxiv.org/abs/2210.08168%20https://ieeexplore.ieee.org/document/10034573/.

[KHS22] Sara Kaviani, Ki Jin Han, and Insoo Sohn. “Adversarial attacks and defenses on AI in medical imaging informatics: A survey”. In: *Expert Systems with Applications* 198.March (2022), p. 116815. ISSN: 09574174. DOI: 10.1016/j.eswa.2022.116815. URL: https://doi.org/10.1016/j.eswa.2022.116815.

[Kim+12] Hak Soo Kim, Samuel B Park, Simon S. Lo, James I. Monroe, and Jason W. Sohn. “Bidirectional local distance measure for comparing segmentations”. In: *Medical Physics* 39.11 (Oct. 2012), pp. 6779–6790. ISSN: 00942405. DOI: 10.1118/1.4754802. URL: http://doi.wiley.com/10.1118/1.4754802.

[Kim+15] Haksoo Kim, James I Monroe, Simon Lo, Min Yao, Paul M Harari, Mitchell Machtay, and Jason W Sohn. “Quantitative evaluation of image segmentation incorporating medical consideration functions”. In: *Medical Physics* 42.6Part1 (May 2015), pp. 3013–3023. ISSN: 00942405. DOI: 10.1118/1.4921067. URL: http://doi.wiley.com/10.1118/1.4921067.

[Kim+22] Beomyoung Kim, Youngjoon Yoo, Chae Eun Rhee, and Junmo Kim. “Beyond Semantic to Instance Segmentation: Weakly-Supervised Instance Segmentation via Semantic Knowledge Transfer and Self-Refinement”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 4268–4277. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00424. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Kim_Beyond_Semantic_to_Instance_Segmentation_Weakly-Supervised_Instance_Segmentation_via_Semantic_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879337/.

[KKS21] Amrita Kaur, Lakhwinder Kaur, and Ashima Singh. “GA-UNet: UNet-based framework for segmentation of 2D and 3D medical images applicable on heterogeneous datasets”. In: *Neural Computing and Applications* 33.21 (Nov. 2021), pp. 14991–15025. ISSN: 0941-0643. DOI: 10.1007/s00521-021-06134-z. URL: https://doi.org/10.1007/s00521-021-06134-z%20https://link.springer.com/10.1007/s00521-021-06134-z.

[Koe+22] Sven Koehler, Lalith Sharan, Julian Kuhm, Arman Ghanaat, Jelizaveta Gordejeva, Nike K. Simon, Niko M. Grell, Florian André, and Sandy Engelhardt. “Comparison of Evaluation Metrics for Landmark Detection in CMR Images”. In: *Informatik aktuell*. Wiesbaden: Springer Vieweg, 2022, pp. 198–203. DOI:

10.1007/978-3-658-36932-3{_}43. URL: https://link.springer.com/10.1007/978-3-658-36932-3_43.

[Kro+21] Florian Kromp, Lukas Fischer, Eva Bozsaky, Inge M. Ambros, Wolfgang Dorr, Klaus Beiske, Peter F. Ambros, Allan Hanbury, and Sabine Taschner-Mandl. “Evaluation of Deep Learning Architectures for Complex Immunofluorescence Nuclear Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 40.7 (2021), pp. 1934–1949. ISSN: 1558254X. DOI: 10.1109/TMI.2021.3069558.

[KSH19] György Kovács, Gheorghe Sebestyen, and Anca Hangan. “Evaluation metrics for anomaly detection algorithms in time-series”. In: *Acta Universitatis Sapientiae, Informatica* 11.2 (2019), pp. 113–130. ISSN: 2066-7760. DOI: 10.2478/ausi-2019-0008.

[Kul12] Brian Kulis. “Metric learning: A survey”. In: *Foundations and Trends in Machine Learning* 5.4 (2012), pp. 287–364. ISSN: 19358237. DOI: 10.1561/2200000019.

[Kum+17] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. “A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology”. In: *IEEE Transactions on Medical Imaging* 36.7 (2017), pp. 1550–1560. ISSN: 1558254X. DOI: 10.1109/TMI.2017.2677499.

[Kum+22] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. “Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda”. In: *Journal of Ambient Intelligence and Humanized Computing* (2022). ISSN: 18685145. DOI: 10.1007/s12652-021-03612-z.

[Kun+22] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Varun Jampani, and Venkatesh Babu Radhakrishnan. “Amplitude Spectrum Transformation for Open Compound Domain Adaptive Semantic Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022), pp. 1220–1227. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i2.20008. URL: https://www.aaai.org/AAAI22Papers/AAAI-8406.KunduJ.pdf.

[LA15] Alexander Lavin and Subutai Ahmad. “Evaluating Real-Time Anomaly Detection Algorithms – The Numenta Anomaly Benchmark”. In: *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, Dec. 2015, pp. 38–44. ISBN: 978-1-5090-0287-0. DOI: 10.1109/ICMLA.2015.141. URL: http://ieeexplore.ieee.org/document/7424283/.

[Lan+15] B. Landman, Igelsias Xu Z., Styner J., Langerak M., and A T. Klein. *Segmentation Outside the Cranial Vault Challenge*. 2015. DOI: 10.7303/SYN3193805. URL: https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn3193805&type=ENTITY.

[Lan+22] Meng Lan, Jing Zhang, Fengxiang He, and Lefei Zhang. “Siamese Network with Interactive Transformer for Video Object Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022), pp. 1228–1236. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i2.20009. URL: https://www.aaai.org/AAAI22Papers/AAAI-702.LanM.pdf.

[Lea+17] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. "Temporal Convolutional Networks for Action Segmentation and Detection". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017, pp. 1003–1012. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.113. URL: https://openaccess.thecvf.com/content_cvpr_2017/papers/Lea_Temporal_Convolutional_Networks_CVPR_2017_paper.pdf%20http://ieeexplore.ieee.org/document/8099596/.

[Lee+22] Ramona Leenings, Nils R. Winter, Udo Dannlowski, and Tim Hahn. "Recommendations for machine learning benchmarks in neuroimaging". In: *NeuroImage* 257.December 2021 (2022). ISSN: 10959572. DOI: 10.1016/j.neuroimage.2022.119298.

[LGL21] Ange Lou, Shuyue Guan, and Murray H. Loew. "DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation". In: *Medical Imaging 2021: Image Processing*. Ed. by Bennett A. Landman and Ivana Išgum. SPIE, Feb. 2021, p. 98. ISBN: 9781510640214. DOI: 10.1117/12.2582338. URL: https://arxiv.org/ftp/arxiv/papers/2006/2006.00414.pdf%20https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11596/2582338/DC-UNet--rethinking-the-U-Net-architecture-with-dual/10.1117/12.2582338.full.

[Li+19a] Jun Huai Li, Ling Tian, Huaijun Wang, Yang An, Kan Wang, and Lei Yu. "Segmentation and Recognition of Basic and Transitional Activities for Continuous Physical Human Activity". In: *IEEE Access* 7 (2019), pp. 42565–42576. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2905575.

[Li+19b] Ruijiang Li, Lei Xing, Sandy Napel, and Daniel L Rubin. *Radiomics and Radiogenomics: Technical Basis and Clinical Applications*. Boca Raton, FL: CRC Press, July 2019. ISBN: 9781351208277. DOI: 10.1201/9781351208277. URL: https://www.taylorfrancis.com/books/9781351208260.

[Li+19c] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. "Selective kernel networks". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (2019), pp. 510–519. ISSN: 10636919. DOI: 10.1109/CVPR.2019.00060. URL: https://sci.bban.top/pdf/10.1109/CVPR.2019.00060.pdf#view=FitH.

[Lic+20] Daniele Liciotti, Michele Bernardini, Luca Romeo, and Emanuele Frontoni. "A sequential deep learning application for recognising human activities in smart homes". In: *Neurocomputing* 396.2019 (2020), pp. 501–513. ISSN: 18728286. DOI: 10.1016/j.neucom.2018.10.104. URL: https://github.com/danielelic/deep-casas/.

[Lin+14] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common objects in context". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS.PART 5 (2014), pp. 740–755. ISSN: 16113349. DOI: 10.1007/978-3-319-10602-1{_}48. URL: https://arxiv.org/pdf/1405.0312.pdf.

[Liu+20] Liangliang Liu, Fang Xiang Wu, Yu Ping Wang, and Jianxin Wang. “Multi-receptive-field CNN for semantic segmentation of medical images”. In: *IEEE Journal of Biomedical and Health Informatics* 24.11 (2020), pp. 3215–3225. ISSN: 21682208. DOI: 10.1109/JBHI.2020.3016306.

[Liu+21] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. “A Review of Deep-Learning-Based Medical Image Segmentation Methods”. In: *Sustainability* 13.3 (Jan. 2021), p. 1224. ISSN: 2071-1050. DOI: 10.3390/su13031224. URL: https://www.mdpi.com/2071-1050/13/3/1224.

[Liu+22] Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. “Single-Domain Generalization in Medical Image Segmentation via Test-Time Adaptation from Shape Dictionary”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022), pp. 1756–1764. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i2.20068. URL: https://www.aaai.org/AAAI22Papers/AAAI-852.LiuQ.pdf.

[Liu+23] Yatong Liu, Yu Zhu, Ying Xin, Yanan Zhang, Dawei Yang, and Tao Xu. “MES-Trans: Multi-scale embedding spatial transformer for medical image segmentation”. In: *Computer Methods and Programs in Biomedicine* 233 (2023), p. 107493. ISSN: 18727565. DOI: 10.1016/j.cmpb.2023.107493. URL: https://doi.org/10.1016/j.cmpb.2023.107493.

[LKC94] T.C. Lee, R.L. Kashyap, and C.N. Chu. “Building Skeleton Models via 3-D Medial Surface Axis Thinning Algorithms”. In: *CVGIP: Graphical Models and Image Processing* 56.6 (Nov. 1994), pp. 462–478. ISSN: 10499652. DOI: 10.1006/cgip.1994.1042. URL: https://www.sci.utah.edu/devbuilds/biomesh3d/FEMesher/references/lee94-3dskeleton.pdf%20https://linkinghub.elsevier.com/retrieve/pii/S104996528471042X.

[LMS14] Miodrag Lovrić, Marina Milanović, and Milan Stamenković. “Algoritmic methods for segmentation of time series: An overview”. In: *Journal of Contemporary Economic and Business Issues* 1.1 (2014), pp. 31–53.

[LN22] Tao Lei and Asoke K Nandi. *Image Segmentation: Principles, Techniques, and Applications*. John Wiley & Sons, 2022, pp. 1–9. ISBN: 9789896540821. URL: http://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203.

[Lou+22] Meng Lou, Jie Meng, Yunliang Qi, Xiaorong Li, and Yide Ma. “MCRNet: Multi-level context refinement network for semantic segmentation in breast ultrasound imaging”. In: *Neurocomputing* 470 (2022), pp. 154–169. ISSN: 18728286. DOI: 10.1016/j.neucom.2021.10.102. URL: https://doi.org/10.1016/j.neucom.2021.10.102.

[LSX21] Peng Lu, Baoye Song, and Lin Xu. “Human face recognition based on convolutional neural network and augmented dataset”. In: *Systems Science and Control Engineering* 9.S2 (2021), pp. 29–37. ISSN: 21642583. DOI: 10.1080/21642583.2020.1836526.

[Lu+19] Yiwei Lu, K. Mahesh Kumar, Seyed Shahabeddin Nabavi, and Yang Wang. “Future Frame Prediction Using Convolutional VRNN for Anomaly Detection”. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, Sept. 2019. ISBN: 978-1-7281-0990-9. DOI: 10.1109/AVSS.2019.8909850. URL: https://ieeexplore.ieee.org/document/8909850/.

[Lun+18] Gabriel Machado Lunardi, Fadi Al Machot, Vladimir A. Shekhovtsov, Vinícius Maran, Guilherme Medeiros Machado, Alencar Machado, Heinrich C. Mayr, and José Palazzo M. de Oliveira. “IoT-based human action prediction and support”. In: *Internet of Things (Netherlands)* 3-4 (2018), pp. 52–68. ISSN: 25426605. DOI: 10.1016/j.iot.2018.09.007. URL: https://doi.org/10.1016/j.iot.2018.09.007.

[Luo+22] Shuai Luo, Yujie Li, Pengxiang Gao, Yichuan Wang, and Seiichi Serikawa. “Meta-seg: A survey of meta-learning for image segmentation”. In: *Pattern Recognition* 126 (2022), p. 108586. ISSN: 00313203. DOI: 10.1016/j.patcog.2022.108586. URL: https://doi.org/10.1016/j.patcog.2022.108586.

[LZW22] Yuang Liu, Wei Zhang, and Jun Wang. “Multi-Knowledge Aggregation and Transfer for Semantic Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022), pp. 1837–1845. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i2.20077. URL: https://www.aaai.org/AAAI22Papers/AAAI-411.LiuY.pdf.

[Ma+21] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. “Loss odyssey in medical image segmentation”. In: *Medical Image Analysis* 71 (2021). ISSN: 13618423. DOI: 10.1016/j.media.2021.102035.

[Mad+09] Messina E. (Editors) Madhavan R. Tunstel E., Robert N. Lass, Evan A. Sultanik, Raj Madhavan, and Edward Tunstel. *Performance Evaluation and Benchmarking of Intelligent Systems*. Boston, MA: Springer US, 2009. ISBN: 978-1-4419-0491-1. DOI: 10.1007/978-1-4419-0492-8. URL: http://link.springer.com/10.1007/978-1-4419-0492-8.

[Mal+22] Priyanka Malhotra, Sheifali Gupta, Deepika Koundal, Atef Zaguia, and Wegayehu Enbeyle. “Deep Neural Networks for Medical Image Segmentation”. In: *Journal of Healthcare Engineering* 2022 (Mar. 2022). Ed. by Chinmay Chakraborty, pp. 1–15. ISSN: 2040-2309. DOI: 10.1155/2022/9580991. URL: https://www.hindawi.com/journals/jhe/2022/9580991/.

[MBL20] Kevin Musgrave, Serge Belongie, and Ser Nam Lim. “A Metric Learning Reality Check”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12370 LNCS (2020), pp. 681–699. ISSN: 16113349. DOI: 10.1007/978-3-030-58595-2{_}41. URL: https://arxiv.org/pdf/2003.08505.pdf.

[McC19] Matthew C. McCallum. “Unsupervised Learning of Deep Features for Music Segmentation”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2019, pp. 346–350. ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8683407. URL: https://ieeexplore.ieee.org/document/8683407/.

[Med+18] Javier Medina-Quero, Shuai Zhang, Chris Nugent, and M. Espinilla. “Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition”. In: *Expert Systems with Applications* 114 (Dec. 2018), pp. 441–453. ISSN: 09574174. DOI: 10.1016/j.eswa.2018.07.068. URL: https://doi.org/10.1016/j.eswa.2018.07.068%20https://linkinghub.elsevier.com/retrieve/pii/S0957417418304937.

[Mei+22] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. “Glass Segmentation using Intensity and Spectral Polarization Cues”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 12612–12621. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.01229. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Mei_Glass_Segmentation_Using_Intensity_and_Spectral_Polarization_Cues_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9880262/.

[Mel+22] Jayesh George Melekoodappattu, Anto Sahaya Dhas, Binil Kumar Kandathil, and K S Adarsh. “Breast cancer detection in mammogram: combining modified CNN and texture feature based approach”. In: *Journal of Ambient Intelligence and Humanized Computing* (Jan. 2022), pp. 1–10. ISSN: 1868-5137. DOI: 10.1007/s12652-022-03713-3. URL: https://link.springer.com/10.1007/s12652-022-03713-3.

[Mes+18] Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D. Plumbley. “Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge”. In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 26.2 (2018), pp. 379–393. ISSN: 23299290. DOI: 10.1109/TASLP.2017.2778423.

[MH19] Mahmut Kaya and Hasan Sakir Bilge. “Deep Metric Learning : A Survey”. In: *Symmetry* 11.9 (2019), p. 1066.

[MHV16] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. “Metrics for polyphonic sound event detection”. In: *Applied Sciences (Switzerland)* 6.6 (2016). ISSN: 20763417. DOI: 10.3390/app6060162.

[Min+06] David Minnen, Tracy L Westeyn, Thad Starner, Jamie A. Ward, and Paul Lukowicz. “Performance Metrics and Evaluation Issues for Continuous Activity Recognition”. In: *Performance metrics for intelligent systems* (2006), pp. 141–148.

[Min+20] L. Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. “Sensor-based and vision-based human activity recognition: A comprehensive survey”. In: *Pattern Recognition* 108 (2020). ISSN: 00313203. DOI: 10.1016/j.patcog.2020.107561.

[Mis+20] Tejaswini Mishra, Meng Wang, Ahmed A. Metwally, Gireesh K. Bogu, Andrew W. Brooks, Amir Bahmani, Arash Alavi, Alessandra Celli, Emily Higgs, Orit Dagan-Rosenfeld, Bethany Fay, Susan Kirkpatrick, Ryan Kellogg, Michelle Gibson, Tao Wang, Erika M. Hunting, Petra Mamic, Ariel B. Ganz, Benjamin Rolnik, Xiao Li, and Michael P. Snyder. "Pre-symptomatic detection of COVID-19 from smartwatch data". In: *Nature Biomedical Engineering* 4.12 (2020), pp. 1208–1220. ISSN: 2157846X. DOI: 10.1038/s41551-020-00640-6.

[MM16] Deepak Mishra and Jiri Matas. *Computer Vision – ECCV 2016 Workshops*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 777–823. ISBN: 978-3-319-48880-6. DOI: 10.1007/978-3-319-48881-3. URL: https://openaccess.thecvf.com/content_ICCVW_2019/papers/VOT/Kristan_The_Seventh_Visual_Object_Tracking_VOT2019_Challenge_Results_ICCVW_2019_paper.pdf%20http://link.springer.com/10.1007/978-3-319-48881-3.

[MM80] Tom M Mitchell and Tom M Mitchell. "The Need for Biases in Learning Generalizations". In: *Rutgers CS tech report* 1.May (1980). URL: https://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf.

[MNA16] Fausto Milletari, Nassir Navab, and Seyed-ahmad Ahmad Ahmadi. "V-Net: Fully convolutional neural networks for volumetric medical image segmentation". In: *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016* (June 2016), pp. 565–571. DOI: 10.1109/3DV.2016.79. URL: https://arxiv.org/pdf/1606.04797%20http://arxiv.org/abs/1606.04797.

[Mod+22a] Seyed Modaresi, Aomar Osmani, Mohammadreza Razzazi, and Abdelghani Chibani. "Multimodal Evaluation Method for Sound Event Detection". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*. IEEE, 2022.

[Mod+22b] Seyed Modaresi, Aomar Osmani, Mohammadreza Razzazi, and Abdelghani Chibani. "Uniform Evaluation of Properties in Activity Recognition". In: *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, 2022.

[Mod+22c] Seyed M.R. Modaresi, Aomar Osmani, Mohammadreza Razzazi, and Abdelghani Chibani. "Evaluation of Early Diagnosis of COVID-19 Algorithms". In: *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* 2022-Octob (2022), pp. 1277–1282. ISSN: 10823409. DOI: 10.1109/ICTAI56018.2022.00193.

[Mos+22] Zohreh Mostaani, RaviShankar Prasad, Bogdan Vlasenko, and Mathew Magimai-Doss. "Modeling of Pre-Trained Neural Network Embeddings Learned From Raw Waveform for COVID-19 Infection Detection". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2022, pp. 8482–8486. ISBN: 978-1-6654-0540-9. DOI: 10.1109/ICASSP43922.2022.9746271. URL: https://ieeexplore.ieee.org/document/9746271/.

[Mot+21] Leonardo Pereira Motta, Pedro Paulo Ferreira Da Silva, Bruno Max Borguezan, Jorge Luis Mac Hado Do Amaral, Lucimar Gonçalves Milagres, Márcio Neves Bóia, Marcos Rochedo Ferraz, Roberto Mogami, Rodolfo Acatauassú Nunes, and Pedro Lopes De Melo. "An emergency system for monitoring pulse oximetry, peak expiratory flow, and body temperature of patients with COVID-19 at home: Development and preliminary application". In: *PLoS ONE* 16.3 March (2021), pp. 1–19. ISSN: 19326203. DOI: 10.1371/journal.pone.0247635.

[MPF21] Tanvir Mahmud, Bishmoy Paul, and Shaikh Anowarul Fattah. "PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images". In: *Computers in Biology and Medicine* 128.November 2020 (2021), p. 104119. ISSN: 18790534. DOI: 10.1016/j.compbiomed.2020.104119. URL: https://doi.org/10.1016/j.compbiomed.2020.104119.

[MS19] Geethu Mohan and M. Monica Subashini. "Medical Imaging With Intelligent Systems: A Review". In: *Deep Learning and Parallel Computing Environment for Bioengineering Systems* (2019), pp. 53–73. DOI: 10.1016/B978-0-12-816718-2.00011-7. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780128167182000117.

[MSK22] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. "Towards a guideline for evaluation metrics in medical image segmentation". In: *BMC Research Notes* 15.1 (2022), pp. 1–8. ISSN: 17560500. DOI: 10.1186/s13104-022-06096-y. URL: https://doi.org/10.1186/s13104-022-06096-y.

[Mu+22] Tianyu Mu, Hongzhi Wang, Chunnan Wang, Zheng Liang, and Xinyue Shao. "Auto-CASH: A meta-learning embedding approach for autonomous classification algorithm selection". In: *Information Sciences* 591 (2022), pp. 344–364. ISSN: 00200255. DOI: 10.1016/j.ins.2022.01.040. URL: https://doi.org/10.1016/j.ins.2022.01.040.

[Mül+22] Dominik Müller, Dennis Hartmann, Philip Meyer, Florian Auer, Iñaki Soto-Rey, and Frank Kramer. "MISeval: A Metric Library for Medical Image Segmentation Evaluation". In: *Studies in Health Technology and Informatics* 294 (May 2022), pp. 33–37. ISSN: 18798365. DOI: 10.3233/SHTI220391. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI220391.

[Nai+21] Ying-Hwey Nai, Bernice W Teo, Nadya L Tan, Sophie O'Doherty, Mary C Stephenson, Yee Liang Thian, Edmund Chiong, and Anthonin Reilhac. "Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset". In: *Computers in Biology and Medicine* 134 (July 2021), p. 104497. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2021.104497. URL: https://doi.org/10.1016/j.compbiomed.2021.104497%20https://linkinghub.elsevier.com/retrieve/pii/S0010482521002912.

[NGC15] Qin Ni, Ana García Hernando, and Iván de la Cruz. "The Elderly's Independent Living in Smart Homes: A Characterization of Activities and Sensing Infrastructure Survey to Facilitate Services Development". In: *Sensors* 15.5 (May 2015), pp. 11312–11362. DOI: 10.3390/s150511312.

[Nie83] Friedrich Nietzsche. *Thus Spoke Zarathustra*. Ernst Schmeitzner, 1883. URL: https://books.googleusercontent.com/books/content?req=AKW5Qaeg8PuhGoyckFpMl95AIHDvF-eTONfFSZdUsWPg2ZA1d4rHDxPRk16gBbLEQv_SA-TI4S4oMGhll92YpNXU2TRwFSIiZeg1ZHBHVXMmUyaMhFqH1TgoqF2Ycha26Klvws1XCjfk_ORm6B0gR9Nc1A92EjNHJiCOF7HnyXjSFTLlB1S2LAYdt9GDkFapTcYlQio2n.

[Nik+21] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Owen Hughes, Joseph R. Ledsam, and Olaf Ronneberger. "Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study". In: *Journal of Medical Internet Research* 23.7 (July 2021), e26151. ISSN: 1438-8871. DOI: 10.2196/26151. URL: https://www.jmir.org/2021/7/e26151.

[NIS04] NIST. *TRECVID 2004 Evaluation (www-nlpir.nist.gov/projects/tv2004)*. 2004. URL: https://www-nlpir.nist.gov/projects/tv2004/index.html.

[NLL22] Houda Najeh, Christophe Lohr, and Benoit Leduc. "Dynamic Segmentation of Sensor Events for Real-Time Human Activity Recognition in a Smart Home Context". In: *Sensors* 22.14 (July 2022), p. 5458. ISSN: 1424-8220. DOI: 10.3390/s22145458. URL: https://www.mdpi.com/1424-8220/22/14/5458.

[NSH20] Aravind Natarajan, Hao Wei Su, and Conor Heneghan. "Assessment of physiological signs associated with COVID-19 measured using wearable devices". In: *npj Digital Medicine* 3.1 (2020). ISSN: 23986352. DOI: 10.1038/s41746-020-00363-7.

[Oke+14] George Okeyo, Liming Chen, Hui Wang, and Roy Sterritt. "Dynamic sensor data segmentation for real-time knowledge-driven activity recognition". In: *Pervasive and Mobile Computing* 10.PART B (2014), pp. 155–172. ISSN: 15741192. DOI: 10.1016/j.pmcj.2012.11.004. URL: http://dx.doi.org/10.1016/j.pmcj.2012.11.004.

[Okt+18] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention U-Net: Learning Where to Look for the Pancreas". In: *arXiv* (Apr. 2018). URL: http://arxiv.org/abs/1804.03999.

[Oli+23] Bekhzod Olimov, Barathi Subramanian, Rakhmonov Akhrorjon Akhmadjon Ugli, Jea-Soo Kim, and Jeonghong Kim. "Consecutive multiscale feature learning-based image classification model". In: *Scientific Reports* 13.1 (Mar. 2023), p. 3595. ISSN: 2045-2322. DOI: 10.1038/s41598-023-30480-8. URL: https://doi.org/10.1038/s41598-023-30480-8%20https://www.nature.com/articles/s41598-023-30480-8.

[OOB11a] Javier Ortiz Laguna, Angel García Olaya, and Daniel Borrajo. “A Dynamic Sliding Window Approach for Activity Recognition”. In: *Practice Nurse*. Berlin, Heidelberg: Springer, 2011, pp. 219–230. ISBN: 978-3-642-22361-7. DOI: 10.1007/978-3-642-22362-4_19. URL: http://link.springer.com/10.1007/978-3-642-22362-4_19%20https://link.springer.com/10.1007/978-3-642-22362-4_19.

[OOB11b] Javier Ortiz Laguna, Angel García Olaya, and Daniel Borrajo. “A Dynamic Sliding Window Approach for Activity Recognition”. In: *Practice Nurse*. Berlin, Heidelberg: Springer, 2011, pp. 219–230. ISBN: 978-3-642-22361-7. DOI: 10.1007/978-3-642-22362-4{_}19. URL: http://link.springer.com/10.1007/978-3-642-22362-4_19%20https://link.springer.com/10.1007/978-3-642-22362-4_19.

[Osm02] A. Osmani. “Learning patterns in multidimensional space using interval algebra”. In: *Lecture Notes in Computer Science*. Vol. 2443 LNAI. Springer, 2002, pp. 31–40. ISBN: 3540441271. DOI: 10.1007/3-540-46148-5{_}4. URL: http://link.springer.com/10.1007/3-540-46148-5_4.

[Osm03] Aomar Osmani. “STCSP : A Representation Model for Sequential Patterns”. In: *Foundations and Applications of Spatio-Temporal Reasoning (FASTR)* (2003). URL: https://www.aaai.org/Library/Symposia/Spring/2003/ss03-03-010.php.

[Pan+22] Junwen Pan, Qi Bi, Yanzhan Yang, Pengfei Zhu, and Cheng Bian. “Label-Efficient Hybrid-Supervised Learning for Medical Image Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (June 2022), pp. 2026–2034. ISSN: 2374-3468. DOI: 10.1609/aaai.v36i2.20098. URL: https://www.aaai.org/AAAI22Papers/AAAI-11780.PanJ.pdf%20https://ojs.aaai.org/index.php/AAAI/article/view/20098.

[Pan15] Arjun Panesar. *Evaluating machine learning models : a beginner’s guide to key concepts and pitfalls*. O’Reilly Media, 2015. ISBN: 9781491932469.

[Pen+22] Cheng Peng, Andriy Myronenko, Ali Hatamizadeh, Vishwesh Nath, Md Mahfuzur Rahman Siddiquee, Yufan He, Daguang Xu, Rama Chellappa, and Dong Yang. “HyperSegNAS: Bridging One-Shot Neural Architecture Search with 3D Medical Image Segmentation using HyperNet”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 20709–20719. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.02008. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Peng_HyperSegNAS_Bridging_One-Shot_Neural_Architecture_Search_With_3D_Medical_Image_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879863/.

[Per+14] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. “Context Aware Computing for The Internet of Things: A Survey”. In: *IEEE Communications Surveys & Tutorials* 16.1 (2014), pp. 414–454. ISSN: 1553-877X. DOI: 10.1109/SURV.2013.042313.00197. URL: http://ieeexplore.ieee.org/document/6512846/.

[Per+16] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. “A benchmark dataset and evaluation methodology for video object segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 724–732. ISSN: 10636919. DOI: 10.1109/CVPR.2016.85. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Perazzi_A_Benchmark_Dataset_CVPR_2016_paper.pdf.

[Pop+07] Aleksandra Popovic, Matías de la Fuente, Martin Engelhardt, and Klaus Radermacher. “Statistical validation metric for accuracy assessment in medical image segmentation”. In: *International Journal of Computer Assisted Radiology and Surgery* 2.3-4 (2007), pp. 169–181. ISSN: 18616429. DOI: 10.1007/s11548-007-0125-1.

[Qin+22] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. “Activation Modulation and Recalibration Scheme for Weakly Supervised Semantic Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022), pp. 2117–2125. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i2.20108. URL: https://www.aaai.org/AAAI22Papers/AAAI-4538.JieQ.pdf.

[Qiu+18] Qiang Qiu, Xiuyuan Cheng, Robert Calderbank, and Guillermo Sapiro. “DCFNet: Deep Neural Network with Decomposed Convolutional Filters”. In: *35th International Conference on Machine Learning, ICML 2018* 9 (2018), pp. 6687–6696. URL: https://arxiv.org/pdf/1802.04145.pdf.

[Qiu+22] Liangdong Qiu, Chongjie Ye, Pei Chen, Yunbi Liu, Xiaoguang Han, and Shuguang Cui. “DArch: Dental Arch Prior-assisted 3D Tooth Instance Segmentation with Weak Annotations”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 20720–20729. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.02009. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Qiu_DArch_Dental_Arch_Prior-Assisted_3D_Tooth_Instance_Segmentation_With_Weak_CVPR_2022_paper.pdf.

[QJE21] Hangwei Qian, Sinno Jialin Pan, and Et.al. “Latent Independent Excitation for Generalizable Sensor-based Cross-Person Activity Recognition”. In: *AAAI* 35.13 (May 2021), pp. 11921–11929. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17416%20www.aaai.org.

[QPM18] Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. “Sensor-Based Activity Recognition via Learning From Distributions”. In: *AAAI Conference on Artificial Intelligence* (2018), pp. 6262–6269.

[QPM21] Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. “Latent Independent Excitation for Generalizable Sensor-based Cross-Person Activity Recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.13 (May 2021), pp. 11921–11929. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17416.

[Qui+18] Bronagh Quigley, Mark Donnelly, George Moore, and Leo Galway. “A Comparative Analysis of Windowing Approaches in Dense Sensing Environments”. In: *UCAmI 2018*. Basel Switzerland: MDPI, Oct. 2018, p. 1245. DOI: 10.3390/proceedings2191245. URL: https://www.mdpi.com/2504-3900/2/19/1245.

[Rei+21] Annika Reinke, Minu D. Tizabi, Carole H. Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, M. Jorge Cardoso, Veronika Cheplygina, Evangelia Christodoulou, Beth Cimini, Gary S. Collins, Keyvan Farahani, Bram van Ginneken, Ben Glocker, Patrick Godau, Fred Hamprecht, Daniel A. Hashimoto, Doreen Heckmann-Nötzel, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Alexandros Karargyris, Alan Karthikesalingam, Bernhard Kainz, Emre Kavur, Hannes Kenngott, Jens Kleesiek, Thijs Kooi, Michal Kozubek, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, David Moher, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, M. Alican Noyan, Jens Petersen, Gorkem Polat, Nasir Rajpoot, Mauricio Reyes, Nicola Rieke, Michael Riegler, Hassan Rivaz, Julio Saez-Rodriguez, Clarisa Sanchez Gutierrez, Julien Schroeter, Anindo Saha, Shravya Shetty, Maarten van Smeden, Bram Stieltjes, Ronald M. Summers, Abdel A. Taha, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, Manuel Wiesenfarth, Ziv R. Yaniv, Annette Kopp-Schneider, Paul Jäger, and Lena Maier-Hein. “Common Limitations of Image Processing Metrics: A Picture Story”. In: *Medical Imaging with Deep Learning* (2021). URL: http://arxiv.org/abs/2104.05642.

[Rei+22] Annika Reinke, Lena Maier-Hein, Evangelia Christodoulou, Ben Glocker, Patrick Scholz, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael Alexander Riegler, Manuel Wiesenfarth, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Ali Emre Kavur, Tim Rädsch, Minu D. Tizabi, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, M. Jorge Cardoso, Veronika Cheplygina, Beth A Cimini, Gary S. Collins, Keyvan Farahani, Bram van Ginneken, Fred A Hamprecht, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles Kahn, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne Martel, Peter Mattson, Erik Meijering, Bjoern Menze, David Moher, Karel G.M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Carole H. Sudre, Ronald M. Summers, Abdel A. Taha, Sotirios A. Tsaftaris, Ben Van Calster, Gael Varoquaux, and Paul F Jaeger. “Metrics Reloaded - A new recommendation framework for biomedical image analysis validation”. In: *Medical Imaging with Deep Learning* (2022). URL: http://www.mauricioreyes.me/Publications/ReinkeMIDL2022.pdf.

[Ren+22] Yuan Ren, Long Yu, Shengwei Tian, Junlong Cheng, Zhiqi Guo, and Yanhan Zhang. “Serial attention network for skin lesion segmentation”. In: *Journal of*

Ambient Intelligence and Humanized Computing 13.2 (2022), pp. 799–810. ISSN: 18685145. DOI: 10.1007/s12652-021-02933-3. URL: https://doi.org/10.1007/s12652-021-02933-3.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *IEEE Access*. Vol. 9. IEEE Computer Society, May 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4{_}28. URL: http://link.springer.com/10.1007/978-3-319-24574-4_28%20http://arxiv.org/abs/1505.04597.

[RK13] Robert J Ross and John Kelleher. "Accuracy and timeliness in ML based activity recognition". In: *Proceedings of the 13th AAAI Conference on Plan, Activity, and Intent Recognition*. Vol. WS-13-13. AAAIWS'13-13. AAAI Press, 2013, pp. 39–46. ISBN: 9781577356240. DOI: 10.5555/2908241.2908247.

[RNN99] Stephen V. Rice, George Nagy, and Thomas A. Nartker. *Optical Character Recognition*. Boston, MA: Springer US, 1999, pp. 507–509. ISBN: 978-1-4613-7281-3. DOI: 10.1007/978-1-4615-5021-1. URL: http://link.springer.com/10.1007/978-1-4615-5021-1.

[Ros+14] André Luis Debiaso Rossi, André Carlos Ponce de Leon Ferreira de Carvalho, Carlos Soares, and Bruno Feres de Souza. "MetaStream: A meta-learning based method for periodic algorithm selection in time-changing data". In: *Neurocomputing* 127 (2014), pp. 52–64. ISSN: 09252312. DOI: 10.1016/j.neucom.2013.05.048. URL: http://dx.doi.org/10.1016/j.neucom.2013.05.048.

[Ros+21] André Luis Debiaso Rossi, Carlos Soares, Bruno Feres de Souza, and André Carlos Ponce de Leon Ferreira de Carvalho. "Micro-MetaStream: Algorithm selection for time-changing data". In: *Information Sciences* 565 (2021), pp. 262–277. ISSN: 00200255. DOI: 10.1016/j.ins.2021.02.075.

[Rot+15] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. "DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (2015). Ed. by Nassir Navab, Joachim Hornegger, William M Wells, and Alejandro Frangi, pp. 556–564. DOI: 10.1007/978-3-319-24553-9{_}68. URL: http://link.springer.com/10.1007/978-3-319-24553-9_68.

[Rot+16] Holger R Roth, Amal Farag, Evrim B Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers. *NIH Pancreas-CT Dataset*. 2016. DOI: 10.7937/K9/TCIA.2016.TNB1KQBU. URL: https://wiki.cancerimagingarchive.net/x/eIlXAQ.

[Roy+23] Santanu Roy, Devikalyan Das, Shyam Lal, and Jyoti Kini. "Novel edge detection method for nuclei segmentation of liver cancer histopathology images". In: *Journal of Ambient Intelligence and Humanized Computing* 14.1 (2023), pp. 479–496. ISSN: 18685145. DOI: 10.1007/s12652-021-03308-4.

[Rue+14] Sylvia Rueda, Sana Fathima, Caroline L. Knight, Mohammad Yaqub, Aris T. Papageorghiou, Bahbibi Rahmatullah, Alessandro Foi, Matteo Maggioni, Antonietta Pepe, Jussi Tohka, Richard V. Stebbing, John E. McManigle, Anca Ciurte, Xavier Bresson, Meritxell Bach Cuadra, Changming Sun, Gennady V. Ponomarev, Mikhail S. Gelfand, Marat D. Kazanov, Ching Wei Wang, Hsiang Chou Chen, Chun Wei Peng, Chu Mei Hung, and J. Alison Noble. "Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: A grand challenge". In: *IEEE Transactions on Medical Imaging* 33.4 (2014), pp. 797–813. ISSN: 02780062. DOI: 10.1109/TMI.2013.2276943.

[SA22] Laxman Singh and Altaf Alam. "An efficient hybrid methodology for an early detection of breast cancer in digital mammograms". In: *Journal of Ambient Intelligence and Humanized Computing* 13.5 (May 2022). ISSN: 1868-5137. DOI: 10.1007/s12652-022-03895-w. URL: https://doi.org/10.1007/s12652-022-03895-w%20https://link.springer.com/10.1007/s12652-022-03895-w.

[SB18] Hela Sfar and Amel Bouzeghoub. "Dynamic Streaming Sensor Data Segmentation for Smart Environment Applications". In: *Neural Information Processing*. Vol. 3316. Springer International Publishing, 2018, pp. 67–77. ISBN: 978-3-540-23931-4. DOI: 10.1007/978-3-030-04224-0{_}7. URL: http://www.springerlink.com/content/r9m8upr0gc4tttqp%20http://link.springer.com/10.1007/978-3-030-04224-0_7.

[Sch+19] Bas Schipaanboord, Djamal Boukerroui, Devis Peressutti, Johan Van Soest, Tim Lustberg, Andre Dekker, Wouter van Elmpt, and Mark J. Gooding. "An Evaluation of Atlas Selection Methods for Atlas-Based Automatic Segmentation in Radiotherapy Treatment Planning". In: *IEEE Transactions on Medical Imaging* 38.11 (2019), pp. 2654–2664. ISSN: 1558254X. DOI: 10.1109/TMI.2019.2907072.

[Sch+20] Oliver Schoppe, Chenchen Pan, Javier Coronel, Hongcheng Mai, Zhouyi Rong, Mihail Ivilinov Todorov, Annemarie Müskes, Fernando Navarro, Hongwei Li, Ali Ertürk, and Bjoern H Menze. "Deep learning-enabled multi-organ segmentation in whole-body mouse scans". In: *Nature Communications* 11.1 (2020), pp. 1–14. ISSN: 20411723. DOI: 10.1038/s41467-020-19449-7.

[Sch07] Sylviane R. Schwer. "Temporal Reasoning without Transitive Tables". In: *arXiv* (June 2007), pp. 1–15. URL: http://arxiv.org/abs/0706.1290.

[Sch87] Jürgen Schmidhuber. "Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook". PhD thesis. Technische Universität München, 1987.

[SEL21] Patrick Schäfer, Arik Ermshaus, and Ulf Leser. "ClaSP - Time Series Segmentation". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. New York, NY, USA: ACM, Oct. 2021, pp. 1578–1587. ISBN: 9781450384469. DOI: 10.1145/3459637.3482240. URL: https://dl.acm.org/doi/abs/10.1145/3459637.3482240%20https:

//dl.acm.org/doi/10.1145/3459637.3482240%20https://sites.google.com/view/ts-clasp.

[Ser+07] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. "Robust Object Recognition with Cortex-Like Mechanisms". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.3 (Mar. 2007), pp. 411–426. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.56. URL: http://platform.almanhal.com/CrossRef/Preview/?ID=2-118837%20http://ieeexplore.ieee.org/document/4069258/.

[Ser+20] Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon. "Sound event detection in synthetic domestic environments". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2020-May (May 2020), pp. 86–90. ISSN: 15206149. DOI: 10.1109/ICASSP40776.2020.9054478. URL: https://ieeexplore.ieee.org/document/9054478/.

[SG16] Tom Sercu and Vaibhava Goel. "Dense Prediction on Sequences with Time-Dilated Convolutions for Speech Recognition". In: *arXiv* (Nov. 2016). URL: http://arxiv.org/abs/1611.09288.

[Sha+21] Allison Shapiro, Nicole Marinsek, Ieuan Clay, Benjamin Bradshaw, Ernesto Ramirez, Jae Min, Andrew Trister, Yuedong Wang, Tim Althoff, and Luca Foschini. "Characterizing COVID-19 and Influenza Illnesses in the Real World via Person-Generated Health Data". In: *Patterns* 2.1 (2021), p. 100188. ISSN: 26663899. DOI: 10.1016/j.patter.2020.100188.

[Shr+20] Nabin K. Shrestha, Francisco Marco Canosa, Amy S. Nowacki, Gary W. Procop, Sherilynn Vogel, Thomas G. Fraser, Serpil C. Erzurum, Paul Terpeluk, and Steven M. Gordon. "Distribution of transmission potential during nonsevere COVID-19 illness". In: *Clinical Infectious Diseases* 71.11 (2020), pp. 2927–2932. ISSN: 15376591. DOI: 10.1093/cid/ciaa886.

[SHV20] Zhao Shuyang, Toni Heittola, and Tuomas Virtanen. "Active Learning for Sound Event Detection". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 28 (2020), pp. 2895–2905. ISSN: 23299304. DOI: 10.1109/TASLP.2020.3029652.

[Sim+19] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. "A large annotated medical image dataset for the development and evaluation of segmentation algorithms". In: *arXiv* (Feb. 2019). URL: http://arxiv.org/abs/1902.09063.

[SJ19] Sangeetha Saman and Swathi Jamjala Narayanan. “Survey on brain tumor segmentation and feature extraction of MR images”. In: *International Journal of Multimedia Information Retrieval* 8.2 (2019), pp. 79–99. ISSN: 2192662X. DOI: 10.1007/s13735-018-0162-2. URL: https://doi.org/10.1007/s13735-018-0162-2.

[Ski+21] Justyna Skibinska, Radim Burget, Asma Channa, Nirvana Popescu, and Yevgeni Koucheryavy. “COVID-19 Diagnosis at Early Stage Based on Smartwatches and Machine Learning Techniques”. In: *IEEE Access* 9 (2021), pp. 119476–119491. ISSN: 21693536. DOI: 10.1109/ACCESS.2021.3106255.

[SL21] Jiaxing Sun and Yujie Li. “MetaSeg: A survey of meta-learning for image segmentation”. In: *Cognitive Robotics* 1.May (2021), pp. 83–91. ISSN: 26672413. DOI: 10.1016/j.cogr.2021.06.003.

[Sma+20] Benjamin L. Smarr, Kirstin Aschbacher, Sarah M. Fisher, Anoushka Chowdhary, Stephan Dilchert, Karena Puldon, Adam Rao, Frederick M. Hecht, and Ashley E. Mason. “Feasibility of continuous fever monitoring using wearable devices”. In: *Scientific Reports* 10.1 (2020), pp. 1–11. ISSN: 20452322. DOI: 10.1038/s41598-020-78355-6.

[SNL13] Ran Shi, King Ngi Ngan, and Songnan Li. “The objective evaluation of image object segmentation quality”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8192 LNCS (2013), pp. 470–479. ISSN: 16113349. DOI: 10.1007/978-3-319-02895-8{_}42.

[Son+22] Jiahuan Song, Xinjian Chen, Qianlong Zhu, Fei Shi, Dehui Xiang, Zhongyue Chen, Ying Fan, Lingjiao Pan, and Weifang Zhu. “Global and Local Feature Reconstruction for Medical Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 41.9 (2022), pp. 2273–2284. ISSN: 1558254X. DOI: 10.1109/TMI.2022.3162111.

[SR22] Nripendra Kumar Singh and Khalid Raza. “Progress in deep learning-based dental and maxillofacial image analysis: A systematic review”. In: *Expert Systems with Applications* 199.March (2022), p. 116968. ISSN: 09574174. DOI: 10.1016/j.eswa.2022.116968. URL: https://doi.org/10.1016/j.eswa.2022.116968.

[SRY22] Dipika Singhania, Rahul Rahaman, and Angela Yao. “Iterative Contrast-Classify for Semi-supervised Temporal Action Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022), pp. 2262–2270. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i2.20124. URL: https://www.aaai.org/AAAI22Papers/AAAI-7097.SinghaniaD.pdf.

[SS10] Tom Schaul and Juergen Schmidhuber. “Metalearning”. In: *Scholarpedia* 5.6 (2010), p. 4650. DOI: 10.4249/scholarpedia.4650.

[Sti+07] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. “The CLEAR 2006 evaluation”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4122 LNCS.December 2005 (2007), pp. 1–44. ISSN: 16113349. DOI: 10.1007/978-3-540-69568-4{_}1.

[Sti+08] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, R. Travis Rose, Martial Michel, and John Garofolo. “The CLEAR 2007 evaluation”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4625 LNCS (2008), pp. 3–34. ISSN: 03029743. DOI: 10.1007/978-3-540-68585-2{_}1.

[Sto+15] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. “Detection and Classification of Acoustic Scenes and Events”. In: *IEEE Transactions on Multimedia* 17.10 (2015), pp. 1733–1746. ISSN: 15209210. DOI: 10.1109/TMM.2015.2428998.

[SWL17] Ahmad Shahi, Brendon J Woodford, and Hanhe Lin. “Dynamic Real-Time Segmentation and Recognition of Activities Using a Multi-feature Windowing Approach”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Vol. 10526. 2017, pp. 26–38. ISBN: 978-3-319-67273-1. DOI: 10.1007/978-3-319-67274-8{_}3. URL: http://link.springer.com/10.1007/978-3-319-67274-8_3.

[SZ15] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015), pp. 1–14. arXiv: 1409.1556. URL: https://arxiv.org/pdf/1409.1556.pdf.

[Sze+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015, pp. 1–9. ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298594. URL: https://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf%20https://onlinelibrary.wiley.com/doi/10.1002/jctb.4820%20http://ieeexplore.ieee.org/document/7298594/.

[Sze+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. “Rethinking the Inception Architecture for Computer Vision”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 2818–2826. ISSN: 10636919. DOI: 10.1109/CVPR.2016.308. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf.

[Sze+17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (Feb. 2017), pp. 11–24. ISSN: 2374-3468. DOI: 10.1609/aaai.v31i1.

11231. URL: http://arxiv.org/abs/1512.00567%20https://ojs.aaai.org/index.php/AAAI/article/view/11231.

[Tag+19] Saeid Asgari Taghanaki, Yefeng Zheng, S. Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. "Combo loss: Handling input and output imbalance in multi-organ segmentation". In: *Computerized Medical Imaging and Graphics* 75 (2019), pp. 24–33. ISSN: 18790771. DOI: 10.1016/j.compmedimag.2019.04.005. URL: https://arxiv.org/pdf/1805.02798.pdf.

[Tak+01] Tak-Chung Fu, Fu-lai Chung, V Ng, and R Luk. "Evolutionary segmentation of financial time series into subsequences". In: *Proceedings of the 2001 Congress on Evolutionary Computation*. Vol. 1. IEEE, May 2001, pp. 426–430. ISBN: 0-7803-6657-3. DOI: 10.1109/CEC.2001.934422. URL: http://ieeexplore.ieee.org/document/934422/.

[Tan+22] Liyao Tang, Yibing Zhan, Zhe Chen, Baosheng Yu, and Dacheng Tao. "Contrastive Boundary Learning for Point Cloud Segmentation". In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 8479–8489. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00830. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Tang_Contrastive_Boundary_Learning_for_Point_Cloud_Segmentation_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9878558/.

[Tan11] O Tange. "GNU Parallel - The Command-Line Power Tool". In: *;login: The USENIX Magazine* 36.1 (Feb. 2011), pp. 42–47. DOI: 10.5281/zenodo.16303. URL: http://www.gnu.org/s/parallel.

[Tar+21] Mehreen Tariq, Sajid Iqbal, Hareem Ayesha, Ishaq Abbas, Khawaja Tehseen Ahmad, and Muhammad Farooq Khan Niazi. "Medical image based breast cancer diagnosis: State of the art and future directions". In: *Expert Systems with Applications* 167.October 2020 (2021), p. 114095. ISSN: 09574174. DOI: 10.1016/j.eswa.2020.114095. URL: https://doi.org/10.1016/j.eswa.2020.114095.

[Tat+18] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. "Precision and recall for time series". In: *Neural Information Processing Systems (NIPS)*. 2018. URL: https://papers.nips.cc/paper/7462-precision-and-recall-for-time-series.

[Tay+18] Ben Taylor, Vicent Sanz Marco, Willy Wolff, Yehia Elkhatib, and Zheng Wang. "Adaptive deep learning model selection on embedded systems". In: *Proceedings of the 19th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems*. New York, NY, USA: ACM, June 2018, pp. 31–43. ISBN: 9781450358033. DOI: 10.1145/3211332.3211336. URL: https://dl.acm.org/doi/10.1145/3211332.3211336.

[TC22] Maryam Taghizadeh and Abdolah Chalechale. “A comprehensive and systematic review on classical and deep learning based region proposal algorithms”. In: *Expert Systems with Applications* 189.October 2021 (2022), p. 116105. ISSN: 09574174. DOI: 10.1016/j.eswa.2021.116105. URL: https://doi.org/10.1016/j.eswa.2021.116105.

[Tem+09] Andrey Temko, Climent Nadeu, Dušan Macho, Robert Malkin, Christian Zieger, and Maurizio Omologo. “Acoustic Event Detection and Classification”. In: *Computers in the Human Interaction Loop* (2009), pp. 61–73. DOI: 10.1007/978-1-84882-054-8{_}7.

[TH15] Abdel Aziz Taha and Allan Hanbury. “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool”. In: *BMC Medical Imaging* 15.1 (2015). ISSN: 14712342. DOI: 10.1186/s12880-015-0068-x. URL: http://dx.doi.org/10.1186/s12880-015-0068-x.

[Tha+18] Rajat Thawani, Michael McLane, Niha Beig, Soumya Ghose, Prateek Prasanna, Vamsidhar Velcheti, and Anant Madabhushi. “Radiomics and radiogenomics in lung cancer: A review for the clinician”. In: *Lung Cancer* 115.June 2017 (2018), pp. 34–41. ISSN: 18728332. DOI: 10.1016/j.lungcan.2017.10.015. URL: https://doi.org/10.1016/j.lungcan.2017.10.015.

[THT14] Abdel Aziz Taha, Allan Hanbury, and Oscar A. Jimenez del Toro. “A formal method for selecting evaluation metrics for image segmentation”. In: *IEEE International Conference on Image Processing (ICIP)*. IEEE, Oct. 2014, pp. 932–936. ISBN: 978-1-4799-5751-4. DOI: 10.1109/ICIP.2014.7025187. URL: http://ieeexplore.ieee.org/document/7025187/.

[Tia+21] Jie Tian, Di Dong, Zhenyu Liu, and Jingwei Wei. *Radiomics and Its Clinical Application: Artificial Intelligence and Medical Big Data*. Academic Press, an imprint of Elsevier, 2021. ISBN: 9780128181010. DOI: 10.1016/C2018-0-02044-7. URL: https://linkinghub.elsevier.com/retrieve/pii/C20180020447.

[Ton+20] Noriyuki Tonami, Keisuke Imoto, Takahiro Fukumori, and Yoichi Yamashita. “Evaluation Metric of Sound Event Detection Considering Severe Misdetections By Scenes”. In: *Detection and Classification of Acoustic Scenes and Events*. 2020, pp. 1–5.

[Tri+19] Darpan Triboan, Liming Chen, Feng Chen, and Zumin Wang. “A semantics-based approach to sensor data segmentation in real-time Activity Recognition”. In: *Future Generation Computer Systems* 93 (2019), pp. 224–236. ISSN: 0167739X. DOI: 10.1016/j.future.2018.09.055. URL: https://doi.org/10.1016/j.future.2018.09.055.

[Tur+19] Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah. “Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis”. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, 2019, pp. 253–257. ISBN: 978-0-578-59596-2. DOI: 10.33682/006b-jx26. URL: http://hdl.handle.net/2451/60771.

[Van19] Joaquin Vanschoren. “Meta-Learning”. In: *Automated Machine Learning: Methods, Systems, Challenges*. Ed. by Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Cham: Springer International Publishing, 2019, pp. 35–61. ISBN: 978-3-030-05318-5. DOI: 10.1007/978-3-030-05318-5{_}2. URL: https://doi.org/10.1007/978-3-030-05318-5_2%20http://link.springer.com/10.1007/978-3-030-05318-5_2.

[VCV22] Gaël Gael Gaël Varoquaux, Olivier Colliot, and Gaël Gael Gaël Varoquaux. “Evaluating machine learning models and their diagnostic value”. In: *Machine Learning for Brain Disorders*. 2022. URL: https://hal.archives-ouvertes.fr/hal-03682454v3%20https://hal.archives-ouvertes.fr/hal-03682454.

[Via18] Kévin Viard. “Modelling and Recognition of Human Activities of Daily Living in a Smart Home”. PhD thesis. Université Paris-Saclay ; Politecnico di Bari. Dipartimento di Ingegneria Elettrica e dell’Informazione (Italia), July 2018. URL: https://tel.archives-ouvertes.fr/tel-01867623.

[Vic+19] Tomas Vicar, Jan Balvan, Josef Jaros, Florian Jug, Radim Kolar, Michal Masarik, and Jaromir Gumulec. “Cell segmentation methods for label-free contrast microscopy: Review and comprehensive comparison”. In: *BMC Bioinformatics* 20.1 (2019), pp. 1–25. ISSN: 14712105. DOI: 10.1186/s12859-019-2880-8.

[VMM22] Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. “You Only Hear Once: A YOLO-like Algorithm for Audio Segmentation and Sound Event Detection”. In: *Applied Sciences (Switzerland)* 12.7 (2022). ISSN: 20763417. DOI: 10.3390/app12073293.

[Wal+14] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. “scikit-image: image processing in Python”. In: *PeerJ* 2 (June 2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: https://peerj.com/articles/453.

[Wan+19] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. “Deep learning for sensor-based activity recognition: A survey”. In: *Pattern Recognition Letters* 119 (Mar. 2019), pp. 3–11. ISSN: 01678655. DOI: 10.1016/j.patrec.2018.02.010. URL: https://doi.org/10.1016/j.patrec.2018.02.010%20https://jd92.wang/assets/files/a10_prl18.pdf%20https://github.com/jindongwang/Deep-learning-activity-recognition.

[Wan+21a] Aiguo Wang, Shenghui Zhao, Chundi Zheng, Jing Yang, Guilin Chen, and Chih Yung Chang. “Activities of Daily Living Recognition with Binary Environment Sensors Using Deep Learning: A Comparative Study”. In: *IEEE Sensors Journal* 21.4 (2021), pp. 5423–5433. ISSN: 15581748. DOI: 10.1109/JSEN.2020.3035062.

[Wan+21b] Guangjie Wang, Hui Liu, Xianpeng Yi, Jinjun Zhou, and Lin Zhang. “ARMS Net: Overlapping chromosome segmentation based on Adaptive Receptive field Multi-Scale network”. In: *Biomedical Signal Processing and Control* 68.January (2021), p. 102811. ISSN: 17468108. DOI: 10.1016/j.bspc.2021.102811. URL: https://doi.org/10.1016/j.bspc.2021.102811.

[Wan+21c] Ze Wang, Zichen Miao, Jun Hu, and Qiang Qiu. "Adaptive Convolutions with Per-pixel Dynamic Filter Atom". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 12282–12291. ISBN: 978-1-6654-2812-5. DOI: 10.1109/ICCV48922.2021.01208. URL: https://openaccess.thecvf.com/content/ICCV2021/papers/Wang_Adaptive_Convolutions_With_Per-Pixel_Dynamic_Filter_Atom_ICCV_2021_paper.pdf%20https://openaccess.thecvf.com/content/ICCV2021/supplemental/Wang_Adaptive_Convolutions_With_ICCV_2021_supplemental.pdf.

[Wan+22a] Chi Wang, Yunke Zhang, Miaomiao Cui, Peiran Ren, Yin Yang, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, and Weiwei Xu. "Active Boundary Loss for Semantic Segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022), pp. 2397–2405. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i2.20139. URL: https://www.aaai.org/AAAI22Papers/AAAI-2277.WangC.pdf.

[Wan+22b] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R. Zaiane. "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer". In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022* 36 (2022), pp. 2441–2449. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i3.20144.

[Wan+22c] Jiacheng Wang, Xiaomeng Li, Yiming Han, Jing Qin, Liansheng Wang, and Zhou Qichao. "Separated Contrastive Learning for Organ-at-Risk and Gross-Tumor-Volume Segmentation with Limited Annotation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.3 (2022), pp. 2459–2467. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i3.20146. URL: https://www.aaai.org/AAAI22Papers/AAAI-8099.WangJ.pdf.

[Wan+22d] Xue Wang, Zhanshan Li, Yongping Huang, and Yingying Jiao. "Multimodal medical image segmentation using multi-scale context-aware network". In: *Neurocomputing* 486 (2022), pp. 135–146. ISSN: 18728286. DOI: 10.1016/j.neucom.2021.11.017. URL: https://doi.org/10.1016/j.neucom.2021.11.017.

[War+06] Jamie A. Ward, Paul Lukowicz, Gerhard Tröster, and Thad E. Starner. "Activity recognition of assembly tasks using body-worn microphones and accelerometers". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.10 (2006), pp. 1553–1566. ISSN: 01628828. DOI: 10.1109/TPAMI.2006.197.

[War+11] Jamie A. Ward, Paul Lukowicz, Hans W. Gellersen, and Ward. "Performance metrics for activity recognition". In: *ACM Transactions on Intelligent Systems and Technology* (2011). DOI: 10.1145/1889681.1889687.

[WGS16] Yuqing Wan, Xueyuan Gong, and Yain Whar Si. "Effect of segmentation on financial time series pattern matching". In: *Applied Soft Computing Journal* 38 (2016), pp. 346–359. ISSN: 15684946. DOI: 10.1016/j.asoc.2015.10.012.

[WM18] Wei Wang and Chunyan Miao. “Activity Recognition in New Smart Home Environments”. In: *Proceedings of the 3rd International Workshop on Multimedia for Personal Health and Health Care - HealthMedia’18*. New York, New York, USA: ACM Press, 2018, pp. 29–37. ISBN: 9781450359825. DOI: 10.1145/3264996.3265001. URL: http://dl.acm.org/citation.cfm?doid=3264996.3265001.

[WOO15] Jie Wan, Michael J. O’Grady, and Gregory M.P. O’Hare. “Dynamic sensor event segmentation for real-time activity recognition in a smart home context”. In: *Personal and Ubiquitous Computing* 19.2 (2015), pp. 287–301. ISSN: 16174909. DOI: 10.1007/s00779-014-0824-x.

[Wu+21] Jiawei Wu, Shengqiang Zhou, Songlin Zuo, Yiyin Chen, Weiqin Sun, Jiang Luo, Jiantuan Duan, Hui Wang, and Deguang Wang. “U-Net combined with multi-scale attention mechanism for liver segmentation in CT images”. In: *BMC Medical Informatics and Decision Making* 21.1 (2021), pp. 1–12. ISSN: 14726947. DOI: 10.1186/s12911-021-01649-w. URL: https://doi.org/10.1186/s12911-021-01649-w.

[WWZ20] Zhaobin Wang, E. Wang, and Ying Zhu. “Image segmentation evaluation: a survey of methods”. In: *Artificial Intelligence Review* 53.8 (2020), pp. 5637–5674. ISSN: 15737462. DOI: 10.1007/s10462-020-09830-9. URL: https://doi.org/10.1007/s10462-020-09830-9.

[WZ20] Fuping Wu and Xiahai Zhuang. “CF Distance: A New Domain Discrepancy Metric and Application to Explicit Domain Adaptation for Cross-Modality Cardiac Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4274–4285. ISSN: 1558254X. DOI: 10.1109/TMI.2020.3016144.

[Xie+22] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. “C 2 AM: Contrastive learning of Class-agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 979–988. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00106. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Xie_C2AM_Contrastive_Learning_of_Class-Agnostic_Activation_Map_for_Weakly_Supervised_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9880221/.

[Xin+21] Hualei Xin, Jessica Y Wong, Caitriona Murphy, Amy Yeung, Sheikh Taslim Ali, Peng Wu, and Benjamin J Cowling. “The Incubation Period Distribution of Coronavirus Disease 2019: A Systematic Review and Meta-analysis”. In: *Clinical Infectious Diseases* 73.12 (2021), pp. 2344–2352. ISSN: 1058-4838. DOI: 10.1093/cid/ciab501.

[Xu+22] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. “Reliable Propagation-Correction Modulation for Video Object Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.3 (2022), pp. 2946–2954. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i3.20200. URL: https://www.aaai.org/AAAI22Papers/AAAI-4288.XuX.pdf.

[XWG20] Zimin Xu, Guoli Wang, and Xuemei Guo. "Sensor-based activity recognition of solitary elderly via stigmergy and two-layer framework". In: *Engineering Applications of Artificial Intelligence* 95.November 2019 (2020), p. 103859. ISSN: 09521976. DOI: 10.1016/j.engappai.2020.103859.

[Yan+20] Tiejun Yang, Yudan Zhou, Lei Li, and Chunhua Zhu. "DCU-Net: Multi-scale U-Net for brain tumor segmentation". In: *Journal of X-Ray Science and Technology* 28.4 (2020), pp. 709–726. ISSN: 08953996. DOI: 10.3233/xst-200650. URL: https://sci.bban.top/pdf/10.3233/XST-200650.pdf#view=FitH.

[Yan+21] Jingdong Yang, Jintu Zhu, Hailing Wang, and Xin Yang. "Dilated MultiResUNet: Dilated multiresidual blocks network based on U-Net for biomedical image segmentation". In: *Biomedical Signal Processing and Control* 68.April (2021), p. 102643. ISSN: 17468108. DOI: 10.1016/j.bspc.2021.102643. URL: https://doi.org/10.1016/j.bspc.2021.102643.

[Yao+20] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. "Video Object Segmentation and Tracking". In: *ACM Transactions on Intelligent Systems and Technology* 11.4 (2020). ISSN: 21576912. DOI: 10.1145/3391743.

[YFF17] Nawel Yala, Belkacem Fergani, and Anthony Fleury. "Towards improving feature extraction and classification for activity recognition on streaming data". In: *Journal of Ambient Intelligence and Humanized Computing* 8.2 (2017), pp. 177–189. DOI: 10.1007/s12652-016-0412-1.

[YFX19] Linjie Yang, Yuchen Fan, and Ning Xu. "Video instance segmentation". In: *Proceedings of the IEEE International Conference on Computer Vision* 2019-Octob (2019), pp. 5187–5196. ISSN: 15505499. DOI: 10.1109/ICCV.2019.00529. URL: https://openaccess.thecvf.com/content_ICCV_2019/papers/Yang_Video_Instance_Segmentation_ICCV_2019_paper.pdf.

[YHL22] Huifeng Yao, Xiaowei Hu, and Xiaomeng Li. "Enhancing Pseudo Label Quality for Semi-supervised Domain-Generalized Medical Image Segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.3 (2022), pp. 3099–3107. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i3.20217. URL: https://www.aaai.org/AAAI22Papers/AAAI-4132.HuifengY.pdf.

[YJ06] Liu Yang and R Jin. "Distance metric learning: A comprehensive survey". In: *Michigan State Universiy* (2006), pp. 1–51. ISSN: 00401951. DOI: 10.1073/pnas.0809777106. URL: http://www.cs.cmu.edu/~liuy/frame_survey_v2.pdf.

[YK16] Fisher Yu and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions". In: *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* (2016). URL: https://arxiv.org/pdf/1511.07122.pdf.

[Zam+20] Masoomeh Zameni, Amin Sadri, Zahra Ghafoori, Masud Moshtaghi, Flora D. Salim, and Et. Al. "Unsupervised online change point detection in high-dimensional time series". In: *Knowledge and Information Systems* 62.2 (Feb. 2020), pp. 719–750. ISSN: 0219-1377. DOI: 10.1007/s10115-019-01366-x. URL: https://doi.org/10.1007/s10115-019-01366-x%20http://link.springer.com/10.1007/s10115-019-01366-x.

[Zha+21] Yan Zhang, Yao Lu, Wankun Chen, Yankang Chang, Haiming Gu, and Bin Yu. “MSMANet: A multi-scale mesh aggregation network for brain tumor segmentation”. In: *Applied Soft Computing* 110 (2021), p. 107733. ISSN: 15684946. DOI: 10.1016/j.asoc.2021.107733. URL: https://doi.org/10.1016/j.asoc.2021.107733.

[Zha+22] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. “TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 12073–12083. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.01177. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Zhang_TopFormer_Token_Pyramid_Transformer_for_Mobile_Semantic_Segmentation_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9880208/.

[Zha+23] Bangcheng Zhan, Enmin Song, Hong Liu, Zhenyu Gong, Guangzhi Ma, and Chih Cheng Hung. “CFNet: A medical image segmentation method using the multi-view attention mechanism and adaptive fusion strategy”. In: *Biomedical Signal Processing and Control* 79.P1 (2023), p. 104112. ISSN: 17468108. DOI: 10.1016/j.bspc.2022.104112. URL: https://doi.org/10.1016/j.bspc.2022.104112.

[Zhe+21] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. “Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2021), pp. 6877–6886. ISSN: 10636919. DOI: 10.1109/CVPR46437.2021.00681.

[Zho+18] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. “Unet++: A nested u-net architecture for medical image segmentation”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11045 LNCS (2018), pp. 3–11. ISSN: 16113349. DOI: 10.1007/978-3-030-00889-5{_}1. URL: https://arxiv.org/pdf/1807.10165.pdf.

[Zho+22] Ziqi Zhou, Lei Qi, Xin Yang, Dong Ni, and Yinghuan Shi. “Generalizable Cross-modality Medical Image Segmentation via Style Augmentation and Dual Normalization”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 20824–20833. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.02019. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Zhou_Generalizable_Cross-Modality_Medical_Image_Segmentation_via_Style_Augmentation_and_Dual_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879047/.

[Zhu+22] Chenming Zhu, Xuanye Zhang, Yanran Li, Liangdong Qiu, Kai Han, and Xiaoguang Han. “SharpContour: A Contour-based Boundary Refinement Approach for Efficient and Accurate Instance Segmentation”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 4382–4391. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.

00435. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Zhu_SharpContour_A_Contour-Based_Boundary_Refinement_Approach_for_Efficient_and_Accurate_CVPR_2022_paper.pdf%20https://ieeexplore.ieee.org/document/9879127/.

[ZLW18] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. “Road Extraction by Deep Residual U-Net”. In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753. ISSN: 15580571. DOI: 10.1109/LGRS.2018.2802944. URL: https://arxiv.org/pdf/1711.10684.pdf.

[ZZ22] Ke Zhang and Xiahai Zhuang. “CycleMix: A Holistic Strategy for Medical Image Segmentation from Scribble Supervision”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 11646–11655. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.01136. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Zhang_CycleMix_A_Holistic_Strategy_for_Medical_Image_Segmentation_From_Scribble_CVPR_2022_paper.pdf%20http://arxiv.org/abs/2203.01475%20https://ieeexplore.ieee.org/document/9879481/.

Titre: Méta-décomposition et processus d'évaluation dans les applications d'apprentissage automatique

Mots clés: Évaluation, Segmentation, Décomposition, Méta-décomposition, Segmentation d'images médicales, Reconnaissance d'activités, Détection d'événements sonores.

Résumé: La segmentation est une étape cruciale dans diverses applications du monde réel telles que l'analyse d'images médicales, la reconnaissance d'activités et la détection d'événements sonores. Elle implique de diviser les données d'entrée en segments plus petits, ce qui induit des modifications dans certaines caractéristiques des données d'entrée. Ce processus introduit au moins deux familles de biais incontrôlables. La première famille de biais est introduite dans le modèle en raison des changements dans l'espace du problème provoqués par la segmentation elle-même. La deuxième famille de biais est causée par le processus de segmentation lui-même, y compris la fixation de la méthode de segmentation et de ses paramètres. Cette thèse présente une nouvelle couche adaptative conçue pour améliorer les modèles de segmentation d'images médicales existants, améliorant ainsi leurs performances. Cette couche adaptative ajuste dynamiquement la taille du champ récepteur en fonction des informations des pixels et de leur voisinage. Ces concepts sont ensuite étendus à des scénarios plus complexes impliquant des types de données hétérogènes, présentant une nouvelle approche de méta-décomposition ou d'apprentissage de la décomposition pour la segmentation. Cette approche atténue les biais implicites tout en permettant une segmentation adaptative pour différents types de données, prenant en compte les variations et les hétérogénéités des données telles que les différences saisonnières dans les activités. Reconnaissant l'impact de la segmentation sur l'espace du problème, la recherche examine les inconvénients des méthodes d'évaluation de pointe, en mettant l'accent sur la nécessité de cadres plus complets qui se concentrent sur des méthodes d'évaluation basées sur des points, négligeant les relations spatiales ou temporelles entre les instances. Pour valider l'efficacité des techniques d'évaluation suggérées et de l'approche de méta-décomposition, des expérimentations approfondies sont menées sur divers ensembles de données réels concrets.

Title: Meta-Decomposition and Evaluation Processes in Machine Learning Applications

Keywords: Evaluation, Segmentation, Decomposition, Meta-Decomposition, Meta-Learning, Medical Image Segmentation, Activity Recognition, Sound Event Detection

Abstract:
Segmentation is a crucial primary step in a variety of real-world applications such as medical image analysis, activity recognition, and sound event detection. It involves partitioning input data into smaller segments, thereby inducing alterations in certain characteristics of the input data. This process introduces at least two families of uncontrollable biases. The first family of biases is introduced to the model due to the changes in problem space made by the segmentation. The second family of biases is caused by the segmentation process itself, including the fixation of the segmentation method and its parameters. This thesis presents a novel adaptive layer designed to augment existing medical image segmentation deep models, enhancing their performance. This adaptive layer dynamically adjusts the receptive field size based on pixel and neighboring information. These concepts are then extended to more intricate scenarios involving heterogeneous data types, presenting a novel meta-decomposition or learning-to-decompose approach for segmentation. This approach mitigates implicit biases while enabling adaptive segmentation for various data types, accommodating data variations and heterogeneities such as seasonal differences in activities. Recognizing the impact of segmentation on the problem space, the research scrutinizes the drawbacks of state-of-the-art evaluation methods, emphasizing the necessity for more comprehensive frameworks, focusing on point-based evaluation methods, neglects spatial or temporal relationships between instances. To validate the efficacy of the suggested evaluation techniques and the meta-decomposition approach, extensive experimentation is conducted across diverse concrete real-world datasets.