

Uncertainty Quantification and Propagation for Land Cover Classification in Remote Sensing

Christoph Schweden

Vollständiger Abdruck der von der TUM School of Engineering and Design der
Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Urs Hugentobler

Prüfende der Dissertation:

1. Prof. Dr.-Ing. habil. Xiaoxiang Zhu
2. Prof. Dr. Göran Kauermann
3. Prof. Dr. Konrad Schindler

Die Dissertation wurde am 16.10.2025 bei der Technischen Universität München
eingereicht und durch die TUM School of Engineering and Design am 13.02.2026
angenommen.

Abstract

The rapid rise of Deep Learning (DL) has thoroughly transformed the field of Remote Sensing (RS), enabling unprecedented advancements in Earth Observation (EO). However, as these models have become increasingly complex, the fundamental need for Uncertainty Quantification (UQ) has emerged to provide a deeper understanding and more reliable interpretation of the models' predictions. This dissertation addresses the critical need for UQ in DL models for RS applications, particularly in Land Use / Land Cover (LULC) classification. As the application spectrum and data variability broaden, moving beyond mere predictive accuracy towards understanding the model's indecisiveness is crucial. Specifically, this dissertation investigates the following research questions:

1. How can human label uncertainty be effectively integrated into the training pipeline of Neural Network (NN) classifiers?
2. Do EO classification models benefit from flexible label uncertainty frameworks incorporating Bayesian reasoning?
3. To what extent can RS models profit from architectural changes targeting effective Out-of-Distribution (OOD) detection?
4. How effective is Conformal Prediction (CP) in providing statistically valid uncertainty estimates for remote sensing applications, and can CP-based strategies enhance Active Learning (AL) efficiency?

The conducted and presented research systematically investigates the integration and utilization of UQ in EO, targeting three separate downstream tasks of UQ. These are namely calibration, i.e. how well the predicted probabilities match the model accuracy, OOD detection, and AL, which utilizes predictive uncertainties to select informative samples. The following four key methodological areas are explored:

1. Incorporating human label uncertainty: By utilizing *distributional labels* derived from expert annotations and modifying a network's output and loss, the published work demonstrates significant improvements in model generalization and calibration while upholding classical performance.
2. Estimating *label embeddings* via Bayesian reasoning: By imposing a three-level UQ scheme, unconstrained label embeddings are estimated based on Bayesian statistics. The study shows drastic improvements in calibration performance and proves Bayesian labels effective for OOD detection.

Abstract

3. Distance-Aware NNs targeting input distance sensitivity: Frameworks like Spectral-Normalized Neural Gaussian Process (SNGP) and Deterministic Uncertainty Estimation (DUE) are shown to substantially enhance OOD detection performance in RS image classification. They effectively transfer input distances to the output space utilizing architectural adaptations and task-oriented approximation techniques.
4. Fine-grained analysis of uncertainties in remote sensing image classification through Conformal Prediction: The framework is demonstrated to provide statistically valid uncertainty estimates through prediction sets, offering more interpretable insights into the model predictions. Furthermore, CP-based strategies significantly enhance AL efficiency, enabling models to achieve high predictive performance with a fraction of the labeled data and simultaneously providing small predictive uncertainties.

Collectively, these contributions advance the State of the Art (SOTA) in UQ for RS, leading to more robust, reliable, and efficient AI systems for EO. This is particularly fulfilled by explicitly addressing uncertainties stemming from data, models, and human annotation.

Zusammenfassung

Der rasante Aufstieg des Deep Learning (DL) hat den Bereich der Fernerkundung grundlegend verändert und beispiellose Fortschritte in der Erdbeobachtung ermöglicht. Da diese Modelle jedoch immer komplexer geworden sind, ist ein grundlegender Bedarf an Unsicherheitsquantifizierung (UQ) entstanden, um ein tieferes Verständnis und eine zuverlässigere Interpretation der Vorhersagen der Modelle zu ermöglichen. Diese Dissertation befasst sich mit der dringenden Notwendigkeit der UQ in DL-Modellen für Fernerkundungs-Anwendungen, insbesondere bei der Bodenbedeckungsklassifizierung. Angesichts des immer breiteren Anwendungsspektrums und der zunehmenden Datenvariabilität ist es von entscheidender Bedeutung, über die reine Vorhersagegenauigkeit hinauszugehen und die Unentschlossenheit des Modells zu verstehen. Konkret untersucht diese Dissertation die folgenden Forschungsfragen:

1. Wie kann die Unsicherheit menschlicher Annotationen effektiv in die Trainingspipeline von neuronalen Netzen (NN) zur Klassifikation integriert werden?
2. Profitieren Erdbeobachtungs-Klassifizierungsmodelle von flexiblen Frameworks für Label-Unsicherheit, die Bayesianische Statistik einbeziehen?
3. Inwieweit können Fernerkundungsmodelle von architektonischen Änderungen profitieren, die auf eine effektive Out-of-Distribution (OOD)-Erkennung abzielen?
4. Wie effektiv ist Conformal Prediction (CP) bei der Bereitstellung statistisch valider Unsicherheitsschätzungen für Fernerkundungsanwendungen, und können CP-basierte Strategien die Effizienz des Active Learning (AL) verbessern?

Die durchgeführte und vorgestellte Forschung untersucht systematisch die Integration und Nutzung von UQ in der Erdbeobachtung und konzentriert sich dabei auf drei separate Downstream-Tasks der UQ. Dabei handelt es sich um die Kalibrierung, d.h. wie gut die vorhergesagten Wahrscheinlichkeiten mit der Modellgenauigkeit übereinstimmen, OOD-Erkennung und AL, bei welcher Vorhersageunsicherheiten zur Auswahl informativer Datenpunkte genutzt werden. Die folgenden vier methodischen Schlüsselbereiche werden untersucht:

1. Einbeziehung der Unsicherheit menschlicher Annotationen: Durch die Verwendung von Labels in der Form von Verteilungen, die aus Expertenannotationen abgeleitet wurden, und die Modifizierung der Modellvorhersage und der Verlustfunktion eines Netzwerks zeigt die veröffentlichte Arbeit signifikante Verbesserungen bei der Modellgeneralisierung und -kalibrierung unter Beibehaltung der klassischen Modell-Performance.

2. Schätzung von Label-Embeddings mittels Bayesianischer Statistik: Durch die Schaffung eines dreistufigen UQ-Schemas werden unbeschränkte Label-Embeddings auf der Grundlage Bayesianischer Statistik geschätzt. Die Studie zeigt drastische Verbesserungen der Kalibrierungsleistung und belegt die Wirksamkeit Bayesianischer Labels für die OOD-Erkennung.
3. Distanzbewusste NNs, die auf die Empfindlichkeit gegenüber Eingabedistanzen abzielen: Frameworks wie Spectral-Normalized Neural Gaussian Process (SNGP) und Deterministic Uncertainty Estimation (DUE) verbessern nachweislich die OOD-Erkennungsleistung bei der Fernerkundungs-Bildklassifizierung. Sie übertragen Eingabedistanzen mithilfe von architektonischen Anpassungen und aufgabenorientierten Approximationstechniken effektiv in den Ausgaberaum.
4. Feinmaschige Analyse von Unsicherheiten in der Klassifizierung von Fernerkundungsbildern durch CP: Das Framework liefert statistisch valide Unsicherheitsschätzungen durch Vorhersagesätze und bietet so besser interpretierbare Einblicke in die Modellvorhersagen. Darüber hinaus verbessern CP-basierte Strategien die AL-Effizienz erheblich, sodass Modelle mit einem Bruchteil der annotierten Daten eine hohe Vorhersagegenauigkeit erzielen und gleichzeitig geringe Vorhersageunsicherheiten bieten.

Zusammen genommen bringen diese Beiträge den Stand der Technik in der UQ für Fernerkundung voran und führen zu robusteren, zuverlässigeren und effizienteren KI-Systemen für die Erdbeobachtung. Dies wird insbesondere durch die explizite Berücksichtigung von Unsicherheiten erreicht, die aus Daten, Modellen und menschlichen Annotationen stammen.

Dedicated to Anna & Emil, with whom I am looking forward to the adventures awaiting

...

Acknowledgments

I can say without a doubt that this dissertation would not have been successful without the support of various people. First and foremost, I want to thank my first supervisor, Prof. Xiao Xiang Zhu. By letting me explore many research topics freely and by providing me with an excellent network of researchers, I truly flourished in driving my research forward. I also want to thank my second supervisor, Prof. Kauermann, for his constant support: Whenever there was an idea to discuss or a draft to revise, he had an open door and an open ear. My project partner Katharina, I want to thank you for a fruitful collaboration; together, we drove label uncertainty quantification research forward. Thanks also to Prof. Schindler, who agreed to serve as an external examiner of my thesis, and Prof. Hugentobler for agreeing to take over the role of the Vorsitz of my defense.

The road to getting a PhD surely is like a roller coaster ride, with many ups and downs and (almost) never following a straight line. Good to have many companions on my side! From the German Aerospace Center (DLR), I want to particularly thank Codruț for constant support on work-related (especially on polishing my rusty coding skills) and non-work-related topics, many much-needed time-outs from work in saunas and beer pubs, and for many great memories. A big thanks also goes out to Teo for traveling together to the most remote conference locations possible and creating unique experiences that will last a lifetime, as well as countless discussions about brewing, fermentation, tea, and so much more. Cédric I would like to thank you for always being open to a well-founded discussion, as well as contributing to many memorable nights. I thank Nikolai for being a great office neighbor and being supportive no matter what. Mohsin introduced me to the broad field of uncertainty quantification, and never said no to a lengthy discussion about gaming or any other topic, really. Thanks also for showing your impressive freestyle cooking skills!

Regarding IT-related issues (of which there were a lot in the very beginning), I want to thank Eike for his continuous support. Also, thank you to Yingjie and Eva-Maria for handling all the other non-IT-related issues. Paul, thanks for making our Berlin trip so worthwhile! A big shoutout also goes out to Nassim for keeping me informed about all the Gossip at DLR and for handling the coffee business seriously. Konrad, I want to thank you for being actively involved in my research, connecting open dots, and for supporting me right before the finish line! Daixin, you were always interested in my ongoing research and happy to discuss, thank you for that. Anthony, I have to thank you again for your unbelievably good Tiramisu! A thank you also goes to Protim for providing me with insights on Conformal Prediction and collaborating on a project. My working group also played a major role in shaping my research: Nina, thanks a lot for the great guidance and constant support. I wish you all the best in your next steps! Jakob, you were always open to discussing ideas, code, and future projects. Thank you

Acknowledgments

for that. Thanks also to Peter and Chandra for actively engaging in my projects. Lastly, my gratitude goes out to Anja and Andrés, for always having an open ear and open door, and for great guidance in the department!

From the future lab, I first and foremost want to thank Peter, not just for being my mentor and collaborating on two key projects with me. Moreover, I thank him for his constant support of my research, for countless meetings, discussions, sketches of ideas, and draft iterations. Thank you heaps! Prof. Shahzad and Prof. Triebel, I want to thank you for your active engagement in my research and in the uncertainty working group. Prof. Roscher and her PhD students Johannes and Lukas, I want to thank you for the many fruitful discussions and for the memorable conference in Bonn! Prof. Bamber, thank you for your interest in my research and the many interesting discussions about aspects of (academic) life we had. Archie, thanks for leading the working group and pushing the joint publication. A big shoutout also goes out to Jörg and Matthias from the science management team, who helped me more than once. Santiago, I want to thank you for giving a talk at DLR and for the nice memories! Ivica, you have been supportive since the beginning. Thank you for that, and all the best to your newly founded family! Nils and Viola, I would like to thank you for the fun times we had on numerous occasions; it was always a pleasure being around you.

From my graduate school, Munich School for Data Science (MuDS), I firstly want to thank Mara and Julia for their unwavering support and excellent management! We had countless events, ranging from summer and winter schools, over scientific exchanges and guest talks, to get-togethers and outdoor activities. They were all terrific and carefully planned, thanks a lot for making this possible! From the PhD students, I would like to particularly thank Martin for always being there and having an open ear, we created a lot of good memories together! Nikita and Jianxiang, thanks for always being open to discussing research. I further would like to thank Alejandra, Benno, Dominik, Henning, Lorenz, Marko, Oleg, Simon, and Vlad (in alphabetical order), for insightful discussions at various MuDS occasions. A huge thanks also goes out to my early mentors, who paved the way towards my PhD journey. Matthias, thank you for supporting me early on and always showing me guidance through my professional career (so far). Janis, thank you for taking me under your hood throughout my Master's and my industry internship, and encouraging me to go down the PhD road. I want to thank Michael for providing insights into the statistics chair and every everyday life of a PhD student, and lastly, thanks to Lukas for providing a fantastic onboarding experience.

Of course, such a journey could not have been possible without the unwavering support of my family. Thank you for always being there for me, encouraging me in troubling times, and showing endless interest in my projects and successes. The biggest shoutout, however, goes out to my dear wife Anna: Without you, none of this would have even remotely worked out the way it did. By constantly supporting, motivating, and encouraging me to drive forward, you made this PhD journey truly unique. Thanks for the endless discussions of my research, even if (or particularly because) it lies outside of your expertise. Roughly 9 months ago, you then gave me the biggest present life has to offer: Our son Emil. Thank you for going the extra mile in spending time with Emil and letting me finish my PhD in a reasonable time!

Contents

Abstract	iii
Zusammenfassung	v
Acknowledgments	ix
Contents	xi
List of Figures	xiii
List of Tables	xv
Acronyms	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Research Questions	3
1.3 Dissertation Structure	4
2 Methodological Background	7
2.1 A Note on Deep Learning	7
2.2 Deep Learning for Remote Sensing	10
2.3 Uncertainty Quantification in a Nutshell	12
3 State of the Art	17
3.1 Land Cover Classification	17
3.2 Uncertainty Quantification in Classification	22
4 Incorporating Label Uncertainty	29
4.1 Distributional Labels Approach for LCZ Classification	30
4.2 Learning from Label Embeddings	33
5 Uncertainty Quantification in Remote Sensing	39
5.1 Distance-Aware Neural Networks	39
5.2 Conformal Score-Based Active Learning	43
6 Conclusion	49
6.1 Limitations of the Current Work	52
6.2 Future Research Directions	54
Bibliography	57

CONTENTS

A Publications	75
A.1 Going Beyond One-Hot Encoding in Classification	75
A.2 Can Uncertainty Quantification Benefit From Label Embeddings?	87
A.3 Can Land Cover Classification Models Benefit From Distance-Aware Architectures?	102
A.4 Assessing Conformal Prediction for Remote Sensing: Application to Local Climate Zone Classification	108
B Related Publications	125

List of Figures

1.1	Classification of the introduced UQ frameworks	5
2.1	Example architecture of a Convolutional Neural Network (CNN).	8
2.2	Topics in RS impacted by DL.	10
2.3	Different types of resolutions associated with satellite imagery.	11
2.4	Broad classification of UQ methods.	13
3.1	Local Climate Zones (LCZ) classification scheme.	18
3.2	Label uncertainty within the So2Sat LCZ42 Label Evaluation Dataset. . .	20
3.3	The Sen2LCZ architecture.	21
4.1	Calibration performance at test time.	32
4.2	Different levels of uncertainty.	33
4.3	Class contributions to the Static Calibration Error (SCE).	36
5.1	Sketch of the SNGP approach.	40
5.2	Data split into In-Distribution (ID) and OOD data.	41
5.3	Explorative Analyses for Conformal Prediction (CP) methods.	44
5.4	Predictive Analyses for Conformal Prediction (CP) methods.	45
5.5	Entropy by pred. set size.	45
5.6	AL results	46
5.7	Entropy vs. pred. set size	46

List of Tables

4.1	Generalisation and Calibration Test Performance	31
4.2	Test Performance for embedding approaches.	36
4.3	OOD Detection Results.	37
5.1	OOD Detection results on the Eurosat dataset.	42
5.2	OOD Detection results on the So2sat LCZ42 dataset.	42
5.3	Odds of label uncertainty	45

Acronyms

AID	Aerial Image Dataset.
AL	Active Learning.
APS	Adaptive Prediction Sets.
AUC	Area under Curve.
AUPR	Area under Precision / Recall Curve.
BNN	Bayesian Neural Network.
CE	Cross Entropy.
CNN	Convolutional Neural Network.
CP	Conformal Prediction.
CV	Computer Vision.
DL	Deep Learning.
DSM	Dempster-Shafer Metric.
DUE	Deterministic Uncertainty Estimation.
DUQ	Deterministic Uncertainty Quantification.
ECE	Expected Calibration Error.
ELBO	Expectation Lower Bound.
EM	Expectation-Maximization.
EO	Earth Observation.
GEE	Google Earth Engine.
GP	Gaussian Process.
GPU	Graphics Processing Unit.
GSD	Ground Sampling Distance.
GUF	Global Urban Footprint.
i.i.d.	independent, identically distributed.
ID	In-Distribution.
iff	if and only if.
ILSVRC	ImageNet Large scale Visual Recognition Challenge.
IoU	Intersection over Union.
KL	Kullback-Leibler.

Acronyms

LAC	Least Ambiguous Set-Valued Classifier.
LCZ	Local Climate Zones.
LDL	Label Distribution Learning.
LS	Label Smoothing.
LULC	Land Use / Land Cover.
MC	Monte-Carlo.
ML	Machine Learning.
MLP	Multi-Layer Perceptron.
MSP	Maximum Softmax Probability.
NN	Neural Network.
OOD	Out-of-Distribution.
OSM	Open Street Map.
p.p.	percentage points.
RAPS	Regularized Adaptive Prediction Sets.
RBF	Radial Basis Function.
ReLU	Rectified Linear Unit.
ResNet	Residual Network.
RFF	Random Fourier Feature.
RGB	Red Green Blue.
RS	Remote Sensing.
SAR	Synthetic Aperture Radar.
SCE	Static Calibration Error.
SNGP	Spectral-Normalized Neural Gaussian Process.
SOTA	State of the Art.
std.	standard deviation.
SVM	Support Vector Machine.
TS	Temperature Scaling.
UCM	US-Merced Land Use.
UQ	Uncertainty Quantification.
VGG	Visual Geometry Group.
VI	Variational Inference.
WRN	WideResNet.
WUDAPT	World Urban Database.

1 Introduction

1.1 Motivation

Remote Sensing (RS) has fundamentally changed humanity’s capacity to monitor and understand Earth’s environment in the past decades. The continuous advancement in sensor technologies and data acquisition modalities has led to a remarkable increase in the size and complexity of acquired satellite data over the years. Remotely sensed data is inherently multi-faceted, characterized by four primary intrinsic dimensions. This intricate and multi-dimensional nature of RS data inherently calls for sophisticated and robust modeling techniques for an effective extraction of information. The advent and rapid propagation of Deep Learning (DL) models have profoundly impacted nearly every facet of data modeling, particularly within Computer Vision (CV). This influence has extended to remotely sensed data in many ways.

In the majority of these applications, existing CV architectures are explicitly adapted to the unique data modalities of RS. This may involve combining different data sources, fusing multi-seasonal or multi-temporal data, or targeting multi-spectral data. The increasing complexity and sheer volume of remotely sensed data, coupled with the widespread integration of DL models into critical applications such as disaster response, climate monitoring, or urban planning, underscore a fundamental shift in the requirements for AI systems in EO. Simply achieving high predictive performance is no longer sufficient. As these models become more profoundly embedded in decision-making processes, the reliability and trustworthiness of their predictions become primary objectives. As current benchmark techniques in AI models for EO, ensembles of models and aggregating multiple probabilistic test predictions are widely considered. Generally, these objectives are targeted by the research field of Uncertainty Quantification (UQ), which this dissertation will cover in great detail with a focus on calibration, OOD detection, and Active Learning.

UQ has emerged as a critical field of research, gaining significant attention with the rise of Machine Learning (ML) methods. At its core, UQ aims to quantify the inherent variability or lack of knowledge accompanying a model’s prediction. Two main types of uncertainties are commonly distinguished: *aleatoric* (or data) uncertainty and *epistemic* (or model) uncertainty. Aleatoric uncertainty is inherent in the data itself, reflecting irreducible randomness or noise in observations, and thus cannot be modeled away. In contrast, epistemic uncertainty arises from the model’s specifications or a lack of knowledge about the true underlying function. It is dependent on model choices and can, in principle, be reduced by utilizing more suitable modeling techniques or acquiring more data. While theoretically distinct, in practice, these two sources of uncertainty are often interdependent.

1 Introduction

In general, UQ provides different measures of indecisiveness, allowing decision-makers to understand the complexity of model predictions and make more informed decisions. As a peculiarity of classification problems, uncertainty in the label space is investigated as part of this dissertation. RS data often contains inherent ambiguities or uncertainties in its labels, which can arise from overlapping class definitions, limited data quality, or the human annotation process itself. UQ methods are crucial for effectively modeling these ambiguities, moving beyond simplistic single-label assumptions. The addressed goals of this work in this regard are twofold: In a first step, human label uncertainty, which can exemplarily be measured by incorporating multiple human labelers, will be directly implemented into the classification pipeline of a NN. As an add-on to this experimental setup, the label uncertainty will be estimated in a broader framework in a next step. For this, a Bayesian framework is utilized to build a continuous embedding space to characterize the label uncertainty more comprehensively.

Several downstream applications directly benefit from the integration of UQ. Calibration ensures that a model’s predicted probabilities align with its actual accuracy, which is crucial given that many neural networks are naturally overconfident. In RS, this framework is, for example, useful for disaster monitoring to avoid false alarms from the model. Out-of-Distribution (OOD) detection, on the other hand, leverages predictive uncertainty to identify data samples that differ significantly from the training distribution, thereby preventing erroneous predictions on unseen or anomalous data. In an EO context, a potential application lies in classifying destroyed buildings after a natural disaster, where the damaged buildings could be targeted as OOD data. Finally, Active Learning (AL) strategically selects the most informative unlabeled data points for human annotation, thereby reducing costly labeling efforts and improving model efficiency by focusing resources where they are most needed. Again, many RS applications are possible. Exemplarily, one can think of a land cover classification task of a previously unseen region, where only little labelled training data is available. These downstream tasks present different ways of evaluating the quality of predictive uncertainty. Hence, various ways of UQ can be benchmarked hereby.

As for the utilization of different methodological frameworks for UQ, both deterministic and probabilistic approaches are investigated in this dissertation. While probabilistic approaches are often leading in public benchmarks with respect to performance, they generally come with a much bigger computational effort. Deterministic approaches, on the other hand, are computationally efficient but often depend on many intrinsic distributional assumptions. While off-the-shelf probabilistic methods such as MC-Dropout and ensembles are included in various benchmark experiments, a special stream of deterministic methods is studied for the task of OOD detection, namely *distance-aware* networks. As an alternative way of UQ, carrying both deterministic and probabilistic components, Conformal Prediction (CP) is considered. It provides statistically valid prediction intervals or sets with guaranteed coverage of the true class or target value. This uniquely different UQ method offers more insight into predictive uncertainty and can be evaluated by the above-mentioned downstream tasks.

1.2 Research Questions

This dissertation systematically investigates the integration and utility of UQ within DL models for EO applications, particularly focusing on land cover classification. The works presented in the following are mainly steered by four research questions, which will be introduced subsequently.

Research Question

How can human label uncertainty be effectively integrated into the training pipeline of Neural Network classifiers, and what impact does this have on model generalization and calibration performance?

This question drives the exploration of novel methods for utilizing the explicit inherent ambiguity present in human annotations. By moving beyond traditional one-hot encoded labels, the corresponding work [1] studies this approach with the aim of enhancing the robustness and reliability of classification models in EO. Particularly targeting calibration, the novel method is compared and combined with classical calibration techniques. Distributional labels are showing competitive results and strong generalization performance with respect to the Cross Entropy (CE) between predictions and ground truth labels, beating baseline models by a margin of 10%. The generalization performance can additionally be upheld against widely-used UQ methods such as *MC Dropout* and ensembles.

Research Question

Do EO classification models benefit from flexible label uncertainty frameworks incorporating Bayesian reasoning?

The next question delves into advanced techniques for representing label uncertainty and assessing their utility for UQ. The accompanying work [2] specifically examines a three-level uncertainty scheme involving Bayesian reasoning, prior information, and particularly suitable distributional embeddings to cover the complex structure of ambiguous labels. Its implications for model calibration and OOD detection are studied in a large benchmark, involving all three levels of the novel UQ framework, whose explicit novelty lies in the utilization of an empirical Bayes approach for dealing with label uncertainty. For the OOD detection task, a variety of common uncertainty metrics are studied to systematically evaluate the different UQ methods. These metrics are not solely based on softmax pseudo-probabilities, but furthermore incorporate the parameters of the distributional embeddings to reflect the complexity of the label underlying structure. Calibration errors were reduced by up to 70% using the novel three-level UQ scheme compared to classical training regimes, while OOD detection performance solely gained 1-2 percentage points (p.p.), yet on a high baseline level.

Research Question

To what extent can RS models profit from architectural changes targeting effective OOD detection?

This question investigates the application of specialized network structures to RS data. The architectures studied in the corresponding publication [3] enforce a sense of continuity of the network mapping between the input distances and the output space. By enforcing this continuity, the output distances become interpretable as they behave in a controllable ratio to the input distances. The goal was to determine how these frameworks enhance the model’s ability to identify samples coming from outside of the training regime. For this, a particular semantic split of the observed land cover classes was enforced. Compared to baseline models, a performance increase of up to 22 p.p. was observed from the distance-aware architectures. With certain additional configurations, small performance gains were enabled, which made a close-to-perfect detection possible.

Research Question

How effective is Conformal Prediction (CP) in providing statistically valid uncertainty estimates for remote sensing applications, and can CP-based strategies enhance AL efficiency?

The last question explores the utility of Conformal Prediction (CP) for EO, a statistical framework providing reliable prediction sets with theoretical coverage guarantees of the true class. Multiple versions of the creation of prediction sets are studied in [4] with regard to their prediction set size, coverage, and predictive power towards aleatoric and epistemic uncertainty. The work also investigates CP’s practicality in optimizing data annotation efforts within RS by considering an AL framework. Many considered performance metrics regarding the CP experiments could be increased by 5 to 10 p.p. compared to the unconformalized baseline model. For the AL experiment, the uncertainty-based querying led to optimal test accuracy with only 30% of training data needed, with consistently lower predictive uncertainty simultaneously. Regarding average accuracy, the CP-based querying outperformed previous UQ-based methods.

1.3 Dissertation Structure

The dissertation is organized as follows: This chapter sets the stage by outlining the fundamental significance of RS and the overall impact of DL within this domain. It introduces the critical need for UQ and poses the specific research questions that guide the subsequent parts. Chapter 2 provides essential contextual information, offering a compact introduction to DL principles, a discussion of their specific application spectrum

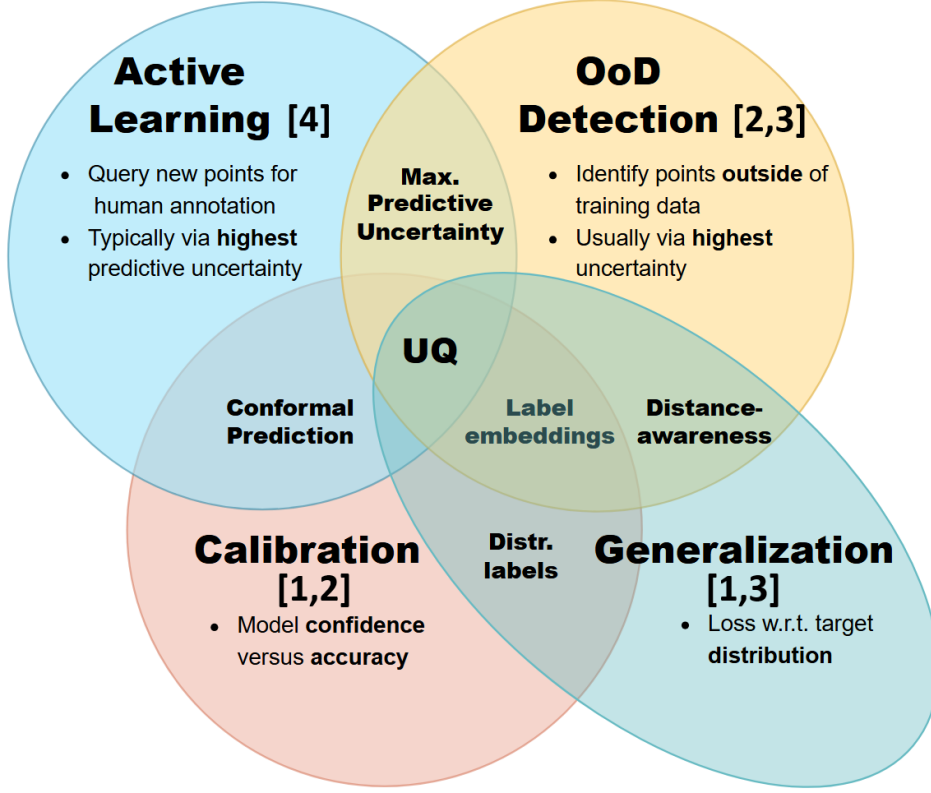


Figure 1.1: Classification of the introduced UQ frameworks with reference to the corresponding publications: Koller *et al.* [1], Schweden *et al.* [2], Koller *et al.* [3], Schweden *et al.* [4].

to remotely sensed data, and a foundational understanding of the field of UQ, including its core concepts and a broad classification of its methods. Following that, Chapter 3 reviews the current advancements in LULC classification, particularly focusing on LCZ classification, and provides an in-depth overview of the State of the Art (SOTA) in UQ within classification tasks.

This context paves the way for the dissertation’s novel contributions [1, 2, 3, 4], which are found in the appendix of this dissertation. For a better classification of the individual proposed UQ frameworks, Figure 1.1 shows how the frameworks fit within the application spectrum of UQ. Firstly, Chapter 4 details the first two core works. Both focus on developing and evaluating methods for directly integrating human label ambiguity into the training pipeline of Neural Network classifiers, demonstrating their impact on model generalization, OOD detection, and calibration performance. Secondly, Chapter 5 presents two further contributions, exploring advanced UQ frameworks that have proven effective for dealing with satellite imagery. It specifically investigates distance-aware neural networks and the CP framework, showcasing their utility in various RS

1 Introduction

applications. The concluding Chapter 6 summarizes the key findings of the dissertation, systematically addressing each of the posed research questions. It discusses the inherent limitations of the current work and outlines promising directions for future research. The appendix includes the full text versions of the peer-reviewed publications that form the empirical and theoretical basis of this dissertation, providing comprehensive details on the methodologies and experimental results.

2 Methodological Background

This chapter is designed to introduce the main concepts used in this work. Starting with a brief introduction to DL, in a next step, the application spectrum of such algorithms to remotely sensed data is shortly discussed. In a last step, the field of UQ is introduced as a whole. Note that methodological details specifically regarding the proposed approaches are listed in the respective chapters and, in more detail, in the accompanying papers in the appendix.

2.1 A Note on Deep Learning

Generally speaking, the field of DL contains a broad spectrum of deep NN architectures designed to solve a variety of tasks. While we give a short introduction in the following, mainly to the concepts used in this work, this falls short of a lot of recent works brilliantly summarizing the whole field of DL. We therefore refer the interested reader to the standard references of Goodfellow *et al.* [5] or Aggarwal *et al.* [6] and can furthermore suggest the application-focused work of Raschka *et al.* [7]. Here, equal importance is assigned to the implementation of the discussed algorithms in Python and specific libraries thereof, the standard open-source software framework for ML and DL implementations.

Let us denote a NN in the following via $h_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, $h_\theta(x) \mapsto y$, underscoring the dependency of the network on the tunable parameters $\theta \in \Theta$, and mapping the input data $x \in \mathcal{X}$ to the output $y \in \mathcal{Y}$. The NN consists of stacked layers of individual *neurons* (and transformations thereof), which are so-called because of their resemblance to neurons in the human brain. An early prototype of modern NN architectures is given by the Multi-Layer Perceptron (MLP). Multiple artificial neurons are here set up in parallel, each taking in several input nodes $x_1, \dots, x_d$ and combining them in a weighted fashion (plus adding a bias term b) to produce a single output $\hat{y}$. Note that both the weights and biases count towards the network's parameters Θ . The weighted sum is furthermore passed through a non-linear *activation function* $\Phi(\cdot)$ to enable adaptation. A single neuron output $\hat{o}$ of an MLP then reads as follows:

$$\hat{o} = \Phi\left(\sum_{j=1}^d w_j \cdot x_j + b\right) \quad (2.1)$$

In fact, the adaptation of an MLP goes as far as stated in the universal approximation theorem [8], namely that any continuous non-linear function can be approximated by the MLP to an arbitrarily small precision (given the MLP is not a polynomial and using infinitely many weights). A chosen number of neurons set up in parallel forms a *layer* of the MLP, and the number of layers determines the depth of the network. What

2 Methodological Background

remains an open question is how to train the parameters of the network. In the supervised setting initially discussed, we have input-output pairs $(x^{(i)}, y^{(i)})$. The predictive precision of the network can be quantified by defining a loss function $\mathcal{L}(h_\theta(x), y)$ measuring the deviation of the model prediction from the output. This loss is targeted to be minimized. Therefore, an optimization task is formed where the loss as an objective function is minimized with respect to the network parameters $\theta \in \Theta$. By construction (using a differentiable activation function $\Phi(\cdot)$), the network is differentiable and hence the gradient $\nabla_\theta h_\theta(x)$ is targeted. Using the common optimization technique *gradient descent*, the loss is iteratively updated by following the path of the steepest gradient. In practice, this is typically not performed on all data at the same time, but rather a *batch* of data $(x^{(1)}, y^{(1)}), \dots, (x^{(n_{\text{batch}})}, y^{(n_{\text{batch}})})$ at a time, yielding

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \frac{1}{n_{\text{batch}}} \sum_{i=1}^{n_{\text{batch}}} \mathcal{L}(h_\theta(x^{(i)}), y^{(i)}). \quad (2.2)$$

Here, η denotes the learning rate, a tunable hyperparameter determining the size of the gradient step. Of particular importance is the special internal structure of NNs, allowing for efficient computation of the above gradient by means of the backpropagation algorithm [9], which is an efficient application of the chain rule to NNs. Although no convergence to the global optimum θ^* is guaranteed, iterative approximation over multiple *epochs* (i.e. iterations over the dataset) often presents convincing performance results.

Most of the modern NN architectures have their foundations in MLPs. Since the application spectrum of this work focuses on CV, i.e., learning from image data using ML or DL methods, we will shortly explain the CNN, the most common architectural basis. The core concepts behind CNNs are depicted in Figure 2.1. What makes the CNN unique compared to previous fully connected networks, such as the MLP, is the concrete utilization of the spatial context in the image.

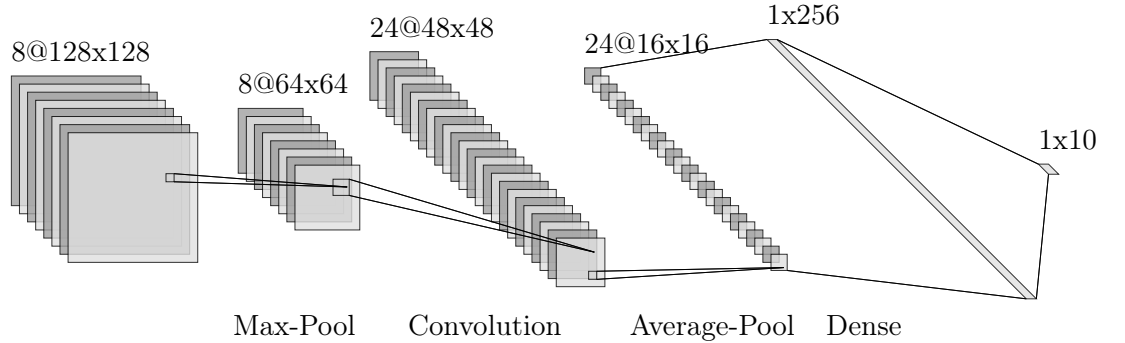


Figure 2.1: Example architecture of a CNN. Own creation via *NN-SVG* [10].

In the example, we look at an image of size 128×128 pixels, with eight separate input channels. In particular for RS, it is common to not restrict the channels to the typical Red Green Blue (RGB) channels but have more fine-grained wavelength spectra. *Pooling* essentially translates to aggregating a quadratic grid of fixed size with respect to a chosen function. While max-pooling selects dominant pixel values, average-pooling equates to a

smoothing layer. The depicted pooling operations reduce the image size (dependent on the grid size and *stride*, i.e. at what step size the grid is moved over the input), yet apply the transformation to each channel individually. Hence, the number of channels stays constant. As for the *convolution*, a grid of learnable parameters, termed *filter*, is applied to the input image by computing the dot product of both tensors. By construction, the filter therefore has the same depth (i.e. number of channels, although after a convolution, these channels do not correspond to the spectral channels anymore).

Different filters are used to capture different spatial patterns, and by increasing the number of distinct filters (as a hyperparameter), the output depth after a convolution increases simultaneously. Activation functions are typically applied after the convolution operations, similarly to the fully connected networks described earlier. Having artificially decreased the image size and increased the depth, what remains is matching the network’s output to the target in order to successfully compute the loss as shown in (2.2). This is done via *dense* layers, which simply utilize MLPs in a prespecified number to transform the input to the desired dimension. Training the parameters of a CNN is again performed via backpropagation, where now the architecture benefits from shared weights within the convolutions across the entire input layer.

As a key milestone, the first framework of a CNN was introduced by LeCun *et al.*. The authors presented the famous *LeNet (-5)* [11] in 1998, exceeding previous advancements in the task of handwritten digit recognition by a large margin. The next big leaps forward were made in the scope of the ImageNet Large scale Visual Recognition Challenge (ILSVRC) [12], an annual contest to beat previous models on the task of large-scale image classification. The dataset in use, *ImageNet* [13], consists of hundreds of thousands of labelled images from thousands of categories, organized by the hierarchical structure of the *WordNet* project [14]. As a first milestone, *AlexNet* was presented as the winner of the 2012 ILSVRC, setting many new standards in the construction of CNNs. These include the use of the Rectified Linear Unit (ReLU) function as activation, Dropout (setting weights to zero randomly with a prespecified probability), weight decay, data augmentation, layer normalization, and the use of a specialized Graphics Processing Unit (GPU) for training.

Utilizing deeper nets (i.e. more layers) and reducing the size of the filters, the Visual Geometry Group (VGG) [15] made another big step towards fine-grained image classification. The *Inception* network architecture [16] furthermore proposed a network within the network. In the inception module, three different filter sizes are utilized in parallel, allowing the network to identify differently shaped and sized objects at the same time. An efficient adaptation using depthwise separable convolution operations, termed *Xception* [17], achieved significant performance gains over previous Inception architectures without any additional parameters. A key architectural design was presented by He *et al.* introducing the Residual Network (ResNet) [18]. With a staggering depth of up to 152 layers, the problem of vanishing or exploding gradients becomes visible, which is tackled via *skip connections* in the framework of ResNets.

These connections effectively add the identity mapping between certain blocks of layers, skipping all the layers in between. For the backpropagation, this enables effective gradient flow [6], preventing them from vanishing or exploding. Noteworthy adaptations

of the initial ResNet architecture include the WideResNet (WRN) [19] (fewer layers, more filters) and the DenseNet [20], which adds direct connections between all layers (hence also skip connections with varying numbers of layers being skipped).

2.2 Deep Learning for Remote Sensing

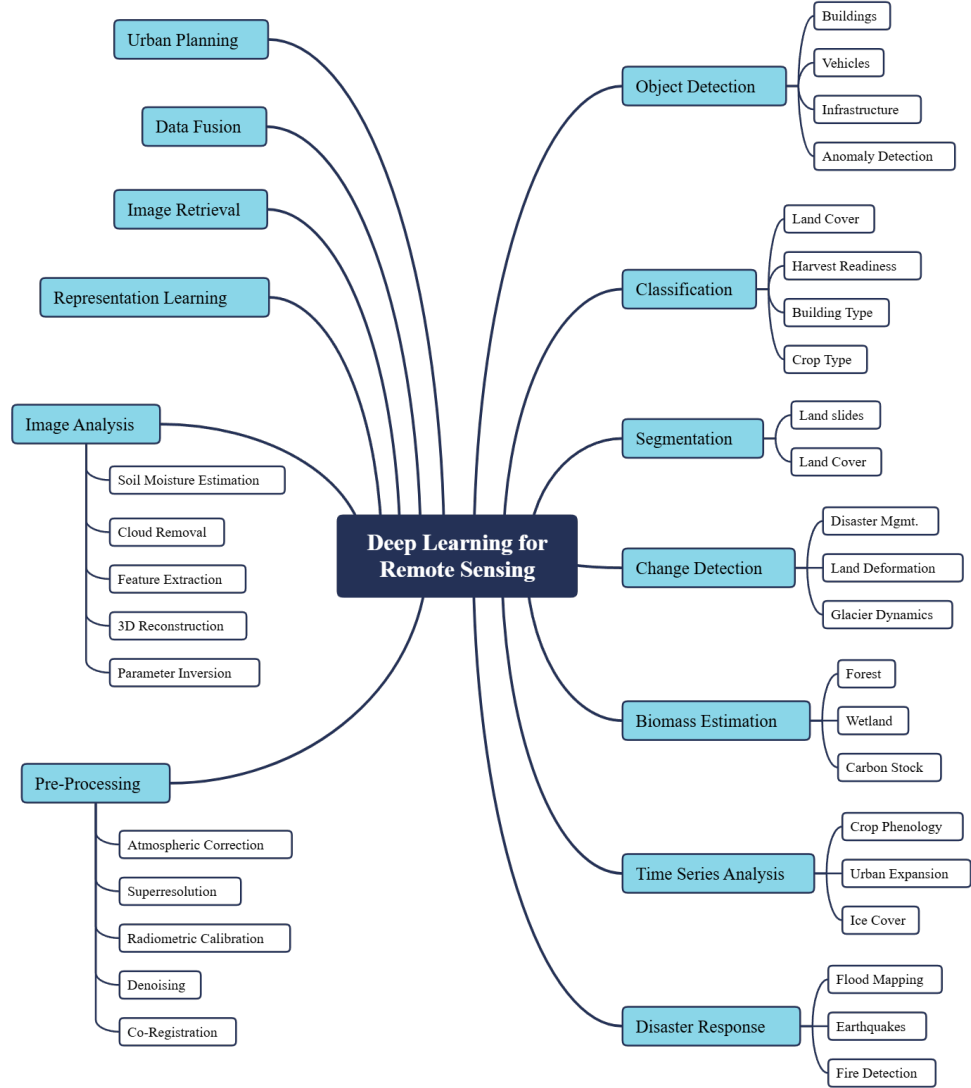


Figure 2.2: Topics in RS impacted by DL (non-exhaustive), own creation.

The rise of DL in almost any aspect of data modeling, and particularly CV, has not halted when it comes to remotely sensed data [21]. For RS applications, this dissertation cannot even begin to describe all tasks impacted by DL; hence, in the following, we will focus on the major aspects. For a quick overview of affected tasks, see Figure 2.2. Ever

since the launch of the first *Landsat* satellite in 1972, sensing the Earth environment remotely shifted its focus from aerial to satellite-based photogrammetry. This was primarily due to longer observation spans and automated coverage. The *Landsat* program has been running ever since and is therefore considered the longest-running acquisition enterprise for satellite imagery [22]. With increasing technological advances and more fine-grained sensor modalities, the acquired satellite data has grown in size and complexity over the years. In particular, remotely sensed data has four main intrinsic data dimensions presented in Figure 2.3. The spatial resolution defines the image quality in terms of its Ground Sampling Distance (GSD), i.e. the length a single pixel of the image covers. Adding to that, the spectral resolution is defined via the number of available wavelength spectra from the satellite’s sensor. These range from gamma rays over the visible range (including the typical RGB spectra in standard imagery) to infrared and radio waves. While the temporal resolution is defined via the revisit time of the satellite, meaning how often it covers the same region of the planet, the radiometric resolution is given by the bit-rate of the acquired images. This multi-faceted type of data naturally requires more complex modeling techniques.

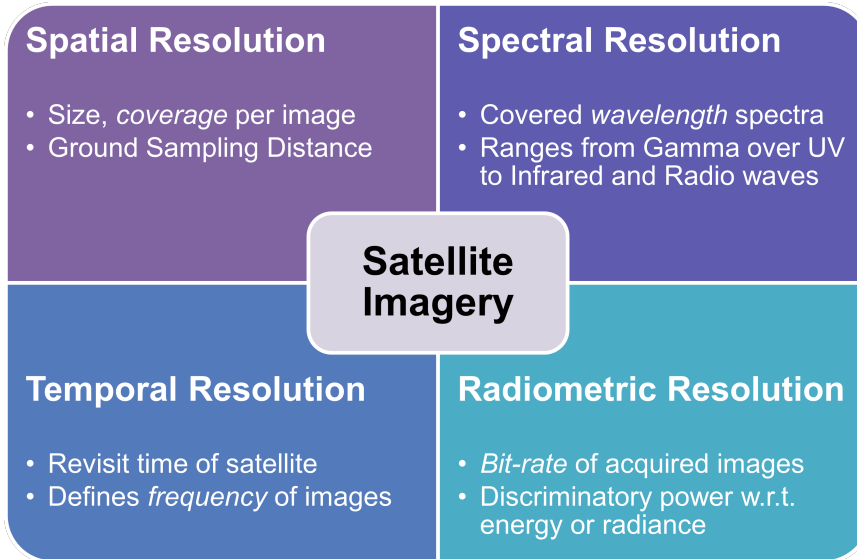


Figure 2.3: Different types of resolutions associated with satellite imagery, own creation.

The first step in the pipeline of utilizing satellite imagery lies in the pre-processing. Several associated tasks, such as denoising or atmospheric correction, which involves removing the distorting effects of the atmosphere to reveal true surface reflectance, can nowadays efficiently be handled by DL algorithms. Taking superresolution as another example, many techniques involving deep NNs have been proposed, see [23] for an extensive survey. Furthermore, geometric correction and registration, the process of precisely aligning images spatially, has been boosted by ML-driven feature matching techniques and DL models that can directly learn and apply spatial transformations, reducing reliance on manual tie points.

2 Methodological Background

Next to the task of processing the satellite data, many downstream tasks have evolved. Starting with the most straightforward task of image classification, i.e., assigning a label on an image-level, the drastic improvements in performance observed in the CV community quickly transmitted to the field of RS [24]. Popular applications, including LULC classification, crop type classification, or building type classification, were advanced at unprecedented speed. Similarly, object detection was impacted, where typical applications include the detection of building footprints or various types of vehicles on land, air, and water. Segmentation tasks, i.e. assigning classes on pixel level (semantic) or delineating individual objects (instance), are no exception, either. Recent trends in these fields were reviewed recently by Hoeser and Kuenzer [25].

Despite the mainstream of image classification and segmentation, many more data modeling tasks are unique to RS. These include, but are not limited to, the estimation of soil moisture or biomass, analysis of time series data for studying e.g. ice cover dynamics or urban expansion, or tasks such as fire detection and monitoring or flood mapping for disaster response. In these applications, oftentimes existing architectures from CV are used as a starting point, and are adapted to different data modalities and peculiarities. This includes, among others, the combination of different sources of RS data, the fusion of multi-seasonal or multi-temporal data, or the reconstruction of three-dimensional data.

2.3 Uncertainty Quantification in a Nutshell

UQ has existed as a research field for quite some time, although recently, it has gained more attention with the rise of ML methods [26]. Most commonly, two main types of uncertainties are distinguished, namely *aleatoric* or data uncertainty, and *epistemic*, or model uncertainty. Epistemic uncertainty arises from the Greek word *epistēmē*, meaning knowledge or understanding, which hints at the variable nature of this type of uncertainty. The aleatoric uncertainty bit, on the other hand, originates from the Latin word *aleatorius*, meaning something pertaining to a gamester. This typically relates to dice games, hence denoting the randomness therein. While aleatoric uncertainty is inherent in the data and thus cannot be modeled away, epistemic uncertainty is dependent on model specifications. It can, in principle, be reduced by utilizing more suitable modeling techniques. Disentangling the two sources of uncertainty is theoretically possible, but in practice, the two sources are interdependent [27]. For a more fine-grained definition of sources of uncertainty and UQ methods, we refer the interested reader to the recent extensive and in-depth survey of Gawlikowski *et al.* [28]. Furthermore, Mena *et al.* [29] present a Bayesian view on the subject, whereas He *et al.* [30] look at the matter from an uncertainty source perspective.

In the most general sense, the aim of UQ lies in quantifying a notion of uncertainty accompanying the model's prediction $P(y|x, \mathcal{D})$, where $\mathcal{D} = \{(x_i, y_i)\}_{i=1, \dots, n}$ denotes the dataset. In a first step, this uncertainty can be separated into the earlier mentioned epistemic and aleatoric parts via the following notation:

$$P(y|x, \mathcal{D}) = \int P(y|x, \theta) p(\theta|\mathcal{D}) d\theta \quad (2.3)$$

Here, the aleatoric part is represented by the conditional distribution of the output y given both the input x and the parameters θ . This represents the variety of the output, steered by the parameters in use and the input. The epistemic bit, on the other hand, is simply the variation of the parameters given the entirety of the data.

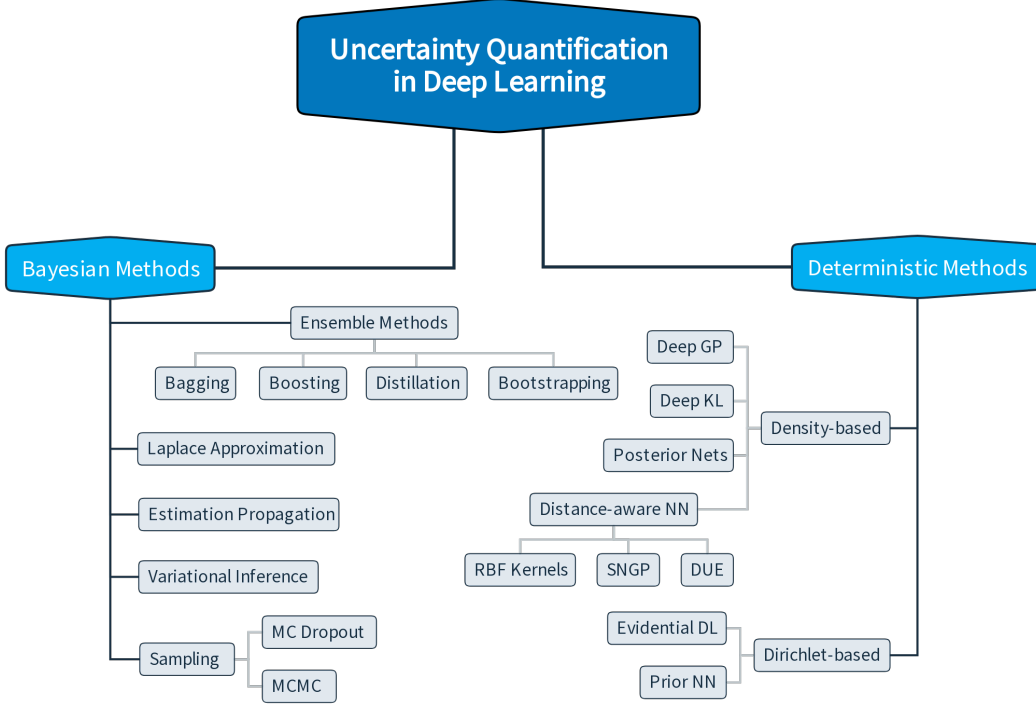


Figure 2.4: Broad classification of UQ methods, own creation.

Two main streams of dealing with this integral (or parts thereof) have been established, namely probabilistic and deterministic methods. See Figure 2.4 for a broad schematic overview. Within the probabilistic methods, *ensemble* methods originating from the work of Dietterich *et al.* [31], have gained a lot of interest recently and achieved many SOTA results. In the simplest setting presented by Lakshminarayanan *et al.* [32], a notion of predictive uncertainty is derived via averaging the predictions of L different models with similar, yet slightly varied hyperparameter settings:

$$P(y|x, \mathcal{D}) \approx \frac{1}{L} \sum_{i=1}^L P(y|x, \theta_i) \quad (2.4)$$

More sophisticated ways of sampling the hyperparameters have been studied [33]. Furthermore, ways of distilling the information from multiple networks [34] as well as bootstrapping [35], bagging, and boosting adaptations [36] have been established, among many others.

2 Methodological Background

In general, the aim of the Bayesian methods lies in approximating the term for the epistemic uncertainty in (2.3) via Bayesian reasoning. For this, first, Bayes’ theorem is utilized to reformulate the term as follows:

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta) \cdot p(\theta) \quad (2.5)$$

Although still intractable, this distribution can be approximated with certain tricks. Most notably, Variational Inference (VI) infers the distribution by minimizing the Kullback-Leibler (KL) divergence to a prespecified family of distributions. This is explicitly done by optimizing the Expectation Lower Bound (ELBO), which matches the KL divergence up to a constant. While Hinton and Van Camp first introduced this technique [37], Barber and Bishop [38] adapted the technique by adding a full covariance matrix to the family of distributions, which was assumed Gaussian. By introducing a stochastic version, Graves [39] turned the approach into a practical tool for NNs. Note furthermore the stream of expectation propagation algorithms first introduced by Hernandez-Lobato and Adams [40] and presented in a stochastic manner by Li *et al.* [41]. Here, the posterior gets approximated through a set of local approximations that are iteratively refined.

Approximate Bayesian methods have become quite successful lately, as they don’t require as much computational complexity as the exact ones. The most prominent example was introduced by Gal and Ghahramani [42] and termed *MC Dropout*. The authors described NNs with dropout layers as approximate Bayesian inference in deep Gaussian Process (GP)s, allowing for easy implementation and quick UQ. This method has been further refined by Kendall and Gal [43], who introduced a sampling technique that accounts for both aleatoric and epistemic uncertainty explicitly. In earlier days, MacKay [44], among others, adapted the Laplace approximation of the Hessian matrix to NNs. Due to the computational complexity, many simplification techniques have been introduced in the following to ease the approximation, see [28] for an extensive history.

Contrary to the probabilistic methods, the deterministic approaches manage their predictions without the need for sampling. For this, certain loss functions or distributional assumptions are placed. Most notably for classification problems, the *Dirichlet* function was chosen by several works to describe the network’s prediction. Since in expectation, every Dirichlet distribution equates to a Categorical distribution, it can be seen as a generalized version thereof. Malinin and Gales [45] utilized the Dirichlet distribution in their *Prior Networks*. Here, the KL divergence was placed as a loss function, and in-distribution samples were equipped with a sharp Dirichlet whereas OOD samples were given a flat Dirichlet, indicating high uncertainty. Using the theory of belief functions [46], Sensoy *et al.* [47] proposed *Evidential Neural Networks*. Here, a notion of uncertainty is calculated based on the individual levels of belief for each class, indicating the unknowingness of the model. See Ulmer *et al.* [48] for a recent extensive survey.

Density-based approaches combine all UQ methods based on the latent output space of a NN. This high-dimensional space is formed after passing the input all the way through the network, before downsampling to the required class dimension is performed. By constructing a Deep Belief Network with GP mappings, Damianou and Lawrence [49] defined Deep GPs, paving the way for methods involving the latent space of a

network. Using normalizing flows to learn class-specific distributions in the latent space, Charpentier *et al.* proposed *Posterior Networks*, which were able to produce new SOTA results on OOD detection benchmark tasks. Combining NNs and a GP [50], *Deep Kernel Learning* [51] was introduced by Wilson *et al.*. This paved the way for many downstream UQ applications.

Using Radial Basis Function (RBF) kernels and a two-sided gradient penalty to enforce sensitivity of the network, Van Amersfoort *et al.* [52] estimated class centroids in the latent space and treated the distance to the centroids as predictive uncertainty quantities, achieving remarkable results. The unique combination of a ResNet with a GP as output mapping was termed *SNGP* [53] by Liu *et al.*, enabling the network to be distance-aware, i.e. mapping distances in the input space effectively to the output space, where a predictive uncertainty can be formed from it. In a similar manner, Van Amersfoort *et al.* [54] proposed to combine the ResNet with an inducing point approximation to the GP and furthermore added spectral batch normalization, achieving SOTA performance at OOD detection at time of publication.

3 State of the Art

3.1 Land Cover Classification

Image classification is a pivotal task in remote sensing. In contrast to semantic segmentation, which assigns different classes on a pixel level, now entire patches of remotely sensed images get assigned one or more classes. As a special case of classification, we look at LULC, which plays a predominant role as the application area spans the global land mass. In the following, we will shortly introduce the most widely-known research datasets for this task (see [55] for a recent and extensive overview). Following that, the LCZ42 dataset, which is the core dataset used in this work and the accompanying publications, will be introduced in more detail. Last, we look at the most common ML and DL models utilized for the introduced datasets.

Popular Datasets

BigEarthNet [56] is a large-scale dataset with 19 classes covering various land cover types and scene categories. The dataset includes more than 500 thousand Sentinel-2 samples, each containing 12 spectral channels. It supports the development of deep learning models for global land cover classification and incorporates a multi-label framework of the CORINE Land Cover database [57]. The dataset has since then been re-released in a multimodal setting [58] and a refined 2.0 version [59], including a pixel-level reference map as well as optimized atmospheric correction. The EUROSAT [60] dataset contains 10 classes representing diverse land cover types. With 27 thousand Sentinel-2 samples (13 channels), EUROSAT is designed for the assessment of large-scale satellite-based land cover classification algorithms.

The RSI-CB dataset [61] includes six categories with a total of 35 sub-classes ranging across various domains such as agriculture, transportation, and construction. It consists of more than 24 thousand samples with RGB channels. As a peculiarity, the dataset has crowdsourced labels utilizing Open Street Map (OSM). Although not strictly limited to LULC, RSI-CB is still widely used as a remote sensing image classification benchmark dataset. The Aerial Image Dataset (AID) [62] consists of 10,000 images referring each to one of 30 classes that cover various land cover types and scene categories. AID is designed for the evaluation of aerial image-based land cover classification algorithms and incorporates high-resolution aerial imagery. It has been re-created on a larger scope with one million samples [63], yet the labeling was automated and therefore cannot be treated equally to AID. The US-Merced Land Use (UCM) dataset [64] includes ca. 2,100 samples corresponding to 21 classes that represent different land use types such as residential, commercial, industrial, and recreational areas. UCM is designed for the evaluation of

land use classification algorithms and incorporates high-resolution aerial imagery, yet the dataset is limited in its size.

The SSL4EO-S12 dataset [65] incorporates multitemporal and multiseasonal imagery from Sentinel-1 and -2 satellites, ensuring diverse and high-quality data. Contrary to other datasets presented, the images are unlabeled and serve as a pre-training stage for downstream LULC classification tasks. For the presented datasets, large performance increases of multiple percentage points can be achieved via pre-training the model using SSL4EO-S12. Building upon the previous dataset, SSL4EO-L [66] includes high-resolution optical and radar satellite imagery from the Landsat satellite missions. This enables the pre-training of classification models based on Landsat data, which offers a wider temporal coverage compared to the Sentinel satellite missions.

Local Climate Zones (LCZs)

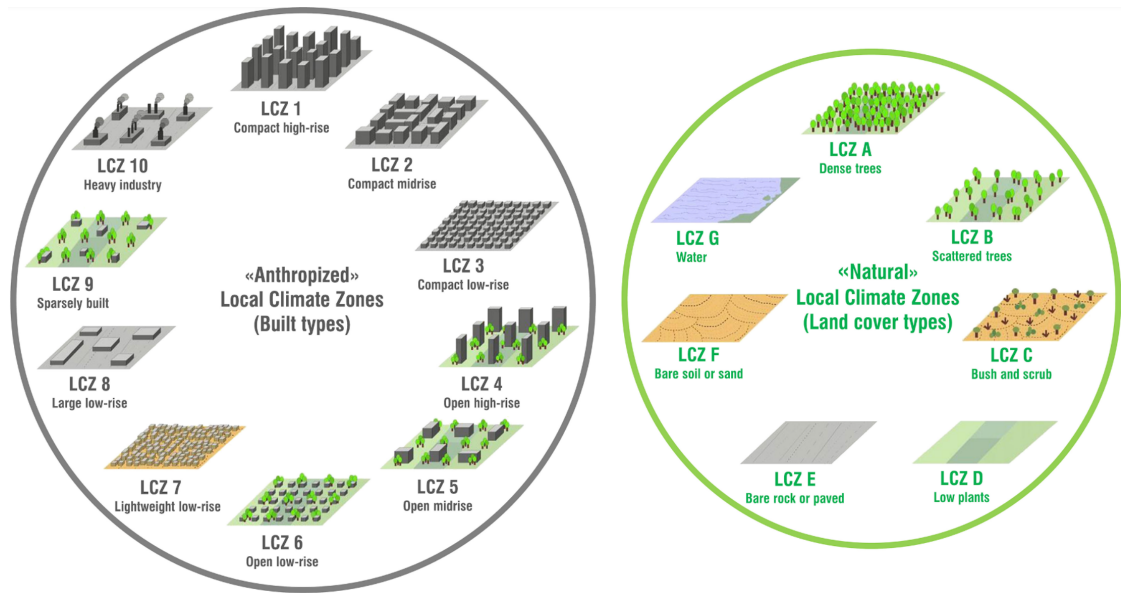


Figure 3.1: LCZ classification scheme as proposed by Stewart & Oke [67]. Figure taken from Han *et al.*[68].

The LCZ scheme is a standardized classification system developed by [69] to categorize urban and natural landscapes based on their surface structure, cover, and morphological characteristics. Designed to improve the understanding of urban microclimates and environmental conditions, the LCZ scheme classifies areas into 17 distinct classes as seen in Figure 3.1. These classes are further divided into 10 building-related and seven vegetation classes suited to describe urban conglomerate and their surroundings. This framework enables consistent comparisons across cities worldwide, supporting urban climate research [70, 71, 72], urban planning [73, 74, 75], and sustainable development efforts [76] by providing a detailed and systematic way to analyze how different urban forms influence local climate conditions.

In order to better understand the development of the global urban footprint, large-scale land cover datasets are needed, and the LCZ scheme is particularly useful for this research question. The So2Sat LCZ42 [77] dataset is an extensive global remote sensing dataset designed for urban land use classification. It was created as part of the So2Sat project [78], which aims to leverage large-scale satellite data to enhance understanding of urban morphology on a global scale. One of the key features of this dataset is its focus on the LCZ classification scheme, which was introduced earlier. The So2Sat LCZ42 dataset comprises multimodal satellite imagery, including Sentinel-1 Synthetic Aperture Radar (SAR) data and Sentinel-2 multispectral optical imagery. This dual modality allows for comprehensive analysis, as SAR can penetrate cloud cover and provide structural information, while optical imagery offers detailed spectral variability valuable for land cover classification.

The publicly available global dataset [79] focuses on 42 urban regions worldwide, covering diverse climates and urban morphologies. The images are provided at 10 to 20 meters GSD (i.e. distance between two adjacent pixels). Each image tile is annotated with one of the 17 LCZ classes and spans a region of 320×320 meters (or 32×32 pixels). Since each tile corresponds to exactly one LCZ, the dataset does not cover the regions entirely but instead focuses on homogeneous areas of sufficient size. Labeling was performed manually, mainly influenced by the labeling methodology utilized by the World Urban Database (WUDAPT) project [80]. The public dataset contains three different splits of the data into training, validation, and testing according to different splitting regimes: Next to an entirely random split, which ensures the independent, identically distributed (i.i.d.) assumption among the individual subsets, two geographically-influenced splits are presented. In one of the splits, each city is split by longitude. Therefore, training and testing data are never from the same part of the city. The other split, also termed *cultural split*, sets aside ten cities from different cultural zones as a whole, increasing the domain gap between training and testing data significantly.

As a peculiarity of the global So2Sat LCZ42 dataset, the authors furthermore evaluated the degree of human label uncertainty associated with the LCZs. For this, a subset of ten European cities was re-labelled independently by ten remote sensing experts. A cultural split of this label evaluation study was recently published [81], accompanying one of the publications that support this thesis. The choice of the cultural split will be discussed in more detail in the corresponding Section 4.1. By providing multiple expert annotations per satellite image, this label evaluation study opens the doors for many research questions to be posed. For some of the questions, the publications included in this thesis are aiming to find answers, while others still leave room for future work. A common starting point for the analysis of label uncertainty is a view of uncertainty on an image level. For this, we first define the majority vote as the class that received the majority of all expert votes cast. In the studied dataset, this majority vote is always uniquely defined. Figure 3.2 (a) displays the so-called *confusion matrix* between the individual expert votes and the majority vote on an image level for the So2Sat LCZ42 label evaluation dataset. A closer look reveals that most of the confusion is centered around certain combinations of classes, while a large portion of class combinations holds little to no ambiguity. Note that the values in the confusion matrix are normalized

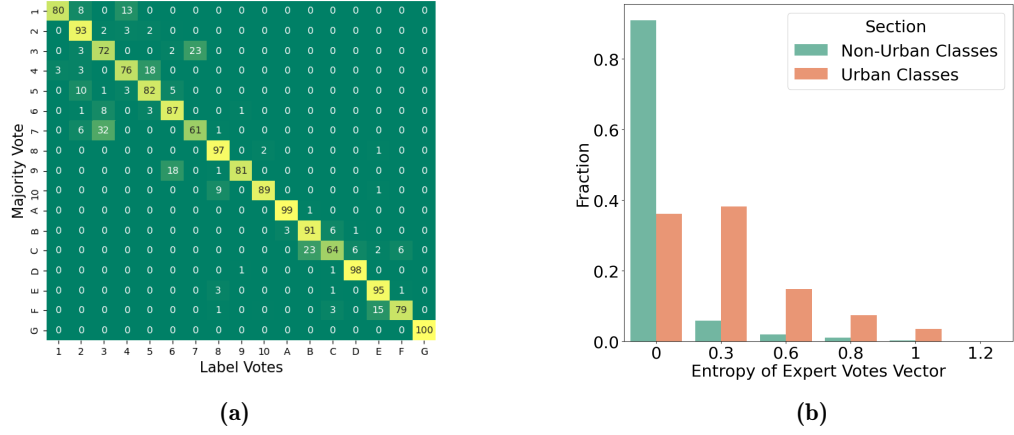


Figure 3.2: Analysis of label uncertainty within the So2Sat LCZ42 Label Evaluation Dataset via (a) confusion matrix between the label votes and the majority vote and (b) the distribution of the Shannon entropy. Own analysis, published in [1].

to one with respect to the majority vote (hence the row sums add up to one). For a more fine-grained quantification of label uncertainty, the Shannon entropy [82] can be utilized as an information-theoretic measure of information. Computed on the empirical distribution of label votes per image, the numbers are displayed in Figure 3.2 (b). The barplot is grouped by the hierarchical split of LCZs into the urban (1-10) and non-urban (A-G) classes, showing a much larger share of samples with higher entropy (hence label uncertainty) for the urban classes. Although this was already partially visible in the confusion matrix, the classwise normalisation did not take into account the class imbalance, which will be more thoroughly discussed in Section 4.1.

Modelling Approaches for LCZ Classification

Before the existence of the large-scale benchmark dataset So2Sat LCZ42, the LCZs were already studied in an automated way by Bechtel *et al.* [83]. The authors utilized several ML models, such as a Support Vector Machine and a random forest, as well as a simple neural network to classify fused satellite imagery from multiple sources into LCZs on a pixel level. Using LANDSAT-8 data, the significant impact of neighboring pixels on LCZ classification was found by Verdonck *et al.* [84]. Although limited to three Dutch cities in their data pool, the authors provide a fine-grained analysis of pixel-level LCZ maps using a random forest classifier and discuss the influence of neighborhood on the results in detail. In a similar work, Yoo *et al.* [85] investigate various specifications of CNNs and random forests in great detail for a set of four megacities, again utilizing LANDSAT-8 imagery. The most promising technique here was found to be a CNN with a large neighborhood (all neighboring pixels in a 3×3 grid with the current pixel in the center), particularly regarding the transferability of models trained on one city and evaluated on another.

Since the availability of high-resolution (10-20m GSD), multispectral and multitemporal Sentinel-1 and -2 data, the analysis of ways to enhance or fine-tune existing LCZ classification methods has intensified. In an early work, Hu *et al.* [86] utilized polarimetric synthetic aperture radar data from the Sentinel-1 mission to extract polarimetric, textural, morphological, and statistical features to feed them into a LCZ classifier. Although polarimetric features by themselves cannot adequately classify images into LCZs, they contribute to the overall best model in combination with morphological features. By combining Sentinel-2 with OSM and Global Urban Footprint (GUF) data, another work [87] studied the effect of training set size and class imbalance on the LCZ classification performance. A key finding was the importance of balancing out the class frequencies instead of using a large-scale, imbalanced dataset right away. Qiu *et al.* [88] investigated a ResNet [18] in combination with multi-seasonal Sentinel-2 data, finding that majority-voting on multi-seasonal predictions was superior over averaging the single-seasonal predictions.

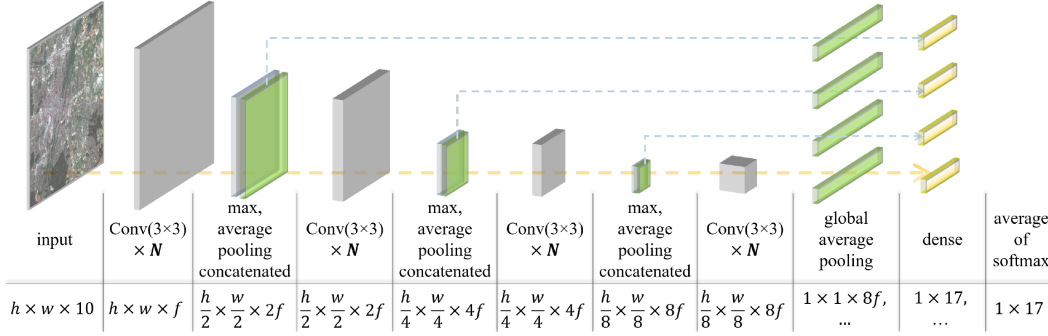


Figure 3.3: The Sen2LCZ architecture as presented by Qiu *et al.* [89].

In a large-scale benchmark study [89], Qiu *et al.* could furthermore present a task-specific network architecture, named Sen2LCZ, which was superior in LCZ classification over many popular NN architectures such as the earlier discussed ResNet (in various depth specifications), the VGG network, DenseNet, and Xception. This was possible with a fraction of the parameters needed, a clever combination of multi-level feature fusion, and a balanced depth and width of the network. The architecture is displayed in Figure 3.3, and furthermore is utilized in the majority of papers accompanying this dissertation. Since then, works on data fusion [90, 91], multi-level attention networks [92], deep transfer learning [93], and utilizing prior knowledge [94], among many others, have been published, see [95] for a recent and extensive review. As a huge step towards a global LCZ map, Zhu *et al.* [96] mapped every global urban conglomerate with more than 300 thousand inhabitants with the LCZ scheme. Both Sentinel-1 and Sentinel-2 data were fused together, and an in-depth classwise analysis as well as a thorough visual inspection are presented.

3.2 Uncertainty Quantification in Classification

Although the characterization of an uncertainty quantity can be an abstract concept, what unites most concepts is a clearly defined downstream task. This task helps to evaluate the quality of the computed uncertainty and allows for benchmarking different UQ methods where applicable. In the following, four separate frameworks for measuring the quality of a predicted uncertainty are presented. Furthermore, we start by introducing the concept of label uncertainty.

Label Uncertainty Quantification

In the context of classification tasks, an often neglected aspect of uncertainty lies within the labels associated with each data instance. Oftentimes, the presence of label uncertainty is confused with label noise [97]. Label noise, either artificially injected or naturally occurring, is equivalent to erroneous labels, which hinders the classification performance of any classifier to a varying degree. Surprisingly, DL models have shown a considerable level of robustness to label noise [98] on several image classification benchmarks. Since its first proper mention, label noise has been studied extensively (see [99] for an early extensive survey and [100] for a survey covering the DL domain specifically). Topics range from discussing various sources of label noise as well as techniques to overcome label noise, ranging from label noise cleaning to noise-tolerant algorithms.

In contrast to noisy labels stands the concept of label uncertainty, referring to situations where the true label of a data instance is inherently ambiguous or uncertain. This ambiguity can stem from multiple sources, including ambiguous or overlapping definitions of the individual classes, limited data quality, inherent ambiguity, and particularly the process of human annotation. The latter, for example, is often present in the medical domain [101], where human judgment is omnipresent but highly ambiguous. A two-level scheme for quantifying label uncertainty was first proposed by Wimmer *et al.* [102]. Label uncertainty is furthermore clearly distinguishable from multi-label frameworks, as it assumes different degrees of uncertainty for different classes. In contrast, most commonly in multi-label classification, multiple labels simply exist in parallel (e.g. because they both appear in an image).

For a long time, there has not been a unifying framework to deal with label uncertainty concretely in classification tasks (aside from treating the uncertainty as label noise or resorting to multi-label classification). This changed with the conceptualization of Label Distribution Learning (LDL) [103]: A distribution over the labels here is formed, which assigns notions of certainty to each class. The distribution quantifies to what degree the data instance is described by the respective class, and is normalized to one as such. Overall, LDL provides a more flexible framework for modeling ambiguous labels that takes into account the degree of confidence for each class. Since the appearance of the label changed drastically in its nature, many new learning regimes for LDL have since been proposed. These include, but are not limited to, Deep LDL with CNNs using the KL divergence [104], the additional inclusion of intrinsic label correlations [105], hierarchical LDL [106], manifold learning [107], and LDL with imbalanced data [108].

Despite first approaches towards quantifying label uncertainty, the community lacked a large-scale benchmark dataset until quite recently. With CIFAR-10H [109], a first benchmark dataset was released. It consists of the official CIFAR10 [110] test dataset labelled by ca. 50 independent annotators. An interesting ablation study was performed by Collins *et al.* [111]. The authors released CIFAR-10S, where every annotator was furthermore allowed to express their uncertainty individually. As the popularity of uncertainty-affixed datasets grew, a large-scale benchmark study [112] focusing on data-centric ML was welcomed.

In RS, label uncertainty has started to be acknowledged, but still plays a minor role. Hechinger *et al.* [113] modelled LCZ classes assuming a single ground truth via a multinomial mixture model. Using the individual expert label votes as input, the model parameters were estimated via a stochastic version of the Expectation-Maximization (EM) algorithm [114]. In a follow-up study [115], label embeddings were learned with the aid of a Bayesian framework. These embeddings serve as the basis for an accompanying work and will be explained in more detail in Section 4.2. Tsutsumida and Kato [116] built a robust land cover map via multi-annotator Google Earth Engine (GEE) data fed to a K-Medoids classifier, which grouped reliable maps. Then, a dynamic classifier selection was utilized in order to receive a unified map. Few-shot classification of Sentinel-2 data was performed by Russwurm *et al.* [117], indicating that humans show limited success against a model-agnostic meta learner. Hence, multiple annotators are not always better, in particular if the ground truth is ambiguous. Wang *et al.* [118] employed ambiguous labels in unsupervised SAR image segmentation, while Garcia *et al.* [119] implemented a Human-in-the-loop for land cover segmentation. Neighbor-Based LDL was used by Li *et al.* [120] to model label ambiguity for aerial scene classification.

Calibration

As the first downstream application of UQ, the concept of calibration is presented. This is opposed to calibrating a sensor, which is a crucial step in the generation and processing pipeline of remotely sensed data. Here, calibration is equivalent to ensuring that the sensor’s outputs reflect the true underlying physical quantities it is intended to measure. Calibrating a NN, on the other hand, refers to essentially balancing the model’s confidence in its prediction and the resulting accuracy. This balancing is desirable as most NNs are naturally overconfident in their predictions [121]. In the ML community, many different streams of calibration techniques have been established recently. Generally, a hold-out calibration dataset is required for the task. Among the most common techniques for calibrating a classifier are Platt scaling [122], Temperature Scaling (TS) [123], and Label Smoothing (LS) [124]. Whereas Platt scaling utilizes logistic regression to calibrate the predictions, TS and LS directly modify the network logits and labels, respectively. While LS was observed to boost calibration at the cost of knowledge distillation from teacher to student [125], other works found it to boost distillation when applied on the teacher side in the presence of label noise [126]. Despite the success of the computationally lightweight TS over other computationally heavy calibration methods [121], an attended version was presented by Mozafari *et al.* [127], specifically serving the cases of highly accurate and

highly inaccurate models. Furthermore, a masked version [128] was presented to overcome the limitations of distilled data, as well as an extension to dropout VI [129] in order to properly account for calibration of the model uncertainty specifically.

Originating from the work of DeGroot *et al.* [130], a common way of visualizing a classifier’s calibration performance is a reliability diagram [131]. The diagram displays the model’s accuracy versus the confidence (typically defined via the highest softmax pseudo-probability) in a bin-wise manner. The bins can be chosen to be either equally sized or in an adaptive way [132]. While the bin size does affect the calibration performance, tuning the bin size is undesirable as it does not reflect the variance of the calibration metric. Hence, ideally multiple bin sizes are evaluated. The calculated mismatch between the two quantities is termed the Expected Calibration Error (ECE), where the expectation arises from both the chosen bin size and the size of the calibration dataset. Since the original work, adaptations towards multidimensional reliability diagrams [133] and class-wise calibration errors [134] have been proposed. On a more general note, Gruber and Buettner [135] introduced proper calibration errors, relating the original errors to proper scores with respective upper bounds and optimal estimation properties.

Despite increased visibility in the RS community, not many works have specifically targeted calibration. Patel *et al.* [136] utilized LS for radar perception and could show that calibrated UQ was partially effective for targeting corrupted input signals, domain shift, and the presence of unknown objects. Using an ECE-based deep ensemble, Conde *et al.* [137] improved on existing ensemble benchmarks regarding accuracy and calibration performance evaluated on the EUROSAT, AID, and RESISC45 datasets earlier introduced. A calibrated version of the focal loss was used by Bai *et al.* [138] for the semantic labeling of RS imagery with NNs. Performance improvements were hereby observed not just in the classical metrics but also with regard to the ECE and visually by means of predictive uncertainty maps. When calibrating DL models for camera trap image sequences, Dussert *et al.* [139] relied on TS and achieved significant performance boosts; similarly, this was observed for crop rotation modeling by Quinton and Landrieu [140]. Using the Laplace approximation, the calibration of multi-label classifiers was successfully targeted by Rewicki and Gawlikowski [141]. In order to make DL models suitable for edge devices, Huang, Park *et al.* [142] distilled the calibration task with the aid of credal sets and conformal prediction.

Out-of-Distribution Detection

Instead of matching the model’s uncertainty with the underlying accuracy as before, in the task of OOD detection, the uncertainty is primarily seen as a tool to express the model’s capability to know where it doesn’t know. In classification, this corresponds to an input that the model cannot classify into one of its known classes. Hendrycks and Gimpel [143] first formalized the task of evaluating OOD detection. They utilized the Maximum Softmax Probability (MSP) of a NN classifier to let the model decide whether a new image is from ID, meaning from a distribution of images the model has already seen during training, or from OOD, which can be characterized by a completely different dataset or by a shifted version of the training dataset. For evaluation purposes, the

authors utilized the Area under Curve (AUC) metrics by treating the OOD detection task as a binary classifier. Hence, the better the model can express its uncertainty towards OOD data via a lower MSP, the higher the AUC metrics.

Although the task of *generalized* OOD detection can be further categorized into sub-tasks [144], the focus of classical OOD detection lies on uncovering samples from a different data regime. In particular, this means that the OOD data has a different set of labels. The model cannot correctly classify the OOD data, but rather only express its uncertainty in the prediction. For the construction of such OOD-aware NNs, different strategies have been proposed. Malinin and Gales proposed Prior Networks [45], which place a Dirichlet distribution onto the label space and equip OOD samples at training time with a flat Dirichlet distribution, hence maximizing its label uncertainty. In a similar manner, Sensoy *et al.* deployed evidential neural networks [47] utilizing subjective logic theory. By doing so, the predictive uncertainty can be directly calculated from the total evidence of the prediction. An adaptation of this method utilizing VI and a mixture of Dirichlet distributions was proposed recently by Ryu *et al.* [145]. Similarly, Joo *et al.* [146] utilized the Dirichlet space in conjunction with the Expectation Lower Bound (ELBO) criterion. Furthermore, Liu *et al.* [147] proposed energy-based models as a powerful class of models, while Postels *et al.* [148] estimated densities in the latent space of the NN. Charpentier *et al.* proposed posterior networks, which do not require OOD data at training time and rely on normalizing flows for a closed-form posterior distribution of the network. Noteworthy is the stream of deterministic UQ methods [53, 52, 54], requiring only a single forward pass for UQ at test time. These methods will be discussed in more detail in Section 5.1.

As opposed to the ML community, RS offers OOD detection frameworks that are closer to the real-world scenario. This phenomenon opens the door for a broader application spectrum, but at the same time poses more challenges to overcome. Gawlikowski *et al.* [149] utilized the earlier mentioned Prior Networks and adapted them to a set of self-defined OOD detection scenarios (split on class level) based on various RS datasets. Energy-based OOD detection was performed by Castillo-Navarro *et al.* [150], where an entirely different RS dataset was used for OOD data. Rolland *et al.* [151] adapted a graph-based NN to be uncertainty-aware for tackling OOD detection with multimodal RS data. For OOD detection based on scene classification and target recognition, Inkawhich *et al.* [152] proposed two adaptations to existing NN architectures: By including a risk minimization framework to the training procedure and simultaneously performing calibration and label correction for the OOD detection part, significant performance increases could be achieved. In addition to the earlier noted calibration experiments, Gawlikowski and Rewicki [141] also performed OOD detection and achieved significantly better results with the Laplace-approximated Bayesian Neural Network (BNN) over the baseline model. In order to enable OOD detection at scale for RS data, Ekim *et al.* [153] introduced a post-hoc method involving the generation of surrogate labels.

Active Learning

Similarly to before, in the task of AL, predictive uncertainties are derived to identify data samples that the model struggles to classify. The major difference lies in the nature of the identified samples: Whereas in OOD detection, identified samples should ideally originate from a shifted or completely separate dataset, AL iteratively adds the identified samples to the initial training dataset. The samples are labelled when added to the dataset, requiring human labelling effort, but at the same time reducing the need for a large initially labelled training dataset. The promise of AL lies in the assumption that samples with a high predictive uncertainty help the model to adapt faster to a training data distribution and, by doing so, simultaneously generalize better to an unseen test dataset. Initially formalized by Cohn *et al.* [154], Schohn and Cohn [155] first introduced the technique for Support Vector Machine (SVM)s. AL has been an active field of research ever since, see [156] for an up-to-date survey.

The process of selecting the samples that get added to the training pool (e.g. via an uncertainty measure) is termed *querying*. Over the years, many sophisticated querying strategies have been established. Nguyen *et al.* [157] discuss many of them in a recent large-scale benchmark study. Out of the variety of querying techniques, noteworthy examples include the Bayesian approach by Houlby *et al.* [158], which is based on the mutual information between new points and the true data distribution. This approach has been refined many times, e.g. via computing the mutual information more efficiently in a batch-wise manner [159] or via introducing a stochastic component to the acquisition step [160]. Besides the Bayesian approaches, also density-based methods [161] and methods based on reinforcement learning [162] have been established, among many others.

AL for remotely sensed data has a long and rich history. After its first successes in the CV community, Mitra *et al.* [163] quickly adopted the SVM approach to multispectral satellite data. Rajan *et al.* [164] successfully established an AL framework to hyperspectral data, which furthermore takes into account changes in the spectral signatures as well as spatial and temporal dependencies. Since then, AL for hyperspectral data has received much attention in the research community [165, 166, 167]. While Tuia *et al.* [168] proposes two novel AL querying strategies for RS image classification, Demir *et al.* [169] utilized batch-mode AL for SVMs, focusing on diversity- and uncertainty-based acquisition functions used for querying.

Conformal Prediction

Differing fundamentally from the previously discussed approaches, CP [170] opens a new point of view on UQ. Instead of a single point or class prediction, CP transforms the network’s output to a prediction interval or set for regression or classification problems, respectively. The ground truth value or class is then contained in the prediction with a pre-specified probability. Two major streams of CP algorithms have been established for NNs: Transductive CP [171] repeats the calibration process for each new data point individually, which is computationally expensive and hence only suited for smaller datasets.

On the other hand, inductive CP [172] requires a hold-out calibration dataset, but is much more scalable for big datasets. The conformalization is then carried out on the hold-out dataset, which needs to be *exchangeable* [173] with the training and testing dataset. This characteristic is similar to the i.i.d. assumption one typically makes in DL, and there exists a direct link between the two [174].

Focussing on classification, the main notion of uncertainty coming from a conformalized model is the size of the prediction set. However, many more measures have been introduced by Vovk *et al.* [175]. In the recent past, CP experienced a second advent with the rise of DL [176]. Sadinle *et al.* [177] introduced Least Ambiguous Set-Valued Classifier (LAC), presenting an adoption of previous inductive CP algorithms with optimized asymptotic and finite sample properties. By introducing Adaptive Prediction Sets (APS) [178], Romano *et al.* set a new benchmark for a CP method that yields minimal prediction set sizes while upholding the theoretical coverage guarantees. An even more optimized prediction set size was achieved via Regularized Adaptive Prediction Sets (RAPS) [179], for which Angelopoulos *et al.* introduced an additional regularization term for the conformity score of APS. Recent advances also utilized CP in combination with TS [180], early stopping mechanisms [181] or Monte-Carlo (MC) approximations [182]. Next to CP beyond exchangeability, also directly optimized prediction intervals [183] were explored.

CP offers a massive benefit for RS applications, where often labels or target values are ambiguous in their nature. Prediction intervals or sets therefore yield better and more explainable insights into a model’s predictive uncertainty. Since the recent successes of CP in CV, many application fields have been presented. Valle *et al.* [184] utilized conformal statistics for a LULC classification framework and analysed prediction set sizes on a class level, potentially helping the downstream task of labelling additional data. Targeting an onboard OOD detection framework, Bhattacharjee and Jung [185] applied the APS algorithms to detect different noise types injected into the EUROSAT dataset. While Harris and Silver [186] applied an ensemble of conformalized models to the task of climate projection, a spatial-aware CP framework for hyperspectral image classification was presented by Liu *et al.* [187]. A broad range of CP applications, including canopy height regression and species classification, was presented by Singh *et al.* [188], while further evaluated downstream tasks include seismic response prediction [189], soil organic carbon estimation [190], birds detection [191], harvest-readiness classification of cauliflower [192], and tropical cyclone track forecasting [193].

4 Incorporating Label Uncertainty

The first content-related part of this dissertation aims to explore ways of including a notion of label uncertainty or ambiguity into the training pipeline of a NN classifier. For this, we investigate the earlier introduced *So2SAT LCZ42* dataset. In general, label uncertainty is an often neglected aspect in DL, and RS applications are no exception. The uncertainty must not be confused with label noise, which is undesirable and should be identified and minimized. For the present data, the ambiguity instead stems from the nature of the labelling process: Ten remote sensing experts, who were trained specifically on the LCZ labelling scheme, independently labelled all present data. While previous works already investigated the overall level of labelling confidence and confusion [77], or focused on a label-based analysis on expert or class level [113], the aim here is twofold:

- We establish novel ways of incorporating label uncertainty into the training of a NN classifier, both by expanding the label space to the Dirichlet space and by utilizing various loss functions
- For the uncertainty-aware classifiers, we evaluate their predictive performance and their ability to express predictive uncertainty by means of various downstream tasks

In the following, we denote vectors in bold and put the running data index in superscript, whereas the subscript is reserved for other indices. We therefore denote the vote count vectors for the images $i = 1, \dots, n$ via $\mathbf{Y} = \mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}$, where $\mathbf{Y}^{(i)} = (Y_1^{(i)}, \dots, Y_K^{(i)})$ counts the votes for each of the $K = 17$ local climate zone classes for image i . By counting the remote sensing experts via $j = 1, \dots, J$, for image i we receive the expert votes via

$$V_1^{(i)}, \dots, V_J^{(i)}, V_j^{(i)} \in \{1, \dots, K\} \quad \forall i = 1, \dots, n. \quad (4.1)$$

We denote the votes as a K -dimensional vector as follows:

$$\mathbf{V}_j^{(i)} = (\mathbb{1}_{\{V_j^{(i)}=1\}}, \dots, \mathbb{1}_{\{V_j^{(i)}=K\}}), \text{ where } \mathbb{1}_{\{V_j^{(i)}=k\}} = 1 \Leftrightarrow V_j^{(i)} = k \quad (4.2)$$

From this formulation, the vote counts are then simply obtained via $Y_k^{(i)} = \sum_j \mathbb{1}_{\{V_j^{(i)}=k\}}$.

Specifically, $Y_k^{(i)} = m$ equates to image i receiving m votes for class k . By letting $M = 10$ represent the number of votes or experts (identical here, since every expert casts exactly one vote), we can deduce that $\sum_{k=1}^K Y_k^{(i)} = M \quad \forall i = 1, \dots, n$. Lastly, let $Y_{max}^{(i)} := \max_j Y_j^{(i)}$ denote the majority voted label, and define the corresponding *one-hot* encoded label via $\mathbf{y}_{max}^{(i)} = (\mathbb{1}_{\{Y_1^{(i)}=Y_{max}^{(i)}\}}, \dots, \mathbb{1}_{\{Y_K^{(i)}=Y_{max}^{(i)}\}})$.

4.1 Distributional Labels Approach for LCZ Classification

In a nutshell, the following approach targets the direct inclusion of the label ambiguity among the experts into the training pipeline. By assigning equal importance to all experts, and by staying in the categorical label space as usual in classification, we form a soft or distributional label. Therefore, instead of taking a *one-hot* label, e.g. by considering the majority vote, we utilize the vector of normalized vote counts:

$$\mathbf{y}_{\text{distr}}^{(i)} = \mathbf{Y}^{(i)} / M \quad (4.3)$$

The version of the dataset used in this work¹ comprises ten European cities plus a total of eight add-on areas needed to provide samples for the class 7 (*lightweight-lowrise*), which does not naturally occur in European cities. For the listing of the considered cities and the geographically separated train-validation-test split, we refer to Tables I and II in the corresponding publication [1]. Note that the imbalance between the individual classes, as can be seen in Table III, is quite severe. The degree of label uncertainty present in the dataset was already illustrated in Figure 3.2, showing notable differences between urban and vegetation classes. Due to this phenomenon, the experiments were solely performed using the samples from the urban classes.

Publication linked to this Section

C. Koller, G. Kauermann, and X. X. Zhu. Going Beyond One-Hot Encoding in Classification: Can Human Uncertainty Improve Model Performance in Earth Observation?. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-11, 2024. doi: 10.1109/TGRS.2023.3336357

Next to the utilization of soft labels, a slight shift with regard to the loss function was performed. Instead of the typical Cross Entropy (CE) loss function, the discrete version of the KL divergence was used. For a batch of data $\{\mathbf{x}^{(i)}, \mathbf{y}_{\text{distr}}^{(i)}\}_{i=1,\dots,m}$, and by denoting the predictive softmax distribution of the network $f_{\theta}(x)$ as $p_{\theta}(y|x)$, the loss is given via

$$\mathcal{L}_{KL}(f_{\theta}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}, \mathbf{y}_{\text{distr}}^{(1)}, \dots, \mathbf{y}_{\text{distr}}^{(m)}) = -\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K y_{\text{distr},k}^{(i)} \cdot \log \frac{y_{\text{distr},k}^{(i)}}{p_{\theta}(\mathbf{y}^{(i)} = k | \mathbf{x}^{(i)})}.$$

To account for zero-entries in the votes leading to $\log(0)$, we set

$$y_{\text{distr},k}^{(i)} \cdot \log \frac{y_{\text{distr},k}^{(i)}}{p_{\theta}(\mathbf{y}^{(i)} = k | \mathbf{x}^{(i)})} \equiv 0 \text{ for } y_{\text{distr},k}^{(i)} = 0.$$

This formulation measures the loss of information when approximating the assumed ground truth distribution with the network. Hence, the goal effectively lies in finding a network that mimics the expert voting behaviour with its predictive distribution.

¹So2Sat LCZ42 Evaluation Data Set (Link to Download: <https://mediatum.ub.tum.de/1659039>)

4.1 Distributional Labels Approach for LCZ Classification

What remains now is to evaluate the newly formulated labels and loss. A generalisation and calibration testbed was chosen for this, as the promise of the model adaptations lies in a more understandable notion of predictive uncertainty. This uncertainty is directly integrated into the calculation of various calibration errors. Similarly to an earlier analysis of human label uncertainty [109], we chose the CE as a measure for generalisation performance. For more information about the individual quantities, we refer to Section D in the corresponding publication [1]. Besides the distributional training earlier introduced, the experiments furthermore included three off-the-shelf calibration methods into the benchmark analysis, namely Label Smoothing (LS) [124], Temperature Scaling (TS) [123], and Monte-Carlo (MC) Dropout [42]. The condensed results are presented in Table 4.1 and show a substantial impact of distributional learning on both generalisation and calibration. For the complete analysis, see Table IV in the complementary paper. All results are stated as the mean $\pm$ the standard deviation (std.) over five different training runs initialized with varying seeds.

	CE One-hot $\downarrow$	CE Distr. $\downarrow$	ECE $\downarrow$	SCE $\downarrow$
One-hot	1.12 ± 0.05	1.38 ± 0.07	9.79 ± 3.18	1.03 ± 0.50
+ LS	1.05 ± 0.01	1.23 ± 0.03	7.33 ± 2.62	1.11 ± 0.25
+ TS	1.00 ± 0.13	1.17 ± 0.07	4.15 ± 2.37	1.44 ± 0.22
+ LS & TS	1.02 ± 0.03	1.18 ± 0.02	3.21 ± 0.96	1.32 ± 0.03
+ MC-Drop	1.12 ± 0.05	1.37 ± 0.06	9.57 ± 3.11	1.04 ± 0.49
+ LS & MC-Drop	1.05 ± 0.01	1.23 ± 0.03	7.11 ± 2.41	1.12 ± 0.24
Distr.	1.06 ± 0.07	1.21 ± 0.07	5.80 ± 1.07	1.21 ± 0.20
+ LS	0.98 ± 0.03	1.08 ± 0.02	8.31 ± 2.46	1.73 ± 0.06
+ TS	0.96 ± 0.09	1.07 ± 0.07	5.89 ± 2.50	1.72 ± 0.15
+ LS & TS	0.95 ± 0.04	1.05 ± 0.04	4.21 ± 1.38	1.58 ± 0.10

Table 4.1: Generalisation and Calibration Test Performance in mean $\pm$ std..

In terms of generalisation, the distributional, uncertainty-aware learning regime clearly outperforms the classical setting. This holds true both for the generalisation towards the majority-voted label and the soft label. The performance gap lies at ≈ 0.07 or 7% in relative terms for the CE to the one-hot encoded label and at ≈ 0.13 or 11% for the CE to the distributional label. It can furthermore be noted that while both LS and TS effectively improve generalisation performance, LS stabilizes the results, showing little to no standard deviation (std.). The latter produces more unstable results, as indicated by the higher std.. MC Dropout seems to have little to no impact on the generalisation performance, which comes as a bit of a surprise. Still, it can potentially be explained by the limited number of dropout layers in the base network.

The results for the calibration performance are a bit more mixed. Note first of all, though, that when ignoring all additional calibration techniques, the sole introduction of distributional labels cuts the Expected Calibration Error (ECE) almost in half and additionally reduces the variation across runs significantly. This becomes apparent in

4 Incorporating Label Uncertainty

Figure 4.1, which displays calibration performance for a single model both for one-hot and distributional label encoding. The visualization is termed as *reliability diagram*, and shows the model’s confidence (i.e. the highest predicted softmax probability) versus the underlying accuracy, binning the results in an equidistant manner (different bin sizes were evaluated and the most consistent one across the seeds was reported). For the one-hot encoded setting, we inspect the typical *overconfidence* of NNs, and CNNs in particular, as was reported previously by Guo *et al.* [121]. This translates to overly high model confidences, which are not adequately reflected in the model’s accuracy. After switching to uncertainty-aware training, the gaps between confidence and accuracy diminish, resulting in a much smaller ECE. In conclusion, the implementation of soft

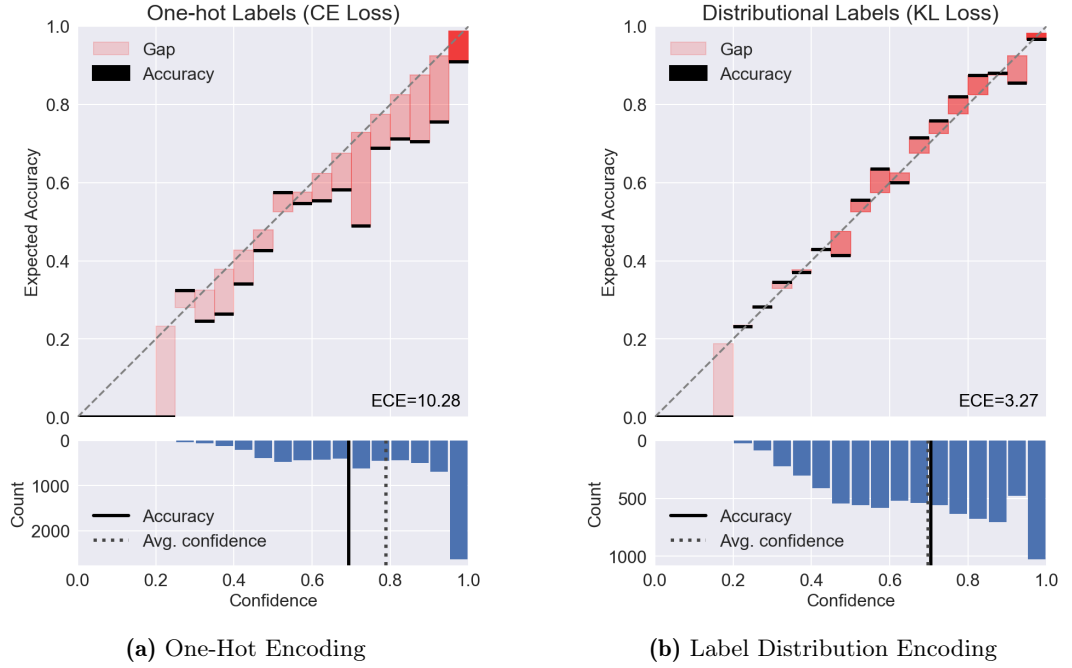


Figure 4.1: Calibration performance at test time (single seed).

labels reflecting the ambiguity of the labelling process has many advantages. While the classical performance metrics such as accuracy, Cohen’s kappa, or weighted average accuracy are more or less unaffected (see Table V in the corresponding publication), the generalization performance with regards to the CE is greatly improved. Furthermore, the calibration performance is heavily impacted, reducing the ECE by almost 50% on average. Although a combination of LS and TS can bypass the distributional learning by a small margin, these results need to be interpreted with care: Since the hyperparameter of the TS method is optimized on the validation dataset, there might exist some overfitting on the test dataset. The overfitting could happen because the two datasets come from the same set of cities. This decision was taken to ensure a geographical separation between the training dataset and the remainder of the data, and due to the fact that there did not exist a suitable split of the ten European cities into separate sets of training, validation,

and testing cities. See Table III in [1] for more insights on the class imbalance between the cities and Table II for the chosen split between training and validation/testing cities.

4.2 Learning from Label Embeddings

In a follow-up study to the previous work, the impact of label uncertainty on the training of a NN classifier was again investigated. The main novelty lies in the introduction of a three-level uncertainty scheme, which can be seen in Figure 4.2. This scheme builds upon the previously formed two-level scheme from Wimmer *et al.* [102], who introduced the scheme for a binary classification problem. Here, we generalize the scheme to multi-class classification and add a third level which incorporates Bayesian reasoning on top of the second level of UQ. In particular, we make the following main contributions:

- By introducing a third level to the pyramid of UQ, we allow for the utilization of prior information about the label space
- The labels embedded via Bayesian reasoning are incorporated into the training regime with three separate modeling techniques
- In terms of calibration performance, the embedding-based approaches can glance, while for OOD detection, the second level of UQ performs best

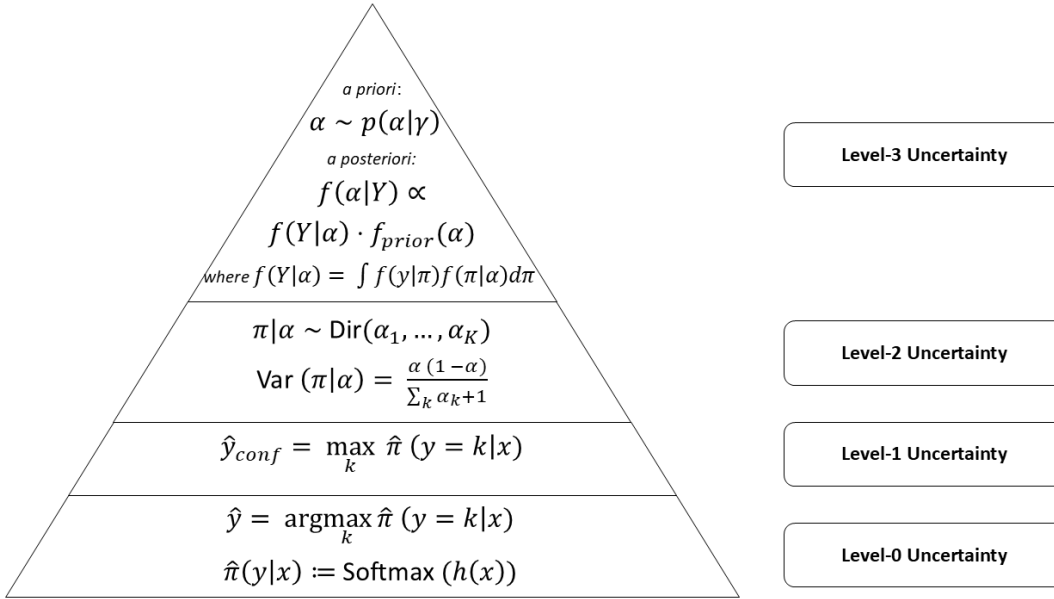


Figure 4.2: Different levels of uncertainty.

Publication linked to this Section

C. Schweden, K. Hechinger, G. Kauermann, and X. X. Zhu. Can Uncertainty Quantification Benefit From Label Embeddings? A Case Study on Local Climate Zone Classification. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-14, 2025. doi: 10.1109/TGRS.2025.3562233

We analyze the UQ pyramid from bottom to top, starting with the classifier’s predictive distribution

$$\hat{\pi}(y|x) := \text{Softmax}(h_{\theta}(x)) = \frac{\exp(h_{\theta}(x))}{\sum_k \exp(h_{\theta,k}(x))} \quad (4.4)$$

which estimates the ground truth probabilities π via the softmax transform of the *logits*, i.e. the unconstrained network output $h_{\theta}(x)$, into pseudo-probabilities. From this notion, we can deduce both the level-0 uncertainty by simply considering the highest probability as point prediction $\hat{y} = \arg \max_k \hat{\pi}_k(y|x)$ and the level-1 uncertainty via the model’s confidence $\hat{y}_{\text{conf}} = \max_k \hat{\pi}_k(y|x)$ in its prediction. The quantities of uncertainty directly based on these pseudo-probabilities are in the following denoted by Maximum Softmax Probability (MSP), which computes $\text{MSP}(\hat{\pi}) := 1 - \max_k \hat{\pi}_k(y|x)$, and the predictive entropy given via

$$H[\hat{\pi}(y|x)] = - \sum_{k=1}^K \hat{\pi}_k(y|x) \cdot \log(\hat{\pi}_k(y|x)) \quad (4.5)$$

For the second level of the UQ pyramid, we follow the approach of earlier works [45, 47, 146] and relax the categorical label space into a Dirichlet space:

$$\pi|\alpha \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad , \quad \alpha_k > 0 \forall k \quad \text{i.e.} \quad f_{\alpha}(\pi) = \frac{1}{B(\alpha)} \prod_{i=1}^K \pi_i^{\alpha_i-1} \quad (4.6)$$

Since every Dirichlet distribution equals a categorical distribution in expectation, we are now given a more flexible modeling toolkit for the label space. A variance can be computed for every Dirichlet parameter (which in expectation equals a class pseudo-probability), representing a concrete notion of uncertainty associated with the parameter value. For practical purposes, the DL model stays as is, meaning the input is passed through the network and a vector of class-wise values is predicted. As a novelty, the logits (i.e. the unnormalized class predictions) of the network are passed through the exponential function to ensure they are greater than zero, hence the model effectiely learns the Dirichlet parameters:

$$\text{Dir}_{\text{post, model}} = \text{Dir}(\exp(f_{\theta}(x)_k))_{k=1, \dots, K} =: \text{Dir}(\hat{\alpha}_{\text{model}, k})_{k=1, \dots, K} \quad (4.7)$$

They can then be directly learned by the network via the exact solution of the KL divergence between two continuous distributions. Regarding UQ measures now based on the learned Dirichlet parameters, the Dempster-Shafer Metric (DSM) [47], the expected

entropy, and the distributional uncertainty were furthermore considered. The latter two measures are thoroughly discussed by Ulmer *et al.* [48]. We refer the interested reader to subsections III C and IV C in the accompanying paper [2] for methodological and implementation details.

As a novelty, we introduce a third level of uncertainty based on Bayesian reasoning. First, we invert the distributional dependency between the observed output Y and the Dirichlet parameters α , which according to the Bayes theorem looks as follows:

$$f(\alpha|Y) \propto f(Y|\alpha) \cdot f_{\text{prior}}(\alpha) \quad (4.8)$$

This allows for assuming a prior distribution of the α parameters. Furthermore, the distributional assumption $f(Y|\alpha)$ can be expressed via the full form of the marginal posterior:

$$f(Y|\alpha) = \int f(y|\pi) f(\pi|\alpha) d\pi \quad (4.9)$$

We can now assume a multinomial distribution for $f(y|\pi)$, which allows for directly solving the above formulation analytically. This is done via integrating out the latent variable π . As a result, we can directly sample from the posterior and flexibly choose a prior. An even simpler case can be considered when choosing a distribution *conjugate* to the Dirichlet as prior, since then the posterior has a closed form. By placing a flat, uninformative prior with constant c , this allows us to directly estimate the posterior parameters $\alpha_{\text{post}} = Y_k + c$ based on the voting counts Y_k with the network. This is done by the so-termed *Simple Dirichlet* method in the following. More details about this can be found in subsection III C of the corresponding publication [2].

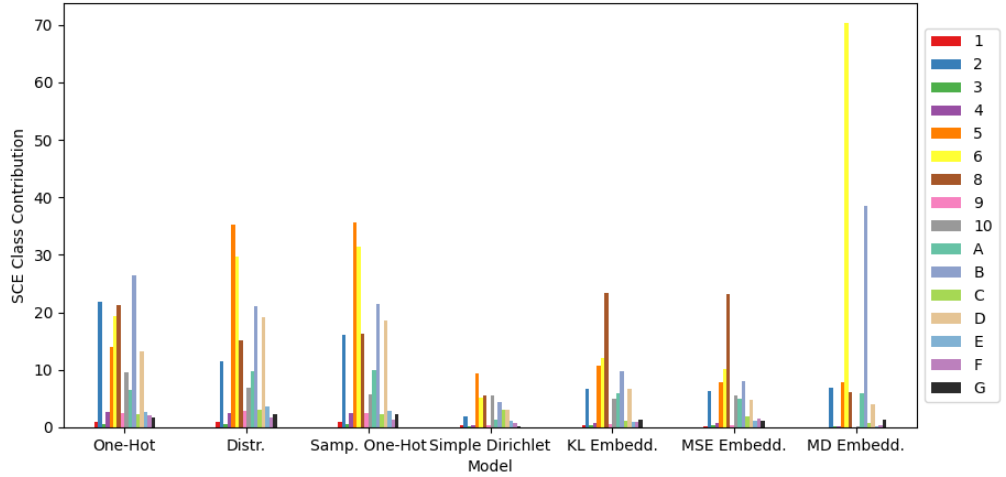
As an addition, this distributional framework allows for the estimation of unconstrained label *embeddings*, which are even more flexible than the earlier estimated Dirichlet parameters α . For this, we simply transform them via $\mathbf{Z} := \log(\boldsymbol{\alpha})$. Now, an uninformative prior such as a normal distribution can be chosen, and the estimation of the posterior is carried out in an *empirical Bayes* manner with the aid of a stochastic Expectation-Maximization (EM) algorithm [114] with MC approximation. For details, we refer to the paper by Hechinger *et al.* [115], which introduces and explains the approximation technique of the embeddings in great detail. Regarding training with the estimated embeddings, three loss functions are considered: Next to the KL divergence between Dirichlets (for which the embeddings are transformed via the exponential function), both the Mean Squared Error (MSE) and the Mahalanobis Distance (MD) are utilized, which incorporate the unbounded embeddings directly. For more computational insights, see Section III C in the accompanying paper [2].

The calibration results of the different levels of uncertainty frameworks on the test set are displayed in Table 4.2. Note that now, all 17 classes are considered (with the exclusion of class 7 due to convergence issues of the EM algorithm). A random split of the data into training, validation, and testing was chosen, resulting in smoother results, as can be seen from the generally low std.s across five training runs. Again, the effect of advanced uncertainty-aware training on accuracy is limited, yet there is no significant decrease over classical learning regimes. The ECE and Static Calibration Error (SCE)

	OA $\uparrow$	ECE $\downarrow$	SCE $\downarrow$	KL-Div. to Emb. $\downarrow$
One-hot	91.10 $\pm$ 0.55	5.12 $\pm$ 0.71	0.58 $\pm$ 0.04	0.39 $\pm$ 0.01
Distr.	91.20 $\pm$ 0.34	6.54 $\pm$ 0.41	0.64 $\pm$ 0.01	0.31 $\pm$ 0.01
Simple Dirichlet	89.46 $\pm$ 0.21	4.34 $\pm$ 0.30	0.17 $\pm$ 0.01	0.48 $\pm$ 0.01
KL Embedd.	89.95 $\pm$ 0.28	1.89 $\pm$ 0.13	0.33 $\pm$ 0.00	0.27 $\pm$ 0.01
MSE Embedd.	90.53 $\pm$ 0.12	2.54 $\pm$ 0.17	0.30 $\pm$ 0.01	0.28 $\pm$ 0.01
MD Embedd.	73.67 $\pm$ 0.36	10.09 $\pm$ 0.78	0.59 $\pm$ 0.03	0.82 $\pm$ 0.02

Table 4.2: Test Performance for embedding approaches in mean $\pm$ std..

now severely benefit from the proposed methods, reducing the quantities by up to 70 %. Note that, surprisingly, the distributional approach earlier introduced in Section 4.1 is not reducing the calibration performance anymore. This is likely due to the large share of vegetation classes in the overall dataset with little to no label uncertainty, diminishing the observed performance gains on the urban classes in Table 4.1. To gain a more detailed look into the class-wise computed SCE, Figure 4.3 shows the individual class contributions to the overall calibration error. The plot reveals that the Simple Dirichlet model shows clear calibration superiority across all classes. At the same time, the two promising embedding-based approaches (ignoring the MD loss) only lack calibration for a handful of classes. The classical learning regimes, on the other hand, have many classes they perform poorly at.

**Figure 4.3:** Class contributions to the SCE.

As a measure of generalization, the KL divergence between the predicted pseudo-probabilities and the softmax-transformed embeddings was considered. Here, the embedding-based approaches obviously glance. Solely the MD-based approach lags behind in

all aspects, which can be partially explained by the complex correlation structure among the different classes present in the dataset. This structure is directly incorporated into the MD loss function, making it hard for the model to produce well-calibrated results and make correct predictions. Additionally, many eigenvalues < 1 were observed in the correlation matrix, leading to numerical instabilities during training when inverting the matrix for the loss calculation.

OOD detection was evaluated as another downstream task. For this, the data was split into the built (1-10) and non-built or vegetation classes (11-17 or A-G). This split has been introduced in an earlier work [149] and furthermore contains little confusion between the two sets, as can be seen in Figure 3.2. Both sets are used as ID and OOD data interchangeably in two separate scenarios. All previously introduced uncertainty metrics are considered for the different UQ frameworks where applicable. For evaluating the performance, the Area under Precision / Recall Curve (AUPR) is presented in Table 4.3, for further results see Section V B in the accompanying paper [2].

ID Data	OoD Data	Setting	AUPR $\uparrow$				
			DSM	MSP	Pred.Ent.	Exp.Ent.	Distr. Unc.
Built-up	Non-built	One-Hot		76.9	77.5		
		Distr.		77.5	73.4		
		Simple Dirichlet	74.9	78.1	74.8	74.9	74.1
		KL Embedd.	61.7	61.6	63.7	61.7	63.0
		MSE Embedd.	57.6	40.8	43.5	58.3	58.8
		MD Embedd.	49.3	56.6	51.4	47.4	50.4
		One-Hot		93.2	94.6		
		Distr.		92.7	94.8		
Non-built	Built-up	Simple Dirichlet	94.2	95.3	95.6	94.2	78.3
		KL Embedd.	95.1	91.6	91.7	95.1	81.8
		MSE Embedd.	57.2	67.5	63.4	39.1	55.5
		MD Embedd.	91.6	88.2	91.8	32.3	33.3

Table 4.3: OOD Detection Results.

It becomes apparent that the setting with vegetation classes as ID data is much easier to model than the opposite case. This is likely due to the uniqueness of the spectral signatures learned by the network: For the vegetation classes, these are characterized by much more variety across the wavelength spectrum, making it easier to distinguish the known vegetation samples from the built OOD classes at test time. Overall, the Simple Dirichlet model, representing the second level of uncertainty in the pyramid in Figure 4.2, seems to provide the best modeling framework. Especially for the vegetation as OOD setting, the embedding approaches are underperforming, potentially because they inject too much label uncertainty into the training process, leaving the network unable to predict expressive uncertainties to disentangle ID and OOD data at test time. For a more detailed discussion, we refer to Section V B in the corresponding publication [2].

5 Uncertainty Quantification in Remote Sensing

As became apparent from Section 2.3, a wide variety of UQ methods have been proposed in the literature. Although many applications to RS data have been investigated and published, a lot of pitfalls exist that hinder methods constructed and tested on benchmark CV datasets from performing well on remotely sensed data. A popular example is the Deterministic Uncertainty Quantification (DUQ) framework [52], which, among other mechanisms, employs a two-sided gradient penalty. Applying this framework to the remotely sensed data discussed here turned out to be unsuccessful, as the gradients soon diminished and the predictive uncertainty of the network became meaningless. This effect, termed *feature collapse*, is known and studied in the literature, but happened to occur quite severely for the studied data and models. In the following sections, we will therefore introduce two frameworks that have proven to work well for satellite imagery, yet we will also show their limitations and downfalls.

5.1 Distance-Aware Neural Networks

As a promising research stream, distance-aware NNs have been established [53, 54]. Being part of the density-based approaches, they target the latent output space of the network. The peculiarity lies in the additional enforcement of a sense of continuity of the network mapping. This allows the network to carry over distances in the semantic input space to the output space, where a larger distance gets associated with a higher predictive uncertainty. In particular, the earlier introduced feature collapse is directly targeted, as the distance awareness involves both an upper and a lower bound for the NN mapping. For classification, let the logit of a neural network $\text{logit}(x) = g \circ h(x)$ be the concatenation of the output mapping g , bridging the gap between the latent space and the output space, and the hidden mapping h , representing the main body of the network. *Input distance awareness* then equates to g being *distance aware* and h (almost) *distance preserving* [194]. The latter is equivalent to a bi-Lipschitz constraint on the hidden mapping h , i.e., there exist constants $0 < L_1 \leq 1 \leq L_2$ such that

$$\underbrace{L_1 \cdot d_X(x, x')}_{\text{Input distance sensitivity}} \leq \underbrace{d_X(h(x), h(x'))}_{\text{Feature space smoothness}} \leq L_2 \cdot d_X(x, x') \quad \forall x, x' \in \mathcal{X}. \quad (5.1)$$

Publication linked to this Section

C. Koller, P. Jung, and X. X. Zhu. Can Land Cover Classification Models Benefit From Distance-Aware Architectures?. *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1-5, 2024. doi: 10.1109/LGRS.2024.3375370

As a popular approach to implementing this continuity, Spectral-Normalized Neural Gaussian Process (SNGP) [53] was proposed and later refined into Deterministic Uncertainty Estimation (DUE) [54]. To get a first glimpse, Figure 5.1 shows the broad idea of the two approaches, which only differ marginally. The distance preservation of the hidden mapping is firstly ensured by the use of a Residual Network (ResNet), which is proven to fulfill the bi-Lipschitz constraint if and only if (iff) all the residual blocks are α -Lipschitz for $0 < \alpha < 1$ [195]. This continuity, in turn, is approximated by applying *spectral normalization* to every weight matrix, i.e. normalizing the weight matrix by its largest singular value (the spectral norm). In practice, the spectral norm gets approximated, which can be done efficiently via the power iteration [196]. Additionally, the batch normalization step gets spectrally normalized as well for the DUE approach. This normalization is quick to implement and depends on a single learnable scale parameter [197].

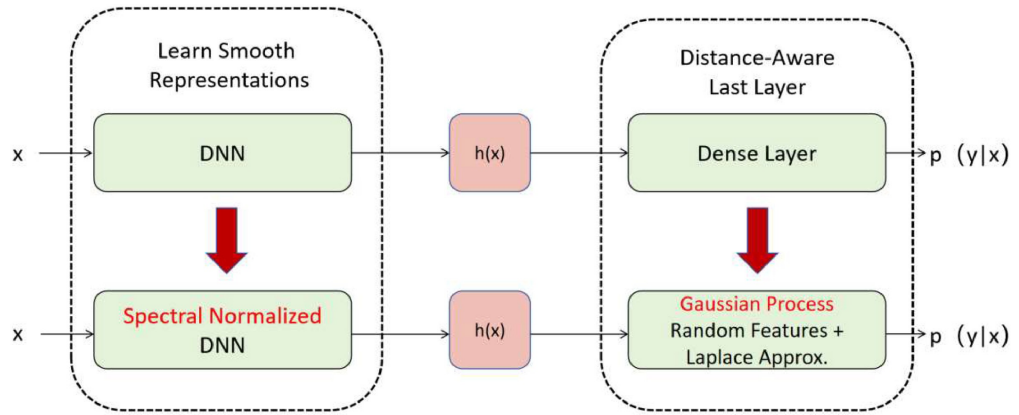


Figure 5.1: Sketch of distance awareness as presented by Linli *et al.* [198].

The distance awareness of the output mapping g is ensured via placing a Gaussian Process (GP) on top of the last hidden layer, which is equipped with an uninformative prior and a RBF kernel. While the SNGP approach utilizes a Random Fourier Feature (RFF) expansion [199] together with a Laplace approximation of the posterior, the DUE framework resorts to an inducing point approximation. We refer to Section II in [3], which is the corresponding publication to this Section, for methodological and implementational details.



Figure 5.2: Data split into ID and OOD data. Own creation, published in [200].

The relevance to RS data stems from the fact that the ResNet is a highly popular class of models for satellite image classification. Furthermore, the implemented changes to the architecture are relatively few, and the computational overhead is limited. To investigate the potential of the distance-aware approaches, two OOD detection scenarios are considered. Next to the earlier described scenario for the LCZ42 dataset (see Section 4.2), the EUROSAT dataset is studied. Here, we manually selected the classes 4 (*industrial buildings*) and 7 (*residential buildings*) to be OOD data, and the remaining eight classes as ID data. As can be seen in Figure 5.2 exemplarily, the split is valid on a semantic level. The reverse split in ID and OOD classes was not considered due to little data support. Results are depicted in Table 5.1, where again the Area under Precision / Recall Curve (AUPR) was chosen as a performance measure with respect to the Dempster-Shafer Metric (DSM), one minus the Maximum Softmax Probability (MSP), and the predictive entropy. The uncertainty metrics are identical to (some of) the previously discussed metrics, see Section 4.2 for details.

Note the abbreviations in the table referring to the addition of spectral normalization (*SN*), a GP output layer (*GP*), spectral batch normalization (*BN*), and the inducing point approximation (*IP*) to the base model. For this, we considered both the 50-layer deep ResNet as well as the WideResNet (WRN) [19], although the 50-layer deep ResNet did not perform sufficiently well. For the WRN, the seemingly best combination lies in the DUE framework, including the inducing point approximation of the GP output layer, and with spectral batch normalization enabled. The performance gaps to the baseline networks are as high as 22 p.p..

For the So2Sat LCZ42 dataset, the results are displayed in Table 5.2. Now, two split scenarios are considered, and since again the AUPR is targeted, the results are directly comparable to earlier experiments with label embeddings (Table 4.3). We see that among the close race, again the inducing points approximation works best. In the case of vegetation classes as OOD, the additional use of spectral normalisation helps, reaching much higher detection performance than with the label embedding framework previously considered. It can be furthermore noted that the overall best results are achieved with the predictive entropy as UQ measure, underscoring its importance as a measure of predictive uncertainty.

ID Data	OoD Data	Method	AUPR $\uparrow$		
			DSM	MSP	Pred.Ent.
Non-built classes	Built-up Classes	WideResNet [200]	0.32	0.59	-
		WideResNet-SN [200]	0.34	0.66	-
		WideResNet-GP [200]	0.36	0.60	-
		WideResNet-GP-IP	0.56	0.81	0.80
		WideResNet-GP-IP-BN	0.54	0.81	0.81
		WideResNet-SNGP [200]	0.35	0.57	-
		WideResNet-SNGP-IP	0.48	0.69	0.67
		WideResNet-SNGP-IP-BN	0.53	0.73	0.73
Non-built classes	Built-up Classes	ResNet50 [200]	0.55	0.55	-
		ResNet50-SN [200]	0.42	0.41	-
		ResNet50-GP [200]	0.54	0.44	-
		ResNet50-SNGP [200]	0.59	0.43	-

Table 5.1: OOD Detection results on the Eurosat dataset.

ID Data	OoD Data	Method	AUPR $\uparrow$		
			DSM	MSP	Pred.Ent.
Built-up classes	Non-built Classes	WideResNet-GP-IP	0.78	0.83	0.85
		WideResNet-GP-IP-BN	0.80	0.82	0.84
		WideResNet-SNGP-IP	0.75	0.82	0.84
		WideResNet-SNGP-IP-BN	0.79	0.85	0.87
Non-built classes	Built-up Classes	WideResNet-GP-IP	0.78	0.92	0.93
		WideResNet-GP-IP-BN	0.81	0.93	0.94
		WideResNet-SNGP-IP	0.73	0.92	0.92
		WideResNet-SNGP-IP-BN	0.79	0.92	0.93

Table 5.2: OOD Detection results on the So2sat LCZ42 dataset.

Overall, large OOD detection performance increases were observed for ResNets when equipped with distance-aware components. Supported by a strong theoretical foundation, these components seemingly help to effectively transfer input distances to the output space. The relevance to RS image classification is significant, since small changes made to existing well-performing architectures have a substantial impact on the UQ performance. Note, however, that the limitation lies in the choice of architecture, since the theory solely supports ResNets. Further, the approach has many intrinsic hyperparameters that need tuning. In particular, the number of inducing points plays a crucial role and was observed to be hyper-sensitive.

5.2 Conformal Score-Based Active Learning

As another rising field of UQ, Conformal Prediction (CP) [170] is investigated. The reason this framework was not included in the schematic overview of UQ methods (Figure 2.4) stems from its special nature: CP is fundamentally a probabilistic UQ approach because its primary contribution is a statistically valid and distribution-free probabilistic coverage guarantee on the prediction sets. The mechanism of generating these sets, once the model and calibration data are fixed, is deterministic, but this determinism is utilized to fulfill the underlying probabilistic promise. As introduced in Section 3.2, the two main streams of CP are composed of transductive [171] and inductive [172] approaches, whereas the latter is much more application-friendly in DL scenarios due to its limited computational overhead.

Publication linked to this Section

C. Schweden, P. Jung, and X. X. Zhu. Assessing Conformal Prediction for Remote Sensing: Application to Local Climate Zone Classification. *IEEE Transactions on Geoscience and Remote Sensing*. Under Review.

In the following, the inductive CP framework is therefore considered for the LCZ classification task. Let the *prediction set* be denoted by $\mathcal{C}(X)$. This set replaces the typical softmax pseudo-probabilities as the model's prediction and contains one or more classes (sometimes none, which is undesirable). The main principle of CP lies in its theoretical coverage guarantee, namely that for a prespecified $\alpha > 0$, the following holds:

$$\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \geq 1 - \alpha, \quad 0 < \alpha \ll 1. \quad (5.2)$$

This essentially means that marginally, the true class Y_{test} is contained in the prediction set at least $1 - \alpha$ % of the time on the previously unseen test set X_{test} . The inductive nature of the investigated approach stems from the way the prediction sets are formed: So-called *conformal scores* $s(x, y) \in \mathbb{R}$ are derived on a hold-out dataset solely used for the conformalization step. This step involves first deriving the threshold $\hat{t} := \frac{[(n_{\text{calib}}+1)(1-\alpha)]}{n_{\text{calib}}}$ based on the chosen sensitivity parameter α . Then, the calibration sets are formed via $\mathcal{C}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{t}\}$. With the assumed *exchangeability* assumption [173], the coverage guarantee, which is fulfilled on the calibration set by construction, carries over to the unseen test set.

The now following approaches are different adaptations to the core pipeline and only differ in their choice of the conformal score function. For the Least Ambiguous Set-Valued Classifier (LAC) [177] method, we set $s(x, y_{\text{true}}) = 1 - \hat{\pi}_{y_{\text{true}}}(x)$, namely one minus the softmax pseudo-probability of the true class. At test time, one minus the MSP is assumed as an approximation. For the Adaptive Prediction Sets (APS) [178], the score is calculated via

$$s(x, y_{\text{true}}) = \sum_{j=1}^{k=y_{\text{true}}} \hat{\pi}_{\text{ord},j}(x), \quad (5.3)$$

incorporating the ordering (descending) of the predicted softmax values. A regularized version thereof, termed Regularized Adaptive Prediction Sets (RAPS) [179], furthermore imposes two regularization terms to keep the prediction sets as small as possible. Additionally, the *top-k* method takes the k highest softmax values as the prediction set, where k gets optimized on the calibration dataset. For more methodological details, see Section III B of the corresponding publication [4].

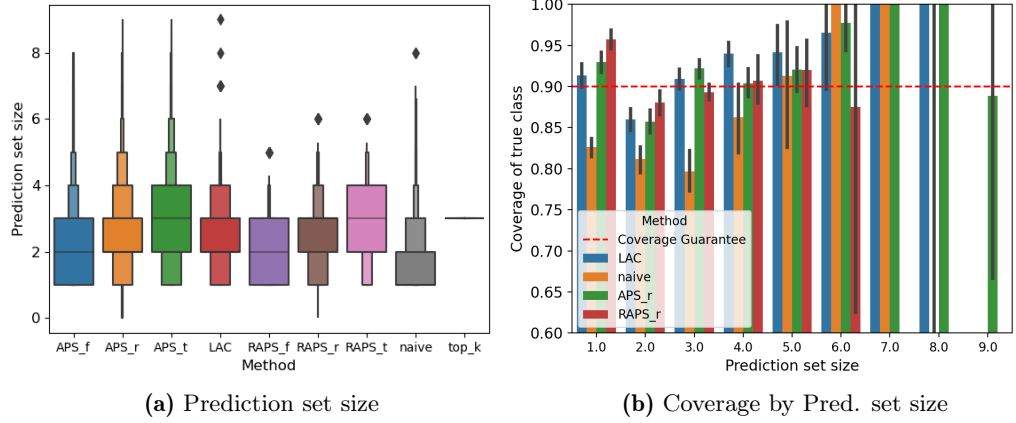


Figure 5.3: Explorative Analyses for CP methods with $\alpha = 0.1$.

For the following analyses, the same geographical split as in Section 4.1 of the LCZ42 dataset was used for a realistic real-world scenario. Similarly, the data was limited to the urban classes due to higher observed degrees of human label uncertainty. As a first step, multiple descriptive analyses are performed, for which exemplary test results are shown for $\alpha = 0.1$ in Figure 5.3. Stemming from the theory, one here expects a marginal coverage probability of 90% for the true class. Simultaneously, a low prediction set size is, of course, favorable. Worth noting is the overall smallest prediction set size by the naive method, i.e., without the explicit conformalization step. Yet without conformalization, the coverage guarantee is not met. Among the CP methods, the RAPS algorithm with randomized setting (*_r*) provides the best tradeoff between small prediction set size and performance.

Further analyses are performed regarding the predictive power of the CP methods with respect to aleatoric and epistemic uncertainty. With regards to the latter, an analysis of model accuracy is conducted separately by prediction set size and presented in Figure 5.4 (a). For this analysis, the underlying assumption is that more uncertain models (with respect to epistemic uncertainty) produce less accurate results. While this trend can be observed, it is noteworthy that the accuracy for the naive method is consistently lower than for the conformalized models for all prediction set sizes. In terms of predictive power towards aleatoric uncertainty, the prediction sets of the individual methods are compared to the label votes cast by the human experts. Both coverage and Intersection over Union (IoU) with respect to the label votes cast are investigated in Figure 5.4 (b), see Section III C in [4] for details on the derivation. Regarding both metrics, the naive method is

5.2 Conformal Score-Based Active Learning

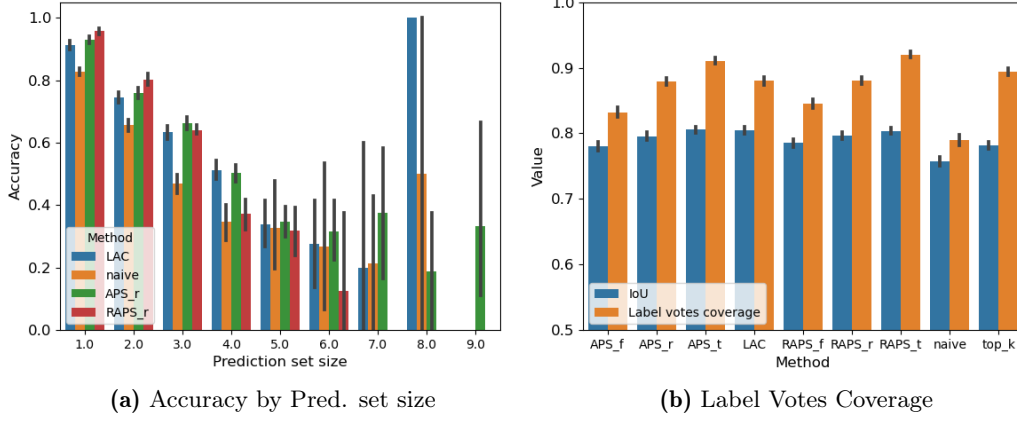


Figure 5.4: Predictive Analyses for CP methods with $\alpha = 0.1$.

again heavily underperforming. We naturally observe higher degrees of coverage for the models producing larger prediction sets, yet this gap in coverage is not entirely reflected in the IoU. Again, the randomized setting of RAPS provides a good solution balancing predictive power and prediction set size. Lastly, the prediction set size is investigated as a direct predictor for the presence of label uncertainty. For this, we look at the entropy of the voting vector as an indicator for label uncertainty (present if > 0).

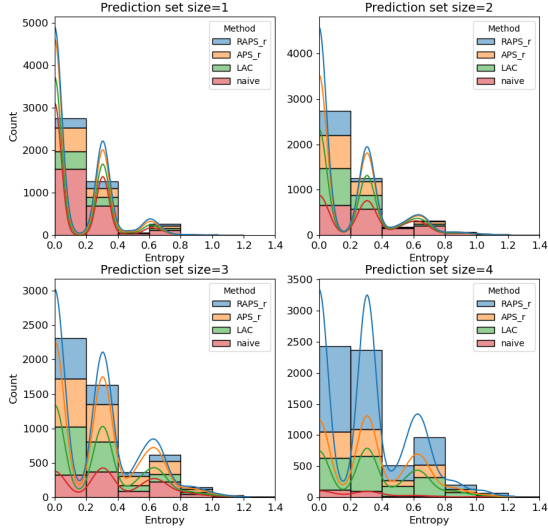


Figure 5.5: Entropy by pred. set size.

Ineff.	APS	LAC	RAPS	naive
1	0.51	0.63	0.95	0.64
2	0.59	0.55	0.21	1.36
3	1.34	0.98	0.85	1.87
4	1.61	1.84	1.41	1.16
5	0.88	1.52	6.00	1.08
6	1.55	1.18	14.00	0.19
7	1.47	3.11		2.67
8	7.00	6.50		8.50
9	3.43	5.00		
10		5.00		

Table 5.3: Odds of label uncertainty

Above, the distribution of entropies by prediction set size (left) as well as the observed odds of the occurrence of label uncertainty (right) are exemplarily shown for $\alpha = 0.05$.

We define the odds via

$$\text{Odds}\left(H(y) > 0\right) := \frac{P\left(H(y) > 0 | \text{Ineff} = k\right)}{P\left(H(y) = 0 | \text{Ineff} = k\right)} \quad (5.4)$$

where Ineff stands for the prediction set size. While most CP methods show an upward trend for the odds of label uncertainty with increasing prediction set size, this is not observed for the naive method. Concluding, CP algorithms give a good indication for both types of uncertainties with their prediction sets.

As a novelty, we furthermore investigate the potential of CP for Active Learning (AL). Note that the conformal scores have been utilized for this task before by Matiz *et al.* [201], yet the focus now lies on taking the prediction sets as input for the querying function. Initially, a model gets pretrained on a randomly selected subset consisting of ca. 10% of the entire dataset to get acquainted with the classification task. Taking the $_t$ setting (always including the borderline label) of the RAPS algorithm and setting $\alpha = 0.05$, the AL loop then looks as follows: First, conformalization is performed utilizing the hold-out calibration set (now a random subset of the entire dataset). After predicting on the remaining training pool, a fixed number (here 1,000) of samples is added to the training pool according to the querying strategy. The model is then retrained on the increased training dataset. After that, the loop repeats itself until a desired training dataset size is reached (here 30 %). For more implementational and methodological details, see Section V in [4].

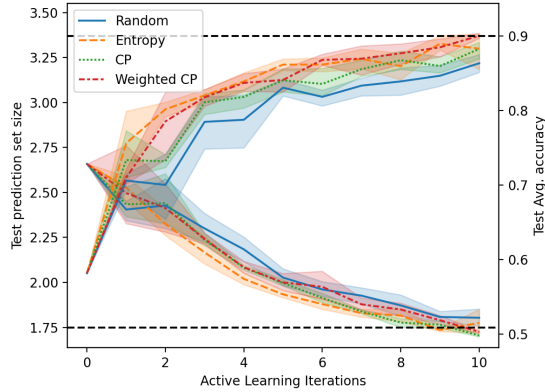


Figure 5.6: AL results

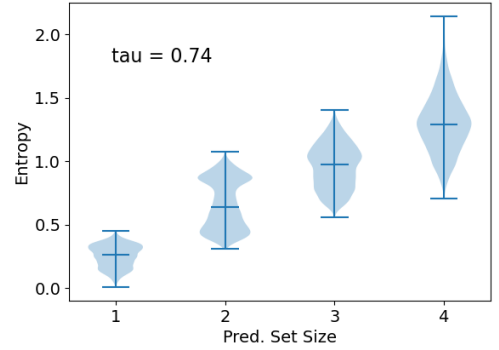


Figure 5.7: Entropy vs. pred. set size

We show the results of different querying strategies in Figure 5.6 from five separate models for each configuration. Next to the random baseline (querying new points randomly), entropy-based querying is considered, utilizing the Shannon entropy of the predicted softmax vector as a query function. For CP-based querying, two scenarios are considered: The regular CP method utilizes the prediction set size as a query function, while the weighted CP approach weights the prediction set size with the inverse class frequencies of the set members. This ensures that additional importance is placed on

the underrepresented classes. The plot displays both the average accuracy (empirical mean of all class-specific accuracies) and the average prediction set size (as a notion of predictive uncertainty) on the hold-out test set. The random baseline gets beaten by a large margin in terms of model performance by all models considered. The convergence of uncertainty-based querying is faster and can reach the test performance of the model trained on all data (indicated by dotted black line) with only ca. 30% of the data. At the same time, the predictive uncertainty shrinks continuously, although the gap to the random baseline diminishes over time.

The gap between entropy- and CP-based querying is marginal, so some more justification for the additional conformalization step needs to be given. Note first that the entropy-based querying is more fine-grained by construction. To undermine this, Figure 5.7 shows the range of observed entropies for the given prediction set sizes before the first AL loop (i.e. based on the entire remaining unlabeled training pool). Since the intervals also overlap, the relation between the two querying strategies is furthermore not unique. What remains beneficial for the CP-based querying, however, is the elegant way of incorporating the class weight via adding the prediction set membership. This eventually led to the overall best performance (though marginally close) in terms of average accuracy. It should also be mentioned that for a classification problem with more than the ten classes considered here, the consideration of prediction set sizes could be interesting both from an AL and an interpretational perspective. For more results and a more fine-grained discussion, see the corresponding publication [4]. Overall, the benefit of CP for remotely sensed data is undisputed. Especially for the studied dataset, a more complex notion of predictive uncertainty given by the prediction sets is advantageous.

6 Conclusion

This dissertation has systematically investigated the integration and utility of UQ frameworks within DL models for EO applications, with a particular focus on Land Use / Land Cover (LULC) classification. The overarching aim has been to move beyond mere predictive accuracy, emphasizing the critical need for understanding and interpreting the inherent uncertainty of complex DL models in real-world scenarios. The proposed approaches address crucial challenges related to model reliability, data ambiguity, out-of-domain expertise, and efficient data acquisition. By tackling different sources of uncertainty, namely data-inherent variability, model specification, and human annotation ambiguity, and evaluating their impact across diverse downstream tasks such as calibration, OOD detection, and AL, a comprehensive perspective on UQ’s utility is provided. The key findings and contributions are summarized in the following, simultaneously linking to the research questions posed in the introductory Chapter 1.

The first research question investigated the effect of human label uncertainty in the training pipeline of a NN classifier. The proposed approach moved beyond traditional one-hot encoded labels by utilizing *distributional labels* derived from normalized expert vote count, assigning a notion of uncertainty to each class that reflects the collective human expert judgment. A further methodological adaptation involved replacing the typical CE loss function with the discrete version of the KL divergence. This change aimed to train the network to more accurately mimic the nuanced expert voting behavior through its predictive distribution.

The impact on generalization performance was notable; uncertainty-aware learning significantly outperformed classical one-hot encoded settings. A performance gap of approximately 7%, was observed for the CE when evaluated against the one-hot encoded label, and about 11%, when evaluated against the distributional label. This indicates that explicitly modeling label ambiguity can indeed lead to more robust and generalizable models. Regarding calibration performance, the introduction of distributional labels alone nearly halved the ECE and substantially reduced the variation across different training runs. Visualizations through reliability diagrams further confirmed that this uncertainty-aware training diminished the characteristic overconfidence often seen in NNs, aligning model confidence more closely with actual accuracy. In comparison to classical calibration methods, while a combination of LS and TS could marginally surpass distributional learning, this outcome requires careful interpretation due to potential overfitting on the validation dataset. Conversely, MC Dropout showed marginal impact on generalization and calibration in the experiments.

Explicitly modeling human label uncertainty, rather than treating it as mere noise or simply disregarding it, transforms a perceived label ambiguity into a valuable source of information for model improvement. Traditional classification often operates under

6 Conclusion

the assumption of a single, unambiguous ground truth label, frequently dismissing any disagreement among annotators (if there even exist multiple) as noise. However, this research highlights label uncertainty as an inherent characteristic of the data, particularly in complex domains like RS. This paradigm shift suggests that in domains where human annotation is intrinsically ambiguous, models can achieve better robustness and calibration by learning from the collective uncertainty of experts.

As for the second research question, the usefulness of more complex label uncertainty modeling frameworks is challenged. The accompanying contribution introduced a novel three-level uncertainty scheme, incorporating Bayesian reasoning and leveraging prior information about the label space. This approach relaxed the categorical label space into a Dirichlet space, providing a more flexible modeling toolkit. The framework also enabled the estimation of unconstrained label embeddings (derived from a logarithmic transformation of Dirichlet parameters), offering even greater flexibility in representing label uncertainty. Three distinct loss functions were explored for training with these estimated embeddings, including the KL divergence (after exponential transformation) and two others that directly utilized the unbounded embeddings.

The calibration performance significantly benefited from these embedding-based approaches, with both ECE and SCE being reduced by up to 70%. Overall, the embedding-based approach using the Dirichlet transformation could show superior performance with regards to calibration and generalization, measured by the ECE and the KL divergence to the embeddings. The *Simple Dirichlet* model, representing the second level of uncertainty in the proposed scheme, demonstrated superior calibration on the class level, as clearly supported by its class-wise SCE contributions. For OOD detection, the *Simple Dirichlet* model consistently exhibited the best performance. The analysis revealed that identifying OOD samples was notably easier when vegetation classes served as ID data, likely attributable to their distinct spectral signatures. While the impact on overall accuracy was limited, there was no significant decrease in classical performance metrics compared to classical methods.

Progressing from distributional labels to a label embedding framework offers a more fine-grained understanding and quantification of uncertainty. Levels 2 and 3 of the uncertainty scheme allow the model to learn not only the distribution of votes but also the inherent uncertainty associated with those distributions. Via the introduced empirical Bayes framework, the label embeddings can be estimated globally, meaning with the inclusion of all observed label vote combinations as likelihood. Eventually, these unconstrained embeddings can be modeled in a greater variety of ways than by constraining oneself to the typically categorical or Dirichlet distribution spaces.

As a third research question, architectural changes to explicitly enhance OOD detection performance were challenged. The publication linked to this question explored *distance-aware* NN frameworks, specifically SNGP and DUE, for remotely sensed data. These architectures enforce a sense of boundedness in the network mapping, which supposedly effectively transfers input distances to the output space, where larger distances are directly associated with higher predictive uncertainty. The boundedness was fulfilled in two ways: In the hidden mapping, it was achieved by utilizing ResNets with approximate spectral normalization applied to their weight matrices. Whereas in the

output mapping, boundedness was realized by placing a GP on top of the last hidden layer, employing either RFF expansion (for SNGP) or inducing point approximation (for DUE).

Significant increases in OOD detection performance were observed. On the EUROSAT dataset, the DUE framework, which incorporates inducing point approximation and spectral batch normalization, yielded performance gains of up to 22 p.p. over baseline models. Similarly, for the So2Sat LCZ42 dataset, the inducing points approximation consistently demonstrated superior performance, with spectral normalization further enhancing detection when vegetation classes were considered OOD. Predictive entropy consistently proved to be the most effective UQ measure in these scenarios. A key advantage of these methods is their computational efficiency; the architectural changes are relatively minor, and the associated computational overhead is limited, making these approaches practical for RS data modeling.

To summarize, it could be shown that OOD detection in RS significantly benefits from the proposed architectural modifications. The core principle of distance-aware NNs is to ensure that semantic distances in the input space are effectively transferred to the output space. Remote sensing data inherently possesses strong spatial and spectral correlations. Semantically similar inputs, such as different urban area types, should therefore map to similar outputs, while semantically distinct inputs, like urban versus vegetation areas, should map to distant outputs. When a model encounters OOD data, it is essentially processing an input that is semantically far from its training distribution. By compelling the network to preserve these distances, distance-aware architectures inherently make the model’s predicted uncertainty more meaningful for OOD detection. The substantial performance gains observed underscore that for RS data, such architectural changes are highly effective for meaningful UQ.

The fourth research question investigated the usefulness of Conformal Prediction (CP), a distinct UQ framework combining deterministic and probabilistic elements, for RS. Its capacity to provide statistically rigorous prediction sets with guaranteed coverage of the true class is a unique approach to expressing a model’s predictive uncertainty. Descriptive analyses revealed that while the *naive* method (without explicit conformalization) yielded the smallest prediction set sizes, they consistently failed to meet the required coverage guarantees. Among the various CP methods, Regularized Adaptive Prediction Sets (RAPS) with a randomized setting demonstrated the most favorable trade-off between achieving small prediction set sizes and maintaining robust performance.

The predictive power of CP methods was further assessed with regard to the two main types of uncertainty. For epistemic uncertainty, models that produced larger prediction sets (indicating higher uncertainty) generally exhibited lower accuracy, confirming CP’s validity in reflecting the model’s knowledge state. Regarding aleatoric uncertainty, CP methods effectively captured data ambiguity, with their prediction sets showing superior coverage and Intersection over Union (IoU) with human expert label votes compared to naive methods. Furthermore, the prediction set size itself served as a direct predictor for the presence of label uncertainty, measured by the odds given a fixed prediction set size. This underscores CP’s ability to reflect inherent data ambiguity.

CP-based strategies also significantly enhanced Active Learning (AL) efficiency. Uncertainty-based querying converged faster than random baselines, achieving the performance level of a model trained on 100% of the data with only approximately 30% of labeled data (and close to target performance with 10-20%), while significantly reducing predictive uncertainty at the same time. Although the performance gap between CP-based and entropy-based querying was marginal, CP-based querying, particularly when incorporating class weighting, led to the overall best average accuracy. The ability to integrate multiple class weights into querying functions represents a notable advantage for addressing class imbalance prevalent in many RS datasets.

The strength of CP lies firstly in its statistically valid prediction sets. Furthermore, they are a good indicator of both model and data uncertainty. This makes CP particularly valuable for human-in-the-loop applications such as AL in EO. Unlike other UQ methods that produce probabilities or abstract uncertainty scores, CP outputs a set of possible classes, which is more directly interpretable. Also, a larger prediction set immediately signals higher uncertainty, directly linking to the human label uncertainty discussed previously. Hence, the prediction set size can also better reflect the inherent ambiguity in the data.

6.1 Limitations of the Current Work

While this dissertation has made significant advancements in UQ for remotely sensed data, it is important to acknowledge several limitations that inherently define the scope of the current work. Most importantly, the two works on label uncertainty require some degree of measurable human annotation uncertainty. Acquiring this richness of uncertainty in practice requires a massive manual labeling effort, which is often not part of the data creation budget. Nevertheless, as was observed from various experiments conducted throughout this dissertation, the additional labeling effort pays off immensely. Not only are generalization and calibration performance greatly affected, but several downstream tasks can benefit largely from the uncertainties derived from specially adapted label uncertainty frameworks.

A fundamental aspect of the inductive CP framework is its reliance on the exchangeability assumption between the calibration, training, and test datasets. In practical RS applications, this assumption is frequently violated due to inherent spatial, temporal, or spectral shifts in data. Although the exchangeability assumption is closely linked to the i.i.d. assumption typically required for DL, the latter is not directly responsible for a successful outcome. This is contrary to the exchangeability assumption, which is crucial for the theoretical coverage guarantee to hold at test time.

A notable limitation is the choice of architecture for distance-aware neural networks. The theoretical foundations of methods like SNGP and DUE only support the Residual Network (ResNet) architecture. This theoretical constraint restricts the direct applicability of these promising UQ methods to other neural network architectures, which might be more suitable for a broad range of RS tasks beyond image classification. Furthermore, a challenge identified within the DUE framework is the hyper-sensitivity of inducing

points. This sensitivity complicates the tuning process, as many other hyperparameters are present within the distance-aware framework. Also, it can potentially compromise the robustness and generalizability of the model when deployed across different datasets.

In the application domain of AL, a marginal gap between entropy- and CP-based querying was observed. This suggests that for certain scenarios, the added complexity of conformalization might require further justification or more pronounced advantages to warrant its adoption over simpler, computationally less intensive methods. The investigation of calibration methods, on the other hand, revealed a potential issue of over-tuning the parameters. When combining LS and TS for calibration, this can occur if the hyperparameters of the methods are optimized on a validation dataset that shares similar origins with the test set, leading to an overly optimized calibration for the unseen test set. In the context of the distributional labels approach, a limited influence of MC Dropout on generalization was observed. The approach showed no improvement, which is potentially attributed to the limited number of dropout layers within the base network, suggesting that its full potential for UQ and generalization improvement was not realized.

A specific data-related challenge encountered was a convergence issue with Class 7 in the LCZ42 dataset. The EM algorithm employed for estimating the label embeddings experienced convergence issues with this class. This required the exclusion of this specific class from some analyses, thereby limiting the adaptability of the findings. Furthermore, the exclusion of the vegetation classes for the distributional label approach needs to be addressed. When a significant portion of the data has unambiguous labels, forcing a distributional representation may not yield additional benefits. This calls for frameworks that can robustly handle both highly uncertain and highly certain labels concurrently. Moreover, the research identified the underperformance of the Mahalanobis distance loss for the embedded labels. This issue was mainly driven by numerical instabilities in the loss calculation, where problems were encountered when inverting the correlation matrix for the loss calculation (primarily due to the presence of many eigenvalues smaller than 1). This highlights the challenge of robustly incorporating complex correlation structures among classes directly into the loss function.

These limitations collectively highlight a central aspect of applying UQ methods to real-world data: bridging the gap between theoretical UQ advancements and their practical application in RS. Many methods are initially developed in Computer Vision (CV) and often build on different characteristics than remotely sensed data inherits. EO data possesses unique attributes, including multi-spectral and multi-temporal dimensions, complex spatial correlations, and measurable human label uncertainty in the case of LCZ42. This implies that while the fundamental principles of UQ are universal, their effective implementation in RS demands careful adaptation and, in some cases, potentially new theoretical developments tailored to the peculiarities of remotely sensed data. The identified limitations therefore underscore that EO is not merely another application domain for UQ but a field that demands specialized UQ methodologies to fully realize their potential.

6.2 Future Research Directions

Building upon the insights gained and the limitations identified in this dissertation, several promising ways for future research emerge. A primary direction involves expanding distance-aware architectures. These have been proven to significantly enhance the quality of predictive uncertainties in LULC classification tasks. Future work should focus on extending the theoretical foundations and practical implementations of distance-aware neural networks beyond their current scope. This would involve adapting these methods to a wider array of architectures, such as U-Nets [202] for semantic segmentation, transformer-based [203] models for sequence data, or graph neural networks [204] for spatial-temporal analysis. Such expansion would significantly broaden their applicability across diverse RS tasks and data modalities, enabling effective UQ in more complex scenarios.

Moreover, the field will greatly benefit from benchmarking UQ frameworks with new RS datasets, particularly with a focus on human label uncertainty. Despite the additional labeling effort required for estimating a notion of label uncertainty, the expected insights and knowledge gains far exceed these efforts: The way for a more unified and generalizable UQ framework for EO data could hereby be established. Regarding this, a recent publication [205] points to a pivotal future direction for more comprehensive benchmarking of UQ methods using new and diverse EO datasets. This will ensure that UQ methods are robust, generalizable, and perform consistently across a wider range of real-world scenarios and downstream tasks.

The label uncertainty frameworks presented in this dissertation can also be easily integrated into existing modeling pipelines. Although requiring a base training set with multiple expert annotations in the first place, notions of uncertainty derived from the uncertainty-aware models are more trustworthy and help sharpen the predictions on unseen data, without requiring much additional training effort. Of course, the application scope of the presented methods lies far beyond LULC. Different data modalities could impact the amount of data or expert annotations needed, for example, because the number of classes is much higher, and hence the complex label ambiguity structure would require more individual annotations per image. Yet, the underlying core methodologies easily transfer to different remotely sensed data sets.

Next research advances could explore methods for dynamically inferring and integrating label uncertainty, moving beyond a fixed number of expert annotations. This could involve developing semi-supervised approaches where the model learns to identify ambiguous samples and then selectively queries human experts for more effective annotations. Alternatively, self-supervised methods could be investigated to infer uncertainty directly from data characteristics, which would then be transferred to human judgment. Furthermore, extending label uncertainty frameworks to multi-modal RS data presents a promising opportunity. Investigating how ambiguities across these diverse data modalities can form a more comprehensive and robust understanding of label uncertainty is crucial.

Additionally, research should explore how label uncertainty can be effectively leveraged within explainable AI frameworks. This would involve providing insights into why

a model is uncertain about a particular prediction, for instance, by visualizing areas of high label ambiguity on a map to guide human interpretation and decision-making processes. Finally, exploring the ethical implications of UQ in critical EO applications is essential. This involves addressing questions such as how uncertainty information should be effectively communicated to non-expert decision-makers and what responsibilities arise when high uncertainty exists in predictions that could impact critical areas like disaster response, defense research, or climate policy. Ensuring that UQ is not just a mere technical enhancement but also a tool for responsible AI deployment is a serious future challenge.

Trustworthy AI for EO relies heavily on solid derivation and interpretation of uncertainties. The presented works are aimed at exploring the most promising UQ techniques in greater detail and making them suitable for deployment in real-world use cases. Necessary steps toward this deployment involve raising awareness about the critical need for UQ to all involved stakeholders, pushing for the implementation of UQ toolkits as mandatory requirements of AI tools in practice, and continuously spreading the word about promising results found in fundamental and applied research.

These proposed potential research directions collectively point towards a single overarching goal: the development of generalizable, trustworthy, and adaptable UQ frameworks for remotely sensed data. The emphasis on expanding architectural compatibility, conducting comprehensive benchmarking with new datasets, unfolding the potential of label uncertainty, the deployment of the proposed methodologies, considering the ethical concerns of UQ, and enabling trustworthy AI, all contribute to this objective. This implies that the next frontier involves moving beyond task-specific solutions towards developing automated and integrated UQ techniques. Such advancements would enable UQ frameworks to intelligently adapt to new datasets and tasks without extensive manual tuning, thereby accelerating the responsible and reliable deployment of trustworthy AI in the rapidly evolving field of EO.

Bibliography

- [1] C. Koller, G. Kauermann, and X. X. Zhu. Going beyond one-hot encoding in classification: Can human uncertainty improve model performance in earth observation? *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024.
- [2] C. Schweden, K. Hechinger, G. Kauermann, and X. X. Zhu. Can uncertainty quantification benefit from label embeddings? a case study on local climate zone classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.
- [3] C. Koller, P. Jung, and X. X. Zhu. Can land cover classification models benefit from distance-aware architectures? *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.
- [4] C. Schweden, P. Jung, and X. X. Zhu. Assessing conformal prediction for remote sensing: Application to local climate zone classification. *IEEE Transactions on Geoscience and Remote Sensing*, Accepted.
- [5] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [6] C. C. Aggarwal et al. *Neural networks and deep learning*, volume 10. Springer, 2018.
- [7] S. Raschka, Y. H. Liu, and V. Mirjalili. *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing Ltd, 2022.
- [8] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
- [9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
- [10] A. LeNail. Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4(33):747, 2019.
- [11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

BIBLIOGRAPHY

- [12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [14] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2015.
- [17] F. Chollet. Xception: Deep learning with depthwise separable convolutions, 2017.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [19] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.
- [21] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.
- [22] A. P. Cracknell. The development of remote sensing in the last 40 years. *International Journal of Remote Sensing*, 39(23):8387–8427, 2018.
- [23] P. Wang, B. Bayram, and E. Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232:104110, 2022.
- [24] M. Mehmood, A. Shahzad, B. Zafar, A. Shabbir, and N. Ali. Remote sensing image classification: A comprehensive review and applications. *Mathematical problems in engineering*, 2022(1):5880959, 2022.
- [25] T. Hoeser and C. Kuenzer. Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends. *Remote Sensing*, 12(10):1667, 2020.

- [26] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.
- [27] M. Valdenegro-Toro and D. S. Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022.
- [28] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.
- [29] J. Mena, O. Pujol, and J. Vitrià. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35, 2021.
- [30] W. He, Z. Jiang, T. Xiao, Z. Xu, and Y. Li. A survey on uncertainty quantification methods for deep learning, 2025. [arXiv:2302.13425](#).
- [31] T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.
- [32] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.
- [33] R. M. Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012.
- [34] A. Malinin, B. Mlodozieniec, and M. Gales. Ensemble distribution distillation. In *International Conference on Learning Representations*, 2020.
- [35] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped dqn. *Advances in Neural Information Processing Systems*, 29, 2016.
- [36] O. Achrack, R. Kellerman, and O. Barzilay. Multi-loss sub-ensembles for accurate classification with uncertainty estimation. In *Advances in Neural Information Processing Systems Workshop on Machine Learning for Engineering Modeling, Simulation, and Design*, 2020.
- [37] G. E. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Sixth Annual Conference on Computational Learning Theory*, pages 5–13, 1993.
- [38] D. Barber and C. M. Bishop. Ensemble learning in bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 168:215–238, 1998.
- [39] A. Graves. Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*, 24, 2011.

BIBLIOGRAPHY

- [40] J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869. PMLR, 2015.
- [41] Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. *Advances in Neural Information Processing Systems*, 28, 2015.
- [42] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.
- [43] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.
- [44] D. J. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.
- [45] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.
- [46] G. Shafer. A mathematical theory of evidence, 1976.
- [47] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.
- [48] D. T. Ulmer, C. Hardmeier, and J. Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*, 2023.
- [49] A. Damianou and N. D. Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215. PMLR, 2013.
- [50] C. K. Williams and C. E. Rasmussen. *Gaussian Processes for Machine Learning*. MIT press Cambridge, MA, 2006.
- [51] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378. PMLR, 2016.
- [52] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020.
- [53] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 2020.
- [54] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.

- [55] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu. Earthnets: Empowering artificial intelligence for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, pages 2–36, 2024.
- [56] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5901–5904, 2019.
- [57] European Environment Agency (EEA) and Copernicus Land Monitoring Service. CORINE Land Cover Change 2012-2018 (raster 100 m), Europe, 6-yearly, 2020. Version 2020_20u1, May 2020. URL: <https://land.copernicus.eu/en/products/corine-land-cover/lcc-2012-2018>, doi: 10.2909/5654b422-af84-4115-ac3c-5d7dea540ebb.
- [58] G. Sumbul, A. De Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl. Bigearthnet-mm: A large-scale, multi-modal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021.
- [59] K. N. Clasen, L. Hackel, T. Burgert, G. Sumbul, B. Demir, and V. Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2025.
- [60] P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2019.
- [61] H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, and L. Zhao. Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors*, 20(6):1594, 2020.
- [62] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.
- [63] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021.
- [64] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010.
- [65] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised

BIBLIOGRAPHY

- learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.
- [66] A. Stewart, N. Lehmann, I. Corley, Y. Wang, Y.-C. Chang, N. A. Ait Ali Braham, S. Sehgal, C. Robinson, and A. Banerjee. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36:59787–59807, 2023.
- [67] I. D. Stewart. *Redefining the urban heat island*. PhD thesis, University of British Columbia, 2011.
- [68] J. Han, N. Mo, J. Cai, L. Ouyang, and Z. Liu. Advancing the local climate zones framework: a critical review of methodological progress, persisting challenges, and future research prospects. *Humanities and Social Sciences Communications*, 11(1):1–18, 2024.
- [69] I. D. Stewart and T. R. Oke. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12):1879–1900, 2012.
- [70] P. J. Alexander and G. Mills. Local climate classification and dublin’s urban heat island. *Atmosphere*, 5(4):755–774, 2014.
- [71] J. A. Quanz, S. Ulrich, D. Fenner, A. Holtmann, and J. Eimermacher. Micro-scale variability of air temperature within a local climate zone in berlin, germany, during summer. *Climate*, 6(1):5, 2018.
- [72] G. Thomas, A. Sherin, S. Ansar, and E. Zachariah. Analysis of urban heat island in kochi, india, using a modified local climate zone classification. *Procedia Environmental Sciences*, 21:3–13, 2014.
- [73] E. Lelovics, J. Unger, T. Gál, and C. V. Gál. Design of an urban monitoring network based on local climate zone mapping and temperature pattern modelling. *Climate Research*, 60(1):51–62, 2014.
- [74] N. Perera and R. Emmanuel. A “local climate zone” based approach to urban planning in colombo, sri lanka. *Urban Climate*, 23:188–203, 2018.
- [75] I. D. Stewart, T. R. Oke, and E. S. Krayenhoff. Evaluation of the ‘local climate zone’ scheme using temperature observations and model simulations. *International Journal of Climatology*, 34(4):1062–1080, 2014.
- [76] J. Yang, S. Jin, X. Xiao, C. Jin, J. C. Xia, X. Li, and S. Wang. Local climate zone ventilation and urban land surface temperatures: Towards a performance-based and wind-sensitive planning proposal in megacities. *Sustainable Cities and Society*, 47:101487, 2019.
- [77] X. X. Zhu et al. So2sat lcz42: a benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020.

- [78] X. X. Zhu et al. So2sat: Big data for 4d global urban mapping $\sim 10^{16}$ bytes from social media to eo satellites. <https://cordis.europa.eu/project/id/714087>, 2017. Accessed: 15.05.2025.
- [79] X. X. Zhu et al. So2Sat LCZ42 Dataset. <https://mediatum.ub.tum.de/1454690>, 2019. Accessed: 15.05.2025.
- [80] J. Ching, G. Mills, B. Bechtel, L. See, J. Feddema, X. Wang, C. Ren, O. Brousse, A. Martilli, M. Neophytou, et al. Wudapt: An urban weather, climate, and environmental modeling infrastructure for the anthropocene. *Bulletin of the American Meteorological Society*, 99(9):1907–1924, 2018.
- [81] X. X. Zhu et al. So2Sat LCZ42 Label Evaluation Dataset. <https://mediatum.ub.tum.de/1659039>, 2023. Accessed: 19.05.2025.
- [82] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [83] B. Bechtel and C. Daneke. Classification of local climate zones based on multiple earth observation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(4):1191–1202, 2012.
- [84] M.-L. Verdonck, A. Okujeni, S. van der Linden, M. Demuzere, R. De Wulf, and F. Van Coillie. Influence of neighbourhood information on ‘local climate zone’ mapping in heterogeneous cities. *International Journal of Applied Earth Observation and Geoinformation*, 62:102–113, 2017.
- [85] C. Yoo, D. Han, J. Im, and B. Bechtel. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157:155–170, 2019.
- [86] J. Hu, P. Ghamisi, and X. X. Zhu. Feature extraction and selection of sentinel-1 dual-pol data for global-scale local climate zone classification. *ISPRS International Journal of Geo-Information*, 7(9):379, 2018.
- [87] C. Qiu, M. Schmitt, P. Ghamisi, and X. X. Zhu. Effect of the training set configuration on sentinel-2-based urban local climate zone classification. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:931–936, 2018.
- [88] C. Qiu, M. Schmitt, M. Lichao, and X. X. Zhu. Urban local climate zone classification with a residual convolutional neural network and multi-seasonal sentinel-2 images. In *10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, pages 1–5. IEEE, 2018.
- [89] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu. Multilevel feature fusion-based cnn for local climate zone classification from sentinel-2 images: Benchmark

BIBLIOGRAPHY

- results on the so2sat lcz42 dataset. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:2793–2806, 2020.
- [90] J. Gawlikowski, M. Schmitt, A. Kruspe, and X. X. Zhu. On the fusion strategies of sentinel-1 and sentinel-2 data for local climate zone classification. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 2081–2084, 2020.
- [91] P. Feng, Y. Lin, G. He, J. Guan, J. Wang, and H. Shi. A dynamic end-to-end fusion filter for local climate zone classification using sar and multi-spectrum remote sensing data. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4231–4234, 2020.
- [92] M. Kim, D. Jeong, and Y. Kim. Local climate zone classification using a multi-scale, multi-level attention network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181:345–366, 2021.
- [93] X. Zhao, J. Hu, L. Mou, Z. Xiong, and X. X. Zhu. Cross-city landuse classification of remote sensing images via deep transfer learning. *International Journal of Applied Earth Observation and Geoinformation*, 122:103358, 2023.
- [94] X. Zhong, H. Li, H. Shen, M. Gao, Z. Wang, and J. He. Local climate zone mapping by coupling multilevel features with prior knowledge based on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.
- [95] F. Huang, S. Jiang, W. Zhan, B. Bechtel, Z. Liu, M. Demuzere, Y. Huang, Y. Xu, L. Ma, W. Xia, et al. Mapping local climate zones for cities: A large review. *Remote Sensing of Environment*, 292:113573, 2023.
- [96] X. X. Zhu, C. Qiu, J. Hu, Y. Shi, Y. Wang, M. Schmitt, and H. Taubenböck. The urban morphology on our planet – global perspectives from space. *Remote Sensing of Environment*, 269:112794, 2022.
- [97] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.
- [98] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.
- [99] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2013.
- [100] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2022.

- [101] Y. Dgani, H. Greenspan, and J. Goldberger. Training a neural network based on unreliable human annotation of medical images. In *15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 39–42. IEEE, 2018.
- [102] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR, 2023.
- [103] X. Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.
- [104] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.
- [105] X. Jia, W. Li, J. Liu, and Y. Zhang. Label distribution learning by exploiting label correlations. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [106] C. Xu and X. Geng. Hierarchical classification based on label distribution learning. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 5533–5540, 2019.
- [107] J. Wang and X. Geng. Label distribution learning by exploiting label distribution manifold. *IEEE Transactions on Neural Networks and Learning Systems*, 34(2):839–852, 2021.
- [108] X. Zhao, Y. An, N. Xu, J. Wang, and X. Geng. Imbalanced label distribution learning. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 11336–11344, 2023.
- [109] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.
- [110] A. Krizhevsky. Learning multiple layers of features from tiny images. *Master’s Thesis, University of Tront*, 2009.
- [111] K. M. Collins, U. Bhatt, and A. Weller. Eliciting and learning with soft labels from every annotator. In *AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 40–52, 2022.
- [112] L. Schmarje, V. Grossmann, C. Zelenka, S. Dippel, R. Kiko, M. Oszust, M. Pastell, J. Stracke, A. Valros, N. Volkmann, et al. Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232, 2022.
- [113] K. Hechinger, X. X. Zhu, and G. Kauermann. Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models.

BIBLIOGRAPHY

- Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1):143–161, 2024.
- [114] G. Celeux, D. Chauveau, and J. Diebolt. Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4), 1996.
 - [115] K. Hechinger, C. Schweden, X. X. Zhu, and G. Kauermann. Human-in-the-loop: Towards label embeddings for measuring classification difficulty. *Statistical Modelling (to appear)*, 2025.
 - [116] N. Tsutsumida and A. Kato. Reducing investigator bias in sampling-based land cover classification by integrating multiple investigators’ maps using a multiple classifier system. *arXiv preprint arXiv:2403.15720*, 2024.
 - [117] M. Russwurm, S. Wang, and D. Tuia. Humans are poor few-shot classifiers for sentinel-2 land cover. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2022.
 - [118] F. Wang, Y. Wu, P. Zhang, Q. Zhang, and M. Li. Unsupervised sar image segmentation using ambiguity label information fusion in triplet markov fields model. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1479–1483, 2017.
 - [119] C. García Rodríguez, J. Vitrià, and O. Mora. Uncertainty-based human-in-the-loop deep learning for land cover segmentation. *Remote Sensing*, 12(22):3836, 2020.
 - [120] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179:145–158, 2021.
 - [121] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
 - [122] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.
 - [123] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
 - [124] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
 - [125] R. Müller, S. Kornblith, and G. Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

- [126] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.
- [127] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, and C. Gagné. Attended temperature scaling: a practical approach for calibrating deep neural networks. *arXiv preprint arXiv:1810.11586*, 2018.
- [128] D. Zhu, B. Lei, J. Zhang, Y. Fang, Y. Xie, R. Zhang, and D. Xu. Rethinking data distillation: Do not overlook calibration. In *IEEE/CVF International Conference on Computer Vision*, pages 4935–4945, 2023.
- [129] M.-H. Laves, S. Ihler, K.-P. Kortmann, and T. Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*, 2019.
- [130] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.
- [131] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, pages 625–632, 2005.
- [132] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, volume 2, 2019.
- [133] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön. Evaluating model calibration in classification. In *22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.
- [134] M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems*, 32, 2019.
- [135] S. Gruber and F. Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35:8618–8632, 2022.
- [136] K. Patel, W. Beluch, K. Rambach, A.-E. Cozma, M. Pfeiffer, and B. Yang. Investigation of uncertainty of deep learning-based object classification on radar spectra. In *IEEE Radar Conference*, pages 1–6. IEEE, 2021.
- [137] P. Conde, R. L. Lopes, and C. Premebida. Improving accuracy and calibration of deep image classifiers with agreement-driven dynamic ensemble. *IEEE Open Journal of the Computer Society*, 2024.

BIBLIOGRAPHY

- [138] H. Bai, J. Cheng, Y. Su, S. Liu, and X. Liu. Calibrated focal loss for semantic labeling of high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:6531–6547, 2022.
- [139] G. Dussert, S. Chamailé-Jammes, S. Dray, and V. Miele. Being confident in confidence scores: calibration in deep learning models for camera trap image sequences. *Remote Sensing in Ecology and Conservation*, 11(1):88–99, 2025.
- [140] F. Quinton and L. Landrieu. Crop rotation modeling for deep learning-based parcel classification from satellite time series. *Remote Sensing*, 13(22):4599, 2021.
- [141] F. Rewicki and J. Gawlikowski. Estimating uncertainty of deep learning multi-label classifications using laplace approximation. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1560–1563, 2022.
- [142] J. Huang, S. Park, N. Paoletti, and O. Simeone. Distilling calibration via conformalized credal inference. *arXiv preprint arXiv:2501.06066*, 2025.
- [143] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017.
- [144] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.
- [145] J. J. Ryu, M. Shen, S. Ghosh, Y. Bu, P. Sattigeri, S. Das, and G. W. Wornell. Improved evidential deep learning via a mixture of dirichlet distributions. *arXiv preprint arXiv:2402.06160*, 2024.
- [146] T. Joo, U. Chung, and M.-G. Seo. Being bayesian about categorical probability. In *International Conference on Machine Learning*, pages 4950–4961. PMLR, 2020.
- [147] W. Liu, X. Wang, J. Owens, and Y. Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.
- [148] J. Postels, H. Blum, Y. Strümpler, C. Cadena, R. Siegwart, L. Van Gool, and F. Tombari. The hidden uncertainty in a neural networks activations. *arXiv preprint arXiv:2012.03082*, 2020.
- [149] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu. An advanced dirichlet prior network for out-of-distribution detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 2022.
- [150] J. Castillo-Navarro, B. Le Saux, A. Boulch, and S. Lefèvre. Energy-based models in earth observation: From generation to semisupervised learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.

- [151] I. Rolland, A. Marinoni, and S. Selvakumaran. Uncertainty-aware graph-based multimodal remote sensing detection of out-of-distribution samples. In *1st Workshop on Complex Data Challenges in Earth Observation*, 2021.
- [152] N. Inkawhich, J. Zhang, E. K. Davis, R. Luley, and Y. Chen. Improving out-of-distribution detection by learning from the deployment environment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2070–2086, 2022.
- [153] B. Ekim, G. A. Tadesse, C. Robinson, G. Hacheme, M. Schmitt, R. Dodhia, and J. M. L. Ferres. Distribution shifts at scale: Out-of-distribution detection in earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2265–2274, 2025.
- [154] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- [155] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning*, volume 2, page 6. Cite-seer, 2000.
- [156] D. Cacciarelli and M. Kulahci. Active learning for data streams: a survey. *Machine Learning*, 113(1):185–239, 2024.
- [157] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.
- [158] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.
- [159] A. Kirsch, J. Van Amersfoort, and Y. Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [160] A. Kirsch, S. Farquhar, P. Atighehchian, A. Jesson, F. Branchaud-Charron, and Y. Gal. Stochastic batch acquisition: A simple baseline for deep active learning. *arXiv preprint arXiv:2106.12059*, 2021.
- [161] Y. Geifman and R. El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.
- [162] Z. Liu, J. Wang, S. Gong, H. Lu, and D. Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *IEEE/CVF International Conference on Computer Vision*, pages 6122–6131, 2019.
- [163] P. Mitra, B. U. Shankar, and S. K. Pal. Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recognition Letters*, 25(9):1067–1074, 2004.

BIBLIOGRAPHY

- [164] S. Rajan, J. Ghosh, and M. M. Crawford. An active learning approach to hyperspectral data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4):1231–1242, 2008.
- [165] J. Li, J. M. Bioucas-Dias, and A. Plaza. Hyperspectral image segmentation using a new bayesian approach with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3947–3960, 2011.
- [166] Z. Wang, B. Du, L. Zhang, L. Zhang, and X. Jia. A novel semisupervised active-learning algorithm for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3071–3083, 2017.
- [167] X. Cao, J. Yao, Z. Xu, and D. Meng. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4604–4616, 2020.
- [168] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.
- [169] B. Demir, C. Persello, and L. Bruzzone. Batch-mode active-learning methods for the interactive classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(3):1014–1031, 2010.
- [170] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*, volume 29. Springer, 2005.
- [171] V. Vovk. Transductive conformal predictors. In *International Conference on Artificial Intelligence Applications and Innovations*, pages 348–360. Springer, 2013.
- [172] H. Papadopoulos, V. Vovk, and A. Gammerman. Conformal prediction with neural networks. In *IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 388–395. IEEE, 2007.
- [173] S. L. Zabell. Predicting the unpredictable. *Synthese*, 90:205–232, 1992.
- [174] T. Fritz, T. Gonda, and P. Perrone. De finetti’s theorem in categorical probability. *Journal of Stochastic Analysis*, 2(4), 2021.
- [175] V. Vovk, V. Fedorova, I. Nouretdinov, and A. Gammerman. Criteria of efficiency for conformal prediction. In *International Symposium on Conformal and Probabilistic Prediction with Applications*, pages 23–39. Springer, 2016.
- [176] A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

- [177] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.
- [178] Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.
- [179] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.
- [180] L. Dabah and T. Tirer. On calibration and conformal prediction of deep classifiers. *arXiv preprint arXiv:2402.05806*, 2024.
- [181] Z. Liang, Y. Zhou, and M. Sesia. Conformal inference is (almost) free for neural networks trained with early stopping. In *International Conference on Machine Learning*, pages 20810–20851. PMLR, 2023.
- [182] D. Bethell, S. Gerasimou, and R. Calinescu. Robust uncertainty quantification using conformalised monte carlo prediction. In *AAAI Conference on Artificial Intelligence*, volume 38 (19), 2024.
- [183] H. Lei and A. Bellotti. Reliable prediction intervals with directly optimized inductive conformal regression for deep learning. *Neural Networks*, 168:194–205, 2023.
- [184] D. Valle, R. Izbicki, and R. V. Leite. Quantifying uncertainty in land-use land-cover classification using conformal statistics. *Remote Sensing of Environment*, 295:113682, 2023.
- [185] P. Bhattacharjee and P. Jung. Onboard out-of-calibration detection of deep learning models using conformal prediction. *arXiv preprint arXiv:2405.02634*, 2024.
- [186] T. Harris and R. Srivier. Quantifying uncertainty in climate projections with conformal ensembles. *arXiv preprint arXiv:2408.06642*, 2024.
- [187] K. Liu, T. Sun, H. Zeng, Y. Zhang, C.-M. Pun, and C.-M. Vong. Spatial-aware conformal prediction for trustworthy hyperspectral image classification. *arXiv preprint arXiv:2409.01236*, 2024.
- [188] G. Singh, G. Moncrieff, Z. Venter, K. Cawse-Nicholson, J. Slingsby, and T. B. Robinson. Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction. *Scientific Reports*, 14(1):16166, 2024.
- [189] P. Hajibabaei, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili. Adaptive conformal prediction intervals using data-dependent weights with application to seismic response prediction. *IEEE Access*, 12:53579–53597, 2024.

BIBLIOGRAPHY

- [190] N. Kakhani, S. Alamdar, N. M. Kebonye, M. Amani, and T. Scholten. Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction. *Remote Sensing*, 16(3):438, 2024.
- [191] N. Zhu, Z. Xi, C. Wu, F. Zhong, R. Qi, H. Chen, S. Xu, and W. Ji. Inductive conformal prediction enhanced lstm-snn network: Applications to birds and uavs recognition. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.
- [192] M. Farag, J. Kierdorf, and R. Roscher. Inductive conformal prediction for harvest-readiness classification of cauliflower plants: A comparative study of uncertainty quantification methods. In *IEEE/CVF International Conference on Computer Vision*, pages 651–659, 2023.
- [193] F. Meng and T. Song. Uncertainty forecasting system for tropical cyclone tracks based on conformal prediction. *Expert Systems with Applications*, 249:123743, 2024.
- [194] J. Z. Liu, S. Padhy, J. Ren, Z. Lin, Y. Wen, G. Jerfel, Z. Nado, J. Snoek, D. Tran, and B. Lakshminarayanan. A simple approach to improve single-model deep uncertainty via distance-awareness. *arXiv preprint arXiv:2205.00403*, 2022.
- [195] P. L. Bartlett, S. N. Evans, and P. M. Long. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv preprint arXiv:1804.05012*, 2018.
- [196] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, 2019.
- [197] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416, 2021.
- [198] Z. Linli, X. Liang, Z. Zhang, K. Hu, and S. Guo. Enhancing brain age estimation under uncertainty: A spectral-normalized neural gaussian process approach utilizing 2.5d slicing. *NeuroImage*, 311:121184, 2025.
- [199] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 2007.
- [200] C. Koller, P. Jung, and X. X. Zhu. Exploring distance-aware uncertainty quantification for remote sensing image classification. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5692–5695, 2023.
- [201] S. Matiz and K. E. Barner. Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification. *Pattern Recognition*, 90:172–182, 2019.
- [202] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

- [203] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [204] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.
- [205] Y. Wang, Q. Song, D. Wasif, M. Shahzad, C. Koller, J. Bamber, and X. X. Zhu. How certain are uncertainty estimates? three novel earth observation datasets for benchmarking uncertainty quantification in machine learning. *arXiv preprint arXiv:2412.06451*, 2024.

A Publications

A.1 Going Beyond One-Hot Encoding in Classification: Can Human Uncertainty Improve Model Performance in Earth Observation?

Reference

C. Koller, G. Kauermann and X. X. Zhu. Going Beyond One-Hot Encoding in Classification: Can Human Uncertainty Improve Model Performance in Earth Observation?. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-11, 2024. doi: 10.1109/TGRS.2023.3336357

Copyright

Article published in the IEEE Transactions on Geoscience and Remote Sensing under a CC-BY-4.0 license. Reproduced with friendly permission from the authors.

Going Beyond One-Hot Encoding in Classification: Can Human Uncertainty Improve Model Performance in Earth Observation?

Christoph Koller¹, Graduate Student Member, IEEE, Göran Kauermann, and Xiao Xiang Zhu², Fellow, IEEE

Abstract—Technological and computational advances continuously drive forward the field of deep learning in remote sensing. In recent years, the derivation of quantities describing the uncertainty in the prediction—which naturally accompanies the modeling process—has sparked interest in the remote sensing community. Often neglected in the machine learning setting is the human uncertainty that influences numerous labeling processes. As the core of this work, the task of local climate zone (LCZ) classification is studied by means of a dataset that contains multiple label votes by domain experts for each image. The inherent label uncertainty describes the ambiguity among the domain experts and is explicitly embedded into the training process via distributional labels. We show that incorporating the label uncertainty helps the model to generalize better to the test data and increases model performance. Similar to existing calibration methods, the distributional labels lead to better-calibrated probabilities, which in turn yield more certain and trustworthy predictions. For reproducibility, we provide our code here https://github.com/ChrisKo94/LCZ_LDL and here https://gitlab.lrz.de/ai4eo/WG_Uncertainty/lcz_ldl.

Index Terms—Calibration, classification, human uncertainty, local climate zones (LCZs), uncertainty quantification (UQ).

I. INTRODUCTION

OVER the past years, deep learning has had a tremendous impact on many research fields across almost all domains, and remote sensing is no exception. Deep neural networks have enhanced the precision and accuracy of task-solving models by large margins. Concurrent advancing computational power of modern hardware has enabled such

complex models to be trained while using continually shrinking time and resources. While the overall performance can be pushed by ever more complex models, the reliability of the resulting predictions is often neglected. Especially for image classification tasks, where prediction takes the form of a probability distribution over a set of possible classes, it is of crucial importance to rely on the model's confidence in its predictions. A common characteristic regarding the labels is label noise [1], which describes the pollution of the labels from potentially various sources.

Another aspect of label quality is ambiguity among the classes, which can occur naturally in the cases of, for example, multilabel classification or head pose estimation. As a proposed solution to deal with label ambiguity, label distribution learning [2] was introduced. The framework combines the idea of labels taking the form of a distribution across the space of the different possible classes, and a suitable loss function to learn this distribution.

Opening up the learning task to distributional labels can be highly beneficial for many applications. For safety-critical fields such as medical image analysis [3] in particular, predictions are required to be well-calibrated, while there is occasionally inevitable human label uncertainty [4]. These aspects also apply to remote sensing data, but are hardly ever discussed. Label ambiguity is a known but rarely tackled problem, where examples of exception are given by using the ambiguity information for synthetic aperture radar (SAR) image segmentation [5] or the application of label distribution learning toward aerial scene classification [6]. In this work, we focus on the classification of local climate zones (LCZs) from satellite images. LCZs are here adopted to allocate urban conglomerates and their environment around the world into 17 different clusters [7], which is helpful, for e.g., identifying potential urban heat islands (UHIs) or urban planning.

In a recently published work [8] the So2Sat LCZ42 dataset was introduced as a new benchmark dataset for which a label confidence of 85% was stated. We study this dataset, which contains satellite images of European cities as well as additional urban areas around the globe. As a peculiarity, for each image, we are presented with ten label votes from human remote sensing experts. These votes do not coincide for many images, which reflect the inherent human uncertainty about the labels in the data and which show the difficulties entangled with the labeling process. Our novel proposal is to embed this human uncertainty during the training of a

Manuscript received 13 February 2023; revised 1 July 2023 and 13 September 2023; accepted 20 October 2023. Date of publication 12 December 2023; date of current version 15 December 2023. This work was supported in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Lab “AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” under Grant 01DD20001; in part by the Helmholtz Association under the Joint Research School Munich School for Data Science—MUDS; in part by the German Federal Ministry for Economic Affairs and Climate Action in the framework of the “National Center of Excellence ML4Earth” under Grant 50EE2201C; and in part by the Munich Center for Machine Learning. (Corresponding author: Xiao Xiang Zhu.)

Christoph Koller is with the Department of Data Science in Earth Observation, Technical University Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany.

Göran Kauermann is with the Department of Statistics, LMU Munich, 80539 Munich, Germany.

Xiao Xiang Zhu is with the Department of Data Science in Earth Observation, Technical University Munich, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2023.3336357

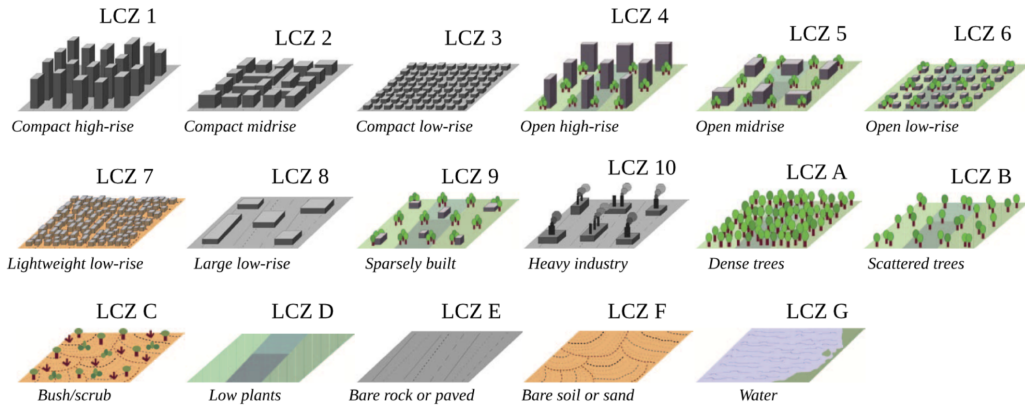


Fig. 1. LCZ classification scheme for the So2Sat LCZ42 dataset as shown by [9]. Classes 1–10 are urban areas, classes A–G are vegetation zones.

neural network classifier and investigate its performance and confidence compared to using traditional one-hot encoded labels. Our experiments send a clear message: Explicitly incorporating the inherent human uncertainty into the training process of the model is highly beneficial for both model performance and the calibration of predictive probabilities.

II. RELATED WORK

Identifying climate zones is a common task when gathering information regarding land cover from satellite images. A popular and widely-used scheme is termed LCZs [7]. The different classes are defined to describe urban areas and their surroundings. Originally published for the evaluation of UHIs [10], [11], the scheme has since then been widely used for various climatological applications [12], [13] and urban planings [14], [15], [16]. The community-driven project termed World Urban Database and Portal (WUDAPT) [17] targets a global high-quality coverage with the LCZ scheme. So far, significant effort has been shown to reach this goal [18], [19], [20], [21]. With the advancements of deep learning in the remote sensing community, more complex models have been established in order to classify processed satellite imagery into LCZs. Data from both the Landsat [22], [23] and the Sentinel satellite missions [24], [25] have been successfully employed for this task. More recently, a new benchmark dataset [8] was introduced to the community. For a scope of 42 global cities, high-quality satellite image patches are freely provided with manually crafted LCZ labels. Recent works based on this benchmark dataset explore various network architectures [26] or fuse multiple data sources [27], [28], [29]. Deep learning models trained on this benchmark have also been applied to achieve global urban LCZ maps for a better understanding of the global urban morphology [30]. As for land cover classification in general, knowledge graphs have been successfully adapted recently [31]. In particular, for zero-shot classification, where previously unseen data is classified, several works have utilized knowledge graphs [32], [33].

The goal of uncertainty quantification (UQ) in the field of deep learning lies in building a model that not only provides a prediction but also a measure of certainty or confidence [34]. Generally, we can distinguish between epistemic uncertainty,

which is caused by the model, and aleatoric uncertainty inherent in the data [35]. Epistemic uncertainty can be reduced by finding a more suitable model or architecture, whereas aleatoric uncertainty is irreducible. Methodological directions to estimate uncertainty quantities include the aggregation of multiple neural networks [36], deterministic networks with distributional assumptions placed on the label space [37], [38], sophisticated use of dropout networks [39], or the Bayesian neural networks [40], [41]. Domain-specific applications in the remote sensing area are still rare [42], [43].

Calibration is a closely related concept, which in the context of classification tasks aims at providing more reliable predictions. In particular, the probabilities derived from a machine learner should adequately describe the certainty inherent in the prediction. Originally proposed for support vector machines [44], Platt scaling describes a parametric method that trains an additional layer to transform the predictions of a classifier into calibrated probabilities. A simplified version termed temperature scaling (TS) has found its way into more recent deep learning-based works [45], [46], [47]. Using only a single parameter, the approach simply scales the softmax probabilities derived from a deep learner in order to limit overconfidence. Another famous technique is termed label smoothing (LS) [48] and helps to overcome overconfident predictions by artificially changing the labels during training. Several recent works [49], [50], [51] have investigated this method since.

As a novelty for the deep learning community, human uncertainty, derived from the labeling stage of the data generation process, has been studied [52]. Specifically, a set of images was labeled by multiple people and therefore received votes for potentially different classes. This rich information was then added to the training process in order to capture human categorization peculiarities [53] or make the underlying classification task more robust [52]. Comparisons to more classic calibration techniques such as LS were drawn in [54]. Though closely linked, the incorporation of this additional information into the labels stands in contrast to the widely studied research direction of investigating label noise [1], [55], [56], which focuses more on identifying mislabeled or anomalous data, or pointing out insufficiently labeled data. In the field of

TABLE I

CITIES AND ADDON AREAS IN THE EVALUATION SUBSET OF THE So2Sat LCZ42 DATASET

Cities		Addon Areas	
Amsterdam	Berlin	Guangzhou	Islamabad
Cologne	London	Jakarta	Los Angeles
Madrid	Milan	Moscow	Mumbai
Munich	Paris	Nairobi	Riodejaneiro
Rome	Zurich		

TABLE II

MANUAL SPLIT OF THE CITIES AND ADDON AREAS OF THE EVALUATION SET INTO TRAINING DATA AND NONTRAINING DATA, WHICH WAS IN A SECOND STEP SPLIT RANDOMLY BY HALF INTO VALIDATION AND TESTING DATA

	Cities		Addon Areas	
Training Data	Amsterdam	Berlin	Guangzhou	Islamabad
	Cologne	London	Jakarta	Nairobi
	Milan	Rome		
Validation / Test Data	Madrid	Munich	Los Angeles	Moscow
	Paris	Zurich	Mumbai	Riodejaneiro

remote sensing, the first advances in investigating human uncertainty have been made recently [57], [58].

III. DATA AND METHODOLOGY

A. Satellite Data

As a basis of this work, the So2Sat LCZ42 dataset [8] is analyzed, comprising approximately 400k labeled image patches of size 32×32 pixels linked each to an area of 320×320 m. The labels follow the classification scheme introduced by [59]: 17 characteristic climate zones, of which ten are defined as urban zones and seven are vegetation zones (see Fig. 1). In the publicly available version,¹ three splits for training, validation, and testing can be chosen: A completely random split of the data, a split at the city level where each city is separated into geographically separated training and testing data, and a third approach that sets aside ten cities from different cultural zones for testing. The labeling process was performed in a manual and labor-intensive process, which is explained in more detail by [8] and largely follows a classic procedure initially used in the World Urban Database (WUDAPT) project [60]. Overall, 13 spectral bands of the Sentinel satellite mission with varying spatial resolutions are available from the Copernicus Hub. For the LCZ data, all four bands with a ground sampling distance (GSD) of 10 m were chosen, as well as the bands with a GSD of 20 m, which were upsampled to 10 m GSD.

As an additional experiment carried out in their work, Zhu et al. [8] launched an evaluation phase to assess the quality of the labeling process. For this, a subset of ten European cities was chosen to serve as reference data for the labelers; nine additional non-European regions were included to ensure a minimum level of class balance. Overall, the evaluation dataset contains roughly 250k satellite image patches, which

were cropped out of polygons from homogeneous regions. The entire list of cities, as well as add-on areas for which the label evaluation was performed, is given in Table I.

In order to geographically separate training and testing data, the corresponding datasets were formed by mutually exclusive subsets of the above cities and addon areas. The split on a city level was specifically chosen not to avoid learning the overall similar data distribution, as all images come from European cities, but rather to train on one set of cities and predict as well as evaluate the method on another set of cities. In Table II, the cities in the respective datasets are listed. Note that for the separation of validation and testing data, a random split of the cities and addon areas was performed, halving the entirety of data points into validation and testing data.

Regarding the choice of cities, next to the geographical separation, a strong emphasis was laid on balancing the occurring class frequencies in between the datasets. For the investigated urban classes (LCZs 1–10), the class frequencies of the different datasets relative to the number of samples in the entire dataset can be found in Table III. Note that for the ground-truth label, here the majority vote $y_{\max}$ as later defined in (3) was taken. Images with incomplete label distributions, that is, those having a majority vote from the urban classes and one or more individual votes from the nonurban classes occur very rarely ($\sim 0.1\%$). The outlier votes have been excluded from the analysis. As can be deduced from Table III, the classes are not perfectly balanced among the training, validation, and test datasets. Yet a significant effort was spent on finding a good split on city level that still retains a moderate level of class imbalance. This imbalance among the datasets adds complexity to the imbalance among the different classes already present in the data.

B. Human Uncertainty

As part of the label evaluation study, a group of ten human remote sensing experts independently cast a label vote for each of the polygons, resulting in a final dataset with ten expert votes for each image. This dataset was made public accompanying this publication and can be downloaded online.² Transforming these expert votes into suitable labels for a classification task can be handled in a variety of ways. Thus, let $\mathbf{Y} = \mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}$ initially be the vote count vectors for the images $i = 1, \dots, n$, where $\mathbf{Y}^{(i)} = (Y_1^{(i)}, \dots, Y_K^{(i)})$ stores the vote counts for each of the K LCZ classes for image i where $K = 17$. By indexing the different remote sensing experts via $j = 1, \dots, J$, for image i we receive the expert votes

$$V_1^{(i)}, \dots, V_J^{(i)}, \quad V_j^{(i)} \in \{1, \dots, K\} \quad \forall i = 1, \dots, n. \quad (1)$$

We define the votes as a K -dimensional vector

$$\mathbf{V}_j^{(i)} = \left(\mathbb{1}_{\{V_j^{(i)}=1\}}, \dots, \mathbb{1}_{\{V_j^{(i)}=K\}} \right), \quad \text{where} \quad \mathbb{1}_{\{V_j^{(i)}=k\}} = 1 \Leftrightarrow V_j^{(i)} = k \quad (2)$$

¹So2Sat LCZ42 dataset (link to download: <https://mediatum.ub.tum.de/1613658>).

²So2Sat LCZ42 evaluation dataset (link to download: <https://mediatum.ub.tum.de/1659039>).

TABLE III
RELATIVE CLASS FREQUENCIES OF THE URBAN CLASSES WITHIN THE TRAINING, VALIDATION AND TESTING SET. TOTALS ARE LISTED IN THE RIGHT AND BOTTOM ENTRIES

	Urban LCZ										Σ
	1	2	3	4	5	6	7	8	9	10	
Training Data	0.83	0.56	0.87	0.61	0.83	0.74	0.65	0.74	0.65	0.94	43k
Validation Data	0.09	0.22	0.07	0.20	0.08	0.13	0.18	0.13	0.17	0.03	8k
Test Data	0.09	0.22	0.07	0.20	0.08	0.13	0.18	0.13	0.17	0.03	8k
Σ	< 1k	10k	2k	1k	11k	18k	3k	9k	1k	3k	59k

and obtain the vote counts via $Y_k^{(i)} = \sum_j \mathbb{1}_{\{V_j^{(i)}=k\}}$. In particular, $Y_k^{(i)} = m$ means that for image i class k received m votes and it holds that $\sum_{k=1}^K Y_k^{(i)} = M$, where $M = 10$ represents the number of votes or experts. A common strategy is to rely on the majority vote of the experts that we define for image i as $Y_{\max}^{(i)} := \max_j Y_j^{(i)}$ to be the class which received the most expert votes. The associated one-hot encoded label is denoted as a K -dimensional vector

$$\mathbf{y}_{\max}^{(i)} = \left(\mathbb{1}_{\{Y_1^{(i)}=Y_{\max}^{(i)}\}}, \dots, \mathbb{1}_{\{Y_K^{(i)}=Y_{\max}^{(i)}\}} \right), \quad \text{where} \\ \mathbb{1}_{\{Y_j^{(i)}=Y_{\max}^{(i)}\}} = 1 \Leftrightarrow Y_j^{(i)} = Y_{\max}^{(i)}. \quad (3)$$

This simplification gives rise to the question of how much information is lost when relying on the majority label decision. It is argued by [8], that majority voting does help to improve label confidence. Yet part of the uncertainty in the voting process is hidden from the classifier when it is presented with the majority vote $\mathbf{y}_{\max}$ in (3). An alternative approach is hence to incorporate the entirety of all votes directly into the label by forming a distributional label. To do so, we directly use the empirical distribution formed by the observed votes and define the soft label for image i in the following discussions via

$$\mathbf{y}_{\text{distr}}^{(i)} = \mathbf{Y}^{(i)} / M. \quad (4)$$

By doing so, the mode of the distributional label coincides with the mode of $\mathbf{y}_{\max}$ (3), but the distributional form allows for a more flexible learning approach, as other classes that were voted for in the evaluation process are also considered.

The ten label votes received from each of the ten independent remote sensing experts inherently store a notion of label uncertainty. This uncertainty is visualized in Fig. 2 by means of a confusion matrix between the individual label votes $V_1^{(i)}, \dots, V_J^{(i)}$ and the majority vote $\mathbf{y}_{\max}^{(i)}$ for all images $i = 1, \dots, n$, as well as a plot showing the entropies of the individual label distributions. In detail, for the given label distribution $\mathbf{y}_{\text{distr}}^{(i)}$ [see (4)], we compute the information theoretic (Shannon) entropy of the distribution given for image i via

$$H(\mathbf{y}_{\text{distr}}^{(i)}) = - \sum_{k=1}^K \mathbf{y}_{\text{distr},k}^{(i)} \log \mathbf{y}_{\text{distr},k}^{(i)} \quad (5)$$

and plot the resulting values in a bar plot grouped by the affiliation of the respective majority vote $\mathbf{y}_{\max}^{(i)}$ [see (3)] to

either the urban or nonurban classes. Whereas zero entropy occurs when all voters agree on a label for an image, maximum entropy is achieved for a uniform distribution across all labels (which cannot occur here because the number of classes exceeds the number of voters). Clearly visible is the higher average entropy of the vote vectors for the urban classes, which corresponds to a higher uncertainty associated with the respective satellite images. We chose to limit our analysis to the urban classes due to the large share of images with the majority vote belonging to nonurban classes that at the same time have zero entropy in the votes.

C. Training Process

In the following, let $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1,\dots,n} \in (\mathcal{X} \times \mathcal{Y})^n$ be the classification data: $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathcal{X}$ are the multispectral LCZ42 image patches, and $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \in \mathcal{Y}$ are the corresponding labels. Furthermore, let $f_{\theta}(x)$ be a neural network classifier based on the parameters stored in θ . Given an input $x \in \mathcal{X}$, the predictive distribution of the network is denoted by $p_{\theta}(y|x)$, and $p_{\theta}(y = k|x)$ returns the estimated probability of x belonging to class k . For training, typically the cross-entropy loss is used in a classification setting. From an information theory perspective, the cross-entropy defines the amount of additional information needed to approximate a sample from the source distribution.

However, in the literature for the recently presented research field of label distribution learning, the Kullback-Leibler (KL) divergence has been established as a loss function [2]. Given a true distribution, the KL divergence measures the information lost when the target distribution is used as an approximation. It measures the dissimilarity between two probability functions and is therefore also termed relative entropy.

If we now assume for the So2Sat LCZ42 data that the distribution formed by the votes of the ten remote sensing experts is the ground truth label distribution, we can measure the dissimilarity between the predicted and the ground truth distribution with the KL divergence. When implemented as a loss function, it is derived for a batch of data $\{\mathbf{x}^{(i)}, \mathbf{y}_{\text{distr}}^{(i)}\}_{i=1,\dots,m}$ via

$$\mathcal{L}_{\text{KL}}(f_{\theta}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}, \mathbf{y}_{\text{distr}}^{(1)}, \dots, \mathbf{y}_{\text{distr}}^{(m)}) \\ = -\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K \mathbf{y}_{\text{distr},k}^{(i)} \cdot \log \frac{\mathbf{y}_{\text{distr},k}^{(i)}}{p_{\theta}(\mathbf{y}^{(i)} = k|\mathbf{x}^{(i)})}. \quad (6)$$

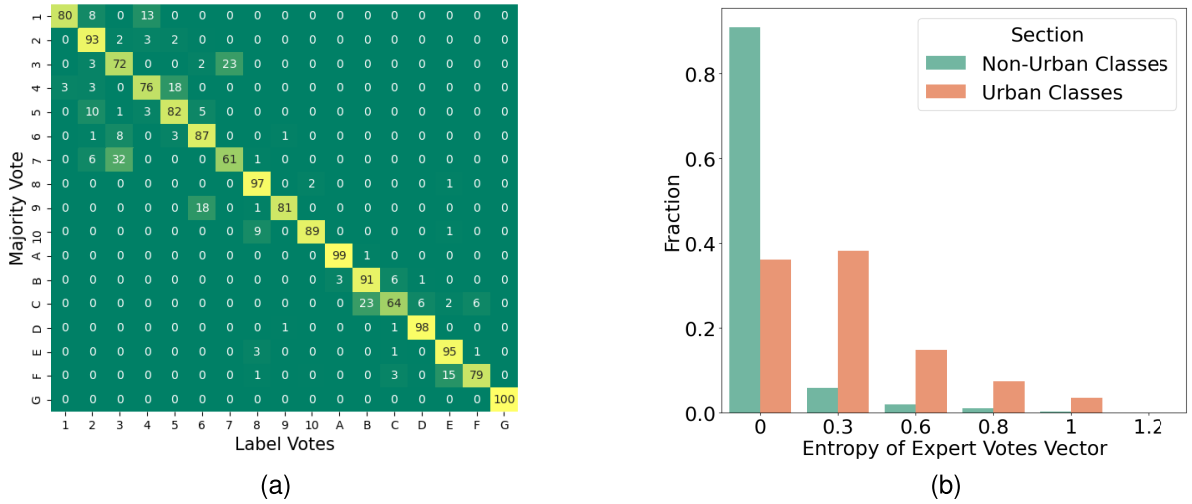


Fig. 2. (a) Confusion matrix of label votes for the evaluation dataset. (b) Entropies of the voting distributions.

Note that for $y_{\text{distr},k}^{(i)} = 0$ we set $y_{\text{distr},k}^{(i)} \cdot \log y_{\text{distr},k}^{(i)} / (p_{\theta}(\mathbf{y}^{(i)} = k | \mathbf{x}^{(i)})) \equiv 0$. This approach is taken to measure the information loss when using the predictive distribution of the neural network to approximate the assumed ground truth distribution over labels. Consequently, training with the KL divergence describes the process of iteratively finding a neural network mimicking human voting behavior.

D. Calibration and Generalization

Neural network classifiers are often prone to overconfidence, namely that predicted probabilities for a class overestimate the percentage of times the algorithm actually yields a correct prediction [46]. In this circumstance, derived uncertainty quantities are not reliable, because the underlying probabilities are ill-defined in the first place. Therefore, one can speak of a frequentist notion of uncertainty when referring to calibration [36]. Deviations from the perfectly calibrated model can be measured with different error rates, of which the expected calibration error (ECE) is the most prominent one. It can be visualized in a so-called reliability diagram [61], [62], which displays the accuracy versus the corresponding confidence in a single plot, averaged over predefined intervals. Following the notation of [46], we denote these bins by $B_m = ((m-1)/M, (m/M)]$, $m = 1, \dots, M$ and define the indices of the images whose confidences $\hat{p}^{(i)}$ fall into the respective bin by I_m , $m = 1, \dots, M$. The needed quantities are then calculated by

$$\text{acc}(I_m) = \sum_{i \in I_m} \mathbb{1}_{\{\hat{y}^{(i)} = y_{\text{max}}^{(i)}\}} \quad \text{and} \quad \text{conf}(I_m) = \sum_{i \in I_m} \hat{p}^{(i)} \quad (7)$$

and displayed in a 2-D plot with respect to the earlier defined intervals. Here, $\hat{y}^{(i)}$ is the one-hot encoded predicted class for input $\mathbf{x}^{(i)}$, $\hat{p}^{(i)}$ the corresponding predicted probability, and one defines $y_{\text{max}}^{(i)}$ as in (3). The ECE is then derived via

$$\text{ECE} = \sum_{m=1}^M \frac{|I_m|}{n} |\text{acc}(I_m) - \text{conf}(I_m)|. \quad (8)$$

A simpler version of the ECE, which only considers the maximum gap between confidence and accuracy out of all bins considered, is the maximum calibration error (MCE). Given the predictions $\hat{\mathbf{y}}^{(i)}$, $i = 1, \dots, n$ and corresponding confidences $\hat{p}^{(i)}$, $i = 1, \dots, n$, the computation is as follows:

$$\text{MCE} = \max_m |\text{acc}(I_m) - \text{conf}(I_m)|. \quad (9)$$

While the MCE gives a first indication whether the evaluated classifier is severely miscalibrated for a certain bin, the same downside appears as for the ECE, namely that the included metrics are not considered class-specific. For this purpose, Nixon et al. [63] introduced the static calibration error (SCE), which measures the accuracies and confidences on a class level. For class k and within bin m , these are depicted by $\text{acc}(m, k)$ and $\text{conf}(m, k)$, leading to the formula for the SCE given by

$$\text{SCE} = \frac{1}{K} \sum_{k=1}^K \sum_{m=1}^M \frac{n_{mk}}{n} |\text{acc}(m, k) - \text{conf}(m, k)| \quad (10)$$

where n_{bk} denotes the number of samples belonging to class k (with respect to the majority vote) within bin b . Here, the class-specific scores are also weighted with respect to the class share within the respective bin.

Presented as a straightforward and effective calibration strategy, Guo et al. [46] introduced TS. The technique can be seen as a simplification to Platt scaling [44] which scales the logits predicted by the classifier by a constant parameter. So given the predicted logit of the neural classifier $f_{\theta}(x)$ for image i via $z^{(i)}$, the corresponding softmax prediction of the scaled logit is given via

$$\text{softmax}(z^{(i)}) = \frac{\exp(z^{(i)}/T)}{\sum_k \exp(z_k^{(i)}/T)}. \quad (11)$$

As described by [46], we optimized the parameter T (termed temperature) with respect to the negative log-likelihood on the validation dataset. Adding to that method, LS represents another quick and easy to implement off-the-shelf calibration

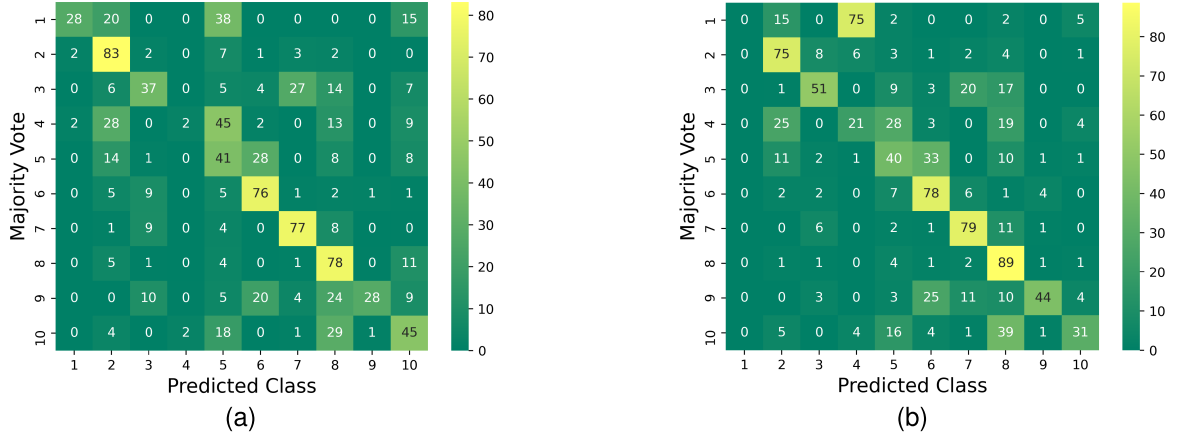


Fig. 3. Exemplary pairwise confusion between predicted class and respective majority vote for the test set. (a) One-hot encoding. (b) Label distribution encoding.

method. As opposed to TS, the method is not applied post-hoc, but uses a hyperparameter α to scale the labels before training. Given the label $y^{(i)}$ for image i , the scaled version is then received via

$$y_{\text{smoothed}}^{(i)} = \alpha \cdot u_K + (1 - \alpha) \cdot y^{(i)} \quad (12)$$

where u_K denotes the uniform distribution over the K classes.

The hyperparameter is not directly optimized, but empirically chosen. As a third calibration option, we utilized Monte Carlo Dropout [39]. Given the already trained networks, we left the dropout layers of Sen2LCZ active during prediction and by doing so created a set of 20 unique predictions. After averaging the softmax vectors of the individual predictions, we proceeded with deriving the calibration and generalization metrics as before. Lastly, a deep ensemble [36] with $k = 5$ ensemble members was considered. As proposed by the authors, five identical models were trained in parallel. Afterward, the calibration and generalization metrics were calculated based on the average of the five predictions.

Further measures of generalization as shown in Table IV are namely are the cross-entropies between the predictive distribution of the network $p_\theta(y|x)$ and the one-hot encoded label based on the majority vote y_{max} or the distributional label y_{distr} . They are termed ‘‘CE One-hot’’ and ‘‘CE Distr’’ and given for image i via

$$\begin{aligned} \text{CE}(y_{\text{max}}^{(i)}, p_\theta(y^{(i)}|x)) \\ = - \sum_{k=1}^K y_{\text{max},k}^{(i)} \log(p_\theta(y^{(i)} = k|x)) \quad \text{and} \end{aligned} \quad (13)$$

$$\begin{aligned} \text{CE}(y_{\text{distr}}^{(i)}, p_\theta(y^{(i)}|x)) \\ = - \sum_{k=1}^K y_{\text{distr},k}^{(i)} \log(p_\theta(y^{(i)} = k|x)) \end{aligned} \quad (14)$$

respectively.

IV. EXPERIMENTS

A. Setup and Settings

For the classification task, we use the model introduced by [26] (Sen2LCZ-Net), which is a modified convolutional neural

network (CNN) using intermediate deep feature representations at multiple stages of the network. These representations are then averaged and pooled at the end before being transformed into the logit space. In particular, we used a network depth of 17 (following from the use of four convolutional layers in each block), a width of 16, a dropout rate of 0.2 (at the end of the second and third block), and activated multilevel feature fusion and double-pooling. Class weights did not lead to improved results, hence they were discarded since also the class imbalance differed largely between the training, validation, and testing set. The Nesterov Adam optimizer implementation of Keras [64] was used for training. An early stopping mechanism was installed, which monitored the validation loss with a patience of 20 epochs. Weights were saved after every epoch if and only if the validation loss decreased. The model has been shown to be superior over many state-of-the-art neural network architectures in extensive benchmark tests on the So2Sat LCZ42 dataset by the authors.

LS was performed with a smoothing parameter of 0.1. TS was implemented as described by [46] via tuning the scaling parameter on the validation set with respect to minimizing the negative log-likelihood. This minimization was performed using the Adam optimizer implementation of Keras [64] with a learning rate of 0.01 and a maximum of 10k iterations (more iterations did not improve the results significantly). It is worth noting that TS scales the logits, but does not change the accuracy of the model, being the reason why we did not include the metric in Table V. Regarding the hyperparameters of Sen2LCZ, we set a batch size of 64 and an initial learning rate of 2×10^{-3} which was gradually reduced by a factor of 0.5 every five epochs.

B. Results

Sen2LCZ-Net was trained on the training set utilizing different labels and loss functions plus additional off-the-shelf calibration methods. The categorical cross-entropy loss incorporates the one-hot encoded labels based on the majority vote of the expert votes, and the KL divergence takes the distribution formed by the empirical distribution of the expert votes. All models were trained using the same training and

TABLE IV

CROSS-ENTROPIES BETWEEN PREDICTED SOFTMAX PROBABILITIES AND LABELS ON THE TEST SET AS WELL AS CALIBRATION ERRORS, AVERAGED OVER FIVE RUNS. CE = CROSS ENTROPY, LS = LABEL SMOOTHING, TS = TEMPERATURE SCALING, MC-DROP = MONTE CARLO DROPOUT. BINNING WAS PERFORMED USING 20 EQUALLY-SIZED BINS. ENSEMBLES TREATED SEPARATELY AS THEY WERE NOT PERFORMED MULTIPLE TIMES

	CE One-hot ↓	CE Distr. ↓	ECE ↓	MCE ↓	SCE ↓
One-hot	1.12 ± 0.05	1.38 ± 0.07	9.79 ± 3.18	23.14 ± 3.97	1.03 ± 0.50
+ LS	1.05 ± 0.01	1.23 ± 0.03	7.33 ± 2.62	19.80 ± 4.82	1.11 ± 0.25
+ TS	1.00 ± 0.13	1.17 ± 0.07	4.15 ± 2.37	15.88 ± 10.60	1.44 ± 0.22
+ LS & TS	1.02 ± 0.03	1.18 ± 0.02	3.21 ± 0.96	11.30 ± 4.26	1.32 ± 0.03
+ MC-Drop	1.12 ± 0.05	1.37 ± 0.06	9.57 ± 3.11	23.10 ± 4.22	1.04 ± 0.49
+ LS & MC-Drop	1.05 ± 0.01	1.23 ± 0.03	7.11 ± 2.41	33.22 ± 26.60	1.12 ± 0.24
Distr.	1.06 ± 0.07	1.21 ± 0.07	5.80 ± 1.07	15.57 ± 4.22	1.21 ± 0.20
+ LS	0.98 ± 0.03	1.08 ± 0.02	8.31 ± 2.46	17.32 ± 4.84	1.73 ± 0.06
+ TS	0.96 ± 0.09	1.07 ± 0.07	5.89 ± 2.50	15.37 ± 3.94	1.72 ± 0.15
+ LS & TS	0.95 ± 0.04	1.05 ± 0.04	4.21 ± 1.38	15.35 ± 1.95	1.58 ± 0.10
One-hot ensemble	0.84	1.03	4.83	16.70	1.74
Distr. ensemble	0.89	1.01	7.96	23.12	1.68

validation data. Identical hyperparameters were used (after grid search), and the same stopping and convergence criteria were set. In particular, the same input in the form of satellite imagery was used in both models.

Exemplary confusion matrices of the predictions on the test set are presented for both models in Fig. 3. Here, the ground truth label is set to be the majority vote of the experts, and the predicted class is defined via the highest predicted probability of the respective network $\hat{p} = \max_k p_\theta(y = k|x)$. First note that for most of the part, both models have the same pitfalls of misclassifying certain classes. This holds in particular for class 3 (compact low-rise) being falsely classified as class 7 (lightweight low-rise), class 4 (open high-rise) being misclassified as class 2 (compact midrise), as well as class 5 (open midrise) being classified as class 6 (open low-rise). Adding to this claim is the strong confusion with class 8 (large low-rise), which is often falsely predicted for images from various classes (with respect to the majority vote).

In a similar manner, images for which experts agreed on class 9 (sparsely built) are often falsely classified as classes 6 and 8. The results regarding the majority-voted class 1 seem rather arbitrary, however, as they largely depend on the data distribution of the few test samples. In particular, there are only around 30 samples of class 1 (with respect to the majority vote) in the test set, of which none could be correctly classified using the distributional approach. Note, however, that the high confusion with class 4 is also visible to some extent in the human label uncertainty. It could furthermore be the case that those particular samples where this confusion occurs in the data have mostly ended up in the test set. Of particular interest is furthermore the comparison between the model confusion and the confusion among voters. With one in every three votes deviating from the majority vote, experts chose to vote for class 3 in cases where the majority of experts settled for class 7. This large confusion is only slightly reflected in model confusion.

The opposite case is matched rather closely in the predictions of both models on the test set. With more than

TABLE V

PERFORMANCE SCORES ON THE TEST SET. ACCURACY AND OTHER RELATED MEASURES WERE DERIVED WITH RESPECT TO THE MAJORITY VOTE. ALL SCORES ARE AVERAGED OVER FIVE RUNS. OA = OVERALL ACCURACY, MAA = MACRO AVG. ACCURACY, WAA = WEIGHTED AVG. ACCURACY, κ = KAPPA SCORE, LS = LABEL SMOOTHING

	OA ↑	MAA ↑	WAA ↑	κ ↑
One-hot	68.4 ± 5.5	42.9 ± 6.4	69.6 ± 2.2	60.2 ± 6.4
+ LS	67.5 ± 2.4	50.3 ± 3.5	69.9 ± 2.0	59.4 ± 2.7
Distr.	67.0 ± 2.2	45.8 ± 3.9	71.0 ± 0.5	58.8 ± 2.4
+ LS	68.6 ± 2.3	43.4 ± 6.1	69.7 ± 2.1	60.4 ± 2.8

one in every six votes deviating from the majority vote, the combinations of classes 4 and 5, as well as 9 and 6, are also found to be confusing for the model in both training cases. However, in the cases of classes 4 and 5, it seems to be significantly less confusing for the model trained with distributional labels. Further deviations of the model confusion from the human confusion can be partially attributed to the previously mentioned arbitrariness of the very low number of affected samples in the data. All other model confusions cannot be directly linked to the confusion in the voting process and therefore result from difficulties entirely attributable to the model.

We can deduce the results of the trained models in Table V. For the distributional label approach, accuracy was again measured with respect to majority voting, which helps to explain the observed minor differences in performance between the two model configurations. Although the distributional approach performs on average better than the regular approach in 3 out of 4 metrics, the macro average accuracy lags behind by a large margin. Regarding the deviance of predictions from the true labels, Table IV shows the cross-entropy between the two distributions on the identical test set as well as the introduced calibration error rates (based on 20 identically-sized bins). When trained with distributional labels, the model can fit better toward the test label distributions when predicting

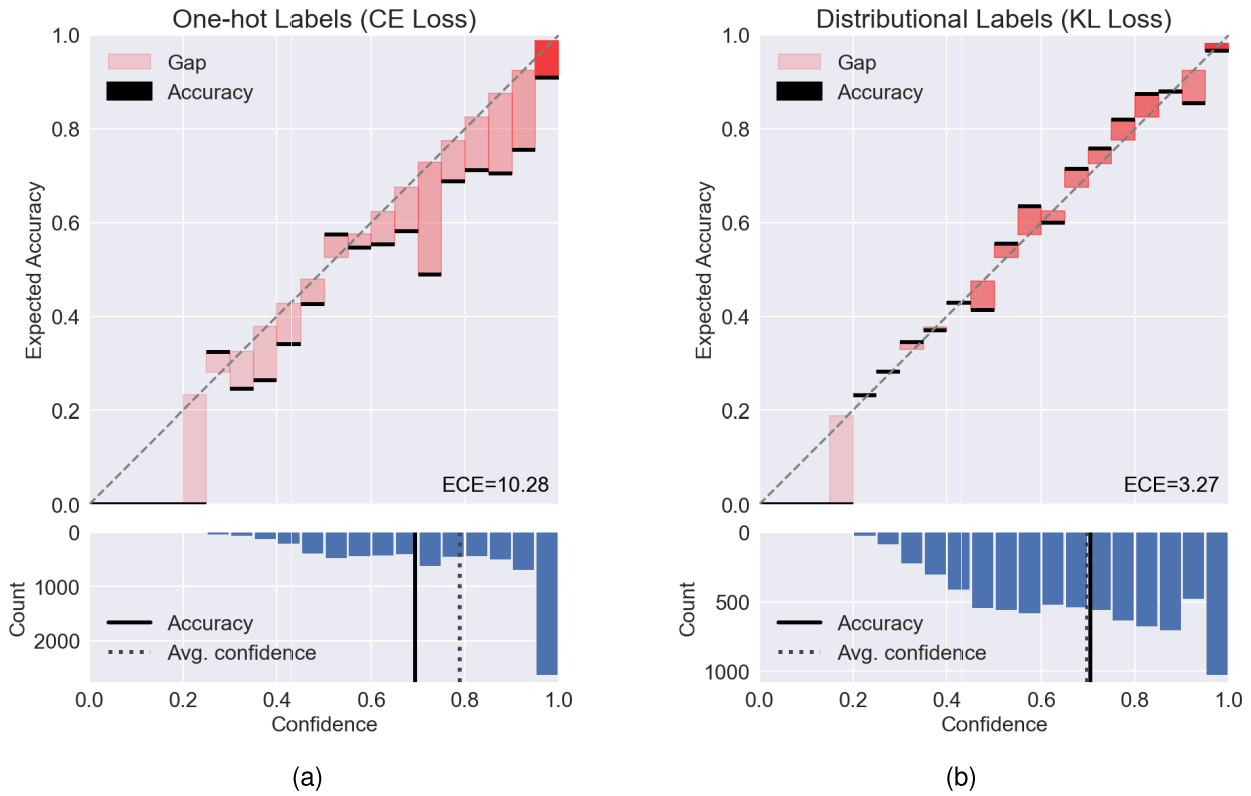


Fig. 4. Exemplary reliability diagrams on the test set. Both visualizations were created using [65]. (a) One-hot encoding. (b) Label distribution encoding.

on the test set by a large margin. This holds true even though the model with distributional labels was trained with the KL divergence as loss.

While LS helps to improve the performance in both settings, TS leads to mixed calibration performance results. The additional use of Monte Carlo Dropout was enforced via averaging the confidences of multiple predictions with activated dropout. While this enabled to bind the ECE to a reasonable level, the generalization performance was nearly unaffected. The use of an ensemble of classifiers, on the other hand, strongly impacted the generalization performance but worsened the calibration ability of the model. Note that we considered the ensemble results separately, as they were not averaged over multiple runs. This would have increased the overall computational demand by a multiple of the number of runs. Overall, all calibration errors benefit from the uncertainty-guided approach, although the overall best errors are achieved with conventional calibration techniques.

Yet more unforeseen is the cross-entropy with respect to the one-hot encoded labels. Although the model trained with the distributional labels uses a different loss function and different labels, it can better approximate the one-hot encoded test labels and underlines the strong generalization performance of the human uncertainty models. While the additional off-the-shelf calibration methods help to marginally improve the cross-entropies, the ECE is negatively affected by individual calibration methods in the distributional setting. An explanation for this could be the fact that the

optimization of the temperature is prone to overfitting on the validation dataset, leading to poor generalization when already considering human uncertainty within the training process. The overall better cross-entropies of the models including human label uncertainty underline a crucial aspect of training with distributional labels: If chosen appropriately, they keep the model more flexible when generalizing to the test data. When explicitly modeling label distributions, all classes with a nonzero entry in the distribution of the label are changing the value of the loss (as compared to the logit of the ground truth class in the case of cross entropy training); therefore the training is more flexible and less prone to overconfidence in its predictions.

This phenomenon is particularly visible in Fig. 4, which displays reliability diagrams [46] of a single run for the two modeling approaches in an uncalibrated setting. Clearly visible is the significantly lower ECE for the model trained with the distributional labels, especially for the regions with higher confidence. In these regions, the confidence is more meaningful and is reflected better by the accuracy, resulting in a better-calibrated model when using distributional labels. Since calibration can be treated as the frequentist notion of uncertainty [36], we can deduce that by incorporating the voting uncertainty into the training process, the resulting model is better calibrated and hence yields a predictive distribution that contains a more feasible sense of uncertainty.

Moreover, the lower part of Fig. 4 shows the overconfidence of the one-hot encoded model, where the average confidence in

the predictions exceeds the overall accuracy by a large margin. On the contrary, the accuracy is almost precisely met by the average confidence when the model has been trained with distributional labels. Note, however, that here the accuracy was again calculated by means of the majority vote when modeling the distributional labels. This can lead to underconfidence (as opposed to overconfidence in the one-hot case) because a set of images can be classified correctly even with relatively low average confidence if the label distributions are multimodal.

V. DISCUSSION

Generally speaking, for every deep learning model it is desirable to find a suitable calibration technique in order to avoid under- or overconfidence of the model in its predictions. Especially for the field of remote sensing, this has often been paid little attention to. Traditional methods such as TS [44], [62] or LS [48] train additional hyperparameters using the validation dataset for this purpose. The approach we take here is vastly different, though it achieves similar goals, and is closely related to the works of [52]. We claim that incorporating human uncertainty about the labels helps to overcome not only the poor calibration of the predicted probabilities but also the inability to generalize well on cities not seen during training. Given the special structure of the studied satellite dataset, the label votes cast by ten remote sensing experts are used to form an empirical distribution over the classes that serve as an approximation of the inherent human label uncertainty. This distribution is then embedded as a distributional label for training. Due to its theoretical properties, KL divergence is implemented as a loss function. In order to explicitly measure the impact of human uncertainty, the model architecture, as well as affiliated hyperparameters, remained unchanged.

The questions we pose while forming the distributional labels are whether this approximation (a) is sufficient and (b) has a potential benefit for modeling purposes and downstream tasks. The latter can be partially answered by the findings of this work: we see a direct improvement in terms of calibration regarding the predictive distribution when incorporating the distributional labels for the uncalibrated baseline model. Here, the ECE was on average almost halved. Also with regard to the overall quality of the approximation, the model benefited from the label distributions. An improvement of approximately 10% on average in the cross-entropy could be seen at test time. This holds for the cross-entropy both between the predicted probabilities and the ground truth label distributions and between the probabilities and the one-hot encoded labels. We see huge potential in downstream tasks that benefit from well-calibrated probabilities, for example in the presence of ambiguous or hand-crafted labels.

As for the first question posed, one has to bear in mind that labeling images is a labor-intensive process. This holds especially true when there is a need for experts in the field of remote sensing. A natural question is therefore whether it would be more beneficial to have a larger number of images or individual votes overall. We see this question as a very promising research direction and will leave it for future work. Adding to that, we would like to note that we see potential of the approach also for more fine-grained applications in

remote sensing such as pixel-level segmentation. In these tasks, label ambiguity is a much more complex problem, as the transition between classes is not always indicated by a clear line. Similarly, a patch that has a single label in a classification setting can have multiple labels in a segmentation setting. A better-calibrated model with more reliable uncertainty estimates would therefore be particularly helpful to represent the underlying ambiguity. To the best of our knowledge, no large-scale dataset with multiple human annotations exists in this domain. Although this would come with a much higher annotation cost, we see many benefits in creating such a dataset for the aforementioned reasons.

In conclusion, we see clear benefits of the proposed approach, namely stronger generalization performance and better calibrated predictive probabilities, hence more trustworthy resulting uncertainties. At the same time, the approach is currently limited to datasets with multiple annotations per image, which are very costly to obtain. Note also that ideally the labeling should be done by domain experts so that the resulting labels reflect reliable human label uncertainty.

VI. CONCLUSION

In this work, we examine the impact of employing human uncertainty in the classification process of satellite images into LCZs. We apply our methodology to the So2Sat LCZ42 data, covering several large European cities as well as additional areas from all over the world. Ten expert label votes are supplied for each satellite image, which comprise a notion of uncertainty within the classification task. Forming a label distribution from these votes allows us to directly implement the human uncertainty linked to the voting process into the network training. To do so, we employ the framework of label distribution learning, which enables the model to better adapt to the uncertainty rooted in the labels.

When label uncertainty is incorporated into training, an improvement in the generalization performance of the model can be measured for the remote sensing dataset studied. The overall loss when generalizing to the test data was reduced on average by a margin of approximately 10%. Off-the-shelf calibration methods such as LS or TS lead to improved performance competitive to the uncertainty-guided approach, yet while requiring additional data for hyperparameter tuning. The ECE, a key measure of calibration quality, is almost halved in the majority of experiments by means of the embedded human label uncertainty when no further calibration technique is applied. The improved calibration and generalization performance of the uncertainty-guided approach can be directly converted to a more feasible notion of predictive uncertainty. This is because the predictive probabilities of the network when generalizing to the test data, yield more reliable estimates of the label distributions associated with the test data. We see promising applications in the remote sensing field that could benefit from the explicit incorporation of human (label) uncertainty into the training process.

REFERENCES

- [1] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

- [2] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.
- [3] S. K. Zhou et al., "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.
- [4] Y. Dgani, H. Greenspan, and J. Goldberger, "Training a neural network based on unreliable human annotation of medical images," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 39–42.
- [5] F. Wang, Y. Wu, P. Zhang, Q. Zhang, and M. Li, "Unsupervised SAR image segmentation using ambiguity label information fusion in triplet Markov fields model," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1479–1483, Sep. 2017.
- [6] J. Luo, Y. Wang, Y. Ou, B. He, and B. Li, "Neighbor-based label distribution learning to model label ambiguity for aerial scene classification," *Remote Sens.*, vol. 13, no. 4, p. 755, Feb. 2021.
- [7] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorological Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.
- [8] X. X. Zhu et al., "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.
- [9] T. Samsonov and K. Trigub, "Towards computation of urban local climate zones (LCZ) from openstreetmap data," in *Proc. 14th Int. Conf. GeoComputation*, Leeds, U.K., 2017, pp. 4–7.
- [10] G. Thomas, A. P. Sherin, S. Ansar, and E. J. Zachariah, "Analysis of urban heat island in Kochi, India, using a modified local climate zone classification," *Proc. Environ. Sci.*, vol. 21, pp. 3–13, Jan. 2014.
- [11] P. J. Alexander and G. Mills, "Local climate classification and Dublin's urban heat island," *Atmosphere*, vol. 5, no. 4, pp. 755–774, 2014.
- [12] I. D. Stewart, T. R. Oke, and E. S. Krayenhoff, "Evaluation of the 'local climate zone' scheme using temperature observations and model simulations," *Int. J. Climatol.*, vol. 34, no. 4, pp. 1062–1080, Mar. 2014.
- [13] J. Quanz, S. Ulrich, D. Fenner, A. Holtmann, and J. Eimermacher, "Micro-scale variability of air temperature within a local climate zone in Berlin, Germany, during summer," *Climate*, vol. 6, no. 1, p. 5, Jan. 2018.
- [14] N. G. R. Perera and R. Emmanuel, "A 'local climate zone' based approach to urban planning in Colombo, Sri Lanka," *Urban Climate*, vol. 23, pp. 188–203, Mar. 2018.
- [15] E. Lelovics, J. Unger, T. Gál, and C. Gál, "Design of an urban monitoring network based on local climate zone mapping and temperature pattern modelling," *Climate Res.*, vol. 60, no. 1, pp. 51–62, May 2014.
- [16] J. Yang et al., "Local climate zone ventilation and urban land surface temperatures: Towards a performance-based and wind-sensitive planning proposal in megacities," *Sustain. Cities Soc.*, vol. 47, May 2019, Art. no. 101487.
- [17] G. Mills, J. Ching, L. See, B. Bechtel, and M. Foley, "An introduction to the WUDAPT project," in *Proc. 9th Int. Conf. Urban Climate*, Toulouse, France, 2015, pp. 20–24.
- [18] O. Danylo, L. See, B. Bechtel, D. Schepaschenko, and S. Fritz, "Contributing to WUDAPT: A local climate zone classification of two cities in Ukraine," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 5, pp. 1841–1853, May 2016.
- [19] B. Bechtel et al., "Mapping local climate zones for a worldwide database of the form and function of cities," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 1, pp. 199–219, Feb. 2015.
- [20] B. Bechtel et al., "Generating WUDAPT level 0 data—Current status of production and evaluation," *Urban Climate*, vol. 27, pp. 24–45, Mar. 2019.
- [21] O. Brousse, A. Martilli, M. Foley, G. Mills, and B. Bechtel, "WUDAPT, an efficient land use producing data tool for mesoscale models? Integration of urban LCZ in WRF over Madrid," *Urban Climate*, vol. 17, pp. 116–134, Sep. 2016.
- [22] C. Yoo, D. Han, J. Im, and B. Bechtel, "Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images," *ISPRS J. Photogramm. Remote Sens.*, vol. 157, pp. 155–170, Nov. 2019.
- [23] M.-L. Verdonck, A. Okujeni, S. van der Linden, M. Demuzere, R. De Wulf, and F. Van Coillie, "Influence of neighbourhood information on 'local climate zone' mapping in heterogeneous cities," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 62, pp. 102–113, Oct. 2017.
- [24] J. Hu, P. Ghamisi, and X. Zhu, "Feature extraction and selection of Sentinel-1 dual-Pol data for global-scale local climate zone classification," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 9, p. 379, Sep. 2018.
- [25] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 151–162, Aug. 2019.
- [26] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multi-level feature fusion-based CNN for local climate zone classification from Sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2793–2806, 2020.
- [27] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Fusing multiseasonal Sentinel-2 imagery for urban land cover classification with multibranch residual convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1787–1791, Oct. 2020.
- [28] J. Gawlikowski, M. Schmitt, A. Kruspe, and X. X. Zhu, "On the fusion strategies of Sentinel-1 and Sentinel-2 data for local climate zone classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 2081–2084.
- [29] P. Feng, Y. Lin, G. He, J. Guan, J. Wang, and H. Shi, "A dynamic end-to-end fusion filter for local climate zone classification using SAR and multi-spectrum remote sensing data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 4231–4234.
- [30] X. X. Zhu et al., "The urban morphology on our planet—Global perspectives from space," *Remote Sens. Environ.*, vol. 269, Feb. 2022, Art. no. 112794.
- [31] X. Hao et al., "Construction and application of a knowledge graph," *Remote Sens.*, vol. 13, no. 13, p. 2511, 2021.
- [32] Y. Li, Z. Zhu, J.-G. Yu, and Y. Zhang, "Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10590–10603, Dec. 2021.
- [33] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 145–158, Sep. 2021.
- [34] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," 2021, [arXiv:2107.03342](https://arxiv.org/abs/2107.03342).
- [35] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021.
- [36] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
- [37] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7047–7058.
- [38] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3183–3193.
- [39] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
- [40] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.
- [41] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for Bayesian uncertainty in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 13153–13164.
- [42] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.
- [43] M. Rußwurm, M. Ali, X. X. Zhu, Y. Gal, and M. Körner, "Model and data uncertainty for satellite time series forecasting with deep recurrent models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 7025–7028.
- [44] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, Mar. 1999.
- [45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, [arXiv:1503.02531](https://arxiv.org/abs/1503.02531).
- [46] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.
- [47] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, and C. Gagné, "Attended temperature scaling: A practical approach for calibrating deep neural networks," 2018, [arXiv:1810.11586](https://arxiv.org/abs/1810.11586).

- [48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [49] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" 2019, *arXiv:1906.02629*.
- [50] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar, "Does label smoothing mitigate label noise?" in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6448–6458.
- [51] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3902–3910.
- [52] J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9617–9626.
- [53] R. M. Battleday, J. C. Peterson, and T. L. Griffiths, "Capturing human categorization of natural images by combining deep networks and cognitive models," *Nature Commun.*, vol. 11, no. 1, pp. 1–14, Oct. 2020.
- [54] C.-B. Zhang et al., "Delving deep into label smoothing," *IEEE Trans. Image Process.*, vol. 30, pp. 5984–5996, 2021.
- [55] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *J. Artif. Intell. Res.*, vol. 11, pp. 131–167, Aug. 1999.
- [56] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," 2017, *arXiv:1705.10694*.
- [57] C. García Rodríguez, J. Vitrià, and O. Mora, "Uncertainty-based human-in-the-loop deep learning for land cover segmentation," *Remote Sens.*, vol. 12, no. 22, p. 3836, Nov. 2020.
- [58] M. Rußwurm, S. Wang, and D. Tuia, "Humans are poor few-shot classifiers for Sentinel-2 land cover," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 4859–4862.
- [59] I. Stewart, "Local climate zones: Origins, development, and application to urban heat island studies," in *Proc. Annu. Meeting Amer. Assoc. Geographers*, Seattle, WA, USA, 2011, pp. 12–16.
- [60] J. Ching et al., "WUDAPT: An urban weather, climate, and environmental modeling infrastructure for the anthropocene," *Bull. Amer. Meteorological Soc.*, vol. 99, no. 9, pp. 1907–1924, Sep. 2018.
- [61] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *J. Roy. Stat. Soc., Ser. D*, vol. 32, nos. 1–2, pp. 12–22, 1983.
- [62] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 625–632.
- [63] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *Proc. CVPR Workshops*, 2019, vol. 2, no. 7, pp. 1–4.
- [64] F. Chollet. (2015). *Keras*. [Online]. Available: <https://keras.io>
- [65] M. Hollemans. (2020). *Reliability Diagrams*. Accessed: Jan. 10, 2022. [Online]. Available: <https://github.com/hollance/reliability-diagrams>



Christoph Koller (Graduate Student Member, IEEE) received the bachelor's degree in mathematics from Technical University Munich (TUM), Munich, Germany, in 2017, and the master's degree in statistics from LMU Munich, Munich, in 2020. He is currently pursuing the Ph.D. degree with the Chair of Data Science in Earth Observation, Remote Sensing Institute, German Aerospace Center (DLR), Weßling, Germany.

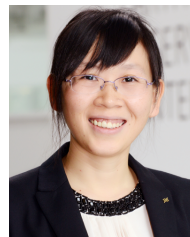
His research interests include uncertainty quantification, calibration, and out-of-distribution detection for machine and deep learning models in the context of remote sensing image classification.



Göran Kauermann received the diploma degree in economic mathematics and the Doctorate of Natural Sciences (Dr.rer.nat.) degree in statistics from Technical University Berlin, Berlin, Germany, in 1991 and 1994, respectively.

He was a Visiting Scholar/Professor at the Department of Statistics, University of Chicago, Chicago, IL, USA, and the Department of Statistics, University of New South Wales, Sydney, NSW, Australia, in 1996 and 2005, respectively. He took the role of an Assistant Professor at the Ludwig-Maximilians-University (LMU) Munich, Munich, Germany, from 1998 to 2000. After his habilitation in statistics in 2000, he served as a Senior Lecturer at the Department of Statistics, the University of Glasgow, Glasgow, U.K., before becoming a Full Professor of statistics at the Department of Economics and Business Administration, University of Bielefeld, Bielefeld, Germany, in 2003. He served as the Dean of the Faculty of Mathematics, Computer Science and Statistics, LMU Munich, from 2019 to 2021. Since 2011, he holds the Chair of Statistics—in Economics, Business Administration and Social Sciences, Department of Statistics, LMU Munich. He has been the Speaker of the Elite Master Program Data Science, LMU Munich, since 2016. His research interests include semi- and nonparametric models, generalized linear models and network data analysis.

Prof. Kauermann served as the Chair of the Deutsche Arbeitsgemeinschaft Statistik ("German Working Group for Statistics") (DAG) from 2005 to 2013 and was elected as a reviewer for Statistics and Economics at the DAG. He served as an (Associate) Editor of the *ASTA—Advances in Statistical Analysis*, *Statistical Modelling: An International Journal*, the *Biometrical Journal*, and the *Journal of the Royal Statistical Society, Series C (Applied Statistics)*.



Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011 and 2013, respectively.

She was the founding Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. From 2019 to 2022, she was a Co-Coordinator of the Munich Data Science Research School (www.mu-ds.de) and the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the PI and the Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich, Germany. Since October 2020, she has been the Director of the Munich Data Science Institute (MDSI), TUM. She is the Chair Professor for Data Science in Earth Observation, TUM. She was a Guest Scientist or Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; the University of Tokyo, Tokyo, Japan; and the University of California, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor at ESA's Phi-Lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, UN's SDGs and climate change.

Dr. Zhu has been a member of Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ, from 2020 to 2023) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, *Pattern Recognition* and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.

A.2 Can Uncertainty Quantification Benefit From Label Embeddings? A Case Study on Local Climate Zone Classification

Reference

C. Schweden, K. Hechinger, G. Kauermann and X. X. Zhu. Can Uncertainty Quantification Benefit From Label Embeddings? A Case Study on Local Climate Zone Classification. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-14, 2025. doi: 10.1109/TGRS.2025.3562233

Copyright

Article published in the IEEE Transactions on Geoscience and Remote Sensing under a CC-BY-4.0 license. Reproduced with friendly permission from the authors.

Can Uncertainty Quantification Benefit From Label Embeddings? A Case Study on Local Climate Zone Classification

Christoph Schweden¹, Graduate Student Member, IEEE, Katharina Hechinger²,
Göran Kauermann, and Xiao Xiang Zhu³, Fellow, IEEE

Abstract—Modern deep learning models have achieved superior performance in almost all fields of remote sensing. An often neglected aspect of these models is the quantification and evaluation of predictive uncertainties. Regarding a classification task, this means that the focus of the analysis solely lies on performance metrics such as accuracy or the loss. On the other hand, a notion of uncertainty indicates the model's indecisiveness among the given classes and is essential to understand where the model struggles to classify the data samples. In this work, three levels of uncertainty are distinguished, starting with the typical softmax pseudo-probabilities as level-1 uncertainty. As a next level, the more flexible Dirichlet framework is utilized as model output space, and hereby also, a Bayesian setting with an uninformative prior is considered. For the level-3 uncertainty, an empirical Bayes setting is incorporated where a latent embedding of the label space is iteratively estimated by the marginal likelihood of the fully parameterized label space. The estimated embeddings are then learned by the network in three different settings: Two regression losses use the embeddings directly, while the closed-form solution of the Kullback–Leibler (KL) divergence uses the embedding parameterized as a Dirichlet distribution. To assess the different levels of uncertainty, the label evaluation subset of the So2Sat LCZ42 dataset, which contains label votes from multiple remote sensing experts, is investigated. The predictive uncertainties are evaluated by means of out-of-distribution (OoD) detection and calibration performance. Overall, the embedding-based approaches show strong performance for calibration, while, for the OoD experiments, the Bayesian Dirichlet setting with an uninformative prior achieves the best performance. In conclusion, embedded labels offer a flexible framework for incorporating uncertain or ambiguous

labels into a supervised training setup. They could be highly beneficial for applications in fields such as urban planning or disaster response.

Index Terms—Calibration, deep learning, label embedding, land cover classification, uncertainty quantification (UQ).

I. INTRODUCTION

IN REMOTE sensing, the advantage of machine and deep learning methods over classical approaches has been shown in practically every aspect and problem setting. Various classes of models and deep neural network (DNN) architectures have been explored recently [2]. DNNs have achieved strong quantitative results in, for example, regression, segmentation, or classification tasks but they often come with a high computational cost of training millions or billions of parameters. Additionally, many DNN architectures represent a closed box that cannot be investigated adequately and quickly. This effectively means that, although solid performance can be achieved, the reason for and how the DNN reached its prediction cannot be explained simply. The first steps toward solving this problem have been summarized recently by [3].

Adding to that, the focus of scientific analyses often lies on the classical performance metrics. These metrics are, for example, represented by the accuracy of a model given a classification task, the mean squared error (mse) loss in the case of a regression problem, or the intersection over union (IoU) in object detection or segmentation tasks. While good performance in a classical view is essential and constantly desired, this view ignores the aspect of uncertainty. A model's uncertainty in its prediction is typically categorized broadly into two streams [4]. For one thing, aleatoric or data uncertainty exists, which naturally appears to some degree in every data source. By definition, this source of uncertainty cannot be reduced as it is inherent in the data.

Furthermore, the uncertainty of the model, termed epistemic uncertainty, quantifies all sources of uncertainty accompanying the modeling process of the data. This includes but not limited to the choice of model or DNN architecture, hyperparameter tuning, regularization or calibration techniques, and loss functions and training regimes in general. In contrast to aleatoric uncertainty, epistemic uncertainty is reducible. This means, in this particular context, that the uncertainty can be reduced by adding more informative data to the model. A reduction is desirable as lower epistemic uncertainty reflects a better

Received 14 December 2024; revised 13 March 2025; accepted 21 March 2025. Date of publication 5 May 2025; date of current version 9 May 2025. This work was supported in part by German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Laboratory “AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” under Grant 01DD20001; in part by the Helmholtz Association through the Joint Research School Munich School for Data Science (MUDS); in part by German Federal Ministry for Economic Affairs and Climate Action in the framework of the “National Center of Excellence ML4Earth” under Grant 50EE2201C; and in part by Munich Center for Machine Learning. (Corresponding author: Xiao Xiang Zhu.)

Christoph Schweden is with the Department of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany.

Katharina Hechinger and Göran Kauermann are with the Department of Statistics, LMU Munich, 80799 Munich, Germany.

Xiao Xiang Zhu is with the Department of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Munich Center for Machine Learning, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2025.3562233

choice of a modeling approach for the data at hand. In practical terms and in the context of remote sensing, one could count sensor noise, occlusion, or varying resolution toward aleatoric uncertainty. In contrast, epistemic uncertainty could be steered by the choice of model, hyperparameter optimization, and regularization, among others.

While lower predictive uncertainty estimates are generally considered better, this conclusion is not always proper. First, note that an overall lower uncertainty would effectively translate to a lower epistemic uncertainty, as the data uncertainty is irreducible for a fixed dataset. Yet, the two sources of uncertainty are not typically independent in practice, and their estimation often relies on distributional and statistical assumptions dependent on the uncertainty quantification (UQ) framework. Furthermore, the disentanglement of the two sources of uncertainty is far from trivial [5].

The next problem arises when trying to directly interpret the estimated uncertainties. Apart from the *lower is better* slogan discussed above, a direct comparison between uncertainties of two different training regimes is not possible. Hence, a quantified uncertainty estimate twice as large as another does not directly translate to one prediction being twice as uncertain as the other prediction. Therefore, uncertainty estimates, in general, can be ranked on an ordinal scale, but relative distances cannot be mapped to a feasible metric space. Due to all these shortfalls, UQ methods are almost always linked to a downstream task. The three most common use cases for the classification setting are calibration, active learning, and out-of-distribution (OoD) detection.

For this present work, the focus lies on evaluating uncertainties by means of OoD detection and calibration analysis. A land cover classification dataset with a particularly interesting label structure is investigated: each satellite image was labeled independently by 10 remote sensing experts. The information gained from the multiple annotations was already successfully used for a deep learning pipeline by [6]. In a recent work, [1] explored the idea of transforming the multiple annotations into a latent label embedding. Here, the aim is to underline the effectiveness of such latent embeddings for a deep learning pipeline. In particular, the objective is to specifically incorporate uncertainty in the label space into our learning regime. This procedure differs from many classical UQ frameworks, which either rely on statistical or distributional assumptions or induce a probabilistic notion into the training.

Our core contributions can be summarized as follows.

- 1) A novel three-level UQ scheme for classification is established, which allows for incorporating prior information into the label space.
- 2) Within the third level of UQ, the embedded labels and their uncertainties are utilized for two downstream tasks, namely calibration and OoD detection, with the aid of three different loss functions.
- 3) For the calibration analysis, strong performance of the embedding-based approaches is observed over classical learning regimes, and for OoD detection, the second level UQ method yields the best overall results.

The article is structured as follows. Section II introduces works related to this one, Section III then introduces the

methodology in use. Following this, Section IV explains the conducted experiments, which are discussed in Section V before Section VI concludes. The code for this work can be found at github.com/ChrisKo94/Learning_Distributions_LCZ

II. RELATED WORK

The often observed *overconfidence* of DNNs almost always causes unreliable uncertainty estimates, which are unfavorable for all downstream tasks. In order to overcome the model's overconfidence, several so-termed *calibration* techniques have been established [7]. These techniques range from scaling the network's output or label space over replacing the model's output mapping with a distance-aware function to making specific changes to the network's architecture. The related calibration errors are several measures that estimate the degree of overconfidence. To do so, the model confidences (generally given by the pseudo-softmax probabilities) are compared against the accuracy of the model in prespecified subsets of the data. The closer the two measures lie together, the lower the calibration error and, hence, the better the model's degree of calibration.

For the task of active learning, data samples are iteratively added to the model and labeled by a human oracle in the process. The querying function deciding which data instances to add usually relies on an uncertainty quantity. Instances with the highest predictive uncertainty get added, which promises to help the model learn most effectively from the data. Ideally, the model can quickly benefit from added data samples that the model struggled to classify initially and can achieve good test performance after a few active learning iterations. Often, multiple querying strategies are benchmarked against one another and must outperform the random baseline, in which instances are queried entirely randomly. Popular recent approaches involve, exemplarily, a stochastic version of batch acquisition [8], conformal prediction-based acquisition [9], or deep Bayesian active learning [10].

Another application of UQ lies in the field of OoD detection. Here, the model initially gets trained on a separate dataset or a specific subset of the overall dataset. In the latter case, the subset is usually defined by a subset of classes easily distinguishable from the rest of the classes. The reason for this lies in the use case for the estimated uncertainty: after being trained on a particular dataset, the model is confronted with data samples from classes or datasets it has never seen before, so-termed OoD data, at test time. The predictive uncertainty here should ideally reflect the complete unknowingness of the model.

This uncertainty is then evaluated against the predictive uncertainty on the initial test dataset, which should generally yield lower uncertainties as the model is already acquainted with the data. Generally, two streams of approaches can be considered, namely those involving OoD data at training time and those that do not require this data for training. For the first, noteworthy advances have been made by energy-based neural networks [11] and prior neural networks [12]. In contrast, for the second approach, distance-aware architectures [13] or kernel-based output mappings [14] have been proposed. For recent advances in the field of remote sensing, see [15], who

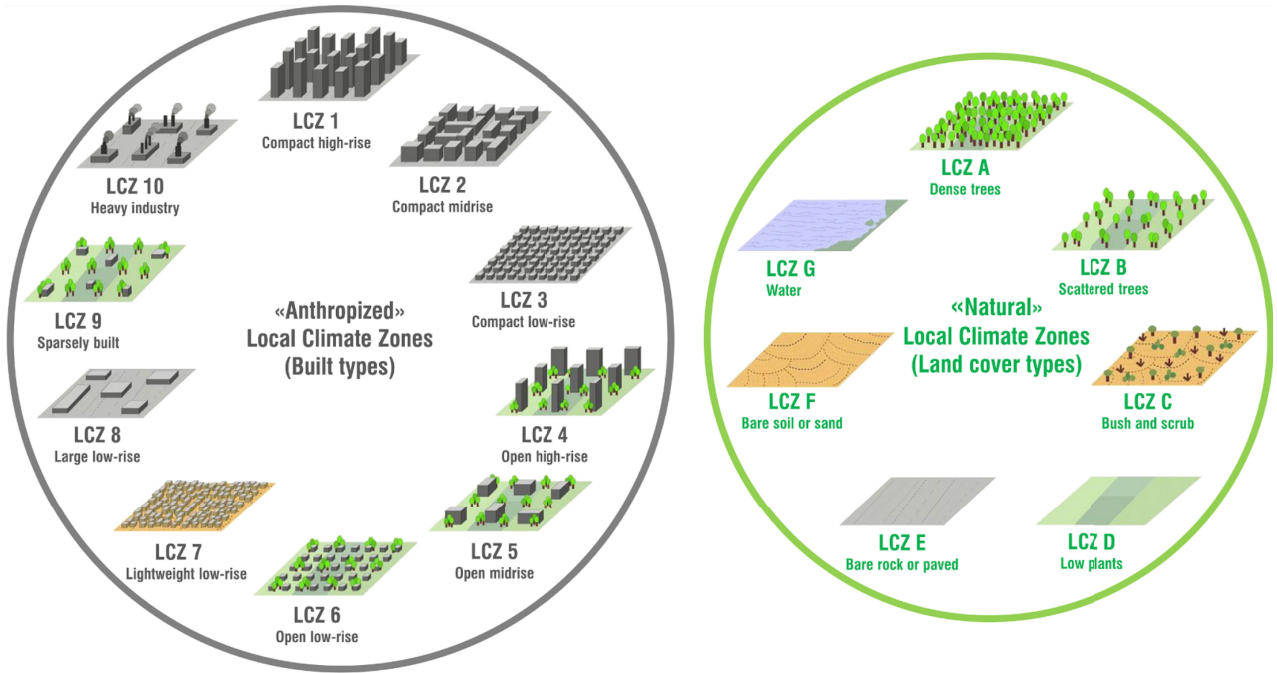


Fig. 1. Labeling scheme of the LCZs as presented in [17].

adapted prior neural networks to satellite imagery, or [16], who evaluated distance-aware neural networks for similar experimental settings.

UQ methods are quite novel in the DNN framework [18] and can generally be distinguished into three main streams. In one of them, single models are replaced by an ensemble of models [19]. The predictions of the individual models are then typically averaged and often lead to increased generalization performance and better model calibration. Next to ensembles, probabilistic methods have proven effective in improving the quality of UQ. Bayesian neural networks (BNNs) [20] are treating the network's weights as distributions, and sample from the network posterior to get a prediction. Another famous example is given by [21], who estimated predictive uncertainties by explicitly accounting for uncertainty via an additional network output node. This output, in turn, yields a posterior prediction from which the uncertainty is sampled.

A second main stream of UQ methods is given by single deterministic networks. These methods are categorized by using a single forward pass through the network for UQ with deterministic network weights. For the uncertainty to effectively measure the distance of a new instance from the training data, distance-aware modules are being built into the network. An early work termed *prior networks* explored the relaxation of the output space into a Dirichlet space and adjusting the loss function to the Kullback–Leibler (KL) divergence accordingly [12]. Similarly, Joo et al. [22] used the analytical solution to the expectation lower bound (ELBO) in a Bayesian belief matching framework, which is equivalent to minimizing the KL divergence. In the framework of deep kernel learning (DKL) [23], recent advances in deterministic UQ were, for example, made by [24], who chose a residual network (ResNet) architecture and modified the network by

adding spectral normalization to the weight matrices and an approximate Gaussian process as output mapping. Furthermore, the use of two-sided gradient penalties in combination with trainable centroids in the DNN's latent space was explored by [25]. Here, the uncertainty was derived by the exponentiated distance between the model's latent prediction and the estimated centroids.

Finally, the third stream of deterministic UQ methods is given by the evidential neural networks [26]. The main idea here lies in treating classification model predictions as evidences embedded in the research stream of subjective logic [27], see [28] for a recent survey on the topic. This very active field of research has since then spread to other domains such as regression [29] or multitask learning [30]. For regression applications in particular, however, [31] argued that evidential regression suffers from overparameterization and should, therefore, instead be treated as a heuristic rather than an exact UQ. Recent exploratory works include the application of the variance-based regression approach to a classification setting [32], the theoretical analysis of the epistemic uncertainty estimates [33], and the use of a mixture of Dirichlet distributions coupled with variational inference [34]. Recently, also hybrid approaches have been proposed, which combine the strengths of BNNs with the efficiency and adaptability of deterministic models [35], [36], [37].

In the context of remote sensing, UQ still plays a minor role overall. Only a few papers focus on the methodological investigation of uncertainties, some of which are presented exemplarily in the following. [1], [6], and [38] explored the label uncertainty arising from multiple annotations in-depth for a land cover classification dataset. For the specific task of OoD detection, [15] investigated different versions of prior networks (see [12]) and showed strong results. Similar perfor-

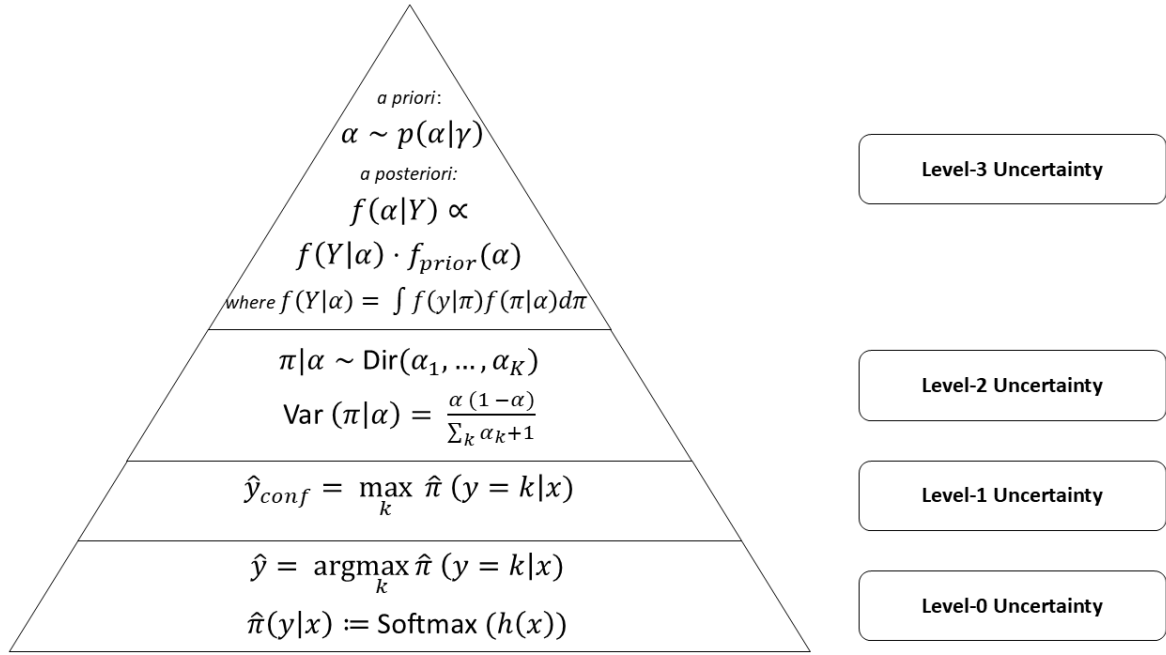


Fig. 2. Visualization of different levels of uncertainty as introduced in this work. Previously, [44] introduced similar levels of uncertainty for binary classification.

mance was achieved with distance-aware architectures [16]. The global estimation of precipitation was supported with statistical data uncertainty by [39]. In the context of cloud removal, [40] proposes a transformer-based architecture with attention and UQ modules, which yields better results and more calibrated probabilities. [41] used an uncertainty-based acquisition function for an active learning framework in the context of hyperspectral image classification. Using a BNN, [42] could effectively use UQ for corn yield prediction. For the analysis of soil moisture, [43] also reverted to uncertainty analysis.

III. METHODOLOGY

A. Data Description

The proposed methodology of learning from so-called *label embeddings* is a generalized framework for dealing with multilabel datasets. More precisely, the framework incorporates information from multiple (expert) annotations per instance by means of a Bayesian statistical model. Since this contribution is extending the work of [1], it also uses the notation introduced there. Consider a labeled (remote sensing) dataset with individual data instances $x_1, \dots, x_n$. For each image $x_i, i = 1, \dots, n$, there exist multiple annotations from independent experts $j = 1, \dots, J_i$. Each expert assigns a class $k = 1, \dots, K$ to each image (or a subset of the entire dataset), which leads to the entirety of expert votes $V_{ij} \in \{1, \dots, K\}$. When aggregating the individual expert votes into a single vector for each image i , this leads to the vote vectors.

$$Y_i = (Y_{i1}, \dots, Y_{iK}), \quad \text{where } Y_{ik} = \sum_{j=1}^{J_i} \mathbf{1}(V_{ij} = k). \quad (1)$$

The vote vectors can be aggregated into soft or hard labels in a variety of ways, which will be explained in more detail in Section IV-A. The dataset studied in this work covers a set of 42 urban conglomerates around the world and is named So2Sat LCZ42 [45]. Each data instance x_i represents a Sentinel-2 satellite image of a homogeneous region of size 320×320 m, which corresponds to 32×32 pixels (10-m ground sampling distance). A labeling scheme of 17 land cover classes was used for this dataset (see Fig. 1), namely the local climate zones (LCZs) [46]. For the purpose of investigating the reliability of this labeling scheme and process, [45] further chose a subset of 10 European cities, which 10 remote sensing experts independently labeled. This dataset is available online.¹ While an overall agreement rate of $\approx 85\%$ was found, there exists a noteworthy degree of disagreement among the individual voters.

Using this extended label information for a DNN was previously studied by [6], who transformed the vote vectors into a soft label via the empirical distribution function. An increased generalization and calibration performance was achieved with adjusted labels and loss function. With this idea in mind, a more general UQ framework with two label embedding approximations will be introduced in the following. For the latter, a different approach is taken: with the aid of statistical machine learning, a latent embedding is estimated, which captures all existing labeling information across the entire dataset. This embedding can be seamlessly transformed into the form of a Dirichlet distribution, which can be considered as a second embedding structure. These embeddings will

¹So2Sat LCZ42 evaluation dataset (link to download: <https://mediatum.ub.tum.de/1659039>)

eventually be integrated into a DNN as a core part of this work.

B. Uncertainty Framework

Before introducing the concept of label embeddings and their estimation, four different levels of uncertainty can be considered in a general classification problem, which are briefly touched in the following. Note that this differentiation builds upon previous works, in particular [44], which stated the levels for binary classification problems but already hinted at multiclass classification problems. The now following concepts are displayed in Fig. 2. Starting from the bottom, level-0 uncertainty is examined first. This level, in fact, holds no uncertainty at all, as it only considers the single class prediction of a neural network. More precisely, for a neural network h with parameters θ , let us denote the output of the network by $h_\theta(x)$ for a given input x . This output is of dimension K , corresponding to the number of classes considered. Typically, this unconstrained output is shaped into pseudo-probabilities by applying the softmax transformation as follows:

$$\hat{\pi}(y|x) := \text{Softmax}(h_\theta(x)) = \frac{\exp(h_\theta(x))}{\sum_k \exp(h_{\theta,k}(x))}. \quad (2)$$

The resulting point prediction is then given by $\hat{y} = \arg \max A_k \hat{\pi}(y|x)$ and in fact holds no notion of uncertainty. For the next level, therefore, the pseudo-probability associated with this point prediction is included, also often termed *model confidence* and given by $\hat{y}_{\text{conf}} = \max_k \hat{\pi}(y|x)$. This directly yields a level of uncertainty, often denoted via $1 - \hat{y}_{\text{conf}}$.

Often, this level of uncertainty is treated as sufficient, or one resorts to an uncertainty quantity based on the entropy of the softmax. Yet, the uncertainty's uncertainty can be inspected even further by relaxing the distributional assumption of the model's output space. Instead of a categorical distribution defined by the pseudo-softmax probabilities, the more general Dirichlet distribution is considered. This was previously performed by [22], who claim that the expanded label space can represent more rich information. Since the expectation of a Dirichlet distribution always yields a categorical distribution, there can be multiple Dirichlets that, in expectation, yield the same categorical distribution but inherit different levels of uncertainty associated with them. Let, therefore, the probability vector π (of which the model's softmax pseudo-probabilities are an estimate of) further be dependent on hyperparameters $\alpha_1, \dots, \alpha_K$:

$$\pi|\alpha \sim \text{Dir}(\alpha_1, \dots, \alpha_K). \quad (3)$$

These hyperparameters can be directly learned by the model, which is a common practice, for example, the *prior networks* [12]. The main advantage of this practice is the variety of possible uncertainty metrics based on the predictive Dirichlet distribution. Many metrics are considered for the experiments in this work; see Section IV-C for more details. As seen in Fig. 2, exemplarily, a notion of uncertainty can be determined for every class prediction given by the variance of the corresponding Dirichlet parameter.

As a novelty, now a third level of uncertainty is introduced. For this, an empirical Bayes framework is utilized to add an additional hierarchy to the existing Dirichlet-categorical distribution. First, a prior distribution of the α 's dependence on some γ is assumed. Then, Bayes' theorem is used to find a distributional form proportional to the posterior of α given the data:

$$f(\alpha|Y) \propto f(Y|\alpha) \cdot f_{\text{prior}}(\alpha). \quad (4)$$

Here, the form of the marginal posterior can be received by integrating out π from the full posterior as follows:

$$f(Y|\alpha) = \int f(y|\pi) f(\pi|\alpha) d\pi. \quad (5)$$

The next step is crucial, as it allows us to estimate the α parameters in a supervised and data-driven manner. In particular, the label information of the entire dataset can be embedded into the label embedding parameters. For this, stochastic estimation techniques can be applied to sample from this marginal posterior and hence estimate α , which will be explained in more detail in the following. Note that this approximation now holds information from the entire dataset since the proportional posterior is dependent on Y and iteratively estimated using all existing label information of the training set.

C. Model Training Procedures

1) *Estimating the Bayesian Embedding*: Given a vote vector Y_i as defined in (1), the overall goal is to estimate a latent variable that incorporates uncertainty among voters about the ground truth label. In particular, the model is then capable of accounting for label ambiguity as opposed to restraining each instance to a single class. Utilizing the idea of a third layer of uncertainty, in the following, an algorithm to estimate the α parameters will be investigated. In order to do this in a data-driven manner, one resorts to an empirical Bayes method to estimate a prior distribution. Therefore, now our estimation approach from [1] is summarized briefly.

First, stepping back to the level-1 uncertainty previously introduced, the distributional assumption of the labels can be viewed slightly differently. Due to the nature of the labeling process, the voted labels can be treated as observations of a multinomial distribution (as opposed to a categorical distribution in the case of softmax pseudo-probabilities) given the sampled probabilities $\pi := (\pi_1, \dots, \pi_K)$, where K denotes the number of classes considered. The likelihood of the vote vector given the latent probabilities is then given as follows:

$$Y|\pi \sim \text{Mult}(\pi, J). \quad (6)$$

Hence, the observed label votes are treated as if they were drawn from the multinomial distribution above. Again, J stands for the number of voters. A posteriori, as stated in (5), this yields

$$\begin{aligned} P(Y = y|\alpha) &= \int f(y|\pi) f(\pi|\alpha) d\pi = \dots \\ &= \frac{J!}{y_1! \dots y_K!} \cdot \frac{B(\alpha + Y)}{B(\alpha)} \end{aligned} \quad (7)$$

for the marginal distribution of the labels given the hyperparameters α after integrating out π . Here, $B(\cdot)$ denotes the beta distribution. As previously outlined, Bayes' theorem as shown in (4) is used next, in order to sample α . The choice of the prior distribution is of uttermost importance here: it can either be specified in advance based on prior knowledge or, if prior knowledge is not available, estimated in a data-driven manner following an empirical Bayes approach. The latter is carried out here.

Note that now, instead of limiting oneself to the constrained α parameters from the conditional Dirichlet distribution in (3), the α 's are transformed into $\mathbf{Z}$ via

$$(z_1, \dots, z_K) = \mathbf{Z} := \log(\alpha) = (\log(\alpha_1), \dots, \log(\alpha_K)) \quad (8)$$

which has the advantage of an unconstrained domain. As a prior for $\mathbf{Z}$, many options are available, but for simplicity and traceability, a normal distribution is chosen. Utilizing the empirical Bayes approach, the parameters of this K -dimensional normal prior distribution are estimated directly from the data using a stochastic version of the EM algorithm (see [47]). Concretely, $\hat{\mathbf{Z}} = (\hat{z}_1, \dots, \hat{z}_K)$ is estimated from the approximate posterior via the mean of the Markov chain Monte Carlo (MCMC) samples from (7). The interested reader is referred to [1] for algorithmic details.

2) *Learning the Embeddings With a Neural Network*: The estimated embeddings can, in the following, be treated as a soft classification output in the form of Dirichlet parameters:

$$\hat{\alpha}_i = (\hat{\alpha}_{i1}, \dots, \hat{\alpha}_{iK}) := \exp(\hat{\mathbf{Z}}_i). \quad (9)$$

Given the estimated embeddings, the posterior Dirichlet distribution for an image i is formed via the following:

$$\begin{aligned} \text{Dir}_{\text{post},i} &:= \text{Dir}(\hat{\alpha}_{i1}, \dots, \hat{\alpha}_{iK}) \\ &=: \text{Dir}(\hat{\alpha}_{\text{embedding},i1}, \dots, \hat{\alpha}_{\text{embedding},iK}). \end{aligned} \quad (10)$$

At the same time, the untransformed output (i.e., the *logits*) of a neural network classifier $\mathbf{h}_\theta(x) : \mathcal{X} \rightarrow \mathbb{R}^K$ is observed, as appearing in (2), when considering a K class classification problem as follows:

$$\mathbf{f}_\theta(\mathbf{x}_i) = (f_\theta(\mathbf{x}_i)_1, \dots, f_\theta(\mathbf{x}_i)_K). \quad (11)$$

To shape the output accordingly, $\exp(\cdot)$ is applied, and the transformation of the network's prediction now feasible for a Dirichlet distribution is given via the following:

$$\hat{\mathbf{A}}_{i,\text{model}} := (\exp(f_\theta(\mathbf{x}_i)_1), \dots, \exp(f_\theta(\mathbf{x}_i)_K)). \quad (12)$$

From this transformed prediction, the network's predictive Dirichlet distribution is formed via

$$\begin{aligned} \text{Dir}_{\text{post},\text{model},i} &= \text{Dir}(\exp(f_\theta(\mathbf{x}_i)_1), \dots, \exp(f_\theta(\mathbf{x}_i)_K)) \\ &=: \text{Dir}(\hat{\alpha}_{\text{model},i1}, \dots, \hat{\alpha}_{\text{model},iK}). \end{aligned} \quad (13)$$

Having these two Dirichlet distributions, the loss function of the neural network can be formulated as the explicit solution of the KL divergence between the predicted Dirichlet of our network and the Dirichlet estimated by the embedding. This loss reads as follows:

$$L(\theta) = \text{KL}(\text{Dir}_{\text{post},\text{model},i} || \text{Dir}_{\text{post},i}) = \dots$$

$$\begin{aligned} &= \log \Gamma(\hat{\alpha}_{\text{model},i0}) - \sum_k \log \Gamma(\hat{\alpha}_{\text{model},ik}) \\ &\quad - \log \Gamma(\hat{\alpha}_{\text{embedding},i0}) + \sum_k \log \Gamma(\hat{\alpha}_{\text{embedding},ik}) \\ &\quad + \sum_k (\hat{\alpha}_{\text{model},ik} - \hat{\alpha}_{\text{embedding},ik}) \\ &\quad \times (\psi(\hat{\alpha}_{\text{model},ik}) - \psi(\hat{\alpha}_{\text{model},i0})). \end{aligned} \quad (14)$$

Here, $\hat{\alpha}_{\text{model},i0} = \sum_k \hat{\alpha}_{\text{model},ik}$ is set and $\hat{\alpha}_{\text{embedding},i0} = \sum_k \hat{\alpha}_{\text{embedding},ik}$. $\psi(\cdot)$ denotes the digamma function, namely $\psi(x) = (d/dx) \ln(\Gamma(x))$. With the above loss formulation (14), it can be proceeded with training a prespecified DNN as usual, meaning backpropagating the loss with stochastic gradient descent (SGD).

The estimated unbounded embeddings $\hat{\mathbf{Z}} = (\hat{z}_1, \dots, \hat{z}_K)$, on the other hand, reveal a generally unbounded but roughly limited from above and below nature. A natural choice of regression loss function is therefore given by the mse. Given the estimated embeddings and the model prediction, which now gets treated as the predicted embeddings as follows:

$$\begin{aligned} \hat{\mathbf{Z}}_{\text{model}} &:= (\hat{z}_{1,\text{model}}, \dots, \hat{z}_{K,\text{model}}) \\ &= \mathbf{f}_\theta(\mathbf{x}_i) = (f_\theta(\mathbf{x}_{i1}), \dots, f_\theta(\mathbf{x}_{iK})). \end{aligned} \quad (15)$$

The mse loss function can be written out via

$$\text{MSE}(\hat{\mathbf{Z}}_{\text{model}}, \hat{\mathbf{Z}}) = \sum_{k=1}^K (\hat{z}_k - \hat{z}_{k,\text{model}})^2. \quad (16)$$

The Mahalanobis distance (MD) as loss function is of further interest. Particularly because this distance measure relies on an empirical estimation of correlation to scale the squared distance between the prediction and the ground truth. Such an estimation of inter-class correlation was measured during the algorithmic estimation procedure of the label embeddings by [1]. This estimated empirical covariance matrix is hence defined by $\hat{\Sigma}_{\text{embedding}}$ and state the MD loss function as follows:

$$\text{MD}(\hat{\mathbf{Z}}_{\text{model}}, \hat{\mathbf{Z}}) = \sqrt{(\hat{\mathbf{Z}}_{\text{model}} - \hat{\mathbf{Z}})^\top \hat{\Sigma}_{\text{embedding}}^{-1} (\hat{\mathbf{Z}}_{\text{model}} - \hat{\mathbf{Z}})}. \quad (17)$$

3) *Simple Dirichlet Embedding*: Building upon a strong foundation of Bayesian statistics, a more simple Dirichlet framework with no prior information and without the use of the estimated embeddings from [1] is investigated. For this, a noninformative prior on the α 's of the form

$$\alpha_{\text{prior},k} = c \quad \forall k = 1, \dots, K \text{ a priori} \quad (18)$$

is chosen for a constant $c \in \mathbb{N}$. Using the same distributional assumption as in (6), the posterior Dirichlet parameters are looking as follows:

$$\alpha_{\text{post},k} = c + y_k \quad \forall k = 1, \dots, K \text{ a posteriori}. \quad (19)$$

Here, y_k denotes the entry of the vote vector $\mathbf{Y}$ as defined in (1) at the k th position. With these posterior Dirichlet parameters, it can be proceeded with training with the loss formulated in (14) as described above.

IV. EXPERIMENTAL SETUP

A. General

For the experiments conducted, two different kinds of data splits were considered. Regarding the generalization and calibration experiments, a random split of the dataset between training and testing data was chosen (80/20). For the OoD detection framework, the instances were separated by two superior classes, namely the built classes (1–10) and the non-built classes (11–17 or A–G). Since there are little to no ambiguities between the two superior classes, the separation is clearly defined for the entire dataset. For the models using the expert label votes, the following decisions were made: As for the *one-hot* models, the single label was set to the majority vote of all experts, which is unique for every data instance. The training was then performed using the single label and the cross-entropy loss function. In the case of the *distributional* models, the empirical distribution of the expert votes was taken, namely Y_i/M , where M denotes the number of experts. The loss was then constructed using the KL divergence. Another version of models, named *sampled one-hot*, was considered to reflect the spread of label votes per image. Instead of combining the label votes into a distribution, here, the label gets sampled from the different label votes. The sampling probabilities are denoted by the empirical share (i.e., if seven experts voted for class 1 and three experts for class 2, the sampling probabilities are 70% and 30%, respectively). The *simple Dirichlet* models take the Dirichlet parameters defined in (19) and use the exact solution of the KL divergence between two Dirichlets as defined in (14) as a loss function. For the KL embedding models, the same loss is used, but now the Dirichlet parameters are computed using the estimated embeddings as described in (9). For the last two classes of models considered, the raw embeddings were taken as targets. Then, the loss function was modified to the mse (16) in case of the mse embedding models and to the MD (17) in case of the MD embedding models. As a baseline model, Sen2LCZ [48] was chosen due to its simplicity and sovereignty over many state-of-the-art convolutional neural networks. Throughout all models, the hyperparameters were left unchanged. Namely, a stepwise decaying learning rate (initial $2 \cdot e^{-04}$, 50% decay after five epochs) was used for a maximum of 30 epochs, while an early stopping mechanism was in place based on the validation loss with a patience of five epochs. A batch size of 128 was considered, and the dataset was randomly shuffled with a tracked seed before training was initialized.

Note that due to numerical instabilities of the embedding algorithm for samples of class 7 (lightweight low-rise), the embeddings were estimated based on 16 classes, and samples with majority votes for class 7 were excluded from the analysis. For cases where label votes for class 7 occurred, these votes were added to class 3 (compact low-rise) instead, as there is a known high ambiguity among those two classes. Hence, the models considered in this work all cover only 16 of the 17 LCZs, and the distributional labels were adjusted accordingly. Note that class 7, mostly referring to slums, does not naturally occur in the European cities. The small number of samples in the complete dataset was, therefore, added from

non-European cities. It is thus argued that the exclusion of class 7 does not play a crucial role in the label embedding framework, as there is little data support in the original dataset. Additionally, this little support, in combination with convergence issues in the embedding estimation, would lead to a skewed interpretation of the data.

For numerical stability, two substantial modifications were made to the training process. First, small ϵ 's were added to the Dirichlet parameters accounting for the cases where they are (near) zero. Furthermore, a temperature [49] to scale the Dirichlet parameters was used because they are too big after being exponentiated (exemplarily, the standard range of the embeddings is $[-10, 10]$, which would be transformed to approximately $[1 \times 10^{-5}, 2.2 \times 10^5]$). A temperature of $t = 3$ was used, and the logits were backtransformed at inference time. This means, loss projection was effectively applied at training time and back projection at inference time.

B. Calibration Analysis

The most common performance metric in a classification setting is the accuracy of a classification model, that is, how often the predicted class matches the ground truth. For our present dataset, a ground truth class can be determined by looking at the majority vote of the individual expert label votes. However, this aggregation, of course, ignores the label uncertainty displayed by the spread of the label votes. Therefore, accuracy is solely considered for a *sanity check*, meaning it is investigated whether the model is generally able to learn the classification scheme. Since a lot of images are not perturbed by label uncertainty at all, the accuracy should furthermore not worsen on these samples when switching to a label embedding learning framework.

Of much bigger interest here is the degree of calibration of a classification DNN. Calibration describes the interdependence between the model's confidence in its prediction and the actual observed accuracy of the model. In practice, this means a model should adequately state its uncertainty about its prediction, which is naturally reflected in the model's confidence. The observed mismatch between the confidence and the accuracy of a trained model can be quantified by different types of *calibration errors*. As a first step, a set of bins splitting up the observable range $[0, 1]$ of confidences and accuracies needs to be set. This is typically done in an equidistant manner, but adaptive methods have also been introduced [50].

The notation of [7] is followed and these bins are denoted by $B_m = (((m-1)/M), (m/M)]$, $m = 1, \dots, M$. In the experiments, different bin sizes were evaluated, and $M = 20$ was chosen as a reference for display. The indices of the images whose confidences $\hat{p}_i$ fall into the respective bin are then given by $I_m, m = 1, \dots, M$. The needed quantities are then calculated via the following:

$$\text{acc}(I_m) = \sum_{i \in I_m} \mathbb{1}_{\{\hat{y}_i = y_{\max,i}\}} \quad \text{and} \quad \text{conf}(I_m) = \sum_{i \in I_m} \hat{p}_i. \quad (20)$$

The accuracy per bin is, therefore, calculated as the proportion of correctly classified samples within a bin, while the confidence is the average predicted confidence for all

samples in the bin. In the formula, $\hat{y}_i$ is the predicted class for input x_i , which is derived by the index of the maximum of the model's predicted embedding vector and the majority vote of image i is referred to as $y_{\max,i}$. $\hat{p}_i$ then corresponds to the highest predicted probability, which was derived by taking the maximum of the posterior mean of the embeddings:

$$\hat{p}_i := \max_k E(\pi_k | \mathbf{Z}) = \max_k \frac{\exp(Z_k)}{\sum_{k'=1}^K \exp(Z_{k'})}. \quad (21)$$

The standard expected calibration error (ECE) is then derived via the following:

$$\text{ECE} = \sum_{m=1}^M \frac{|I_m|}{n} |\text{acc}(I_m) - \text{conf}(I_m)|. \quad (22)$$

Further versions of the ECE include the maximum calibration error (MCE), which evaluates the maximum gap between confidence and accuracy out of all bins considered.

$$\text{MCE} = \max_m |\text{acc}(I_m) - \text{conf}(I_m)|. \quad (23)$$

Additionally, the static calibration error (SCE) [50] evaluates calibration performance on a class level. For class k and within bin m , define accuracy and confidence via $\text{acc}(m, k)$ and $\text{conf}(m, k)$, respectively. The SCE is then given by the following:

$$\text{SCE} = \frac{1}{K} \sum_{k=1}^K \sum_{m=1}^M \frac{n_{mk}}{n} |\text{acc}(m, k) - \text{conf}(m, k)| \quad (24)$$

where n_{bk} denotes the number of samples belonging to class k (with respect to the majority vote) within bin b .

Lastly, the KL divergence between the mean of the Dirichlet distribution (i.e., a pseudo-softmax probability distribution) predicted by the model and the mean of the Dirichlet distribution based on the embedding can be interpreted as a measure of generalization performance.

C. OoD Detection

At test time, the model's predictive uncertainty is first evaluated by means of OoD detection. Therefore, two types of samples are presented to the model, namely samples from the classes it was trained on (in-distribution or ID) and samples from the remaining classes (OoD). Then, the model predicts a label embedding for each sample. It is unimportant to look at performance measures such as accuracy, especially since the model cannot correctly predict the OoD data anyway, as it comes from classes unknown to the model.

Instead, a measure of uncertainty accompanying the prediction is focused and this uncertainty is used to test the model's performance to distinguish between ID and OoD data. This is done by means of a binary classifier, which is presented with all predictive uncertainties as scores, and classifies the samples into ID and OoD (0 and 1) according to this score. Ideally, a well-calibrated model will produce low uncertainty scores for ID data and higher uncertainty scores for OoD samples, facilitating effective classification. In classical neural networks,

however, the tendency toward highly confident model predictions with little uncertainty values is often observed. This is observed even when the model is fed OoD data, which, in theory, should make the model unsure and, therefore, lead to high uncertainty.

For our dataset at hand, the classes are split into ID and OoD classes by using the intrinsic hierarchical class scheme. This scheme differentiates between built and non-built classes. The latter is composed of the classes A through G (or 11 through 17 if numbered continuously), while the built classes are defined by the initial 10 classes, 1 through 10. Both of the two hierarchical classes can be used for ID and OoD classes, respectively. The hierarchical split between built and non-built classes is justified by the structural differences between these two sets of classes: While the built classes refer to areas of anthropization (see Fig. 1, the non-built classes describe different land cover types. Furthermore, [15] already used this ID/OoD split for an earlier work.

Regarding the uncertainty score, various quantities can be considered. First, the focus lies on measures that take the pseudo-softmax probabilities

$$\hat{\pi}_k := \frac{\hat{\alpha}_{\text{model}, k}}{\sum_{k'=1}^K \hat{\alpha}_{\text{model}, k'}} \quad (25)$$

and the estimated posterior parameters $\hat{\alpha}_{\text{model}, k}$ as defined in (13) as starting point. A widely used measure based on the distribution of the softmax pseudo-probabilities is the (Shannon) entropy. For the predictive distribution $\hat{P}(y|\pi)$, this reads as follows:

$$\begin{aligned} H[\mathbb{E}_{P(\pi|x)}[P(y|\pi)]] \\ &= - \sum_{k=1}^K \hat{\pi}_k \cdot \log(\hat{\pi}_k) \\ &= - \sum_{k=1}^K \frac{\hat{\alpha}_{\text{model}, k}}{\sum_{k'=1}^K \hat{\alpha}_{\text{model}, k'}} \cdot \log\left(\frac{\hat{\alpha}_{\text{model}, k}}{\sum_{k'=1}^K \hat{\alpha}_{\text{model}, k'}}\right). \end{aligned} \quad (26)$$

As thoroughly explained by [28], this quantity measures the total uncertainty present in the prediction. The authors furthermore disentangle the total quantity into aleatoric and epistemic uncertainty by means of the expected entropy. This term is derived by taking the expectation of the entropy with respect to the predictive distribution; see [28] for details of the derivation. The expected entropy, which supposedly measures the aleatoric bit of the uncertainty, is quantified via the following:

$$\begin{aligned} \mathbb{E}_{P(\pi|x)}[H[P(y|\pi)]] &= \dots \\ &= \sum_{k=1}^K \hat{\pi}_k \left(\psi(\hat{\alpha}_{\text{model}, k} + 1) - \psi\left(\sum_{k'=1}^K \hat{\alpha}_{\text{model}, k'}\right) \right) \end{aligned} \quad (27)$$

where ψ again denotes the digamma function $\psi(x) = (d/dx) \ln(\Gamma(x))$. Of further interest can be the so-termed *distributional uncertainty*, which [28] depicts via the difference between the total and the aleatoric uncertainty. This quantity measures the amount of information about $\hat{\pi} := (\hat{\pi}_1, \dots, \hat{\pi}_K)$ given y and conditioned on x and is also termed mutual

TABLE I
CALIBRATION AND GENERALIZATION METRICS FOR ALL EVALUATED MODELS. ACCURACY METRICS INCLUDE OA, MAA, WAA, AND KAPPA STATISTIC (κ). CALIBRATION METRICS INCLUDE ECE, MCE, SCE, AND KL DIVERGENCE TO EMBEDDING (KL Div.). BOLD VALUES INDICATE THE BEST-PERFORMING MODEL FOR EACH METRIC

	OA $\uparrow$	MAA $\uparrow$	WAA $\uparrow$	κ $\uparrow$	ECE $\downarrow$	MCE $\downarrow$	SCE $\downarrow$	KL-Div. to Emb. $\downarrow$
One-hot	90.3	59.8	88.4	88.4	4.4	18.8	0.6	0.40
Distr.	91.4	80.7	91.1	89.8	6.6	23.4	0.6	0.32
Sampled One-hot	90.7	76.7	90.2	89.0	6.8	23.3	0.7	0.32
Simple Dirichlet	89.6	72.4	88.7	87.6	4.1	90.0	0.2	0.48
KL Embedd.	89.9	66.9	89.3	88.0	1.7	10.2	0.3	0.27
MSE Embedd.	90.5	72.1	89.9	88.8	2.7	12.6	0.3	0.28
MD Embedd.	73.7	31.3	66.7	67.9	10.2	76.3	0.6	0.83

information. It reads as follows:

$$\begin{aligned}
I[y, \pi|x] &= H[\mathbb{E}_{p(\pi|x)}[P(y|\pi)]] - \mathbb{E}_{p(\pi|x)}[H[P(y|\pi)]] = \dots \\
&= -\sum_{k=1}^K \hat{\pi}_k \left(\log \hat{\pi}_k - \psi(\hat{\alpha}_{\text{model}, k} + 1) \right. \\
&\quad \left. + \psi\left(\sum_{k'=1}^K \hat{\alpha}_{\text{model}, k'} + 1\right) \right). \quad (28)
\end{aligned}$$

See [28] for a more detailed derivation. Next to the aleatoric and distributional bit of the uncertainty, multiple measures have been established to estimate the epistemic uncertainty. Most notably, the maximum softmax probability (MSP) can be considered, which is given via

$$\text{MSP}(\hat{\pi}) := \max_k \hat{\pi}_1, \dots, \hat{\pi}_K. \quad (29)$$

This measure is, for example, used by the works of [51] and [25]. Besides that, the Dempster–Shafer metric (DSM) was used by [26]. This measure is founded on the Dempster–Shafer theory of evidence (see [52] for details) and can be computed via the following:

$$\text{DSM}(\hat{\pi}) := \frac{K}{\sum_{k=1}^K (\hat{\alpha}_{\text{model}, k} + 1)}. \quad (30)$$

Lastly, the overall so-termed *evidence* of the predicted Dirichlet distribution can be considered as a measure for epistemic uncertainty as follows:

$$\text{Evidence}(\hat{\pi}) := \sum_{k=1}^K \hat{\alpha}_{\text{model}, k}. \quad (31)$$

V. RESULTS AND DISCUSSION

A. Calibration Results

Now, the quality of the predicted uncertainties is investigated by means of the pseudo-softmax probabilities for all models previously introduced. For these experiments, there is no split between ID and OoD data, rather the models trained on all classes are evaluated on a test bed. In Table I, the

calibration and generalization results for the different models are presented.

Note first that the typical one-hot encoded model achieves the second-highest values in the classical performance metrics. Next to the classical overall accuracy (OA), the macro average accuracy (MAA), which takes the average of the class-wise accuracies, and the weighted average accuracy (WAA), which weighs the class-wise accuracies by the class share in the dataset, are displayed. The κ statistic [53] furthermore describes the inter-rater agreement between the prediction and the label. This phenomenon should receive little attention because there is no kind of uncertainty presented to the model. Therefore, the model solely focuses on classifying the single ground truth class correctly. The slight performance increase of the model trained with the soft labels based on the empirical distribution of the votes (*Distr.*) can be potentially explained by smoother training enabled via the natural regularization of the labels.

When now focusing on the calibration metrics as introduced earlier, the clear advantage of the embedding-based approaches becomes apparent. The ECE of the KL embedding-based model is reduced by more than 50% compared to traditional models such as the one-hot or distributional label-based models. The message is similar for the mse loss in comparison with embedding-based models. These strong calibration results imply the practical usage of the embeddings for remote sensing classification tasks since a high degree of reliability is thereby enforced. Solely for the MD loss, the results are not convincing. It is argued that the underlying correlation structure of the embeddings might be too complex to enable smooth training and can, on the other hand, hinder calibration performance. The look at MCE reveals a picture similar to that of the ECE. Note that the MCE depends highly on the choice of number and width of bins, therefore the high MCE of the simple Dirichlet model is not problematic. The SCE is also generally much lower for the embedding-based models, yet here the simple Dirichlet model can also shine and beats the embedding-based models by a small margin. Note, however, that the gap to the traditional models remains extremely high. As another performance metric of interest, the KL divergence between the models' predictions and the categorical distribution defined over the latent embeddings of the test set (i.e., the

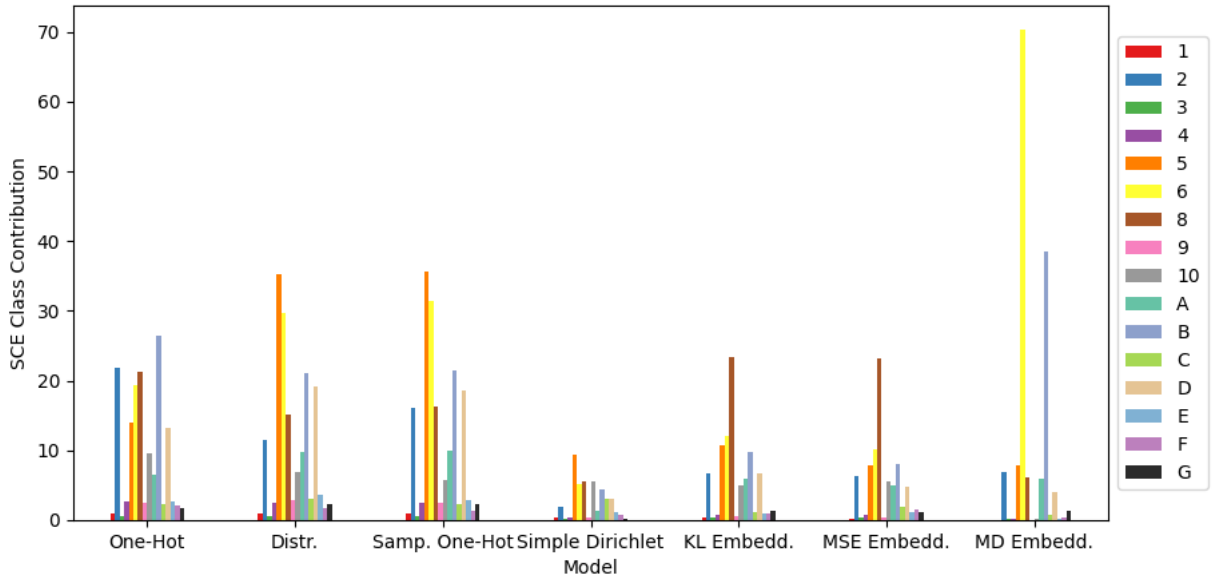


Fig. 3. SCE contributions per class for all models. Classes: 1 (compact high-rise), 2 (compact midrise), 3 (compact low-rise), 4 (open high-rise), 5 (open midrise), 6 (open low-rise), 8 (large low-rise), 9 (sparsely built), 10 (heavy industry), A (dense trees), B (scattered trees), C (bush and scrub), D (low plants), E (bare rock or paved), F (bare soil or sand), and G (water).

expectation of the Dirichlet) is considered. This metric can be seen as a measure of generalization performance toward the unseen embedding at test time. With regard to this metric, the embedding-based approaches clearly outperform all other models by a large margin.

For a closer look at calibration performance, Fig. 3 shows the individual class contributions to the overall SCE by method. In more detail, the individual summands of the sum over $k = 1, \dots, K$ in (24) are visualized and explain the individual class contributions to the overall SCE. The SCE is a more fine-grained calibration measure than the ECE, as it derives calibration performance on a class level. In the analysis, now the entirety of the softmax probability distribution is considered. This closer look reveals that despite a few outliers (e.g., class 8 and class E), the Bayesian embedding approaches, with the exception of the seemingly too vague Mahalanobis embedding approach, lead to increased calibration performance on a class level. The construction of the Bayesian embeddings backs this up: via integrating label uncertainty among samples across the entire dataset, a more realistic and uncertainty-aware embedding is created. With this embedding, the entire softmax probability vector better reflects the label uncertainty present in the data. Regarding the outliers, the low calibration performance can be partially explained by a small share of the respective classes in the data (hence little support for effective and confident learning) and small confusion to other classes (hence spatially separated embeddings leading to suboptimal learning). For an in-depth analysis of the class votes, see [38].

B. OoD Detection

The OoD detection performance for all models described in Section IV-A is shown in Table II. As indicated by the first two columns, all models are trained in two different scenarios:

for once, the models are trained using randomly 80% of the data samples where the majority vote was cast for a built class (1–10), of which 25% are reserved for validation at the end of each epoch. The remaining 20% is reserved for the testing stage. At this stage, OoD data is presented to the model belonging to the non-built classes (A–G), with respect to the class voted for by the majority of the experts. Identically, the models in the second scenario are trained with parts of the data belonging to the non-built classes (with respect to the majority vote) and are then presented with OoD data from built classes. Based on the initial label evaluation study performed by [45], these two superior classes, namely built and non-built, are close to mutually independent. This means almost no votes were cast for the other overarching class, different from the overarching class the majority voted class belongs to.

In a first step, all uncertainty metrics presented in Section IV-C are computed for the ID and OoD test samples. For all predicted uncertainties, the corresponding areas under the curve (AUCs) are displayed wherever computable. In more detail, for the softmax-based approaches (*one-hot*, *distributional*, and *sampled one-hot*), uncertainty metrics that are based on Dirichlet parameters cannot be computed. These metrics are namely the DSM, the *expected entropy*, the overall *evidence*, and the *distributional uncertainty*. Note that these AUCs are based on a binary classifier, which is constructed to differentiate between ID and OoD data samples only based on the predictive uncertainties that are presented at test time. Both the area under the precision–recall curve (AUPR), which compares the precision and recall of the binary classifier, and the area under the receiver operator curve (AUROC), which compares the recall and the false positive (FP) rate of the binary classifier, are investigated for a varying decision threshold. In more detail, given the true positives (TPs), true negatives (TNs), FPs, and false negatives (FNs) of the classifier results,

TABLE II

OOD DETECTION RESULTS FOR ALL MODELS AND A RANGE OF DIFFERENT UNCERTAINTY METRICS. TWO SCENARIOS ARE CONSIDERED, WHERE BUILT AND NON-BUILT CLASSES ARE USED FOR ID AND OoD, RESPECTIVELY, AND VICE VERSA. AREAS UNDER CURVES ARE CALCULATED FOR A BINARY CLASSIFIER BUILT UPON THE PREDICTIVE UNCERTAINTIES. THE OoD UNCERTAINTIES ARE LABELED WITH 1; THE ID UNCERTAINTIES ARE LABELED WITH 0. THE BINARY CLASSIFIER THEN TRIES TO SEPARATE THE TWO KINDS OF UNCERTAINTIES BY ESTIMATING A THRESHOLD. AUROC = AREA UNDER PRECISION-RECALL CURVE. AUPR = AREA UNDER PRECISION-RECALL CURVE, DSM = DEMPSTER-SHAFER METRIC (30), MSP = 1 - MAXIMUM SOFTMAX PROBABILITY (29), PRED.ENT. = PREDICTIVE ENTROPY (26), EXP.ENT. = EXPECTED ENTROPY (27), EVID. = EVIDENCE (31), DISTR.UNC. = DISTRIBUTIONAL UNCERTAINTY (28). THE BEST RESULTS ARE INDICATED IN BOLD

ID Data	OoD Data	Setting	AUROC $\uparrow$						AUPR $\uparrow$					
			DSM	MSP	Pred.Ent.	Exp.Ent.	Evidence	Distr. Unc.	DSM	MSP	Pred.Ent.	Exp.Ent.	Evid.	Distr. Unc.
Built-up	Non-built	One-Hot		78.8	79.8					76.9	77.5			
		Distr.		82.4	78.4					77.5	73.4			
		Sampled One-Hot		79.0	73.3					74.0	69.9			
		Simple Dirichlet	81.7	82.7	80.9	81.7	18.3	79.3	74.9	78.1	74.8	74.9	33.8	74.1
		KL Embedd.	70.0	66.1	68.3	70.0	30.0	66.5	61.7	61.6	63.7	61.7	37.4	63.0
		MSE Embedd.	44.7	35.5	36.0	59.9	55.3	53.1	57.6	40.8	43.5	58.3	58.3	58.8
		MD Embedd.	45.8	51.8	47.1	48.0	54.2	53.9	49.3	56.6	51.4	47.4	50.2	50.4
Non-built	Built-up	One-Hot		96.2	96.7					93.2	94.6			
		Distr.		95.3	96.4					92.7	94.8			
		Sampled One-Hot		94.7	95.6					92.4	94.1			
		Simple Dirichlet	96.7	96.6	96.7	96.7	3.3	66.4	94.2	95.3	95.6	94.2	31.1	78.3
		KL Embedd.	95.6	93.4	93.3	95.6	4.4	74.5	95.1	91.6	91.7	95.1	31.1	81.8
		MSE Embedd.	56.9	70.4	64.3	31.2	43.1	52.4	57.2	67.5	63.4	39.1	48.7	55.5
		MD Embedd.	92.2	92.7	94.0	9.7	7.8	9.8	91.6	88.2	91.8	32.3	31.9	33.3

the precision is computed as $[TP/(TP + FP)]$, the recall via $[TP/(TP + FN)]$, and the FP rate via $[FP/(FP + TN)]$. In order to make the AUCs expressive, the size of the OoD dataset and the ID test set was balanced (namely, whichever dataset was larger was cropped to the size of the smaller one). If this balancing is ignored, the resulting plot would be skewed as there are far more non-built data samples than from the built classes.

Looking at the results of non-built or vegetation as OoD detection, the difficulty of the task becomes immediately visible. Apparently, for the given space of considered models and uncertainty metrics, there is an upper limit of $\approx 83\%$ in AUC. Overall, the performance of the embedding-based approaches is lacking a bit behind the other models for all uncertainty metrics considered. Exemplarily, the KL embedding model only achieves 66.1% AUROC for the MSP uncertainty compared to the distributional model with 82.4%. This can potentially be explained by a high level of data-specific uncertainty in the embedding space for the built classes. By learning this data-specific uncertainty during training, the predictive ID uncertainty remains data-specific and hinders the binary classifier from successfully differentiating between ID and OoD uncertainty. Note that this phenomenon is not surprising, as the estimated voting uncertainty is purposely fed, which is highly data-specific, to the model via the Bayesian embeddings. Hence, the differentiation between training and test data (from a different data regime) is less obvious. The big advantage of such uncertainty-aware training lies within the better-calibrated uncertainties, as was seen in the calibration analysis.

For the opposite case where the built-up classes are OoD, the simple Dirichlet model achieves an AUROC of 96.7%

compared to the KL embedding model's 95.6% for the DSM uncertainty (see Table II). Regarding the AUPR, the gap is less narrow; here the simple Dirichlet achieves 95.6% and whereas the best embedding-based model (MD embedding) only achieves 91.8%. This overall best result emphasizes the advantage of Dirichlet embedding models, particularly with the uninformative prior. The uninformativeness a priori ensures training without too strong a focus on the ID data (as opposed to the Bayesian embedding models). This then helps to better deal with OoD samples at test time. Strong results can be hereby achieved across many of the uncertainty metrics considered. Note that three of them can only be calculated if a predictive Dirichlet distribution is formed by the model. This strong performance of the simple Dirichlet model across AUROC and AUPR for various uncertainty metrics underlines the robustness of the proposed approach. Fig. 4(a) shows the relatively clear separation of ID and OoD uncertainties (MSP as uncertainty metric) of this model. While most of the ID uncertainties are centered near zero and sparsely scattered outwards, the OoD uncertainties are mostly taking much higher values. This leads to a generally good separability of the two datasets. On the right side, the corresponding ROC curve [see Fig. 4(b)] is displayed, showing a moderate increase of the curve for varying the decision threshold.

In the opposite case, where built samples are presented as OoD data, a much higher level of AUC performance can be seen. This can be attributed to the clear separation of the individual non-built classes. The clear separation, in turn, yields very confident models. When presented with built samples, which are fundamentally different in their nature,

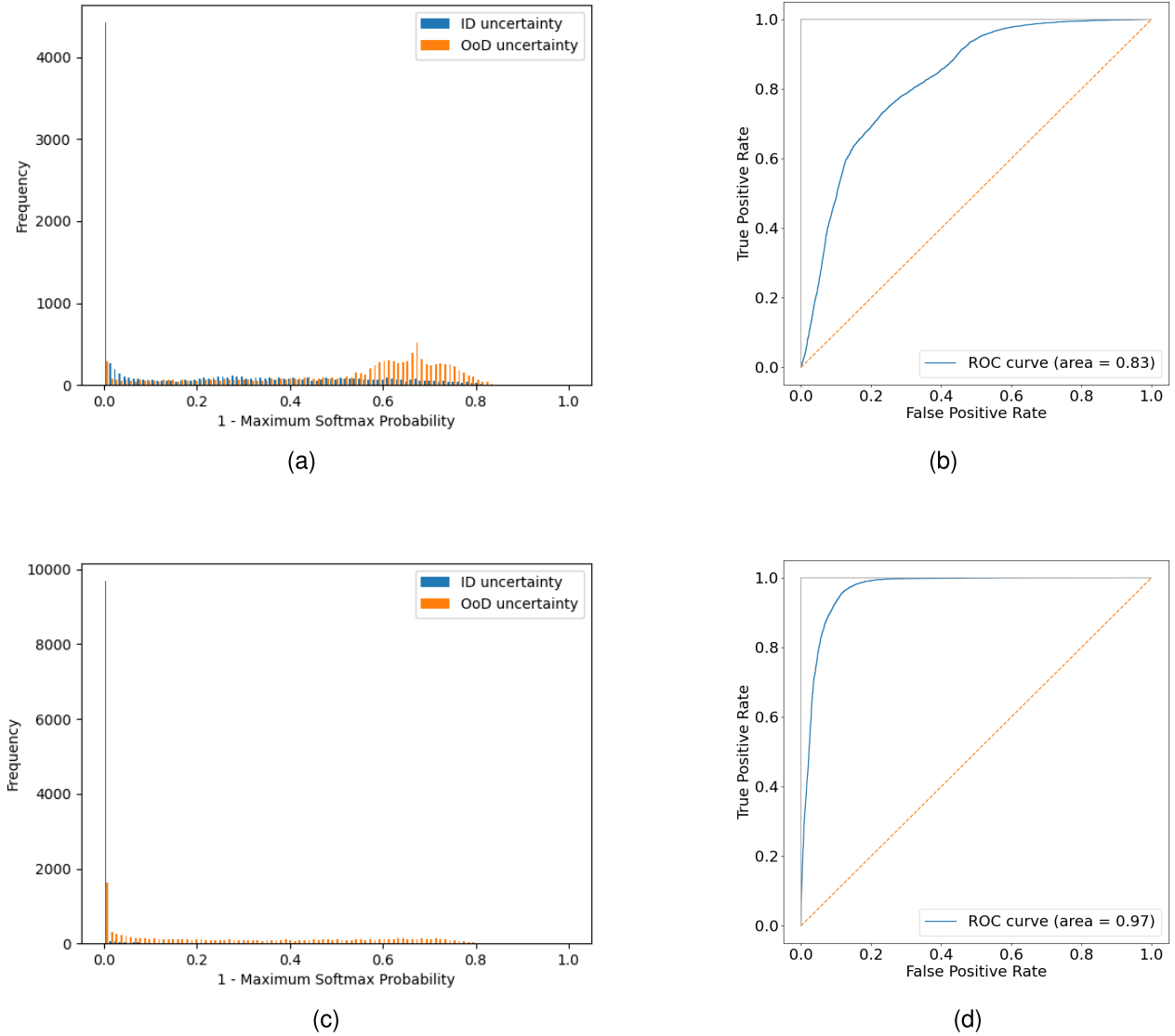


Fig. 4. Exemplary (a) and (c) histograms of predictive uncertainty scores and (b) and (d) ROC curves for binary OoD classifier. The two plots show the interdependence between the grouped uncertainty quantities and the performance of the binary OoD detector.

a small predictive uncertainty is sufficient to effectively distinguish this sample from the confident (hence near-zero) ID uncertainties. This phenomenon becomes clearly visible in Fig. 4(c), where a large spread of OoD uncertainties stands against a pile of near-zero ID uncertainties in the case of the best-performing model (simple Dirichlet). This is then reflected in a near-perfect ROC curve and AUROC metric depicted in Fig. 4(d). Again, the uninformativeness a priori seemingly helps to keep the model cautious of OoD samples at test time.

An interesting observation of this OoD setting is that the Bayesian embedding-based approaches are performing similarly well to the other models. It can be concluded that the overall lower inter-class uncertainty of the non-built classes, which is visible in both the label distributions and the corresponding estimated embeddings, helps sufficiently well to

make the training OoD-aware. Especially when solely looking at distribution-based uncertainty metrics such as expected entropy and distributional uncertainty, the Bayesian embedding models generally achieve higher detection performance than the Dirichlet-embedding models, although they are trained with data-specific uncertainty.

Generally speaking, a lot of uncertainty metrics that were considered might not be suitable for the present OoD detection task or the investigated models. The *expected entropy*, theoretically measuring the aleatoric part of the uncertainty, can only give sense-making results for the simple Dirichlet model and the KL embedding. This is because all results below 50% can be discarded; they are worse than random guessing. Following this, the *distributional uncertainty* is also mismatched, likely because it includes the expected entropy as a term. The performance result using the *evidence* as

uncertainty metric is also underperforming. It is believed this is due to the fact that this quantity is unnormalized and does not align with the unconstrained embedding space on which it is based. Even for the somewhat contained Dirichlet space, this metric is not suitable.

VI. CONCLUSION

This study introduced a three-level uncertainty scheme, including a Bayesian embedding setup that demonstrated superior calibration across multiple metrics. For the third-level uncertainty, two types of label embeddings were investigated: a Bayesian setup to estimate an unconstrained label embedding as well as a Dirichlet setup with an uninformative prior. A core idea of the Bayesian embedding approach is the loosening of the restrictions of typical label spaces defined, for example, Dirac or categorical distributions. These embeddings have then been incorporated into a deep learning framework through various transformations and adapted loss functions. In particular, the embeddings were, in one case, utilized as Dirichlet parameters of a predictive distribution, which was then learned by means of the explicit solution of the KL divergence between Dirichlet distributions. In other cases, the untransformed embeddings were directly learned by the network with the aid of two popular regression loss functions.

In the studied dataset LCZ42, label uncertainty is present in the form of multiple expert annotations per image. The aim of the Bayesian embedding approaches is, therefore, to project this uncertainty into the embedding space and, in the following, make the network learn in an uncertainty-aware manner. Similarly, the uncertainty is projected into the Dirichlet space for the Dirichlet embedding. Two separate test beds were constructed to evaluate the approaches, namely an OoD detection framework that evaluates the model's predictive uncertainties by means of different uncertainty metrics and a calibration framework that examines the pseudo-softmax probabilities. In the calibration analysis, the Bayesian embedding approaches, with the partial exception of the MD loss, can show significant improvements in terms of calibration performance. The most common performance metrics are reduced by more than 50% compared to traditional learning regimes. In particular, the KL embedding model reduced the ECE by over 60% compared to the one-hot model, as shown in Table II. In the OoD detection experiments, the simple Dirichlet approach could show strong OoD detection results compared to models learning with a constrained label space. Regarding the AUROC metric, for example, the Dirichlet model achieved 95.6% as opposed to the 94.8% achieved by the distributional model. Here, the uninformative prior seemingly helped best to differentiate unseen OoD data from the ID data as seen during training. The Bayesian embedding models, which were trained with data-specific uncertainty, could not sufficiently well detect OoD data.

Though the introduced techniques can capture the label uncertainty arising from the expert labeling pipeline in a broad range, these techniques are, of course, limited to the number and diversity of experts considered. While the 10 experts provide a solid basis for label uncertainty, the question arises as to whether more labels per image or more images are

desirable. [54] discusses this question in an NLP context and concludes that more labels can sharpen the latent class borders better than in the opposite case. For satellite image classification, answering this question is an open issue. Next to label-based embeddings, an interesting direction would be to incorporate semantic information of the different LCZ classes into an embedding, potentially with an unsupervised, semi-supervised, or self-supervised approach. This stream is left for future research. Additionally, in terms of generalization of the proposed framework, it would be wise to broaden the scope to other datasets with a similar label structure. Unfortunately, to the best of our knowledge, there does not exist a similar large-scale dataset in the remote sensing domain. For an extensive analysis of the proposed framework, which lies out of the scope of this journal, the interested reader may be referred to [55] for an overview of many datasets with noisy and ambiguous labels across domains.

REFERENCES

- [1] K. Hechinger, C. Koller, X. Xiang Zhu, and G. Kauermann, "Human-in-the-loop: Towards label embeddings for measuring classification difficulty," 2023, *arXiv:2311.08874*.
- [2] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [3] A. Höhl et al., "Opening the black-box: A systematic review on explainable AI in remote sensing," 2024, *arXiv:2402.13791*.
- [4] E. Hullermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, 2021.
- [5] M. Valdenegro-Toro and D. S. Mori, "A deeper look into aleatoric and epistemic uncertainty disentanglement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1508–1516.
- [6] C. Koller, G. Kauermann, and X. X. Zhu, "Going beyond one-hot encoding in classification: Can human uncertainty improve model performance in Earth observation?" *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5400211.
- [7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Intl. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 1321–1330.
- [8] A. Kirsch, S. Farquhar, P. Atighehchian, A. Jesson, F. Branchaud-Charron, and Y. Gal, "Stochastic batch acquisition: A simple baseline for deep active learning," 2021, *arXiv:2106.12059*.
- [9] S. Matiz and K. E. Barner, "Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification," *Pattern Recognit.*, vol. 90, pp. 172–182, Jun. 2019.
- [10] A. Kirsch, J. V. Amersfoort, and Y. Gal, "BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 7026–7037.
- [11] W. Liu, X. Wang, D. J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2020, pp. 21464–21475.
- [12] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 7047–7058.
- [13] J. Z. Liu et al., "A simple approach to improve single-model deep uncertainty via distance-awareness," *J. Mach. Learn. Res.*, vol. 24, no. 42, pp. 1–63, 2023.
- [14] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," 2021, *arXiv:2102.11409*.
- [15] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616819.
- [16] C. Koller, P. Jung, and X. X. Zhu, "Can land cover classification models benefit from distance-aware architectures?" *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

- [17] M. Migliari, E. Briche, J. Despax, L. Chesne, and O. Baverel, "A cross-analysis matrix comparing multi-site local climate zone trends: Application to identify sustainable built morphologies improving summer daytime urban microclimate," *Sustain. Futures*, vol. 7, Jun. 2024, Art. no. 100162.
- [18] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1513–1589, Jul. 2023.
- [19] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2016, pp. 6405–6416.
- [20] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Cham, Springer, 2012.
- [21] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision," in *Proc. Adv. Neural Inf. Process. Syst.*, Mar. 2017, pp. 5580–5590.
- [22] T. Joo, U. Chung, and M.-G. Seo, "Being Bayesian about categorical probability," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4950–4961.
- [23] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep Kernel learning," in *Proc. 19th Int. Conf. Artif. Intell. Statist.*, 2016, pp. 370–378.
- [24] J. Z. Liu et al., "A simple approach to improve single-model deep uncertainty via distance-awareness," *arXiv:2205.00403*, 2022.
- [25] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9690–9700.
- [26] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2018, pp. 9690–9700.
- [27] A. Jsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Cham, Switzerland: Springer, 2018.
- [28] D. T. Ulmer, C. Hardmeier, and J. Frellsen, "Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation," *Trans. Mach. Learn. Res.*, 2023. [Online]. Available: <https://openreview.net/forum?id=xqS8k9E75c>
- [29] A. Amini, W. Schwarting, A. P. Soleimany, and D. Rus, "Deep evidential regression," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 14927–14937.
- [30] D. Oh and B. Shin, "Improving evidential deep learning via multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 7895–7903.
- [31] N. Meinert, J. Gawlikowski, and A. Lavin, "The unreasonable effectiveness of deep evidential regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 9134–9142.
- [32] R. Duan, B. Caffo, H. X. Bai, H. I. Sair, and C. Jones, "Evidential uncertainty quantification: A variance-based perspective," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 2121–2130.
- [33] M. Jürgens, N. Meinert, V. Bengs, E. Hüllermeier, and W. Waegeman, "Is epistemic uncertainty faithfully represented by evidential deep learning methods?" 2024, *arXiv:2402.09056*.
- [34] J. J. Ryu et al., "Improved evidential deep learning via a mixture of Dirichlet distributions," 2024, *arXiv:2402.06160*.
- [35] D. T. Chang, "Hybrid Bayesian neural networks with functional probabilistic layers," 2021, *arXiv:2107.07014*.
- [36] S. Jose, S. Louis, S. Dascalu, and S. Liu, "Transfer learning-based hybrid approach for Bayesian network structure learning," *Int. J. Artif. Intell. Tools*, vol. 31, no. 7, Nov. 2022, Art. no. 2260003.
- [37] S. I. Allec and M. Ziatdinov, "Active and transfer learning with partially Bayesian neural networks for materials and chemicals," 2025, *arXiv:2501.00952*.
- [38] K. Hechinger, X. X. Zhu, and G. Kauermann, "Categorising the world into local climate zones: Towards quantifying labelling uncertainty for machine learning models," *J. Roy. Stat. Soc. Ser. C, Appl. Statist.*, vol. 73, no. 1, pp. 143–161, Jan. 2024.
- [39] J. Dong et al., "Statistical uncertainty analysis-based precipitation merging (SUPER): A new framework for improved global precipitation estimation," *Remote Sens. Environ.*, vol. 283, Dec. 2022, Art. no. 113299.
- [40] P. Ebel, V. S. F. Garnot, M. Schmitt, J. D. Wegner, and X. X. Zhu, "UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2086–2096.
- [41] X. Liao, B. Tu, J. Li, and A. Plaza, "Class-wise graph embedding-based active learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522813.
- [42] Y. Ma, Z. Zhang, Y. Kang, and M. Özdoğan, "Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach," *Remote Sens. Environ.*, vol. 259, Jun. 2021, Art. no. 112408.
- [43] J. Liu et al., "Uncertainty analysis of eleven multisource soil moisture products in the third pole environment based on the three-corned hat method," *Remote Sens. Environ.*, vol. 255, Mar. 2021, Art. no. 112225.
- [44] E. Hüllermeier, "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?" in *Proc. Uncertainty Artif. Intell.*, Jan. 2022, pp. 2282–2292.
- [45] X. X. Zhu et al., "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [Software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.
- [46] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorological Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.
- [47] G. Celeux, D. Chauveau, and J. Diebolt, "Stochastic versions of the em algorithm: An experimental study in the mixture case," *J. Stat. Comput. Simul.*, vol. 55, no. 4, pp. 287–314, Nov. 1996.
- [48] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multi-level feature fusion-based CNN for local climate zone classification from sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2793–2806, 2020.
- [49] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [50] J. Nixon et al., "Measuring calibration in deep learning," in *Proc. CVPR workshops*, Jun. 2019, pp. 1–9. [Online]. Available:
- [51] B. Charpentier, D. Zügner, and S. Günnemann, "Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 1356–1367.
- [52] G. Shafer, *A Mathematical Theory of Evidence*, vol. 42. Princeton, NJ, USA: Princeton Univ. Press, 1976.
- [53] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.
- [54] C. Gruber, K. Hechinger, M. Aßenmacher, G. Kauermann, and B. Plank, "More labels or cases? Assessing label variation in natural language inference," in *Proc. 3rd Workshop Understand. Implicit Underspecified Lang.*, 2024, pp. 22–32.
- [55] L. Schmarje et al., "Is one annotation enough?—A data-centric image classification benchmark for noisy and ambiguous label estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 33215–33232.

A.3 Can Land Cover Classification Models Benefit From Distance-Aware Architectures?

Reference

C. Koller, P. Jung and X. X. Zhu. Can Land Cover Classification Models Benefit From Distance-Aware Architectures?. *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1-5, 2024. doi: 10.1109/LGRS.2024.3375370

Copyright

Article published in the IEEE Geoscience and Remote Sensing Letters under a CC-BY-4.0 license. Reproduced with friendly permission from the authors.

Can Land Cover Classification Models Benefit From Distance-Aware Architectures?

Christoph Koller¹, Graduate Student Member, IEEE, Peter Jung, and Xiao Xiang Zhu², Fellow, IEEE

Abstract—The quantification of predictive uncertainties helps to understand where the existing models struggle to find the correct prediction. A useful quality control tool is the task of detecting out-of-distribution (OOD) data by examining the model's predictive uncertainty. For this task, deterministic single forward pass frameworks have recently been established as deep learning models and have shown competitive performance in certain tasks. The unique combination of spectrally normalized weight matrices and residual connection networks with an approximate Gaussian process (GP) output layer can here offer the best trade-off between performance and complexity. We utilize this framework with a refined version that adds spectral batch normalization and an inducing points approximation of the GP for the task of OOD detection in remote sensing image classification. This is an important task in the field of remote sensing, because it provides an evaluation of how reliable the model's predictive uncertainty estimates are. By performing experiments on the benchmark datasets Eurosat and So2Sat LCZ42, we can show the effectiveness of the proposed adaptations to the residual networks (ResNets). Depending on the chosen dataset, the proposed methodology achieves OOD detection performance up to 16% higher than previously considered distance-aware networks. Compared with other uncertainty quantification methodologies, the results are on the same level and exceed them in certain experiments by up to 2%. In particular, spectral batch normalization, which normalizes the batched data as opposed to normalizing the network weights by the spectral normalization (SN), plays a crucial role and leads to performance gains of up to 3% in every single experiment. For reproducibility, the code can be found here: https://github.com/ChrisKo94/DUE_Land_Cover.

Index Terms—Distance awareness, land cover classification, out-of-distribution (OOD), spectral normalization (SN), uncertainty quantification.

Manuscript received 8 November 2023; revised 3 January 2024; accepted 23 January 2024. Date of publication 10 April 2024; date of current version 12 April 2024. This work was supported in part by the Helmholtz Association through the Joint Research School “Munich School for Data Science (MUDS)”; in part by the German Federal Ministry of Education and Research (BMBF) through the Framework of the International Future AI Lab “Artificial Intelligence for Earth Observation (AI4EO): Reasoning, Uncertainties, Ethics and Beyond” under Grant 01DD20001; and in part by the German Federal Ministry of Economics and Technology through the Framework of the “National Center of Excellence ML4Earth” under Grant 50EE2201C. (Corresponding author: Xiao Xiang Zhu.)

Christoph Koller is with the Chair of Data Science in Earth Observation, Technical University Munich, 80333 Munich, Germany, and also with the Remote Sensing Institute, German Aerospace Center, 82234 Weßling, Germany.

Peter Jung is with the Institute of Optical Sensor Systems, German Aerospace Center, 12489 Berlin, Germany, and also with the Communications and Information Theory Group, Technical University of Berlin, 10623 Berlin, Germany.

Xiao Xiang Zhu is with the Chair of Data Science in Earth Observation, Technical University Munich, 80333 Munich, Germany, and also with the Munich Center for Machine Learning, 80992 Munich, Germany (e-mail: xiaoxiang.zhu.ieee@gmail.com).

Digital Object Identifier 10.1109/LGRS.2024.3375370

I. INTRODUCTION

OVER the last decade, the urgency for having a notion of uncertainty along with a model's prediction has arisen in many research communities. Many different techniques have been invented and developed by the machine and deep learning community [1]. Out of the developed approaches, many uncertainty quantification techniques solely focus on cleverly transforming the output of one or multiple deep learning models. In their work, Lakshminarayanan et al. [2] train multiple neural networks in parallel and achieve state-of-the-art performance on many computer vision and regression tasks. The combination of the individual model predictions also yields an expressive notion of model uncertainty. Another famous example is given by Monte Carlo dropout [3], where the model architecture is left unchanged during inference time with the exception of leaving the existing dropout mechanism enabled. Again, the resulting predictions give a sense of model uncertainty accompanying the prediction itself.

More methodologically advanced approaches are covered exemplarily by the broad field of Bayesian neural networks [4]. Here, the individual weights of a neural network architecture are not treated as fixed parameters, but rather as distributions. While tracing the exact likelihood of such networks becomes infeasible and computationally expensive, uncertainties can be modeled much more fine-grained by averaging over the posterior weight distributions. Many extensions of the core idea have been introduced, for example, a sampling-free variational inference approach [5], or a method for inference on the subnetwork level [6].

Contrastive to the Bayesian approaches, deterministic methods do not place distributional assumptions over parts of the network and do not rely on sampling from distributions or predictions. Especially, for the downstream task of detecting out-of-distribution (OOD) data, this class of networks has been shown to yield competitive performance [7]. The approaches can generally be divided into generative and discriminative ones. Generative approaches utilize an explicit likelihood from a generative model, whereas discriminative approaches use regularized predictions to form their predictive uncertainties.

Regarding the applicability and usefulness of such uncertainty quantification approaches in the field of remote sensing, little work has been done so far. Landgraf et al. [8] modified the loss function to incorporate predictive uncertainties during training and showed convincing results on semantic segmentation tasks. Regarding OOD detection, Gawlikowski et al. [9] used a Dirichlet prior network that yielded strong performance on a range of remote sensing image classification datasets

with respect to OOD detection. The little attention paid to uncertainty quantification in deep neural networks for remote sensing motivates this work, which is focused on employing so-called distance-aware network architectures by ensuring Lipschitz continuity of the network mapping from the input to the output space.

II. METHODOLOGY

We now present the general framework of distance-aware uncertainty quantification as initially presented by Liu et al. [10] and refined by van Amersfoort et al. [11]. The distance awareness here refers to the model's capacity to adequately project distances in the input space onto the output space or prediction of the model. As explained in [10], this is equivalent to the existence of a notion of uncertainty $u: \mathcal{F}(\mathcal{X}) \rightarrow \mathbb{R}_+$ satisfying this adequate projection property. This uncertainty measure operates on the output space of the network $\mathcal{F}$, and $\mathcal{X}$ here describes the input space of the network. For a new point $x \in \mathcal{X}$ and the training data $\mathcal{X}_{\text{train}} \subset \mathcal{X}$, we first denote a fitting metric by $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$, e.g., of the form $d_X(x_1, x_2) = \|x_1 - x_2\|_{\mathcal{X}}$, $x_1, x_2 \in \mathcal{X}$ with an arbitrary yet suitable norm. With this metric, for the network to be distance-aware, the uncertainty measure u needs to satisfy

$$\forall x \in \mathcal{X}: u(x) = v(\mathbb{E}_{x' \sim \mathcal{X}_{\text{train}}} d(x, x')) \quad (1)$$

for a given monotonically increasing function $v: \mathbb{R}_+ \rightarrow \mathbb{R}_+$.

A. Hidden Mapping Distance Preservation

Consider a neural network classifier, which yields the well-known unscaled prediction $\text{logit}(x) = g \circ h(x)$, where h describes the hidden mapping, i.e., all layers up until the last hidden representation and g describes the output mapping onto the label space. In this scenario, the distance awareness can be separately defined for both mappings. Regarding the hidden mapping h , Lakshminarayanan et al. [10] state that the distance awareness is here equivalent to fulfilling a bi-Lipschitz constraint of h . This means that there need to exist constants $0 < L_1 \leq 1 \leq L_2$, such that

$$L_1 \cdot d_X(x, x') \leq d_X(h(x), h(x')) \leq L_2 \cdot d_X(x, x') \quad (2)$$

holds for all $x, x' \in \mathcal{X}$.

Several regularization techniques have been established to enforce this constraint. For example, the constraint can be fulfilled for differentiable functions if the norms of the gradients are bounded. This can be enforced by adding a penalty term $\mathcal{P}_L$ (the subscript L stands for loss) of the form

$$\mathcal{P}_L := \lambda \cdot [\|\nabla_x f_\theta(x)\|_2^2 - L_2]^2 \quad (3)$$

to the loss function, where $\nabla_x f_\theta(x)$ denotes the gradient of the neural network mapping with respect to x , L_2 is the earlier mentioned upper Lipschitz constant, and λ denotes a hyperparameter steering the impact of the gradient penalty on the loss. This procedure is called two-sided gradient penalty and has been successfully applied to generative adversarial networks (GANs) [12] and radial basis function (RBF) networks [13]. Alternatively, the Lipschitz constant L_2 can be controlled by

normalizing the weight matrices of every layer. Specifically, the Lipschitz constant of a fully connected layer $f_k(x) = Wx$ can be represented via the following:

$$\|f\|_{\text{Lip}} = \sup \sigma(\nabla f_k(x)), \quad \text{where} \\ \sigma(A) := \max_{x: x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2 \leq 1} \|Ax\|_2. \quad (4)$$

This spectral norm is equivalent to the largest singular value of A . Since most common activation functions have Lipschitz norm 1 (or predefined L), the Lipschitz constant of a neural network can be computed by the multiplication of the Lipschitz constants of the individual layers. Miyato et al. [14] made use of the normalization of every layer weight matrix via spectral norm (pointwise division) with respect to the l_2 norm for GANs. This proved to be superior over other weight clipping or regularization techniques (e.g., [15], [16]) and over the gradient penalty earlier introduced.

B. ResNet Hidden Mapping

A key neural network architecture for remote sensing image classification is the class of residual networks (ResNets). Since the now following approach is tailored toward ResNets, it is highly relevant for the remote sensing field. For ResNets [17], there exists a more profound way to control the bi-Lipschitz constraint of the hidden mapping h . The hidden mapping is here given via $h(x) = h_l \circ h_{l-1} \circ \dots \circ h_1(x)$, where $h_j(x) = x + \gamma_j(x)$, $j = 1, \dots, l$, denotes the individual residual blocks. If and only if all of the residual mappings $\gamma_j(x)$ $j = 1, \dots, l$ of the residual blocks are α -Lipschitz for $0 < \alpha < 1$ on $\mathcal{X}$, the bi-Lipschitz condition in (2) holds. Because the concatenation of α -Lipschitz functions is again α -Lipschitz [18], the individual Lipschitz conditions are sufficient for the bi-Lipschitzness of the entire hidden mapping h . For the l residual blocks contained in h , we are then left with the following Lipschitz constants:

$$0 < L_1 = (1 - \alpha)^l < 1 < L_2 = (1 + \alpha)^l < \infty. \quad (5)$$

Given the dataset spans over all of $\mathbb{R}^n$ with respect to the Euclidean norm, the condition $\alpha < 1$ of the bi-Lipschitzness is equivalent to bounding the spectral norm (i.e., largest singular value λ) of the weight matrix W_j of γ_j . In more realistic settings, the spectral norm needs to be approximated. The power iteration presented by Behrmann et al. [20] is most often used in practice for this approximation due to its fast convergence. In order to bound the spectral norm for each layer $j = 1, \dots, l$, the respective weight matrix of the convolutional layer is updated via

$$W_j = \begin{cases} c \cdot W_j / \hat{\lambda}, & \text{if } c < \hat{\lambda} \\ W_j, & \text{otherwise} \end{cases} \quad (6)$$

where $\hat{\lambda}$ is the approximation for λ obtained by the power iteration. Also, c denotes a hyperparameter to practically control the exact upper bound depending on the input data.

As a novelty compared with a previous study [19], we here follow the convention of [11] and apply the spectral normalization (SN) also to the batch normalization. The Lipschitz constant of the batch normalization operator is given by

TABLE I

EXPERIMENTAL RESULTS ON THE EUROSAT DATASET. AUROC AND AUPR ARE DERIVED FROM A BINARY CLASSIFIER FED WITH ID AND OOD DATA FROM THE TEST SET. AUROC = AREA UNDER RECEIVER OPERATOR CURVE, AUPR = AREA UNDER PRECISION RECALL CURVE, DSM = DEMPSTER-SHAFFER METRIC, MSP = $1 - \text{MAXIMUM SOFTMAX PROBABILITY}$, AND PRED.ENT. = PREDICTIVE ENTROPY. BEST RESULTS BY UNCERTAINTY MEASURE IN BOLD, AND OVERALL BEST RESULTS ARE UNDERLINED

ID Data	OoD Data	Method	Loss ↓	AUROC ↑			AUPR ↑		
				DSM	MSP	Pred.Ent.	DSM	MSP	Pred.Ent.
Non-built classes	Built-up Classes	WideResNet [19]	0.93	0.32	0.65	-	0.32	0.59	-
		WideResNet-SN [19]	0.94	0.17	0.74	-	0.34	0.66	-
		WideResNet-GP [19]	1.10	0.26	0.65	-	0.36	0.60	-
		WideResNet-GP-IP	0.14	0.77	0.90	0.90	0.56	0.81	0.80
		WideResNet-GP-IP-BN	0.14	0.76	0.90	0.90	0.54	0.81	0.81
		WideResNet-SNGP [19]	1.38	0.21	0.63	-	0.35	0.57	-
		WideResNet-SNGP-IP	0.15	0.68	0.83	0.83	0.48	0.69	0.67
		WideResNet-SNGP-IP-BN	0.14	0.54	0.73	0.86	0.53	0.73	0.73
Non-built classes	Built-up Classes	ResNet50 [19]	3.27	0.50	0.62	-	0.55	0.55	-
		ResNet50-SN [19]	4.24	0.43	0.42	-	0.42	0.41	-
		ResNet50-GP [19]	3.82	0.65	0.45	-	0.54	0.44	-
		ResNet50-SNGP [19]	3.58	0.52	0.43	-	0.59	0.43	-

$\max_i |(\eta_i / (\text{Var}(x)_i)^{1/2})|$ [21], with η being the learnable scale parameter and i being the index describing the batch size. This normalization differs from plain SN, because it is input data-driven and does not alter the network mapping by a single hyperparameter.

C. Output Mapping Distance Awareness

The distance awareness of the output mapping g can be handled via modeling g as a Gaussian process (GP) [22] on the hidden mapping output space $\mathbb{H} := \{h(x): x \in \mathcal{X}\}$. This process is generally specified by a mean function $m(h)$ and a covariance function $k(h, h')$. Then, an uninformative prior with a mean of 0 and an RBF kernel is placed on the latent process. The likelihood gets built by exposing the GP to all training data; after that the process gets optimized a posteriori. Due to computational and analytical intractability, the spectral normalization + Gaussian process (SNGP) approach [10] proposes the following: 1) a random Fourier feature (RFF) expansion [23] of the initial GP followed by 2) a Laplace approximation for the posterior. We refer the interested reader to [24]. Note that the SNGP approach proved to effectively increase the model's predictive uncertainty quality in an OOD detection setting for remote sensing image classification [19].

Novel to this work is the usage of the deterministic uncertainty estimation (DUE) framework proposed by van Amersfoort et al. [11], which will be investigated by various experiments in the following. For DUE, the GP gets placed directly onto the last layer of the hidden mapping. Then, the K -means algorithm is used to identify m inducing points (defined via the centroids found by the algorithm) in the hidden mapping feature space. The GP is then evaluated only on these points, which keeps the GP nonparametric (as opposed to the RFF expansion in the case of SNGP) [11]. For optimization, the expectation lower bound (ELBO) between the induced

GP and the full likelihood is minimized, and the loss gets backpropagated as usual using stochastic gradient descent (SGD).

III. EXPERIMENT

For the experimental section, we investigate different networks for OOD detection in the context of land cover classification for remote sensing images. In particular, we evaluate the predictive uncertainties on unseen classes for which we differentiate between built-up classes and nonbuilt classes and train on one set and predict on another and vice versa. We closely follow the code implementation of van Amersfoort et al. [11], which can be found in [25]. We expand the experiments conducted by Koller et al. [19] by adding another deterministic uncertainty quantification technique, DUE. It is here denoted by GP-IP, which stands for a GP output mapping with the inducing points (IPs) approximation by van Amersfoort et al. [11]. We compare GP-IP to SNGP output mapping as well as SN and GP individually and the aforementioned batch normalization (BN). Two core networks are considered: ResNets with a depth of 50 (ResNet50) and wide ResNets with a depth of 28 (WideResNet). The OOD detection is carried out by a binary classifier defined by the predictive uncertainties, on which the area under the receiver operator curve (AUROC) and under the precision recall curve (AUPR) are derived. The Dempster-Shafer metric (DSM) and $1 - \text{the maximum softmax probability (MSP)}$ have already been previously used as uncertainty metric in [19]; here, the predictive entropy of the softmax prediction (Pred.Ent.) is added specifically for GP-IP approaches.

Adding to the experimental setup presented by Koller et al. [19], we conduct OOD detection on another remote sensing image classification benchmark dataset, namely, on the So2Sat LCZ42 [26] dataset. This dataset

TABLE II

EXPERIMENTAL RESULTS ON THE SO2SAT LCZ42 DATASET. AUROC AND AUPR ARE DERIVED FROM A BINARY CLASSIFIER FED WITH ID AND OOD DATA FROM THE TEST SET. AUROC = AREA UNDER RECEIVER OPERATOR CURVE, AUPR = AREA UNDER PRECISION RECALL CURVE, DSM = DEMPSTER-SHAFFER METRIC, MSP = 1 - MAXIMUM SOFTMAX PROBABILITY, AND PRED.ENT. = PREDICTIVE ENTROPY. BEST RESULTS BY UNCERTAINTY MEASURE IN BOLD, AND OVERALL BEST RESULTS ARE UNDERLINED

ID Data	OoD Data	Method	Loss ↓	AUROC ↑			AUPR ↑		
				DSM	MSP	Pred.Ent.	DSM	MSP	Pred.Ent.
Built-up classes	Non-built Classes	WideResNet-GP-IP	0.12	0.81	0.87	0.87	0.78	0.83	0.85
		WideResNet-GP-IP-BN	0.11	0.82	0.86	0.87	0.80	0.82	0.84
		WideResNet-SNGP-IP	0.11	0.80	0.87	0.88	0.75	0.82	0.84
		WideResNet-SNGP-IP-BN	0.11	0.82	0.88	0.89	0.79	0.85	0.87
Non-built classes	Built-up Classes	WideResNet-GP-IP	0.09	0.84	0.92	0.93	0.78	0.92	0.93
		WideResNet-GP-IP-BN	0.08	0.86	0.93	0.94	0.81	0.93	0.94
		WideResNet-SNGP-IP	0.08	0.81	0.93	0.93	0.73	0.92	0.92
		WideResNet-SNGP-IP-BN	0.10	0.84	0.92	0.93	0.79	0.92	0.93

contains labeled Sentinel-2 imagery of 42 urban conglomerates around the world. The labels follow the local climate zone (LCZ) scheme from [27]. We split the 17 classes into built-up (seven) and nonbuilt (ten) classes for the OOD detection task. This split is widely used, and a label evaluation study performed by Zhu et al. [26] showed little to no label ambiguity between the two label sets. We use the training set of the publicly available cultural split of the dataset and randomly separate it into training and testing with 70% and 30%, respectively. The EuroSat dataset [28] also contains labeled Sentinel-2 patches, with a ten-class land cover labeling scheme. It was split similarly, with only two of the ten classes being specified as built-up classes. For additional details and class descriptions regarding the Eurosat dataset, see [19] for details.

Regarding the core network and the SNGP approach, we use the same hyperparameters as initially stated by Koller et al. [19] with the exception of having a width of $k = 5$ for the WideResNet (WRN). Although a width of 1 outperformed the full WRN network in the previous work [19], we saw stronger performance with a width of 5 in the case of the DUE approach. The width is a key parameter for the WideResNet, as it controls how many filters are trained in parallel for each convolution. We chose a width of 5, because it clearly outperformed networks with less width and was still running in a reasonable amount of time. The strong performance may be due to the nature of the GP output mapping: before, the GP was evaluated completely; now, a limited number of IPs keep the output mapping more flexible and allow for a more complex hidden mapping.

The number of IPs plays a crucial role in the DUE approach. We experimented with multiple values ranging from 10 to 200 and found 50 to be the best-performing number. Note that the number of IPs is interdependent on the batch size and on the number of classes. Experimentally, we found a batch size of 64 fitting for our layout. This is because a larger batch size not only scales the computational effort of a WRN exponentially but also requires more IPs, which has a similar effect. The additional hyperparameters were chosen

as described in [19]. The predictive uncertainties for the SNGP approaches are computed via the DSM and the maximum softmax probability (see [19] for details). For the inducing point (-IP) methods, we follow the original authors' approach and additionally use the predictive entropy of the softmax vector [11]. Regarding the computational cost, we observed an increased computational demand from the GP output layer of roughly 1.3–1.5 times the demand of the core network itself. Due to very fast convergence, the spectral (batch) normalization has little to no effect on the computational cost.

IV. DISCUSSION

The performance gain of DUE over previous methods is immediately visible in Table I: all combinations using the Gaussian process with IPs (GP-IPs) beat previous baselines by a large margin for the task of detecting OOD data within the Eurosat dataset. Again, the DSM performs poorly, whereas the other two uncertainty metrics achieve more or less similar OOD detection results. Overall, the predictive entropy seems to work best. Interestingly, SN seems to hinder the model from effectively detecting OOD data, particularly for the area under the precision recall curve (AUPR). Despite the high value for c , the unrestricted hidden mapping performs slightly better than the restricted one. On the other hand, the data-driven spectral batch normalization has positive effects on the performance in most settings.

For the So2Sat LCZ42 dataset, an overall higher OOD performance can be observed, and the gap between the AUROC and the AUPR has narrowed; see Table II. This can be partially attributed to the larger size of the dataset. But, the more clear class separation could potentially also play a role. Adding to that, the patch size is only 25% of the size of the Eurosat patches. Due to poor performance on the Eurosat dataset, which lay far from the performance of the DUE approach, we did not perform LCZ42 experiments for the original SNGP approach. The balance between built-up and nonbuilt classes is relatively even, which allowed us to perform the OOD detection task in both ways. Interestingly, the detection of built-up classes as OOD samples works a lot better than

detecting nonbuilt classes when having seen only built-up classes during training. Overall, the earlier findings regarding spectral (batch) normalization can also be applied here.

In general, we achieve similar performance results as in [9] for the LCZ42 dataset. Note, however, that the DUE approach does not need any kind of OOD data during training. This result is interesting, since the approach taken in [9] constructs a framework that explicitly maximizes the loss toward OOD data, and the framework incorporates OOD data for this maximization already during training. The approach of DUE [11] and similar approaches, on the other hand, do not incorporate this data. Instead, they design the network to automatically map points further away, which lie further from the training data in the input space. This design seems to work reasonably well for land cover classification applications; we, therefore, leave further exploration of this important model architecture class for future research.

V. CONCLUSION

A model's ability to express the uncertainty about its prediction can be measured by means of OOD detection. Deterministic approaches with single forward passes achieved superior performance on machine learning benchmark datasets. By constructing such a network architecture to be distance-aware, unseen data points at test time are automatically mapped further away from the training data. This is then reflected by a higher uncertainty in the prediction. The SNGP approach [10] performed very well with little changes to the existing ResNets, and clear advantages of the approach could be shown for the field of remote sensing image classification [19]. Building upon the refined approach of [11], we introduced spectral batch normalization and a GP output layer, which is only realized on a limited number of IPs.

We then revisited the OOD detection task and showed strong results over two land cover benchmark datasets. In particular, for the Eurosat dataset, a performance gain in OOD detection of more than 10% points was observed. Even higher performance numbers were achieved on the So2Sat LCZ42 dataset. Overall, the IPs approximation works more reliably than the previous approximation using Laplace and RFF expansion. We believe this is due to the flexibility provided by the non-parametric properties of the GP and the reduced complexity due to the limited amount of IPs. Similarly, the data-driven spectral batch normalization shows more positive effects than the plain SN. These results emphasize the effectiveness of the proposed deterministic single forward pass uncertainty quantification framework for remote sensing image classification.

REFERENCES

- [1] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. Suppl. 1, pp. 1513–1589, Oct. 2023, Art. no. 1589.
- [2] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
- [3] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1050–1059.
- [4] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.
- [5] J. Schmitt and S. Roth, "Sampling-free variational inference for neural networks with multiplicative activation noise," in *Proc. DAGM German Conf. Pattern Recognit.*, 2021, pp. 33–47.
- [6] E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato, "Bayesian deep learning via subnetwork inference," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2510–2521.
- [7] J. Postels et al., "On the practicality of deterministic epistemic uncertainty," 2021, *arXiv:2107.00649*.
- [8] S. Landgraf, M. Hillemann, K. Wursthorn, and M. Ulrich, "U-CE: Uncertainty-aware cross-entropy for semantic segmentation," 2023, *arXiv:2307.09947*.
- [9] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3140324.
- [10] J. Liu, Z. Lin, S. Padhy, D. Tran, T. B. Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7498–7512.
- [11] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," 2021, *arXiv:2102.11409*.
- [12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," 2017, *arXiv:1704.00028*.
- [13] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9690–9700.
- [14] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
- [15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 214–223.
- [16] G.-J. Qi, "Loss-sensitive generative adversarial networks on Lipschitz densities," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1118–1140, May 2020.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [18] P. L. Bartlett, S. N. Evans, and P. M. Long, "Representing smooth functions as compositions of near-identity functions with implications for deep network optimization," 2018, *arXiv:1804.05012*.
- [19] C. Koller, P. Jung, and X. Xiang Zhu, "Exploring distance-aware uncertainty quantification for remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 5692–5695.
- [20] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen, "Invertible residual networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 537–582.
- [21] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree, "Regularisation of neural networks by enforcing Lipschitz continuity," *Mach. Learn.*, vol. 110, no. 2, pp. 393–416, Feb. 2021.
- [22] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
- [23] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Adv. Neural Inf. Process. Syst.*, pp. 1–8, 2007.
- [24] J. Zhe Liu et al., "A simple approach to improve single-model deep uncertainty via distance-awareness," 2022, *arXiv:2205.00403*.
- [25] J. Van Amersfoort. (2021). *Code for on Feature Collapse and Deep Kernel Learning for Single Forward Pass Uncertainty*. [Online]. Available: <https://github.com/y0ast/DUE>
- [26] X. X. Zhu et al., "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [Software and data sets]," *IEEE Geosci. Remote Sens. Mag. (replaces Newsletter)*, vol. 8, no. 3, pp. 76–89, Sep. 2020.
- [27] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorological Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.
- [28] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Obser. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

A.4 Assessing Conformal Prediction for Remote Sensing: Application to Local Climate Zone Classification

Reference

C. Schweden, P. Jung and X. X. Zhu. Assessing Conformal Prediction for Remote Sensing: Application to Local Climate Zone Classification. *IEEE Transactions on Geoscience and Remote Sensing*. Accepted.

Copyright

Article submitted to IEEE Transactions on Geoscience and Remote Sensing. Reproduced with friendly permission from the authors.

Assessing Conformal Prediction for Remote Sensing: Application to Local Climate Zone Classification

Christoph Schweden *Student Member, IEEE*, Peter Jung, Xiao Xiang Zhu *Fellow, IEEE*

Abstract—While neural networks are at the forefront of solving a variety of tasks in remote sensing, these black-box models often lack a clear sense of uncertainty. Currently experiencing a rebirth in the shine of deep learning is the statistically founded uncertainty framework of conformal prediction (CP). Under certain weak assumptions, this framework provides theoretical coverage guarantees for the ground truth value or class. The coverage guarantee holds for the predictions of the CP framework, which take the form of a prediction set containing one or more classes for classification or segmentation problems, or a prediction interval in the case of regression problems. Additionally, the size of the prediction set or interval also serves as an intuitive indicator of uncertainty.

We investigate this uncertainty framework for a large-scale land cover classification dataset termed *So2Sat LCZ42*, which comes with ten independent expert annotations per annotated satellite image patch. First, we evaluate different adaptations of CP calibration methods on the dataset. Next to the analysis of coverages and prediction set sizes, we also evaluate the methods with respect to the explainability of the conformalizer's uncertainty towards aleatoric and epistemic uncertainty. Additionally, we construct an active learning (AL) framework based on CP prediction sets and show promising results in terms of expressiveness regarding predictive uncertainty. CP-based AL achieves significantly better test performance than random querying and yields lower predictive uncertainties than the model trained on the full training set, with only a fraction of the data needed.

Index Terms—Conformal Prediction, Uncertainty Quantification, Land Cover Classification, Calibration, Deep Learning

I. INTRODUCTION

Uncertainties can arise from many different sources. In remote sensing, these can, for example, be various types of sensor noise during data curation or uncertainties in the data processing pipeline. When remote sensing data is then labeled in the next step, this assignment can also contain uncertainty. We can think of measurement uncertainty regarding the target value or ambiguity among labels for a classification setting. Generally speaking, all of these components form the aleatoric or data uncertainty. This type of uncertainty is inherent in the data and cannot be modeled away. The epistemic or model uncertainty, on the other hand, arises from

model misspecification or lack of adaptation. Since the era of deep learning, many frameworks for quantifying the two sources of uncertainty have been established. While there are both frequentist and Bayesian approaches, which both yield competitive performance, ensemble techniques have also been established that combine the predictions of multiple ensemble members. For classification problems, there exists a straightforward experiment to evaluate the uncertainties: By defining the uncertainty as a counterpart to the model's confidence (typically given by the highest softmax pseudo-probability), one can evaluate the uncertainty against the model's accuracy in a binwise manner. This is usually referred to as calibration and can give a first indication about the quality of the predictive uncertainties.

A novel research direction in the context of Uncertainty Quantification (UQ) has been opened by [1], who introduced *Conformal Prediction (CP)*. Broadly speaking, any predictor can be modified using CP such that the true value or class is predicted with a pre-specified confidence. In more detail, this means that the modified predictor does not yield a single point prediction anymore but instead predicts a set of classes in the case of classification or a prediction interval in the case of regression. A core distinction can be made between transductive and inductive conformal prediction: While transductive [2] approaches require to evaluate the non-conformity measure on all possible subsets of labels within the training set, inductive [3] methods solve this issue by requiring a separate hold-out calibration set on which the measure is evaluated. Here, it is essential that the calibration set is *exchangeable* [4] with both the training and the testing set. This is equivalent to the fact that any finite permutation of a subset coming from one of the three sets (or a mix thereof) follows the same joint probability distribution as the original subset. Note that this assumption is similar to the *i.i.d.* (independent, identically distributed) assumption of deep learning frameworks, and one can link the two assumptions using De Finetti's theorem [5]. Furthermore, the conformalized model also inherits a notion of predictive uncertainty next to the theoretically proven coverage guarantee of the true class or point. This notion is typically stated by the size of the prediction set in the classification case or, for regression problems, by the width of the prediction interval.

For the evaluation of predictive uncertainties, one usually has to construct a downstream experiment that explicitly uses the quantified uncertainties. This is required because different standalone uncertainty estimates are rarely comparable in a meaningful way. One such downstream task is active learning (AL), where a model is initially trained with a relatively small subset of the existing training dataset. Once trained, the model's

C. Schweden is with the Department of Data Science in Earth Observation, Technical University of Munich, Munich, Germany. C. Schweden is also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany.

P. Jung is with the Institute of Optical Sensor Systems, German Aerospace Center, Berlin, Germany and with the Communications and Information Theory Group, Technical University of Berlin, Berlin, Germany.

X.X. Zhu is with the Department of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Munich Center for Machine Learning, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

predictive uncertainties on the remainder of the training dataset are utilized to sequentially add data to the initial dataset pool. In an early work, [6] describe this process of *querying* data, which is the key concept in AL. Querying here means nothing other than handing certain identified data points to a so-called human oracle, which then reveals the ground truth label or target value. The model is then retrained on these newly added labeled data points. The overall goal of doing this iteratively is to quickly adapt the model by feeding it with a few representative data samples at a time. Here, the quality of the predictive uncertainties is essential as it determines the representativeness of the predicted data points. What remains crucial is to evaluate the uncertainty-based querying strategies against the random baseline, i.e., adding random points to the training dataset pool. Overall, typically the accuracy on a hold-out testing dataset is evaluated after each retraining step of the model. In the proposed experimental setting, predictive uncertainty is further considered as an objective variable. To summarize, we make the following main contributions:

- We evaluate the promising CP framework by means of various adapted score functions on a large-scale land cover benchmark dataset.
- We investigate the prediction sets with respect to the coverage as well as their explainable power towards aleatoric and epistemic uncertainty.
- We build a novel CP score-based Active Learning benchmark and show promising results that beat the random baseline by a large margin, both with respect to accuracy and predictive uncertainty.

In the following work, we proceed as follows: First, Section II gives an overview of current and past research in the fields of UQ, CP, and AL, as well as applications of UQ to the field of remote sensing. Then, Section III introduces the dataset and the core neural network utilized, before the general CP framework for classification is introduced. Furthermore, the means of evaluating CP methods are introduced, and an AL framework based on CP is explained. Next, Section IV shows experimental results on the *So2Sat LCZ42* dataset, namely a benchmark of different CP methods with respect to coverage guarantees, prediction set size, and expert label votes coverage as well as an analysis on class level. Lastly, Section V presents the AL experimental results based on a CP-based querying function, before Section VI concludes.

II. RELATED WORK

A variety of streams have evolved to tackle UQ in deep learning; see [7] for a recent overview. Although creating a significant computational overhead, ensembles of networks [8] have been established as a hard-to-beat baseline in many applications. Using a Bayesian setting, on the other hand, allows for relaxing the solution space of a network by placing prior distributions on the weights [9] [10]. A prominent Bayesian approximation technique is stated by *MC Dropout* [11], which uses multiple forward passes through a network with dropout to yield predictive uncertainties. Deterministic networks have been studied as another stream of UQ methods. Here, the uncertainties are based on a single forward pass through a

deterministic network. Popular approaches are *SNGP* [12], which combine spectral normalization of weight matrices and an approximate Gaussian Process (GP) as output layer, and *DUE* [13], which introduces spectral batch normalization and enriches the aforementioned GP with a more stable inducing point approximation. As a substream of the deterministic approaches, Dirichlet-based approaches have been introduced. These utilize Dirichlet distributions to loosen the typical categorical output space in classification problems. Prominent representatives are the *Prior Networks* [14] and the *Evidential Networks* [15].

Since its introduction by [1], conformal prediction has been developed methodologically and applied to many research fields. In particular for deep learning models, many adaptations have been introduced recently [16]. A first milestone approach for classification tasks is given by the *least ambiguous set-valued classifier (LAC)* [17]. Here, the foundation was laid for constructing prediction sets that satisfy the theoretical coverage guarantees and, at the same time, are least ambiguous (i.e., as small as possible). As an advancement of LAC, *adaptive prediction sets (APS)* were introduced soon after by [18]. Via generalizing the quantile function of the LAC approach and adding a sub-training routine to the calibration step, the authors showed increased conditional coverage across various datasets. Through adding a regularization mechanism, [19] constructed *Regularized Adaptive Prediction Sets (RAPS)* and were able to further reduce the prediction set sizes while still fulfilling the theoretical coverage guarantees. The authors further introduced a more straightforward prediction set constructor, namely the *Top-k* algorithm, which simply considers the k most likely predictions where k is again estimated in such a way that it fulfills the pre-specified coverage guarantee. Recent methodological advances include, exemplarily, the adaptation of conformal prediction to cases beyond the exchangeability assumption [20] or linking CP with an early stopping mechanism [21]. Works on fusing CP with Monte Carlo Dropout [22] or with temperature scaling [23] were also presented recently.

AL was introduced to machine learning algorithms by [25], who adopted the framework to Support Vector Machines (SVMs). Since then, many works, particularly involving deep learning architectures, have been published; see [26] for a recent survey. In the field of remote sensing, AL also plays a pivotal role. An early application using SVMs was presented by [27], who used AL for the task of segmenting multispectral images. Further works focused on hyperspectral image classification [28], remote sensing image classification in a more general sense involving multiple algorithms and datasets [29], batch-mode AL [30], and recently also on a multi-label classification framework [31]. AL in combination with conformal prediction has also been investigated. [32] initially proposed a transductive CP framework for querying, and [33] added an eigendecomposition for the indication of information to the framework. An inductive approach, on the other hand, was proposed by [34] for an ensemble of neural networks. [35] constructed a testbed for image classification using notions of informativeness, diversity, and information density in a single query function. The same authors later

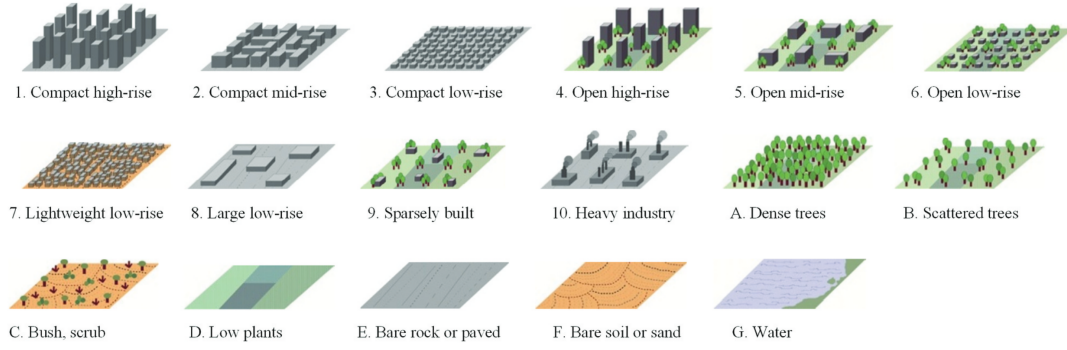


Fig. 1: Local Climate Zone Classification scheme as presented by [24]

also proposed another query function based on a constrained linear regression model, which further incorporates a notion of uncertainty [36].

In the field of remote sensing, uncertainties continue to play an important role. Systematic quantification and evaluation, however, are often not in focus. The uncertainty evaluation downstream task of Out-of-Distribution detection was studied both by [37], who utilized modified versions of prior networks, and by [38], where distance-aware neural networks were considered. Calibration can be seen as another way of evaluating uncertainties in deep learning [39]. [40] investigated the calibration and generalization performance of a remote sensing dataset, which has been independently labeled by ten experts, a continuation of the research stream on human label uncertainty [41]. The expert labels of this dataset were also studied by [42], who utilized a multinomial mixture model and a stochastic version of the Expectation-Maximization algorithm, assuming no label ambiguity. Recently, conformal prediction has also found potential applications in remote sensing. [43] studied the effect of label noise on the calibration results of various CP methods, while [44] investigated three large-scale use-cases for CP in remote sensing image analysis. Further CP applications include Land Use Land Cover (LULC) classification [45], soil organic carbon estimation [46], birds and drone detection [47], as well as harvest readiness classification [48] and tropical cyclone tracks forecasting [49].

III. METHODOLOGY

A. Data & Model

The dataset that will be studied in this work follows the classification scheme of the *Local Climate Zones (LCZs)* introduced by [50]. This 17-class scheme describes urban conglomerates and their surroundings and can be further categorized into urban and vegetation classes. A dataset with global coverage of 42 cities was introduced to the community by [51], who manually labeled co-registered Sentinel-1 and Sentinel-2 patches according to the LCZ labeling scheme. Note that the dataset does not cover the entire cities, but focuses on homogeneous regions which can be attributed to a single class. Nonetheless, significant ambiguity exists among the classes. This ambiguity can be attributed to the characterization of the different classes, which exemplarily depend on the density and height of the buildings in the case of the urban classes.

This ambiguity motivates the use of conformal prediction, as a prediction set with coverage guarantee can help to better explain the model's uncertainty when dealing with ambiguous classes.

Next to the global dataset, a subset of the data was also investigated to determine a notion of label confidence. For this, ten remote sensing experts were asked independently to label the identical data coming from ten European cities. By doing so, a label confidence (i.e., average agreement rate among the experts) of ca. 85% was estimated. We denote the individual label votes of the ten experts for image i via

$$\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iK}) \text{ where } Y_{ik} = \sum_{j=1}^M \mathbf{1}(V_{ij} = k) \quad (1)$$

and $V_{ij} \in \{1, \dots, K\}$ denotes a vote from expert j on image i for class k ($K = 17$ denotes the number of LCZ classes). We furthermore normalize the vote vectors yielding the empirical distribution of the votes given by

$$\mathbf{y}_{\text{distr}}^{(i)} = \mathbf{Y}^{(i)} / M \quad (2)$$

where $M = 10$ denotes the number of voters. The distributional (soft) label defined via the empirical distribution (2) will be used for training the models later on in order to be label uncertainty-aware. Although moderate ambiguity exists among the experts, for each image, the majority-voted class is uniquely defined. We denote this class by y_{maj} and treat it as the ground truth class where applicable.

The 17 classes can be divided into ten building and seven non-building classes, see Figure 1. Since the agreement of experts for the non-building classes was much higher on average (and those classes are generally easier to classify and yield very little ambiguity), we decided to solely target the building classes with our CP framework in order to focus on the more expressive classes in terms of label uncertainty. Little to no ambiguity exists among building and non-building classes (see [51] for details). Therefore, we manually excluded the non-building samples from the dataset without having to sacrifice label information from the building classes. We are thus left with ca. 60 thousand labeled Sentinel-2 images, which we will utilize in the following experiments. As shown by [52], the relatively small patch size of 32 by 32 pixels of these data samples does not require a particularly deep neural network for

effectively learning the labels. The authors introduced *Sen2LCZ*, a lightweight convolutional neural network that outperformed classical deep architectures such as a ResNet [53] or a VGG [54]. Due to this fact and later effective usages of this model (see, e.g., [37]), we choose this model as the core network. It is noteworthy, however, that ResNets are theoretically *distance-aware* [12], which makes them particularly suitable for Out-of-Distribution detection tasks [38], [43]. Previous advancements by [40] led to our choice of the *Kullback-Leibler (KL-)* Divergence as loss function in combination with the soft labels earlier defined. Hyperparameters and model specifications, such as early stopping and learning rate decay, were chosen similarly to previous works. For a more realistic testing scenario, we decided to use the train, calibration and test split proposed by [40]: Training data comes from a different set of European cities than calibration and test data, but since the latter two are coming from the same cities, exchangeability between calibration and test data is assured. Nevertheless, there exists a significant generalization gap to bridge between training and testing.

B. General CP Framework for Classification

We now introduce the general framework for inductive conformal prediction in the case of a deep neural network classifier. Let $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1, \dots, n} \in (\mathcal{X} \times \mathcal{Y})^n$ be the data consisting of data points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathcal{X}$ and labels $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \in \mathcal{Y} = \{1, \dots, K\}$ coming from the sets $\mathcal{X}$ and $\mathcal{Y}$. We learn the mapping between labels and data points using a neural network $f_\theta(x)$ depending on the parameters stored in θ . Given an input $x \in \mathcal{X}$, the network yields $p_\theta(Y|x)$ as predictive distribution. In more detail, $\hat{\pi}_k(x) := p_\theta(Y = k|x)$ returns the estimated probability of x belonging to class k . We furthermore denote the confidence of the network via $\hat{p} = \max_k \pi_k(x)$ and the corresponding class prediction as $\hat{y}$. The core principle of conformal prediction for classification lies within the definition of a prediction set $\mathcal{C}(X)$. This prediction set can contain one or more labels from $\mathcal{Y}$, as opposed to the single class prediction $\hat{y}$, or can be empty. The CP framework now theoretically guarantees a coverage guarantee of its prediction sets on the unseen test set, meaning that the set of true test labels Y_{test} lies within the (set of) prediction sets $\mathcal{C}(X_{\text{test}})$ within a certain probability:

$$\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \geq 1 - \alpha, \quad 0 < \alpha \ll 1 \quad (3)$$

Here, α denotes a pre-specified confidence level, which steers the error rate of the coverage guarantee. To construct the prediction sets, we set aside a calibration set different from the training and testing sets. The weak assumption mentioned earlier for the coverage guarantee to hold is the *exchangeability* [4] of those three sets. Concretely, any finite permutation of a subset of the union of the three sets has the same joint probability distribution as the initial subset. As a next step, we calculate *conformal scores* $s(x, y) \in \mathbb{R}$ on the calibration set of size n_{calib} that supposedly quantify the likeliness of the model's prediction. Based on the scores calculated from the calibration set, a threshold is set via $\hat{t} := \frac{\lceil (n_{\text{calib}} + 1)(1 - \alpha) \rceil}{n_{\text{calib}}}$. Then, we utilize the exchangeability assumption to carry over

this likelihood behavior to the unseen test data. Concretely, we derive the predictive scores on the test set and define the prediction sets as follows:

$$\mathcal{C}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{t}\} \quad (4)$$

The now following versions of CP for classification problems with neural networks solely differ in their choice of conformal scores. For the *least ambiguous set-valued classifier (LAC)* [17] algorithm, we set $s(x, y_{\text{true}}) = 1 - \hat{\pi}_{y_{\text{true}}}(x)$, namely one minus the softmax value of the true class, which we know for the calibration set. The threshold $\hat{t}$ is then calculated as defined above and denotes the remaining uncertainty for true class predictions in $(1 - \alpha)\%$ of all calibration data points. This uncertainty is taken as the upper bound for (the maximum of) the predicted softmax probabilities on the test set. Due to the assumed exchangeability, the coverage guarantee of $(1 - \alpha)\%$ is fulfilled.

As an adoption of this method, [18] introduced *Adaptive Prediction Sets (APS)*. The goal here lies in optimizing the conditional coverage, referring to coverages within certain classes or subgroups. The authors propose as a score function

$$s(x, y_{\text{true}}) = \sum_{j=1}^{k=y_{\text{true}}} \hat{\pi}_{\text{ord}, j}(x) \quad (5)$$

which includes all ordered softmax values $\hat{\pi}_{\text{ord}, 1}(x), \dots, \hat{\pi}_{\text{ord}, K}(x)$ from highest to lowest until reaching the ground truth class (known for the calibration set). Again, the quantile is formed as above, and the prediction set is formed via (4). Building upon APS, [19] proposed an improved version termed *Regularized Adaptive Prediction Sets (RAPS)*. The authors imposed an additional regularization term to enforce even smaller prediction set sizes. With $\lambda \geq 0$ and $k_{\text{reg}} \geq 0$ being the regularization hyperparameters, they change the scoring function to

$$s(x, y_{\text{true}}) = \sum_{j=1}^{k=y_{\text{true}}} \hat{\pi}_{\text{ord}, j}(x) + \lambda(\sigma_x(y_{\text{true}}) - k_{\text{reg}})^+ \quad (6)$$

where $\sigma_x(y)$ denotes the ranking of the true label y_{true} within the softmax pseudo-probabilities. The prediction sets are then formed similarly as before via (4).

Lastly, we consider the *top-k* conformal calibration method, which takes the classes with the k highest softmax values as the prediction set. Note that k gets optimized on the calibration dataset to satisfy the theoretical coverage guarantees. In more detail, k gets chosen as the $\lceil (n_{\text{calib}} + 1)(1 - \alpha) \rceil$ largest value of the ranks of the true classes within the ordered softmax values of the predictions on the calibration dataset.

C. Evaluation Metrics for Conformal Prediction

The overall coverage serves as the key metric for conformal prediction. It measures the percentage of prediction sets that actually contain the true label. For an open-world application of CP, the coverage also represents the key promise of the framework: Even if the single class prediction is incorrect, we can rely on the prediction set to restrict the space of possible classes for a given sample. Typically, coverage and

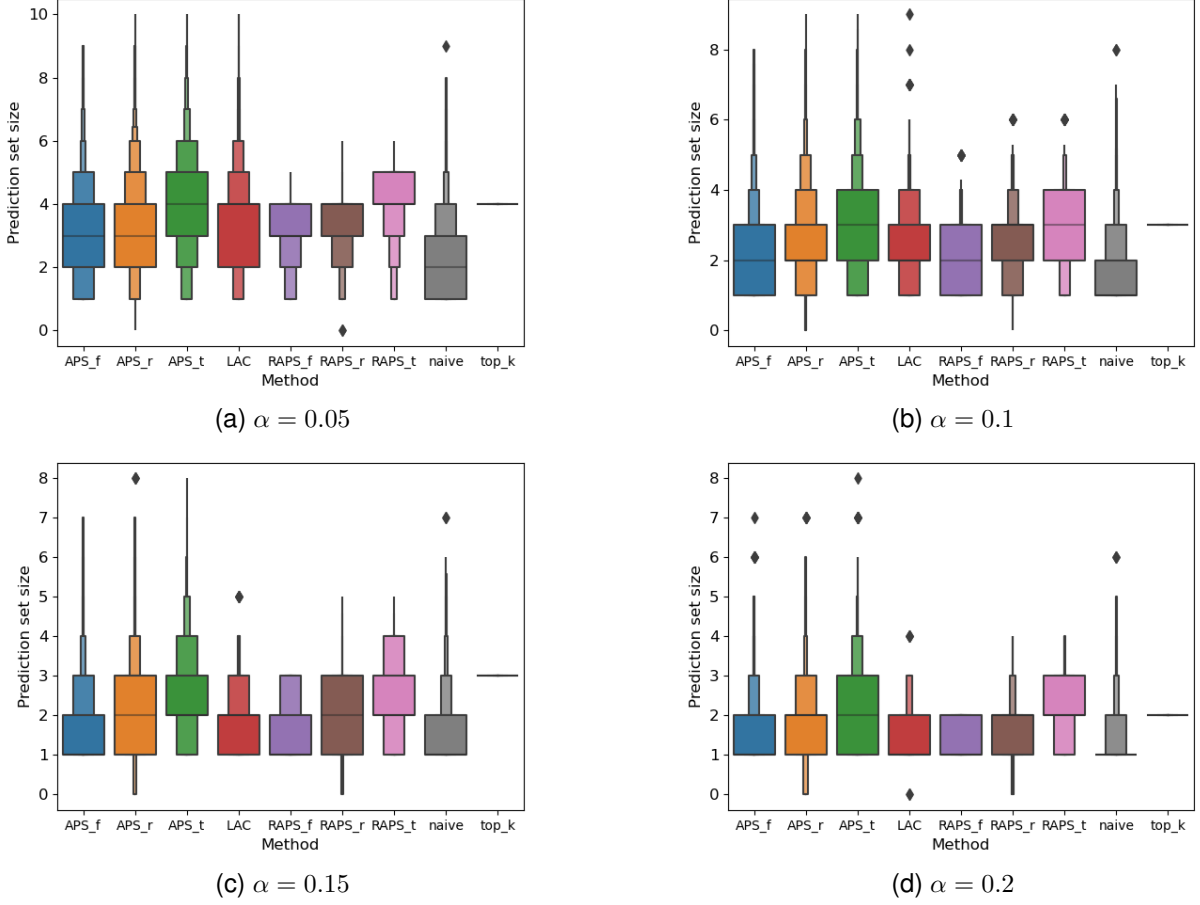


Fig. 2: Prediction Set Sizes by values of α for the different conformalization methods.

the following metrics are computed during inference on the held-out test set of size n_{test} , for which the measured coverage reads as follows:

$$\text{Cov} := \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\{y_i \in \mathcal{C}(x_i)\}} \quad (7)$$

Ideally, the measured coverage fulfills the theoretical coverage guarantee and additionally produces small prediction sets. The average prediction set size, also termed *inefficiency* [55], is used as a metric for this. On the test set, it is calculated via

$$\text{Ineff} := \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}(x_i)| \quad (8)$$

As a peculiarity of the studied dataset, we furthermore investigate the degree to which the different conformal prediction algorithms cover the votes cast by the human experts during the labeling stage in their prediction sets. For this, fix $i \in \{1, \dots, n_{\text{test}}\}$ and test input image x_i and define $V_i := \cup_{j=1}^M \{V_{ij}\} \subset \{1, \dots, K\}$ as the set of classes that experts voted for. We then look at the coverage of all labels that received votes, denoted by

$$\text{Cov}_{\text{lab}, i} := \frac{|V_i \cap \mathcal{C}(x_i)|}{|V_i|}. \quad (9)$$

Although the label votes coverage yields the percentage of all cast label votes that are covered by the prediction set, it does not penalize large prediction sets, which naturally include more labels voted. For this reason, we also investigate the *Intersection over Union (IoU)* between the two sets, namely the prediction set and the set of labels voted by the experts:

$$\text{IoU}_{\text{pred.-vot.}, i} := \frac{|V_i \cap \mathcal{C}(x_i)|}{|V_i \cup \mathcal{C}(x_i)|} \quad (10)$$

Neither metrics yet take into account the actual number of votes cast per class. In order to specifically include this information, we modify the above metrics to focus on the realizations of Y_{ik} instead of just the support. Again, we can first investigate the coverage of the label votes by measuring the percentage of all votes covered by the prediction set:

$$\text{Cov}_{\text{lab.-vot.}, i} := \frac{\sum_{k=1}^K Y_{ik} |V_i \cap \mathcal{C}(x_i) \cap \{k\}|}{\sum_{k=1}^K Y_{ik} |V_i \cap \{k\}|}. \quad (11)$$

For the IoU, the two terms of the fraction can be investigated separately. For the intersection, we can use the above formulation, which measures all votes cast that lie in the prediction

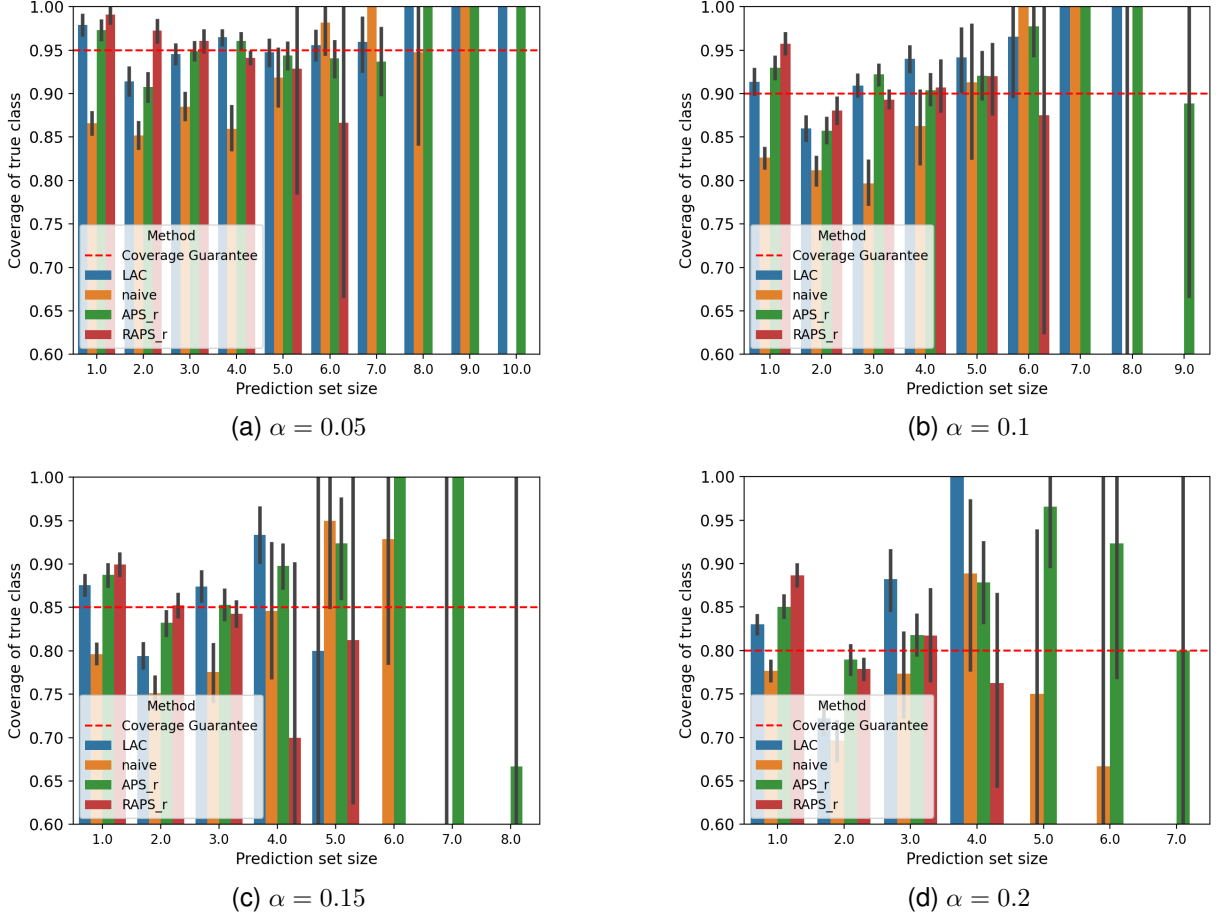


Fig. 3: Coverages of chosen (conformalized) predictors grouped by prediction set sizes, separately for different α values.

set. For the union, we need to consider how to treat classes that are in the prediction set and have no votes cast by the experts. Since we do not rank the members of a prediction set by likelihood, we assign equal probability to all members and use this uniformity in the calculation of the union. Hence, the IoU considering all individual votes reads as follows:

$$\text{IoU}_{\text{lab.-vot.},i} := \frac{\sum_{k=1}^K Y_{ik} |V_i \cap \mathcal{C}(x_i) \cap \{k\}|}{\sum_{k=1}^K (Y_{ik} |V_i \cup \mathcal{C}(x_i) \cap \{k\}| + |(\mathcal{C}(x_i) \setminus V_i) \cap \{k\}|)} \quad (12)$$

IV. CP EXPERIMENTS

A. Coverage and Inefficiency Analysis

First, we compare the prediction set sizes produced by the different CP algorithms introduced in Section III-B for the test samples. In the following, *naive*, *top_k*, and *LAC* describe the respective methods in the visualizations, while *APS* and *RAPS* can be further distinguished. For these two methods, one can decide whether the label that exceeds the threshold is always included (ending *_t*), is always excluded (ending *_f*), or is decided randomly (ending *_r*). Note that solely the random configuration achieves the exact coverage guarantee

theoretically. Figure 2 displays the range of the prediction set sizes of the test set for the different methods, separately for different chosen confidence levels α . Due to the nature of conformalization, lower α values force the conformalized model to produce larger prediction sets in order to achieve the desired coverage. At first sight, the naive method can glance. This is despite the fact that the method does not incorporate any conformalization step, but rather sums up the ranked softmax probabilities until the threshold $1 - \alpha$ is reached. On average, it yields the smallest prediction set sizes when evaluated on the test set. This hints at a relatively large degree of confidence, as most often, one or two class softmax pseudo-probabilities are sufficient to reach the desired threshold. Furthermore note that there exists no theoretical guarantee about the coverage of the true label within the prediction set, hence the softmax pseudo-probabilities cannot be treated as trustworthy estimates of the coverage probabilities.

At the same time, the spread of the set sizes predicted by the naive method is relatively high. APS methods and the LAC conformalizer yield a similar spread (although APS has more outliers), while the RAPS methods are less widely spread in their results. This underscores the importance and effectiveness of regularizing the prediction set sizes in order to have reasonable set sizes even in cases where high uncertainty

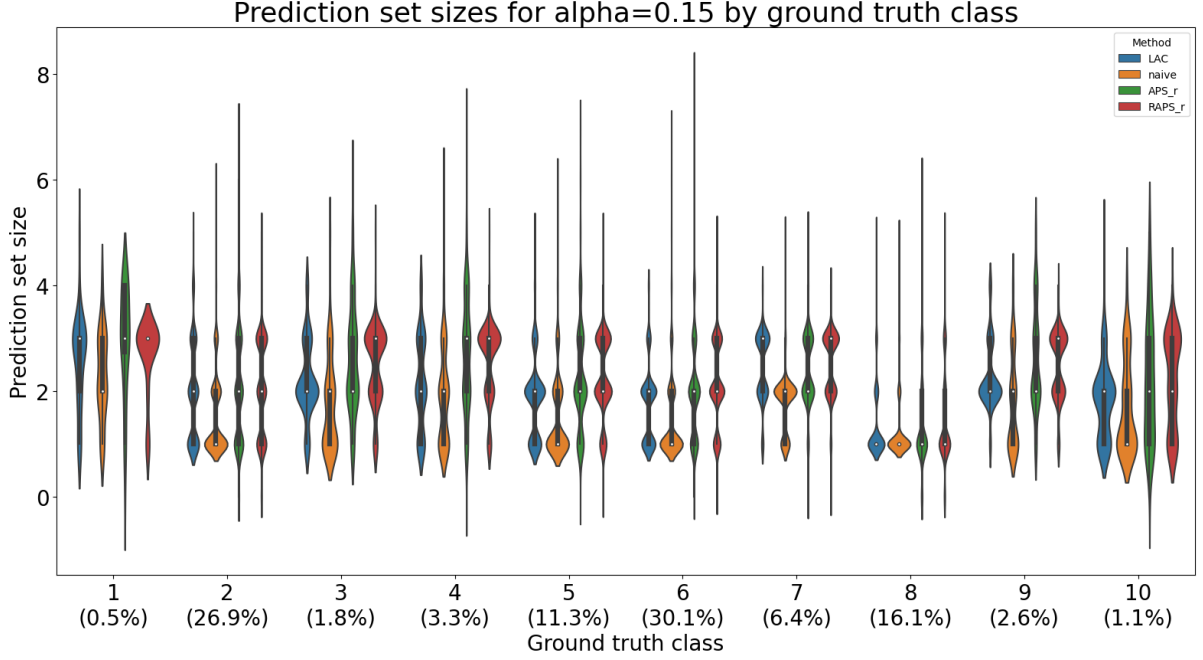


Fig. 4: Inefficiencies per ground truth class for all methods. Relative frequencies of the classes are listed below the respective x-axis ticks.

exists in the softmax pseudo-probabilities. Note, however, that the random versions of APS and RAPS yield empty prediction sets in some cases, which is undesirable. For $\alpha = 0.15$, which represents a realistic scenario by matching the observed label confidence of the dataset with the desired coverage guarantee, the RAPS_f method yields superior performance as it matches closely with the naive method in terms of average prediction set size while showing a smaller spread of the sizes.

Inefficiency by itself, however, does not adequately describe the quality of the predictive uncertainty of a (conformalized) predictor. The theoretical coverage guarantee of the conformalization environment should also be marginally met. To investigate this in more detail, we look at the coverage (9) for each method for varying α values. Note that the $_t$ and $_f$ versions of APS and RAPS were discarded since they do not fulfill the coverage guarantees in theory. Furthermore, the top-k method was intentionally left out due to prediction set sizes far above average. Figure 3 displays the coverages as a barplot with the standard error as a grey bar for the leftover methods by prediction set size (separately for each α value).

Since the coverage guarantee is only given marginally, it is of no surprise that even for the conformalized methods, the guarantee per prediction set size is not always met. Yet, the error bar crosses the coverage line in almost any case.

In this investigation, now the clear downside of the unconformalized *naive* method becomes apparent: Although often producing very small prediction sets, the coverage is often significantly below the theoretical coverage guarantee. The randomized versions of APS and in particular of RAPS can hereby glance, underscoring the effectiveness of the produced

predictive uncertainties by means of the prediction set sizes. Note that in particular for the prediction set size of one, the coverage coincides with the accuracy and furthermore gives a notion of calibration for the naive method: Since the confidence (i.e. highest softmax probability) in this setting lies at least at the threshold $1 - \alpha$, a well-calibrated model should match this value with its accuracy. Here, an apparent mismatch is visible for the naive method, hinting at an overconfident baseline model. Note that the margin is smaller for higher α values, hence the overconfidence is particularly pronounced in the very high confidence regions. After conformalization, the calibration drastically improves. Thus, prediction sets of size one become more trustworthy and better calibrated. However, they become less likely because conformalization tries to balance overconfidence and fulfilling the coverage guarantee.

Even though the coverage guarantee of the proposed conformalizers only holds marginally, it is still of great value to investigate the inefficiencies on a class level. To do so, we can look at the inefficiencies of all methods grouped by the underlying ground truth class (see e.g. Figure 4 for $\alpha = 0.15$). Unsurprisingly, the different (conformalized) methods yield similar patterns of the prediction set sizes on class level, only differing in their spread. Yet more interestingly, underrepresented classes do not necessarily yield larger prediction set sizes. Seemingly, some classes are associated with higher uncertainty, regardless of their share in the dataset.

B. Conformal Prediction Uncertainty as Indicator

In contrast to many other UQ methods, conformal prediction offers a very clear and precise notion of predictive uncertainty

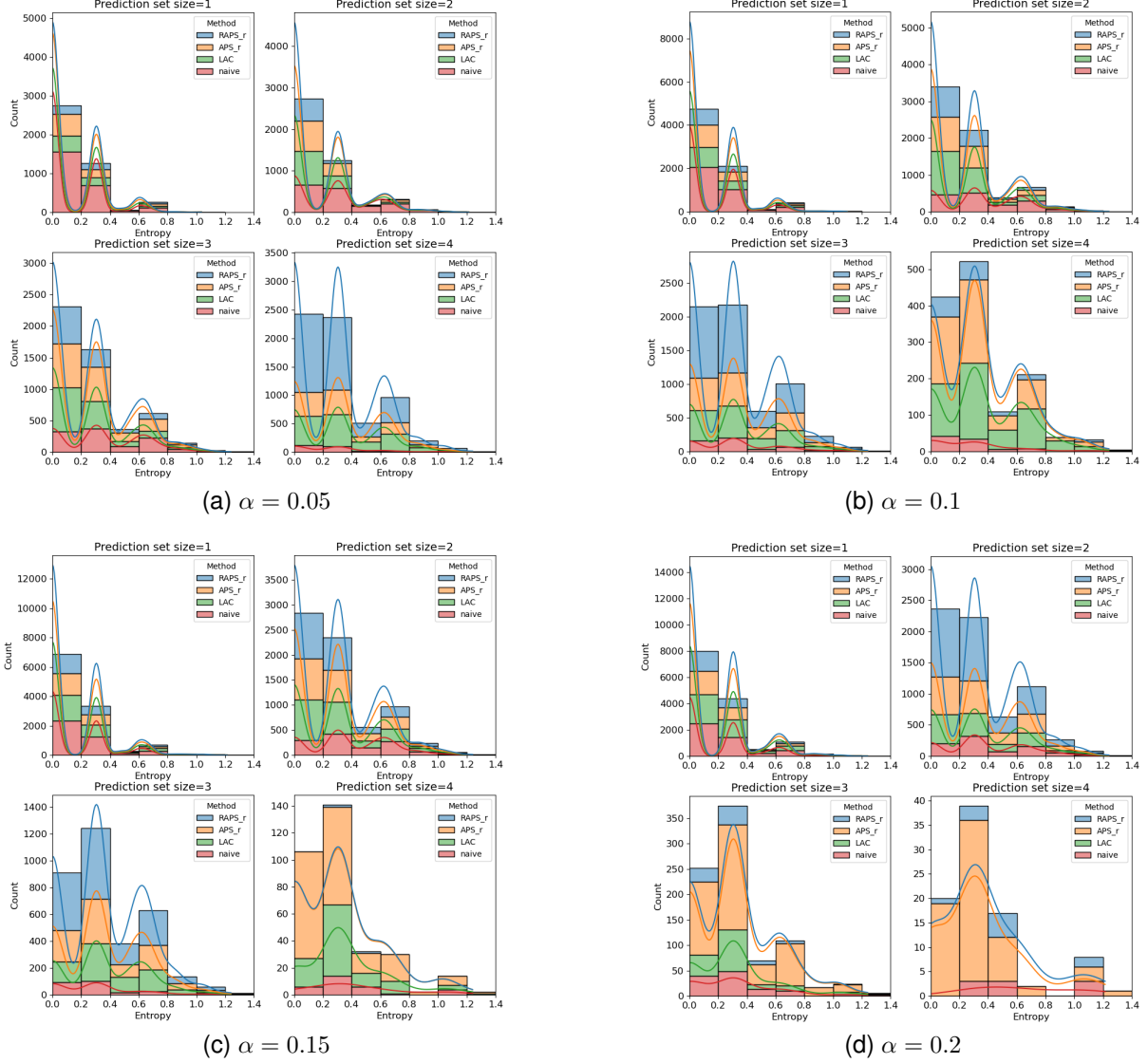


Fig. 5: Label uncertainty (by means of Shannon entropy) grouped by prediction set size, separately for each α value.

via the prediction set size. Naturally, this uncertainty includes both aleatoric and epistemic uncertainty. The hope of conformalization lies in refining a classifier’s uncertainty. While coverage checks the validity of the prediction set size, the question can be raised whether it can also successfully hint at other uncertainty quantities. In the following, we therefore want to investigate its indicative power with respect to the underlying label uncertainty and to a notion of epistemic uncertainty by means of predictive performance.

We first start by comparing the prediction set size with the underlying degree of aleatoric label uncertainty, which we measure via the Shannon entropy of the soft label (2) defined via

$$H(\mathbf{y}_{\text{distr}}^{(i)}) = - \sum_k \mathbf{y}_{\text{distr},k}^{(i)} \cdot \log \mathbf{y}_{\text{distr},k}^{(i)}. \quad (13)$$

Visually, we inspect the interdependency between the total uncertainty of the prediction sets and the aleatoric uncertainty

of the labels in Figure 5. For each prediction set size separately, we look at the distribution of the entropies in a stacked barplot. Note first of all, as was already observed in previous experiments, that the unconformalized naive method yields generally lower prediction set sizes. What also becomes apparent is the larger share of medium and high entropy values for the conformalized methods, e.g., in the $\alpha = 0.05$ setting. The effect is less visible for higher α values since the overall average prediction set size decreases. Yet, the trend is still visible as the red bars of the naive method are diminishing for higher prediction sets.

In order to better answer the question of whether higher prediction sets of conformalized methods can effectively hint at larger underlying label uncertainty, we state the conditional probabilities of the presence of label uncertainty given a fixed prediction set size in Table I below. For the methods previously discussed, the table presents the observed probabilities of a

α	Method	Ineff.	$P(H(y) = 0 \text{Ineff} = k)$	$P(H(y) > 0 \text{Ineff} = k)$	Odds($H(y) > 0$)
0.05	APS	1	66.39 %	33.61 %	0.51
		2	62.91 %	37.09 %	0.59
		3	42.74 %	57.26 %	1.34
		4	38.37 %	61.63 %	1.61
		5	39.16 %	60.84 %	1.55
		6	53.24 %	46.76 %	0.88
		7	40.51 %	59.49 %	1.47
		8	12.50 %	87.50 %	7.00
		9	22.58 %	77.42 %	3.43
0.05	LAC	1	61.33 %	38.67 %	0.63
		2	64.57 %	35.43 %	0.55
		3	50.58 %	49.42 %	0.98
		4	35.25 %	64.75 %	1.84
		5	39.65 %	60.35 %	1.52
		6	45.79 %	54.21 %	1.18
		7	49.66 %	50.34 %	1.01
		8	24.32 %	75.68 %	3.11
		9	13.33 %	86.67 %	6.50
		10	16.67 %	83.33 %	5.00
0.05	RAPS	1	51.17 %	48.83 %	0.95
		2	82.66 %	17.34 %	0.21
		3	54.00 %	46.00 %	0.85
		4	41.57 %	58.43 %	1.41
		5	14.29 %	85.71 %	6.00
		6	6.67 %	93.33 %	14.00
0.05	naive	1	60.94 %	39.06 %	0.64
		2	42.45 %	57.55 %	1.36
		3	34.89 %	65.11 %	1.87
		4	46.35 %	53.65 %	1.16
		5	48.16 %	51.84 %	1.08
		6	16.36 %	83.64 %	5.11
		7	27.27 %	72.73 %	2.67
		8	10.53 %	89.47 %	8.50

TABLE I: Inefficiency Analysis for the most promising CP methods (RAPS and APS in randomized version), for $\alpha = 0.05$.

sample having no degree of label uncertainty as well as the complementary probability of having label uncertainty, both conditional on varying prediction set sizes (all based on the relative frequencies observed on the hold-out test set). The last column furthermore states the resulting odds of observing label uncertainty for a given α , method, and prediction set size, given by

$$\text{Odds}(H(y) > 0) := \frac{P(H(y) > 0 | \text{Ineff} = k)}{P(H(y) = 0 | \text{Ineff} = k)} \quad (14)$$

First of all, we observe that generally, across methods, a higher prediction set size increases the likelihood (expressed by the odds) of the presence of label uncertainty. This likelihood differs from method to method: While for the naive method, the odds are relatively small for prediction sets smaller than 5, the odds for the APS method are already quite large beginning with prediction set size 3 and higher (with the exception of prediction set size 6, which might have a small support). Interestingly, the odds for the RAPS method are only starting to drastically increase after prediction set size 3. Therefore, the indication for this method lags a bit behind the other methods. The LAC

method, on the other hand, can be better suited for this task and presents large odds for prediction sets of size 4 and larger (again with a minor exception of class 7 due to insufficient data support).

To understand on a more fine-grained level to what extent the labels voted by the ten experts are covered by the prediction sets produced by the various conformalizers, we investigate the two quantities introduced in (9) and (10). The results are combined and displayed in Figure 6. First, note the small differences across methods for the smallest investigated α value. Yet, there still exists a significant gap in label votes coverage between the unconformalized *naive* method and the conformalized methods. This gap is observed across the different values of α , indicating a lack of indication for label uncertainty for the unconformalized method. Unsurprisingly, the overall numbers decrease for large α values, since this leads to smaller prediction set sizes in general. As expected, the conformalized configurations of APS and RAPS, which always include the borderline label, achieve higher coverages than the configurations that never include this label or which decide randomly (the latter being the only case where the coverage guarantee is exactly met). More interestingly, the IoU is also positively impacted by the inclusion of the borderline label, in

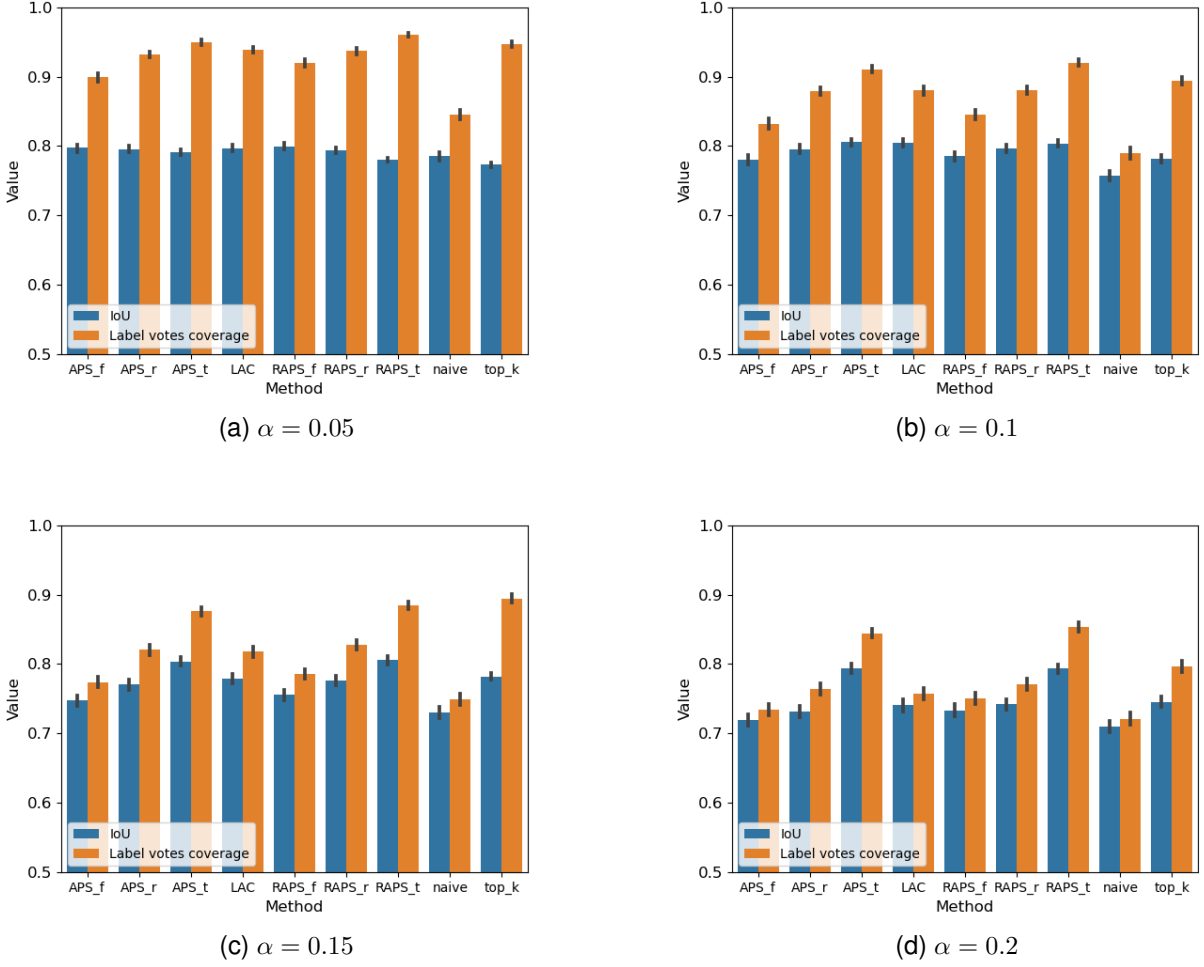


Fig. 6: Label votes coverages and IoUs by values of α .

particular for larger α values. Hence, one can conclude that the borderline label often carries valuable information about the intrinsic label uncertainty. It is noteworthy here that the overall highest IoU is already achieved at the $\alpha = 0.15$ level, hence forcing larger prediction sets with even smaller α values is counterproductive for indicating label uncertainty.

As a further analysis, the interdependency between prediction set size and epistemic uncertainty was studied. We investigate this dependency by looking at the accuracy of the classifier with respect to the prediction set size. Underlying is the claim that an uncertain model is more often incorrect in its prediction than a certain one. Although the theoretical coverage guarantee makes no statement about the accuracy, one would still expect the conformalized model to have higher accuracy at predictions with a smaller prediction set size. Consequently, a large prediction set size should, if expressing higher uncertainty correctly, yield a lower accuracy. This uncertainty expressed by the conformalized model is then compared to the epistemic uncertainty reflected by the accuracy. The results are presented in Figure 7. One can note that especially for small prediction set sizes, the conformalized

models always yield higher accuracies, hence predictions from these models have become more trustworthy through conformalization. In conclusion, even though the naive method produces the smallest prediction sets on average, the accuracy of a given prediction set size is almost always higher after conformalization. Therefore, the conformalizer’s uncertainty yields a better indicator for the model’s uncertainty than the uncertainty from an unconformalized model.

V. ACTIVE LEARNING EXPERIMENTS

The overall goal of the proposed AL experiments is to check the quality of different uncertainty quantification methods. If the uncertainties are valid, adding uncertain (i.e., expressive) samples helps the model generalize the unseen test data better. Hence, a fast convergence of the test performance metrics is desired. In particular, the generalization performance should exceed the random baseline which increases training pool with randomly queried data. The uncertainties are used to iteratively query new data (with high predictive uncertainties), which are then added to the initial training data set and labeled by the human oracle. The model is then retrained on the new pool of

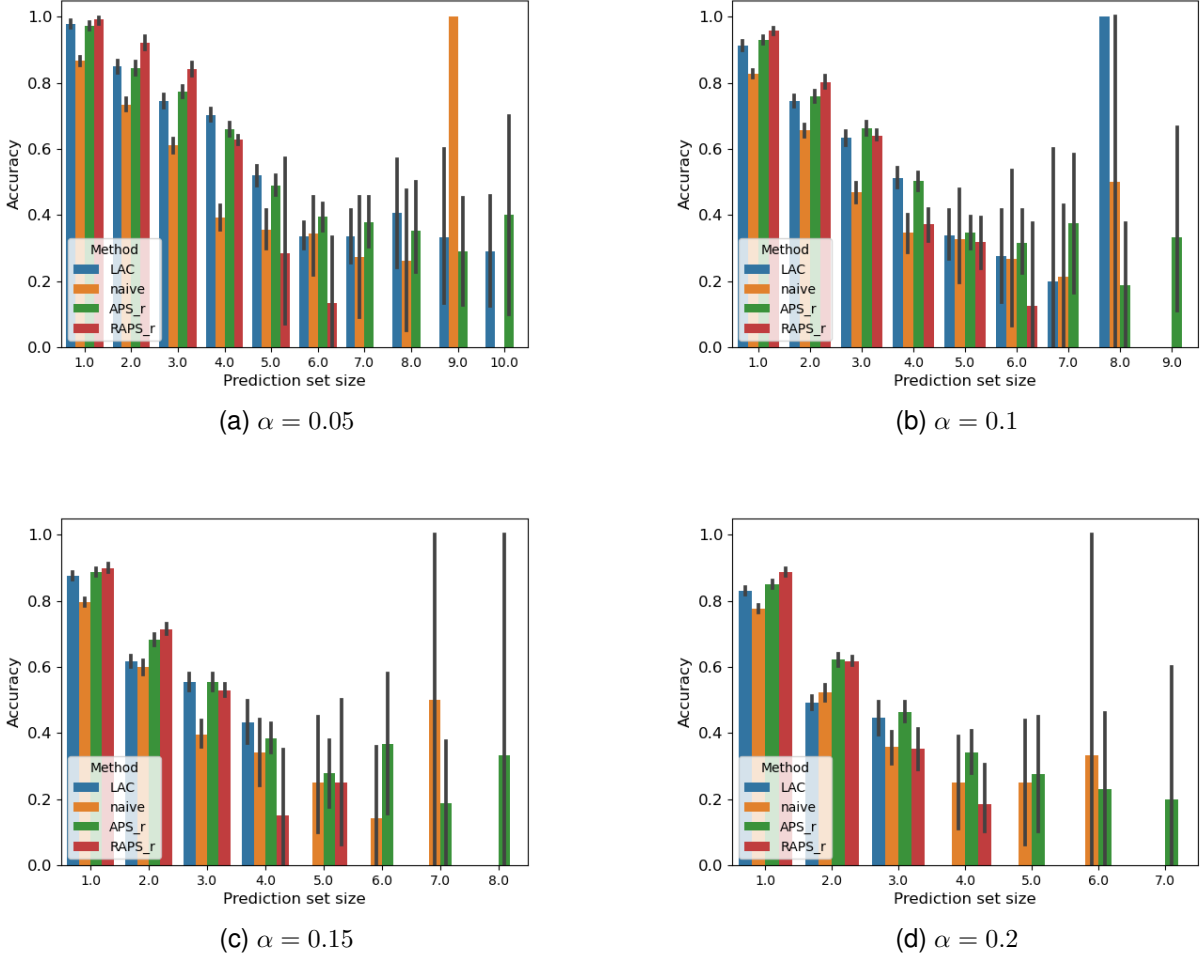


Fig. 7: Accuracy grouped by prediction set size for (conformalized) predictors, separately for different α values.

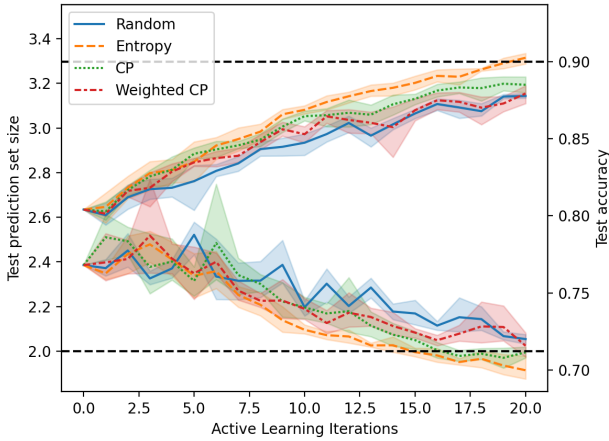
training data before uncertainties are used to query data again. We verify the quality of the uncertainties by measuring the performance of the model on the hold-out test set after every loop.

We first briefly describe the general AL loop. An initial training set is formed via a random subset of the entire available training data. Next to the usual test set, a calibration set is formed, where exchangeability should be guaranteed between the three separate sets. After initializing a model and training it on the initial training set, we use the model to predict the remaining training data. Based on these predictions, and using multiple querying strategies, we select data points from the remaining training pool that get added to the initial training data. After that, the model is retrained on the increased training data pool and again used for inference. Also, at this stage, we evaluate the model’s test performance and store it for later.

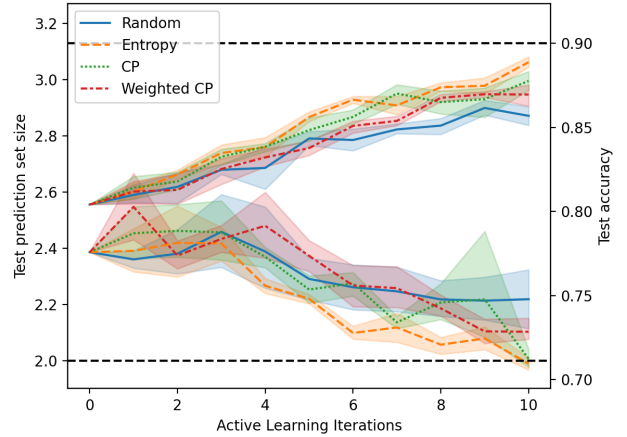
As a novelty, we introduce inefficiency-based AL. While previous works such as [35] incorporated the score function of a conformalizer for querying, we now shift the focus towards the estimated prediction sets. In particular, this inefficiency-based approach does not directly depend on the softmax

pseudo-probabilities anymore, which might reflect over- or underconfidence of the underlying model. Therefore, before each inference step, we conformalize our classifier and utilize the resulting threshold for estimating the prediction sets for the remaining training data. Then, we query the points that get added to the training set with respect to the prediction set size. Since large prediction sets indicate a high degree of uncertainty in the prediction, we add those points to our training set to facilitate the model’s generalization ability on the test set. We test this with a slightly shifted interest: While still evaluating the test accuracy and ensuring the theoretical coverage guarantee of the true label, the focus now additionally lies more on lower and more expressive predictive uncertainty.

This is measured via the average prediction set size on the test set. Suppose inefficiency-based AL leads to a faster convergence of the average test prediction set size (to, e.g., a lower bound). In that case, we can deduce that uncertainty-aware AL using conformal prediction can generate lower predictive uncertainties while guaranteeing a solid level of coverage of the true label. Suppose, furthermore, or instead, the inefficiency-based AL leads to a faster increase in test accuracy.



(a) 500 samples per AL iteration



(b) 1000 samples per AL iteration

Fig. 8: Random Split setting with varying no. of samples per AL iteration.

In that case, we can deduce that adding more uncertain samples to the training pool helps the model to adapt quicker to unseen data.

As described earlier, we now test the quality of the predictive uncertainty by means of an AL downstream task. For this, we randomly split the entire pool of available data into a large training pool (70%) as well as reasonably sized calibration and test data pools (15% each). Out of the large training pool of approximately 35k satellite patches, we randomly select 5k samples, which serve as an initial training pool. In contrast, the remaining samples are denoted as the unlabeled training pool. A randomly initialized model is then initially trained on the 5k training samples (pre-training stage). In each AL iteration, the separately trained models predict on the unlabeled training pool and select a fixed number of data points through their respective querying strategies. As a novelty of this work, we investigate the predictive uncertainty of conformalization methods as a querying strategy. We choose the *RAPS* conformalizer with $\alpha = 0.05$, with the inclusion of the borderline label to ensure sufficiently large prediction set sizes. After each training step (and initially after the pre-training stage), the model is then conformalized by means of the hold-out calibration dataset, and the prediction sets are computed for the unlabeled training pool. These prediction sets are then used for querying in two distinct ways: In the first setting, the points are selected solely based on their prediction set sizes (*CP*). Meanwhile, in a second setting, the prediction set size gets weighted with the inverse class frequencies of the set members (*weighted CP*) to foster the inclusion of underrepresented classes.

Next to a baseline model, which samples the points that get added to the training data randomly, we also investigate softmax entropy-based querying. This technique is widely used in the machine learning community [56] [57] [58] and utilizes the Shannon entropy of the predicted softmax vector for querying (higher is preferred, which supposedly represents higher predictive uncertainty). The promise of AL lies in the fast generalization ability of a model once presented with predominantly uncertain samples. Therefore, we limit ourselves

in the following experiments to adding a total of 10k samples to the initial training dataset, corresponding to roughly one-third of the available unlabeled training pool. In two separate scenarios, 500 and 1000 samples are added to the training data per step, respectively. We present the results in Figure 8. In particular, test accuracy and, more importantly, prediction set sizes for the test set predictions are reported in each iteration for five models per configuration (mean $\pm$ standard deviation), each randomly initiated with a different seed.

Multiple conclusions can be drawn from these results. First and foremost, we see clear advantages of the uncertainty-based querying methods over the random querying, which justifies the methodological framework and computational overhead (in the case of conformalization). At the same time, we see equal or in certain cases even slightly better test performance of uncertainty-based models for a moderate share of data used (ca. 40 %) compared to the full model, which gets trained on all available training data right away. Its performance is indicated by the black dashed line. Furthermore, the overall convergence for the higher number of samples per iteration appears to be smoother. Yet for this setting, a slightly higher variation exists overall across the different models for each configuration, in particular for the *CP* and weighted *CP* configurations.

Regarding the superior performance of the entropy-based querying, we investigated the following: Since the prediction set size is limited to whole positive numbers and bounded from above by the number of classes, it is more tightly bound than the entropy of the softmax. The latter is also bounded from above and below, but can take any value in between the bounds on a continuous scale. We visualized the realisations of the two quantities for the first prediction of the pre-trained model on the remaining unlabeled training data pool in Figure 9. Note the core message of this plot, namely that querying points with respect to maximum entropy will include points solely with the maximum prediction set size, but not vice versa. This is due to the large spread of the entropies given the highest prediction set size.

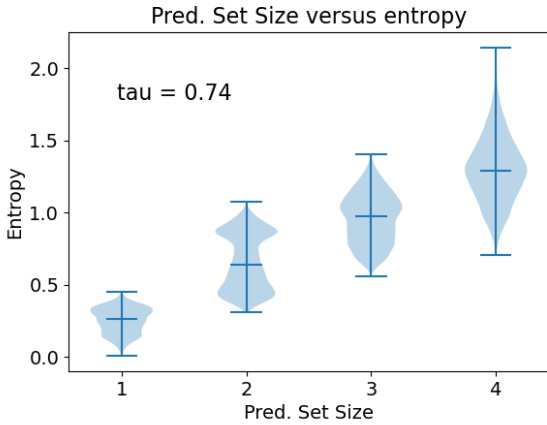


Fig. 9: Softmax Entropies and Prediction Set Sizes for the unlabeled training pool.

Lastly, we try to explain the slightly underperforming weighted *CP* method. Note first that the prediction set gives no internal ordering to the classes it contains. Therefore, solely when a prediction set contains an underrepresented class, the weight will drastically increase, and the sample is more likely to be queried, regardless of whether the underrepresented class is the true class or not. Note, however, that the ground truth label is not known for the unlabeled training pool until the samples are added to the initial training data. Therefore, we cannot utilize this information at the querying stage. Second, the class imbalance is quite large in the dataset, leading to huge weights for underrepresented classes. This can, of course, be targeted and optimized, but lies outside the scope of this work.

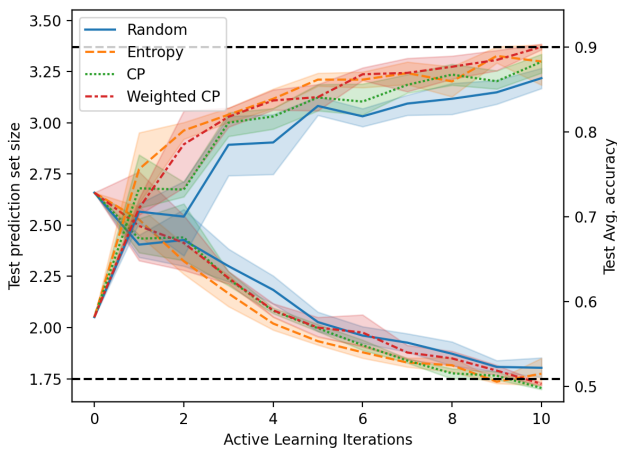


Fig. 10: Random Split setting with 1000 samples per AL iteration.

One advantage of the weighted *CP* approach becomes apparent in Figure 10, where now the balanced accuracy (averaging all ten class-specific accuracies with equal weight) is investigated. The balanced accuracy is particularly useful

for non-uniform class cardinality yet uniform class importance [59], as is the case for the present dataset. We observe similar trends in predictive uncertainty, but a clear advantage of the weighted *CP* approach over the regular *CP* and the random approach is visible for the balanced accuracy. Therefore, the weighting can be seen as a valuable add-on to the *CP*-based querying, which specifically targets balanced accuracy across all classes. Note that the gap to the entropy-based querying is again marginal, underscoring the advantage of the entropy-based method discussed earlier.

VI. CONCLUSION

We have studied the framework of conformal prediction, which constructs prediction sets consisting of multiple classes instead of a single class for a given classifier by means of statistical properties. These prediction sets then fulfill theoretical guarantees about the inclusion of the ground truth label and provide more insights into the model's predictive uncertainty. We have utilized this uncertainty quantification toolbox for the large-scale land cover classification dataset *LCZ42*, for which we have multi-label data coming from ten independent expert annotators. As a baseline for comparison, we define an unconformalized method of constructing prediction sets simply by adding up softmax probabilities (highest to lowest) until the desired inclusion is achieved. This method is then compared to more sophisticated conformalized methods such as *LAC* or *RAPS*. Although the baseline often yields smaller prediction sets overall, the coverage guarantee can not always be fulfilled, and the accuracy with regard to the prediction set size is not as trustworthy as for the conformalized methods. Adding to that, the conformalized methods also manage to cover more of the intrinsic label uncertainty, hence presenting an intuitive interpretation of predictive uncertainty. Both findings support the hypothesis that the conformalizer's predictive uncertainty presents a good indicator for both aleatoric uncertainty, measured by the intrinsic label uncertainty, and epistemic uncertainty, under the assumption that higher model uncertainty should, on average, align with lower predictive uncertainty.

As a further study, we evaluated the *CP* framework by means of AL. By utilizing the prediction set size as a query function, we managed to successfully beat the random baseline and to achieve test performance equivalent to that of a full model, using only a fraction of the data. Note that entropy-based querying could even beat the *CP*-based querying, which underscores the importance of uncertainty in the AL framework. Overall, we conclude that conformal prediction offers a promising toolkit for expressing a classifier's predictive uncertainty in an explainable manner. We could show that a conformalized classifier can reach better coverages as well as more trustworthy statements about the uncertainty associated with its prediction. We see huge potential for conformal prediction in the broader field of modeling remote sensing data, which often requires an explainable and trustworthy notion of predictive uncertainty.

ACKNOWLEDGEMENT

The work is jointly supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001), by the Helmholtz Association under the joint research school Munich School for Data Science - MUDS, by the German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" (grant number: 50EE2201C) and by the Munich Center for Machine Learning.

REFERENCES

- [1] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.
- [2] V. Vovk, "Transductive conformal predictors," in *Artificial Intelligence Applications and Innovations: 9th IFIP WG 12.5 International Conference, AIAI 2013, Paphos, Cyprus, September 30–October 2, 2013, Proceedings 9*. Springer, 2013, pp. 348–360.
- [3] H. Papadopoulos, V. Vovk, and A. Gammerman, "Conformal prediction with neural networks," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, vol. 2. IEEE, 2007, pp. 388–395.
- [4] S. L. Zabell, "Predicting the unpredictable," *Synthese*, vol. 90, pp. 205–232, 1992.
- [5] T. Fritz, T. Gonda, and P. Perrone, "De finetti's theorem in categorical probability," *arXiv preprint arXiv:2105.02639*, 2021.
- [6] D. Angluin, "Queries and concept learning," *Machine learning*, vol. 2, pp. 319–342, 1988.
- [7] J. Gawlikowski *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.
- [8] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [9] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for bayesian uncertainty in deep learning," *Advances in Neural Information Processing Systems*, vol. 32, pp. 13 153–13 164, 2019.
- [10] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *ICML*, 2015.
- [11] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.
- [12] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," *Advances in neural information processing systems*, vol. 33, pp. 7498–7512, 2020.
- [13] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," *arXiv preprint arXiv:2102.11409*, 2021.
- [14] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 7047–7058.
- [15] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 3183–3193.
- [16] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.
- [17] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 223–234, 2019.
- [18] Y. Romano, M. Sesia, and E. Candes, "Classification with valid and adaptive coverage," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3581–3591, 2020.
- [19] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik, "Uncertainty sets for image classifiers using conformal prediction," in *International Conference on Learning Representations*, 2021.
- [20] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023.
- [21] Z. Liang, Y. Zhou, and M. Sesia, "Conformal inference is (almost) free for neural networks trained with early stopping," in *International Conference on Machine Learning*. PMLR, 2023, pp. 20810–20851.
- [22] D. Bethell, S. Gerasimou, and R. Calinescu, "Robust uncertainty quantification using conformalised monte carlo prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38 (19), 2024.
- [23] L. Dabah and T. Tirer, "On calibration and conformal prediction of deep classifiers," *arXiv preprint arXiv:2402.05806*, 2024.
- [24] L. Zhou, Z. Shao, S. Wang, and X. Huang, "Deep learning-based local climate zone classification using sentinel-1 sar and sentinel-2 multispectral imagery," *Geo-Spatial Information Science*, vol. 25, no. 3, pp. 383–398, 2022.
- [25] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *ICML*, vol. 2, no. 4. Citeseer, 2000, p. 6.
- [26] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM computing surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.
- [27] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern recognition letters*, vol. 25, no. 9, pp. 1067–1074, 2004.
- [28] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 4, pp. 1231–1242, 2008.
- [29] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.
- [30] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, 2010.
- [31] J. Lin, T. Yu, and Z. J. Wang, "Rethinking crowdsourcing annotation: partial annotation with salient labels for multilabel aerial image classification," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–12, 2022.
- [32] S.-S. Ho and H. Wechsler, "Query by transduction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 9, pp. 1557–1571, 2008.
- [33] V. Balasubramanian, S. Chakraborty, and S. Panchanathan, "Generalized query by transduction for online active learning," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 1378–1385.
- [34] T. Löfström, U. Johansson, and H. Boström, "Effective utilization of data in inductive conformal prediction using ensembles of neural networks," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2013, pp. 1–8.
- [35] S. Matiz and K. E. Barner, "Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification," *Pattern Recognition*, vol. 90, pp. 172–182, 2019.
- [36] —, "Conformal prediction based active learning by linear regression optimization," *Neurocomputing*, vol. 388, pp. 157–169, 2020.
- [37] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.
- [38] C. Koller, P. Jung, and X. X. Zhu, "Can land cover classification models benefit from distance-aware architectures?" *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.
- [39] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.
- [40] C. Koller, G. Kauer mann, and X. X. Zhu, "Going beyond one-hot encoding in classification: Can human uncertainty improve model performance in earth observation?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.
- [41] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9617–9626.
- [42] K. Hechinger, X. X. Zhu, and G. Kauer mann, "Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 73, no. 1, pp. 143–161, 2024.
- [43] P. Bhattacharjee and P. Jung, "Onboard out-of-calibration detection of deep learning models using conformal prediction," *arXiv preprint arXiv:2405.02634*, 2024.

- [44] G. Singh, G. Moncrieff, Z. Venter, K. Cawse-Nicholson, J. Slingsby, and T. B. Robinson, "Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction," *Scientific Reports*, vol. 14, no. 1, p. 16166, 2024.
- [45] D. Valle, R. Izbicki, and R. V. Leite, "Quantifying uncertainty in land-use land-cover classification using conformal statistics," *Remote Sensing of Environment*, vol. 295, p. 113682, 2023.
- [46] N. Kakhani, S. Alamdar, N. M. Kebonye, M. Amani, and T. Scholten, "Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction," *Remote Sensing*, vol. 16, no. 3, p. 438, 2024.
- [47] N. Zhu, Z. Xi, C. Wu, F. Zhong, R. Qi, H. Chen, S. Xu, and W. Ji, "Inductive conformal prediction enhanced lstm-snn network: Applications to birds and uavs recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.
- [48] M. Farag, J. Kierdorf, and R. Roscher, "Inductive conformal prediction for harvest-readiness classification of cauliflower plants: A comparative study of uncertainty quantification methods," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 651–659.
- [49] F. Meng and T. Song, "Uncertainty forecasting system for tropical cyclone tracks based on conformal prediction," *Expert Systems with Applications*, vol. 249, p. 123743, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417424006092>
- [50] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bulletin of the American Meteorological Society*, vol. 93, no. 12, pp. 1879–1900, 2012.
- [51] X. X. Zhu *et al.*, "So2sat lcz42: a benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 3, pp. 76–89, 2020.
- [52] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multilevel feature fusion-based cnn for local climate zone classification from sentinel-2 images: Benchmark results on the so2sat lcz42 dataset," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2793–2806, 2020.
- [53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [55] V. Vovk, V. Fedorova, I. Nouretdinov, and A. Gammerman, "Criteria of efficiency for conformal prediction," in *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*. Springer, 2016, pp. 23–39.
- [56] P. Sebastiani and H. P. Wynn, "Maximum entropy sampling and optimal bayesian experimental design," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 1, pp. 145–157, 2000.
- [57] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2008, pp. 1–8.
- [58] Z. Qiu, D. J. Miller, and G. Kesidis, "A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 4, pp. 917–933, 2016.
- [59] A. Gupta, N. Tatbul, R. Marcus, S. Zhou, I. Lee, and J. Gottschlich, "Class-weighted evaluation metrics for imbalanced data classification," 2021. [Online]. Available: <https://openreview.net/forum?id=PBfaUXYZzU>

B Related Publications

The following publications have been further published, but are not directly linked to this dissertation. They represent intermediate results or collaborations outside of the main PhD projects.

- **C. Koller**, M. Shahzad and X. X. Zhu. Uncertainty-Guided Representation Learning in Local Climate Zone Classification. *2022 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2022, pp. 183-186. doi: 10.1109/IGARSS46834.2022.9883897
- K. Hechinger, **C. Schweden**, X. X. Zhu and G. Kauermann, Human-in-the-loop: Towards Label Embeddings for Measuring Classification Difficulty. *Statistical Modelling (accepted)*, 2025.
- **C. Koller**, P. Jung and X. X. Zhu. Exploring Distance-Aware Uncertainty Quantification for Remote Sensing Image Classification. *2023 IEEE International Geoscience and Remote Sensing Symposium IGARSS* 2023, pp. 5692-5695. doi: 10.1109/IGARSS52108.2023.10281435
- Y. Wang, Q. Song, D. Wasif, M. Shahzad, **C. Koller**, J. Bamber and X. X. Zhu. How Certain are Uncertainty Estimates? Three Novel Earth Observation Datasets for Benchmarking Uncertainty Quantification in Machine Learning. *arXiv preprint arXiv:2412.06451*, 2024.