Interpretable Spatial Machine Learning for Environmental Modelling

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Kerstin Rau

aus Wangen im Allgäu

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	01.10.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Thomas Scholten
2. Berichterstatter/-in:	Jun. Prof. Dr. Nicole Ludwig

Disclaimer: this thesis uses Federico Marotta's `kaobook` template based on Felix Dangel's doctoral thesis. I am grateful for Corinna Gall, Frank Schneider, Tobias Rentschler, Joanna Silwa, Martin Meßmer and Katharina Eggensperger for their help to improve the text. For readability, grammar, spelling, and writing style, generative AI tools, ChatGPT-4o and DeepL, were used.

Acknowledgments

I would like to sincerely thank Prof. Dr. Thomas Scholten, who saw potential in me during my state examination and gave me the opportunity to do a PhD. Thank you also for your support, great supervision and your openness, for asking unconventional questions, and for always being willing to discuss unconventional ideas. I am especially grateful for the freedom and trust you gave me, whether it was attending conferences and field trips (even when not directly related to my research), engaging in science communication projects, or supporting me in teaching.

A very special thanks goes to my second supervisor Prof. Dr. Philipp Hennig, who was also an exceptional and reliable support throughout this time. It never felt like is was "just" the second supervision, you were always fully involved and present when it mattered.

To both of you: thank you for your understanding and support during and after my illness. Your trust and encouragement, without questions or conditions, made a real difference, and I am truly grateful for that.

Many thanks to Prof. Dr. Nicole Ludwig, who kindly agreed to review my dissertation, and to Dr. Harald Neidhardt for being part of my examination committee.

Much appreciation to the Cluster of Excellence – Machine Learning for Science, which initially funded my position and helped connect the community, and to the Tübingen AI Center, which made my contract extensions possible. I would also like to thank the State Authority for Geology, Mineral Resources and Mining (LGRB) in Baden-Württemberg, and in particular Wolfgang Fleck (thanks again for the German Soil Survey Guidelines KA6 Book), Frank Baumann, and Michael Blaschek, for providing the data used in this thesis and for sharing their expertise in selecting the study areas. A big thank you to Frank Schneider, who always supported me with great motivation on all kinds of questions, especially programming-related ones, and helped make my code both functional and elegant. Special thanks also to Katharina Eggensperger, who took on the optimization of the neural networks used in this work.

My heartfelt thanks go to Franziska Weiler, who supported me not only with administrative matters, often on very short notice, but also took the time to listen to all my challenges and struggles. She was always there to support me with her perspective and sound advice when I needed it most.

I would like to extend my deepest thanks to both of my research groups: my soil science group and the MoML group, for creating such a welcoming and supportive work environment, first online and later in person. Thank you for the lunch breaks and all the great social events. I always felt comfortable and included, and I'm truly grateful for that. A very special thank you goes to Dr. Tobias Rentschler, who was my first office mate and made my start easier by showing me everything. He always understood my struggles and even let me join his fieldwork, thank you for that! Thanks also to Dr. Corinna Gall, with whom I could share all of my thoughts. We've shared many joyful, and thankfully only a few difficult, moments together, and I truly appreciated having you by my side throughout. I'm also very grateful to Nicolas Riveras-Muñoz, for the constant emotional support, the adventurous field trip in Chile and as a Spanish teacher. Many thanks to Dr. Steffen Seitz, who patiently answered my countless organizational questions, even though he wasn't officially my supervisor.

Kudos to my friends from university, sports, and beyond (yes, you know who you are!), who were always there for me during this long journey. I truly appreciate your presence and support more than words can say. Thank you for standing by me through all the stressful and impulsive phases, and for your endless patience.

Thank you!

Kerstin Rau
Tübingen, Juni 30, 2025

Abstract

Machine learning (ML) algorithms, and in particular Artificial Neural Networks (ANNs), have emerged as powerful tools for Digital Soil Mapping (DSM), especially in areas where conventional mapping approaches are limited by high costs, labor intensity, and data accessibility constraints. The ability of ANNs to learn complex, nonlinear relationships from large, multi-source datasets makes them well-suited for the prediction of spatial soil properties across diverse landscapes. While their predictive performance has been widely recognized, ANNs often lack mechanisms for uncertainty quantification (UQ), a critical limitation, particularly when models are applied beyond their original training domain. This thesis addresses the need for interpretable and spatially explicit uncertainty measures in ANN-based soil prediction by implementing the Last-Layer Laplace Approximation (LLLA), a Bayesian Deep Learning (DL) technique that approximates model uncertainty at low computational cost.

The first study focuses on the application of LLLA within a classification task of 41 soil units in a heterogeneous landscape of southern Germany. A simple fully connected multilayer perceptron (MLP) is trained on a carefully selected subset of soil regions, excluding certain areas to simulate extrapolation scenarios. The results demonstrate that LLLA effectively identifies areas where the ANN is overconfident, particularly in zones distant from the training domain or underrepresented in the input data. The uncertainty maps generated provide valuable insight into the spatial structure of model confidence and highlight regions requiring further data collection, thereby supporting more informed sampling strategies.

The second study explores model transferability by applying a trained ANN to a separate but geographically similar region without retraining. This scenario reflects a common challenge in DSM: leveraging legacy datasets to predict in data-sparse environments. Despite moderate accuracy losses, the ANN successfully generalises certain patterns, particularly those associated with alluvial features, suggesting that characteristic soil-landscape relationships are partially transferable. However, the model exhibits a bias towards overrepresented soil units from the training region. The integration of LLLA enables the identification of low-confidence predictions and supports spatially resolved interpretation of model limitations.

Collectively, the two studies illustrate the potential of combining ANNs with Bayesian UQ through the LLLA, as a post-hoc and computationally efficient approach, to enhance the reliability and interpretability of soil predictions. The results underline that uncertainty is not just a by-product, but an essential part of predictive modelling, especially in extrapolation tasks. This thesis contributes a practical and computationally feasible framework for uncertainty-aware soil mapping, which also can be implemented post-hoc in existing ANN workflows. It further supports a shift in DSM practice towards transparent, uncertainty-informed spatial predictions that can guide stakeholders in environmental planning, land management, and resource allocation.

Zusammenfassung

Algorithmen des maschinellen Lernens, insbesondere künstliche neuronale Netze (ANN), haben sich als leistungsstarke Werkzeuge für die digitale Bodenkartierung (DSM) erwiesen, insbesondere in Gebieten, in denen herkömmliche Kartierungsansätze durch hohe Kosten, Arbeitsaufwand und begrenzte Datenverfügbarkeit eingeschränkt sind. Die Fähigkeit von ANNs, komplexe, nichtlineare Beziehungen aus großen und aus mehreren Quellen stammenden Datensätzen zu erlernen, macht sie gut geeignet für die Vorhersage räumlicher Bodeneigenschaften in verschiedenen Landschaften. Während ihre Voraussagekraft weithin anerkannt ist, fehlt es ANNs oft an Mechanismen zur Quantifizierung der Unsicherheit, was eine kritische Einschränkung darstellt, insbesondere wenn die Modelle über ihren ursprünglichen Trainingsbereich hinaus angewendet werden. Diese Arbeit befasst sich mit dem Bedarf an interpretierbaren und räumlich expliziten Unsicherheitsmaßen in der ANN-basierten Bodenvorhersage durch die Implementierung der Last-Layer Laplace Approximation (LLLA), einer Bayes'schen Deep-Learning-Technik, die die Modellunsicherheit bei minimalen Rechenkosten approximiert.

Die erste Studie konzentriert sich auf die Anwendung von LLLA im Rahmen einer Klassifizierungsaufgabe von 41 Bodentypen in einer heterogenen Landschaft in Süddeutschland. Ein einfaches, voll vernetztes mehrschichtiges Perzeptron wird auf einer sorgfältig ausgewählten Teilmenge von Bodeneinheiten trainiert, wobei bestimmte Bodeneinheiten ausgeschlossen werden, um Extrapolationsszenarien zu simulieren. Die Ergebnisse zeigen, dass LLLA effektiv Bereiche identifiziert, in denen das ANN zu selbstsicher ist, insbesondere in Regionen, die weit vom Trainingsgebiet entfernt oder in den Eingabedaten unterrepräsentiert sind. Die erstellten Unsicherheitskarten geben einen wertvollen Einblick in die räumliche Struktur der Modellzuverlässigkeit und heben Regionen hervor, die eine weitere Datenerhebung erfordern, was wiederum fundiertere Probenahmestrategien unterstützt.

In der zweiten Studie wird die Übertragbarkeit des Modells untersucht, indem ein trainiertes ANN ohne erneutes Training auf eine andere, aber geografisch ähnliche Region angewendet wird. Dieses Szenario spiegelt eine häufige Herausforderung im DSM wider: die Nutzung vorhandener Datensätze zur Vorhersage in datenarmen Umgebungen. Trotz leichter Genauigkeitsverluste verallgemeinert das ANN erfolgreich bestimmte Muster, insbesondere solche, die mit Auenböden verbunden sind, was darauf hindeutet, dass charakteristische Boden-Landschafts-Beziehungen teilweise übertragbar sind. Allerdings weist das Modell eine Tendenz zu überrepräsentierten Klassen aus der Trainingsregion auf. Die Integration von LLLA ermöglicht die Identifizierung von Vorhersagen mit geringer Konfidenz und unterstützt die räumlich aufgelöste Interpretation der Modellgrenzen.

Zusammengenommen veranschaulichen die beiden Studien das Potenzial der Kombination von neuronalen Netzen mit der Bayes'schen Unsicherheitsquantifizierung, insbesondere durch die LLLA, als post-hoc und rechnerisch effizienter Ansatz, um die Zuverlässigkeit und Interpretierbarkeit von Bodenvorhersagen zu verbessern. Die Ergebnisse unterstreichen, dass Unsicherheit nicht nur ein Nebenprodukt, sondern ein wesentlicher Bestandteil der Modellierung ist, insbesondere bei Extrapolationsaufgaben. Diese Arbeit liefert einen praktischen und rechnerisch machbaren Rahmen für eine auf Unsicherheit basierende Bodenkartierung, die auch nachträglich in bestehende ANN-Workflows implementiert werden kann. Sie unterstützt eine Verlagerung der DSM-Praxis hin zu transparenten, auf Unsicherheiten basierenden räumlichen Vorhersagen, die den Akteuren bei der Umweltplanung, dem Landmanagement und der Ressourcenzuteilung als Orientierung dienen können.

Table of Contents

Notation

The notation is influenced by Bishop [19] and Goodfellow et al. [61].

Tensor, Vectors, Matrices, Numbers

x	Vector of the input features
x_n	Datum n of the input vector
x_i	i-th component of vector x
y	Output vector
y_n	Datum n of the output vector
y_{nk}	k-th component of n-th output vector
$\mathcal{D}$	Dataset
w	Weight vector/parameter vector
w_i	i-th weight parameter
$w_{ji}^{(l)}$	Weight from neuron i in layer l-1 to neuron j in layer l
b	Bias vector
$b_j^{(l)}$	Bias term for neuron j in layer l
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^N$	N-dimensional real vector space
M	Number of input features
C	Number of outputs/classes
D	Total number of parameters
N	Number of data points/samples

Functions and Operators

$f(\cdot)$	Function
$f_w(\cdot)$	Multilayer Perceptron with weights w
$f_w(x)$	Outputs of the MLP function with weights w
∇_w	Gradient operator with respect to w
∇_w^2	Hessian operator (second derivative) with respect to w
$\arg\min$	Argument of the minimum
$\max(a, b)$	Maximum of a and b
$\exp(\cdot)$	Exponential function
$\ln(\cdot)$	Natural logarithm
$\sum_{i=1}^{n}$	Summation from $i = 1$ to n
$\int$	Integration operator

Probability and Statistics

$p(\cdot)$	Probability density/mass function
$p(A\|B)$	Conditional probability of A given B
$\mathcal{N}(\mu, \sigma^2)$	Normal (Gaussian) distribution with mean μ and variance σ^2
$\mathcal{N}(\mu, \mathbf{\Sigma})$	Multivariate normal distribution with mean vector μ and covariance matrix $\mathbf{\Sigma}$
μ	Mean
σ^2	Variance
$\mathbf{\Sigma}$	Covariance matrix

Artificial Neural Networks

H	Number of neurons in hidden layer
L	Number of layers
$h(\cdot)$	Hidden layer activation function (*e.g.* ReLU)
$\sigma(\cdot)$	Output layer activation function (*e.g.* softmax)
z	Pre-activation value (weighted sum)
a_k	Activation value for neuron k
$\text{ReLU}(z)$	Rectified Linear Unit function
$E(w)$	Cross-Entropy loss
$\ell(x_n, y_n; w)$	Loss for single sample
$r(w)$	Regularization term
w_{MAP}	Maximum a posteriori estimate of weights
$L(D; w)$	Likelihood function
MAP	Maximum a posteriori

Uncertainty quantification

$p(w\|D)$	Posterior distribution of weights given data
$p(w)$	Prior distribution of weights
$p(D\|w)$	Likelihood of data given weights
$p(y = i\|x, D)$	Predictive probability for class i

Soil Forming Factors

S	Soil as a function of forming factors
cl	Climate
o	Organisms (vegetation, fauna, in McBratney et al. [121] also human activity)
r	Topography
p	Parent material
t	Time
$S_{c/a}$	Soil class or attribute at a location
s	Other properties of the soil
c	Climatic properties
a	Age
n	Spatial position

Acronyms & Abbreviations

E.g. or *e.g.*	For example (*exempli gratia*)
Etc. or *etc.*	And so on (*et cetera*)
I.e. or *i.e.*	That is (*id est*)
ANN	Artificial Neural Network
BF	Black Forest
BNN	Bayesian Neural Network
CART	Classification and Regression Trees
CNN	Convolutional Neural Network
DL	Deep Learning
DSM	Digital Soil Mapping
KNN	k-Nearest Neighbors
LGRB	State Authority for Geology, Mineral Resources and Mining, Baden-Württemberg, Germany
LA	Laplace Approximation
LLLA	Last-Layer Laplace Approximation
GBM	Gradient Boosting Machine
GP	Gaussian Process
MC	Monte Carlo
ML	Machine Learning
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
NDVI	Normalized Difference Vegetation Index
OOD	Out-of-distribution
ReLU	Rectified Linear Unit
RF	Random Forest
SJ	Swabian Jura
SVM	Support Vector Machine
UQ	Uncertainty Quantification
WRB	World Reference Base

1. Overview

1.1 Introduction

Artificial Neural Networks (ANNs) have become essential tools in a wide range of scientific disciplines due to their ability to model complex nonlinear relationships in large datasets [192]. One of their promising applications is in Digital Soil Mapping (DSM), which involves using models and environmental data to predict and map soil properties across landscapes. In this context, machine learning (ML) in general has shown considerable potential for accurately predicting soil properties over large regions. Over the past two decades ML methods and specifically ANNs have increasingly supported and automated what was conventional a labor intensive task in soil science [13, 66, 83, 121]. With the growing availability of open source software and computational resources, these techniques and their implementations have become more accessible [47]. Furthermore, the publication of extensive open access datasets, such as those containing climate information, digital elevation models, and remote sensing data, has made it easier to apply these methods in various geospatial contexts [57, 121].

Despite significant progress, a major challenge remains: many models carry intrinsic uncertainty that is not explicitly tracked or quantified. This means that uncertainty arises not only from the data, such as measurement errors, laboratory testing variability, or inaccuracies in satellite observations, but also from the model's own structure and assumptions, which can behave unpredictably outside the training domain. As a result, predictions in extrapolation settings are often unreliable [79, 125, 183]. In particular, ANNs, often regarded as black-box models, tend to produce overly confident predictions when confronted with unknown or underrepresented data, due to a lack of built-in uncertainty estimation and an overreliance on learned patterns from the training distribution [25, 77, 135]. Their outputs can appear reliable even in regions where the model has no empirical support, underscoring the need for methods that explicitly account for uncertainty. Otherwise this overconfidence can be especially problematic in high-stakes applications, where predictions based on ANN-generated maps can lead to costly or damaging decisions [68].

As a result to improve the reliability of ML models in DSM, it is therefore not only important to achieve high predictive accuracy but also to provide well-calibrated uncertainty estimates. Crucially, these estimates should not be limited to a single summary value for the entire map, but should be spatially explicit to reveal where the model is confident and where it is uncertain across the landscape. This gives us spatial uncertainty that describes how uncertainty varies across an area. A promising solution lies in probabilistic modeling approaches, particularly those based on Bayesian inference, which offer a principled framework for uncertainty quantification (UQ). By accounting for uncertainty in a model's parameters, Bayesian methods help capture epistemic uncertainty, the

uncertainty due to limited data or model knowledge. Among recent Bayesian approaches, the Last-Layer Laplace Approximation (LLLA) stands out as a computationally efficient method. LLLA can be applied post-hoc to trained ANNs, meaning it does not require model retraining [104]. This makes it especially useful for improving uncertainty calibration in existing models with low additional cost.

In this thesis, the use of LLLA, to improve the spatial uncertainty calibration of ANNs in the context of DSM is investigated. This approach helps identify locations, where predictions may be unreliable, providing valuable guidance for targeted data collection and contributing to more interpretable and robust soil maps. It also enhances the reliability of transferring ANNs to new environments. By integrating LLLA with established soil modelling practices, this research aims to support the development of more reliable, transparent and trustworthy ANNs, especially in domains where decision making depends on both accuracy and interpretability. Poorly calibrated ANNs can produce misleadingly confident predictions, potentially resulting in harmful or costly decisions. In contrast, well-calibrated ANNs provide a realistic assessment of uncertainty, helping stakeholders distinguish between areas of high and low confidence and make informed choices.

To structure this thesis and guide the development of the research, I define a set of focused research questions (RQ). In response, this thesis offers three key contributions that reflect both theoretical insights and practical advancements in the application of LLLA for ANNs to DSM.

(RQ1) *How effectively can the Last-Layer Laplace Approximation (LLLA) quantify and correct the uncertainty in Artificial Neural Networks (ANNs) predictions for soil mapping?*

(RQ2) *To what extent can Artificial Neural Networks (ANNs) trained on well-sampled regions generalise to soil predictions on geographically similar but distant, data-sparse regions via model transfer?*

(RQ3) *What role does spatial uncertainty play in the reliability and practical application of digital soil maps?*

1.2 Outline

This thesis is structured into three main parts: a conceptual and technical foundation (Part I), the achieved research contributions (Part II), and a conclusion discussing the implications and an outlook (Part III). Together, they are investigating how interpretable ANNs can advance the field of DSM while taking uncertainty into account, with a particular focus on data-sparse regions.

Part I introduces the conceptual foundations. It motivates the integration of ML techniques into soil science, provides a brief overview of the history and current state of the art. A central focus of this part is the challenge of interpretability and reliability in ANNs, which are often still regarded as "black boxes" despite their predictive power. In detail, I examine the following aspects.

Chapter 2 begins with an overview of soil classification and the general principles of DSM. It introduces the scorpan model as a foundational framework for DSM and outlines how it informs various modeling approaches. The chapter highlights the growing use of ML in soil prediction, particularly ANNs, driven by their scalability and strong predictive performance. It presents the most commonly used ML models for DSM, outlining their respective strengths and weaknesses. In addition, it critically examines key limitations of ANNs, especially in terms of interpretability and reliability of their predictions.

Chapter 3 discusses, after a general introduction of ANNs, the importance of UQ in spatial models. It introduces the LLLA as a scalable Bayesian method that can be integrated into ANNs to provide principled uncertainty estimates. The chapter also contrasts LLLA with other UQ techniques, highlighting its advantages in terms of computational efficiency and compatibility with existing ANN architectures.

Part II of this thesis presents its two central contributions, both focused on applying ANNs to predict soil unit maps, with a particular emphasis on quantifying model uncertainty using the LLLA method. Each contribution explores a different use case, demonstrating how uncertainty in ANNs can be explicitly modeled, visualized, and interpreted to gain insight not only into the mapped area, but also into the behavior and reliability of the model itself.

Chapter 4 introduces the LLLA for predictive uncertainty estimation of ANNs applied to soil classification tasks. ANNs lack calibrated uncertainty estimates and tend to produce overconfident predictions, particularly in regions geographically distant from the training domain or sparsely covered by training data. We address this issue using the LLLA, which augments the final layer of a trained ANN with a Gaussian posterior. This yields predictions with uncertainty statements without requiring changes to the training process or reducing model accuracy. Our application focuses on soil classification in southern Germany, where we exclude two regions with their specific soil units during training and include them only in the prediction phase. By incorporating LLLA uncertainty estimates, the model not only delivers predictions but also provides a measure of confidence. The setup with the excluded soil units allows us to evaluate whether the model appropriately expresses low confidence in unfamiliar regions and to explore additional patterns that

may reveal the model's strengths and limitations. This enables stakeholders to assess model reliability, identify knowledge gaps, and prioritize data acquisition in regions of low predictive certainty.

Disclaimer 1.1 Chapter 4 is based on the peer-reviewed journal publication with the following co-author contributions:

K. Rau, K. Eggensperger, F. Schneider, P. Hennig, and T. Scholten. "How can we quantify, explain, and apply the uncertainty of complex soil maps predicted with neural networks?" *Science of The Total Environment* 944 (2024). eprint: https://doi.org/10.1016/j.scitotenv.2024.173720 [151]

	Ideas	Experiments	Analysis	Writing
K. Rau	50 %	65 %	70 %	75 %
K. Eggensperger	5 %	20 %	5 %	10 %
F. Schneider	5 %	15 %	0 %	5 %
P. Hennig	20 %	0 %	5 %	0 %
T. Scholten	20 %	0 %	20 %	10 %

Chapter 5 extends the Bayesian Deep Learning (DL) approach LLLA for the extrapolation of soil unit predictions, addressing challenges arising from uneven soil sampling distributions. The available soil data is often regionally very dense due to previous sampling, and the aim is to use this data to make well-founded predictions for similar but unsampled or insufficiently sampled regions. To explore this, we train an ANN on a region and transferring it to an unseen but similar region without additional retraining in south Germany. To quantify predictive uncertainty, we again apply LLLA to the final layer of the network. By applying LLLA, we obtain spatially explicit uncertainty estimates, highlighting regions where predictions are reliable and where model uncertainty increases. This allows for assessing the suitability of the model for extrapolation to new areas and identifying locations where additional sampling could enhance model performance.

Disclaimer 1.2 Chapter 5 is based on the peer-reviewed journal publication with the following co-author contributions:

K. Rau, K. Eggensperger, F. Schneider, M. Blaschek, P. Hennig, and T. Scholten. "Quantifying spatial uncertainty to improve soil predictions in data-sparse regions". *SOIL* 11.2 (2025). eprint: https://doi.org/10.5194/soil-11-833-2025 [150]

	Ideas	Experiments	Analysis	Writing
K. Rau	60 %	80 %	70 %	80 %
K. Eggensperger	0 %	5 %	5 %	0 %
F. Schneider	0 %	5 %	0 %	0 %
M. Blaschek	10 %	10 %	0 %	10 %
P. Hennig	10 %	0 %	5 %	0 %
T. Scholten	20 %	0 %	20 %	10 %

Part III provides a comprehensive summary of the studies, linking back to the research questions introduced in Section 1.1. It outlines how the proposed methodology addresses key challenges in applying ANNs to DSM tasks through the integration of UQ using LLLA. In addition, it explores the potential implications of this approach for improving model transparency, guiding targeted sampling efforts, and informing decision-making in soil monitoring and mapping. The chapter concludes with an outlook on future directions, including methodological improvements such as benchmarking, expanded applications in diverse landscapes, and the integration of additional data sources to further support robust and scalable soil prediction systems.

Part I.

Background & Motivation

2. Digital Soil Mapping and its Challenges

DSM is a modern scientific field used to create detailed maps of soil classes or soil properties (like texture, acidity, organic matter, *etc.*) across large areas using computers, data, and models. It emerged as a sub-discipline of soil science in response to the limitations of conventional soil surveys and the growing need for timely, high-resolution soil information. Unlike conventional methods, which rely heavily on extensive and costly field surveys, DSM combines soil samples taken from specific locations with environmental data. These data include remote sensing data, such as optical imagery (e.g., multispectral or hyperspectral data) and Synthetic Aperture Radar data, along with climate data (e.g., temperature, precipitation), digital elevation models, terrain derivatives (e.g., slope, aspect, curvature), and land use or land cover information. Using this information, scientists build geostatistical or ML models that learn how soil properties are related to the environment. These models are then applied to predict soil classes or properties in areas where no samples were taken, producing continuous maps of the landscape. This underscores one of DSM's key advantages: the shift from conventional polygon-based maps to high-resolution, pixel-based maps that provide a much more accurate representation of soil variability across the landscape.

2.1 From Soil Formation to Soil Classification

Soils are an essential component of the terrestrial environment, underlying many of the natural and human systems we depend on, such as plant growth, water filtration, carbon storage and cycling, and food production. Most individuals encounter soil early in life, through play, gardening, or outdoor exploration, but as attention shifts toward to large-scale soil applications in agriculture and ecosystem regulation, its presence becomes less visible. Despite this, soil continues to play a vital, though often overlooked, role in sustaining food systems, regulating environmental processes, and supporting biodiversity throughout our lives.

While the practical use of soils, *i.e.* agriculture, has been integral since the first human settlements, the systematic scientific investigation into soil formation processes only emerged in the 19th century. A foundational contribution came from Dokuchaev [46] in 1883, who first conceptualized soil as a natural body formed through the interaction of five key environmental factors: climate, organisms, relief (topography), parent material, and time. These are now widely recognized as the soil-forming factors. This concept was further advanced by Jenny [90], who introduced a quantitative formulation of these interactions, described in Definition 2.1. Jenny's approach established a framework, that still serves as the basis for DSM today. However, even prior to the development of DSM, this conceptual model guided classical soil surveyors and cartographers. When delineating soil units in the field, their mapping decisions were based not only on direct observations but also on an understanding of the surrounding landscape and its formative influences.

Definition 2.1 (CLORPT model by Jenny [90]) Let S denote soil as a function of various forming factors. This approach to understanding soil formation was further researched to a quantitative and analytical formulation:

$$S = f(cl, o, r, p, t, ...) \tag{2.1}$$

where:

- cl refers to climate,
- o refers to organisms (including vegetation, microbes, fauna),
- r refers to relief (topography),
- p refers to parent material (rock or sediment),
- t refers to time.

This model emphasizes that soil is not a static substance but a dynamic product of interacting natural factors. The dots in the ellipsis indicates that additional variables may also influence the function.

The work of Dokuchaev [46] not only represents a foundational advancement in the scientific understanding of soils, he introduced a systematic classification framework based on these soil forming factors, rather than focusing solely on physical or chemical properties like texture or fertility. Another central concept emerging from his work is the differentiation of soils into zonal, intrazonal, and azonal categories, based on the degree to which their development reflects climatic and geographical influences.

These principles laid the groundwork for modern soil classification systems, including the World Reference Base for Soil Resources (WRB) and the German soil classification system, the latter being specifically designed to accommodate national soil conditions and officially applied in environmental and agricultural contexts [7, 48]. These modern systems employ a hierarchical structure in which soil types are distinguished based on moisture conditions, diagnostic horizon sequences and the associated pedogenic processes that occur in the topsoil and subsoil. These types are further subdivided into subtypes, soil varieties, and subvarieties, allowing for detailed differentiation based on diagnostic features.

For the purposes of this study, I focus on soil units, a categorization level that groups soils according to shared morphological, physical, or developmental characteristics specified by the *State Authority for Geology, Mineral Resources and Mining Baden-Württemberg (LGRB)*. With this conceptual and historical foundation in place, we can now turn our attention to DSM, which builds upon these classical principles by integrating them with modern geospatial technologies and data-driven approaches.

2.2 Foundations and Advances in Digital Soil Mapping

DSM has evolved from conventional soil mapping by adopting a quantitative, data-driven approach grounded in pedological theory. The evolution of DSM as a scientific field began in the 1970s, driven by the growing availability of soil and environmental data and the emergence of computer technologies that enabled spatial analysis [187]. As computational speed and data collection capacity improved, DSM methods became more attractive and feasible. By the early 2000s, the demand for high-resolution soil information in land use planning, environmental monitoring, and precision agriculture further increased interest in DSM, particularly in countries like Germany, where authorities began integrating these methods into their workflows [13].

Foundational contributions to the field came from studies such as Cialella et al. [37], Lagarcherie and Holmes [108], McKenzie and Austin [124], and Moore et al. [128], which demonstrated the potential of spatially predicting soil properties using environmental features. The formal operational definition (Definition 2.2) of DSM was introduced by McBratney et al. [121], who generalised the soil-forming factors model by Jenny [90] by incorporating "soil" itself as a factor and explicitly considering spatial location. At the same time, Scull et al. [161] introduced the concept of predictive soil mapping in his review, thereby consolidating the conceptual and methodological foundations of the field.

Definition 2.2 (scorpan model by McBratney et al. [121]) Let $S_{c/a}$ denote a soil class or attribute at a location. Then,

$$S_{c/a} = f(s, c, o, r, p, a, n) \tag{2.2}$$

where:

- s refers to other properties of the soil,
- c refers to climatic properties,
- o refers to organisms, vegetation or fauna or human activity,
- r refers to topography or landscape attributes,
- p refers to parent material or lithology,
- a refers to age, the time factor,
- n refers to space, spatial position.

This conceptual foundation is operationalized through the typical DSM workflow, in that case ML methods, in particular ANNs, as illustrated in Figure 2.1. The input layers represent various environmental features that influence soil formation and are selected to capture the soil-forming factors. For example, climate data may include variables such as temperature and precipitation; topographic attributes like slope, elevation, and aspect are used to capture relief; vegetation indices such as Normalized Difference Vegetation Index (NDVI) represent the influence of organisms; geological and geomorphological indicators serve as proxies for parent material; and remote sensing-derived variables are often used to approximate soil texture and moisture conditions. Together, these variables serve as predictor features for the model.

The central part of the figure depicts an ML model, in that case an ANN, that learns the relationship between these environmental inputs and known soil properties or classes derived from ground-truth observations. The ANN consists of an input layer with one neuron for each feature, a series of hidden layers where complex, nonlinear relationships between features are modeled, and an output layer that generates predictions in the form of soil classes or continuous soil property values. More technical details regarding the ANN structure are provided in Section 3.1.

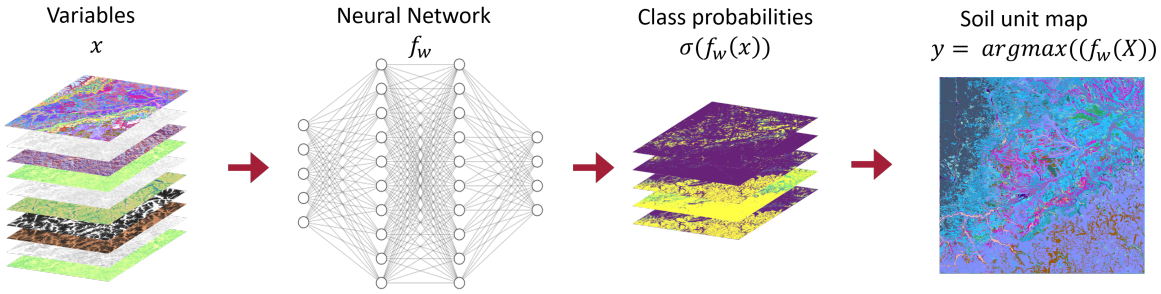

Figure 2.1: Illustration of a typical DSM workflow using ANNs: On the left, multiple input layers represent various environmental features (*e.g.* terrain attributes and calculated indices). These are fed into an ANN, which processes the data through interconnected layers. The model outputs class probabilities distribution for each pixel (middle right), which are then compiled into a final classified map (far right) showing different soil units across the study area.

The prediction output layers shown at third position of the figure represent intermediate results generated by the model during the analysis process. In the case of classification, these layers typically express the model's current estimations as likelihoods or probabilities for each possible soil unit at a given location.

Building on this intermediate predictions produced by the model, the final output, shown on the right side of the figure, is a synthetic soil map created by assigning each pixel to the soil unit with the highest predicted likelihood. In the case of an ANN performing soil classification, this step typically involves applying the $argmax$ function to the output probabilities, which selects the class (*i.e.*, soil unit) with the highest predicted value for each pixel. This figure thus illustrates the DSM workflow, which consists of collecting environmental features, training a ML model, predicting soil units across unsampled areas, and ultimately generating a spatially continuous soil map.

Notably, even before the term DSM was formally established, such methodologies were already being applied to both continuous soil properties and categorical soil classes. Early efforts often began with geostatistical approaches such as kriging [62, 82], but soon expanded to include more diverse methods. For example, the work of McBratney [122], who demonstrated the use of fuzzy k-means clustering for quantitative soil classification, or Cialella et al. [37], who applied classification trees to predict soil drainage classes. Environmental correlation methods also gained recognition through studies such as Moore et al. [128], which linked topographic features with soil properties.

The adoption of decision trees and ensemble algorithms marked a shift toward more robust predictive modeling: for instance, Henderson et al. [78] applied rule-based data mining techniques to conduct nationwide soil prediction in Australia. These advances coincided with the GlobalSoilMap

initiative, through which regional nodes across Africa, the Americas, Asia, Europe, and Oceania compiled soil profile databases and features to produce the first generation of continental-scale soil property grids by the mid-2010s [74]. In parallel, ISRIC – World Soil Information launched SoilGrids, a global mapping platform that integrates ML with remote sensing inputs to estimate soil properties worldwide [144]. As larger and more comprehensive soil datasets became available, the focus in DSM increasingly shifted from conventional geostatistical methods toward ML approaches.

2.3 Machine Learning Methods in Digital Soil Mapping

Building on this shift, ML has become a central component of DSM. While the broader DSM framework includes a variety of techniques, such as conventional statistical methods (*e.g.*, kriging, regression) and rule-based models, this section focuses specifically on the development and application of ML algorithms within the context of DSM.

Since the late 2000s, ML algorithms (*e.g.* Random Forests (RFs), Gradient Boosting Machines (GBMs), Support Vector Machines (SVMs), *etc.*) became common in DSM studies [183]. These methods often outperform conventional approaches by providing higher prediction accuracy and the ability to assess the importance of features. The rapid adoption of ML in DSM in the 2010s was driven by advances in computing power, the emergence of cloud computing, and the increasing availability of large, freely accessible datasets, many of them from remote sensing technologies [140]. Since then, the field has evolved to include probabilistic models and increasingly sophisticated ML algorithms that enable improved predictions of both soil properties and soil classes. These advances have led to improved spatial resolution and mapping accuracy [183].

This development has been documented in numerous review studies evaluating DSM methods and their results [34, 65, 106, 117, 126, 140, 180, 182]. Many of these review papers highlight the growing interest and increasing prevalence of ML-based approaches in soil science. Some focus specifically on the state of the art in certain countries or continents, for example Behrens and Scholten [14] in Germany, Taghizadeh-Mehrjardi et al. [176] in Iran, Dharumarajan et al. [45] in India, and Nenkam et al. [133] in Africa.

ML methods offer several key advantages over conventional statistical models, primarily due to their ability to capture complex, nonlinear relationships between soil properties and environmental features. These patterns are often difficult to detect using conventional techniques [80]. To address this, the following section presents a concise overview of key ML approaches that have been increasingly adopted in DSM.

2.3.1 Modeling Techniques and Algorithm Trends

While not intended as a formal literature review, this section offers a brief insight into the development and growing use of various ML techniques in the context of DSM. For a comprehensive examination, readers are encouraged to consult the following review articles: Chen et al. [34], Grunwald [65], Lagacherie [106], Ma et al. [117], Minasny and McBratney [126], Padarian et al. [140], Wadoux [180], and Wadoux et al. [182].

Over time, numerous ML algorithms have been explored for their suitability in various soil mapping tasks. Among these, Random Forests (RFs), developed by Breiman [25], have emerged as one of the most widely used algorithm in DSM for both continuous and categorical soil predictions. Grimm et al. [63] was one of the pioneers to use RFs in DSM. Its popularity stems from its robustness, capability to model nonlinear relationships, and its internal mechanism for estimating feature importance, which aids in model interpretability [183]. For example, Hengl et al. [79] used RFs within the SoilGrids project to predict a wide range of soil properties at a 250 m spatial resolution globally. More broadly, Padarian et al. [140] observed that RFs consistently performed well across diverse soil prediction tasks. Heung et al. [80] likewise found that RFs offered one of the best trade-offs between prediction accuracy and model interpretability for soil taxonomic unit tasks.

Cubist [148], a rule-based regression model that combines linear models with decision trees, has shown good performance particularly in small to medium datasets. Khaledian and Miller [96] observed that Cubist performed comparably to RFs and better than Multiple Layer Regression or ANN when applied to relatively small DSM datasets under 100 samples. This advantage is attributed to Cubist's ability to generate interpretable rules and local linear models, which enhances model transparency while also maintaining lower computational demands. Similarly, Malone et al. [118] successfully employed Cubist for predicting soil pH across Australian landscapes.

Support Vector Machines (SVMs) have been applied in DSM tasks that benefit from their ability to model complex boundaries and reduce overfitting [40]. Kovačević et al. [102] used SVMs to classify soil types in Mediterranean regions. Heung et al. [80] reported that SVMs with radial basis function kernels yielded high classification accuracy (up to 72%) when predicting soil taxonomic classes.

Though relatively simple, k-Nearest Neighbor (KNN) methods have proven useful in DSM, particularly when modeling categorical outputs [41, 49]. Heung et al. [80] found KNN to perform comparably to SVMs in classification accuracy. However, its performance is sensitive to data density and feature scaling. KNN's main advantages lie in its simplicity and effectiveness in datasets where similar environments yield similar soil types. Mansuy et al. [119] produced high-resolution digital soil maps of Canada's managed forests using the KNN method. The study achieved reliable mapping of key soil properties, improving over conventional coarse-resolution maps and offering valuable tools for forest management and environmental monitoring.

Decision tree methods like Classification and Regression Trees (CART) and its ensemble variants are valued because the tree structure is easy to

visualize and interpret [26]. Han et al. [72] proposes a high-resolution soil moisture estimation framework for China using the CART algorithm, integrating multisource data in a "pyramid" model structure to address spatial data inconsistencies and model generalisability issues. The framework showed stable performance and higher accuracy than conventional methods.

Gradient Boosting Machines (GBM), *e.g.*, XGBoost or LightGBM, have recently gained popularity due to their high predictive accuracy in many fields [52]. Chen et al. [34] reported that gradient boosting models often outperformed RFs in DSM studies conducted under the GlobalSoilMap framework, particularly for predicting soil pH and organic carbon at depths up to 2 meters. Their capacity to handle interactions and rank features also contributes to their appeal in DSM research.

One of the earliest applied statistical models in DSM is Multiple Linear Regression (MLR). While limited in handling nonlinearity and multicollinearity, it remains useful for its transparency and ease of interpretation. Khaledian and Miller [96] compared MLR with several ML methods and noted that MLR generally underperformed in predictive accuracy but offered advantages when interpretability was a priority.

Artifical Neural Networks (ANNs) represents a class of flexible, nonlinear models capable of capturing complicated interactions between soil properties and environmental features. Although ANNs were previously underused due to high computational complexity and interpretability concerns, they have re-emerged as a leading method in DSM research with the advent of high-performance computing, large data sets and better optimisation algorithms. Padarian et al. [140] noted increasing use of DL models in spectral soil analysis and soil property prediction. In addition, Heung et al. [80] evaluated ANN for soil classification and found performance on par with advanced tree-based models, particularly when ample training data were available. Earlier, Behrens et al. [13] demonstrated the potential of ANNs for regional soil unit prediction in Western Germany, underscoring their applicability even in earlier DSM efforts.

Given the increasingly successful applications of ANNs in DSM and their suitability for modeling nonlinear, high-dimensional relationships, ANNs are expected to have a growing advantage, particularly in the future, where data availability will no longer be a major constraint, as more and more data is generated and become publicly accessible [180]. Wadoux et al. [183] already found that advanced ML methods often outperform simpler approaches due to their ability to capture complex, nonlinear patterns.

However, it is important to acknowledge that certain precautions must be taken to avoid producing misleading results. Compared to the other models, ANNs, as one of the advanced ML methods, are typically considered "black box" models [183]. This means that their internal workings are not easily interpretable, making it difficult to understand how specific inputs influence the outputs. Additionally, ANNs typically do not provide inherent mechanisms for UQ. Together, the lack of interpretability and the absence of uncertainty tracking represent significant disadvantages of using ANNs. In the following section, I address this issue and, in Part II discuss potential solutions to improve the UQ of ANNs.

2.3.2 Uncertainty Quantification

With the increasing use of DSM, evaluating model performance and reliability has become a key concern. While many studies have focused on improving the predictive accuracy of ML models (*e.g.* [29, 95, 171]), the quantification and communication of uncertainties remain underdeveloped. The validation of model results and the estimation of explicit uncertainty are increasingly recognised as essential tasks in DSM, as emphasised by Belkadi and Drias [16]. Wadoux et al. [182] identified UQ as one of the key challenges in pedometrics, especially in contexts with complex soil processes, sparse observations, and high spatial variability.

However, despite its importance, UQ is still relatively rare in DSM. According to Piikki et al. [143], only 35 % of DSM studies report uncertainty estimates, with Wadoux et al. [183] and Padarian et al. [140] reporting even lower figures of 30% and 24%, respectively. This lack of coverage of uncertainty is particularly pronounced in ANNs: several studies that use ANNs, such as Behrens et al. [13], Heung et al. [80] and Aitkenhead and Coull [5], ignore uncertainty metrics completely. This is despite the fact that DL models, especially Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) , have outperformed conventional methods in several studies. Hybrid approaches that combine conventional models with DL techniques show promise for improving both performance and interpretability [16].

To fill this gap, a number of UQ methods have been proposed. A widely used technique is bootstrapping [167], in which multiple models are trained on randomly resampled subsets of the training data and the variability of their predictions is used to construct empirical confidence or prediction intervals. This method is commonly used in combination with various ML models, including CNNs [139], RFs, SVMs and Cubist [60], as well as GBMs [35, 71]. Another established approach is quantile regression [98], which estimates conditional quantiles of the response variables, thus capturing the entire prediction distribution. This technique has been successfully applied in DSM by Vaysse and Lagacherie [178] and Poggio et al. [145].

Several methods have been developed specifically for UQ in ANNs. Bayesian neural networks (BNNs) model a posteriori distribution over weights, enabling probabilistic predictions [112]. Ensemble approaches, as demonstrated by Guevara et al. [67], aggregate the results of multiple independently trained networks to estimate uncertainty. Monte Carlo (MC) dropout, introduced as a Bayesian approximation, retains dropout layers at inference time to generate a distribution of predictions [87]. These methods are explained in more detail in Section 3.2, with a focus on their mechanisms, advantages and limitations.

Despite these advances, the application of UQ methods in ANN-based DSM remains limited. ANNs are inherently complex and heavily parameterised and lack natural mechanisms for uncertainty estimation. As a result, most ANN-based DSM studies report only deterministic results. Methods such as BNNs and MC dropout offer promising solutions, but are often computationally intensive and difficult to implement. A notable exception is Wadoux [181], which used bootstrapping in combination with variance estimation as a practical UQ approach for ANNs.

In summary, while interest in UQ within DSM is growing, especially for interpretable models such as RF and linear regression, its integration into DL frameworks remains limited. To support more transparent and reliable ground predictions, there is an urgent need for broader application of UQ methods for ANNs.

Deep Learning Models and their Uncertainty Measurements

3.

As shown in the previous chapter, ANNs have become essential tools in DSM, with their conceptual basis rooted in the early mathematical formalization of neural activity [123]. The growing interdisciplinary relevance of ANNs is reflected in the fact that the 2024 Nobel Prize in Physics was awarded to researchers who advanced ML through the integration of physics-based frameworks, underscoring the increasing importance of these models in contemporary data-driven science. Inspired by the functioning of biological neurons [153, 188], ANNs are capable of learning complex patterns in large and high-dimensional datasets, often beyond human interpretability. While early developments were rooted in biologically motivated structures, modern ANNs now encompass a broad class of models with little or no connection to biological systems.

Initially applied in domains such as image recognition and classification [61], their flexibility has enabled applications across various disciplines, including spatial sciences, particularly when working with large datasets and without requiring extensive prior knowledge of the data. As already mentioned in Section 2.3, ANNs are also increasingly used in DSM, where they support the spatial modelling of soil properties. In the following sections, I describe the architecture of ANNs, discuss methods to assess their predictive uncertainty, and introduce the LLLA, a Bayesian approach that we apply to produce spatially explicit and calibrated uncertainty estimates for our case studies.

3.1 General Structure of Artificial Neural Networks

ANNs are composed of multiple layers of interconnected units called neurons, typically organized into an input layer, one or more hidden layers, and an output layer. Each neuron in the input layer corresponds to a feature, in our case environmental variables, in the dataset, while the output layer contains neurons representing the target outputs, in our case, soil units. The hidden layers in between capture nonlinear relationships in the data by passing weighted inputs through activation functions. This layered architecture enables the network to learn and model complex, nonlinear relationships between input features and target outputs [19, 75]. Various extensions to the basic ANN architecture have been developed to address specific data types and challenges: Convolutional Neural Networks (CNNs) for spatial or image data [165], Dropout layers to prevent overfitting by randomly deactivating neurons during training [166], and Batch Normalization to stabilize and speed up learning [88]. For sequential data, RNNs and Long Short-Term Memory Network are commonly used [84].

In this work, however, I focus on the Multilayer Perceptron (MLP), a fully connected feed-forward architecture where each neuron in one layer connects to every neuron in the next layer. The feed-forward nature implies that data flows unidirectionally from input to output, without cycles. MLPs are particularly effective for tabular or low-dimensional spatial data common in DSM workflows [61].

Given a dataset $\mathcal{D} := \{(x_n \in \mathbb{R}^M,\ y_n \in \mathbb{R}^C)\}_{n=1}^N$ with weights $w \in \mathbb{R}^D$, a standard MLP $f_w : \mathbb{R}^M \to \mathbb{R}^C$ with $f_w(x) \in \mathbb{R}^C$, with one hidden layer can be mathematically formulated as:

$$f_w(x)_k = \sigma\left(\sum_{j=1}^{H} w_{kj}^{(2)} \cdot h\left(\sum_{i=1}^{M} w_{ji}^{(1)} x_i + b_j^{(1)}\right) + b_k^{(2)}\right), \tag{3.1}$$

where x is the input vector representing the features, M is the dimension of x, *i.e.* the number of features, $h(\cdot)$ is a nonlinear activation function, k indexes the output layer and $\sigma(\cdot)$ is the activation function for the output layer. The parameters w are the weights and b the biases across both layers, while H is the number of neurons in the hidden layer. In conclusion, a MLP is essentially a nonlinear function that maps inputs x_n to the outputs y_n. This function is built by stacking layers of linear transformations (matrix multiplications) followed by nonlinear activation functions.

One of the most widely used activation functions for in DL is the Rectified Linear Unit (ReLU), introduced by Householder [86]. It is defined as:

$$\mathrm{ReLU}(z) = \max(0, z), \tag{3.2}$$

where z is the output of a neuron before applying the ReLU activation function. The ReLU activation function, which will be used as the activation function h in this thesis, returns 0 for any negative input and passes positive values through unchanged. Despite its simplicity, it avoids the vanishing gradient problem, allowing better learning in deep models while also being computationally efficient, even in large networks [17, 105]. For the sake of completeness, it should be noted that other activation functions exist, such as sigmoid, tanh, Leaky ReLU, and Exponential Linear Unit function [169]. However, since these are not used in our work, they are not discussed in detail.

In the case of multiclass classification tasks, which is the focus of this thesis, the activation function of the last layer σ used is the softmax function. It is applied to the output layer of the ANN to transform the raw output scores (logits) into a probability distribution. Each output value lies between 0 and 1, and the sum of all outputs equals 1, making the results interpretable as class probabilities. The softmax function is defined as follows [27]:

$$\sigma(a_k) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}, \tag{3.3}$$

where a_k is the vector of raw values for all classes.

Now that the model architecture and activation function have been defined, the focus is on the training process used to optimize the MLP for the target task. The training data consists of input vectors x_n along with their corresponding labels y_n, forming a supervised learning setup. This means the goal is to train a model to learn a mapping from input data to known target outputs using labeled examples. During training, each input is passed through the network in a forward pass to produce a prediction $f_w(x_n)$. This prediction is then compared to the true label using a loss function, that we want to minimise. In supervised multiclass classification, the cross-entropy loss is typically used:

$$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{C} y_{nk} \ln f_w(x_n)_k, \tag{3.4}$$

where N is the number of training examples, C is the number of classes and y_{nk} represents the one-hot encoded class label. The objective is to minimize this loss by adjusting the trainable parameters of the network, *i.e.* the weights w and biases b. These parameters define how input information is transformed as it propagates through the network layers. The network is trained using backpropagation, a method that calculates how the loss changes with respect to each weight and bias in the network. This information tells the model which direction and how much to adjust each parameter to reduce the loss. These adjustments are then then used to update the parameters using an optimization algorithm such as variants of Adam [97, 115]. This process is repeated over several epochs, *i.e.*, it runs through the entire training data set, often using mini-batches of training data, *i.e.*, subsets that are processed together by the ANN. Through this iterative training process, the MLP learns to approximate the underlying function that maps inputs to outputs.

Hyperparameters (*e.g.* learning rate, batch size, number of layers) are not learned during training. Instead, they must be manually set or tuned. The standard approach involves splitting the dataset into three distinct subsets:

- a **training set** for optimizing the model's parameters (weights and biases),
- a **validation set** for model selection and hyperparameter tuning,
- and a independent **test set** for final performance evaluation.

The validation set is used to assess the model's generalisation performance during training and to guide adjustments to hyperparameters. This process aims to strike a balance between underfitting and overfitting while minimizing computational cost. Underfitting happens when the model is too simple to capture patterns in the data, while overfitting occurs when the model learns noise in the training data instead of learning generalizable patterns. Several techniques are commonly used for this task. For example, Grid Search tests all possible combinations of hyperparameters, while Random Search samples them randomly for quicker results [18, 159]. More advanced methods, such as Bayesian Optimization, model the search space and use past results to predict

which hyperparameter combinations are most likely to perform well [56].

To further prevent overfitting in ANNs, regularization techniques are essential. Two common methods include weight decay and early stopping. Weight decay penalizes large weights in the model, encouraging simpler, more generalisable representations. Early stopping, on the other hand, monitors performance on a validation dataset and halts the training process when the validation error stops decreasing. This ensures that the model does not continue learning noise from the training data, which helps maintain better generalisation to unseen data [19, 69]. Once the optimal configuration is found, the final model is evaluated on the test set to estimate its performance on unseen data.

3.2 Uncertainty Quantification of Artificial Neural Networks

While most popular performance metrics offer a snapshot of model accuracy, they do not fully convey predictive uncertainty, *i.e.*, how confident the model is in its predictions. But this is crucial for decision-making, risk management, model interpretation, and evaluating the trustworthiness of the predictions. However, a major limitation of standard ANNs is that they do not inherently provide this uncertainty. Once trained, they behave deterministically: given the same input, they always produce the same output, a single point prediction. Additionally, unlike other ML models like RFs, ANNs are often seen as "black boxes" because they offer little transparency about how predictions are made. They don't give much insight into what's happening inside the model. This lack of interpretability, combined with the absence of uncertainty estimates, makes it difficult to understand or trust the model's decisions.

Uncertainty can generally be categorized into two types, aleatoric and epistemic uncertainty. Aleatoric uncertainty occurs due to inherent noise in the data, it is irreducible and reflects randomness that cannot be eliminated. In contrast, epistemic uncertainty is based on the model's lack of knowledge, often due to limited training data or out-of-distribution (OOD) inputs [94]. When we address uncertainty in ML, our primary goal is typically to reduce epistemic uncertainty. However, aleatoric uncertainty will always be present to some extent, as it is intrinsic to the data itself.

An important point to highlight is the common misinterpretation of softmax outputs, as defined in Equation (3.3), as indicators of model uncertainty. Although the softmax function generates values that look similar to probability distributions, these values do not provide calibrated uncertainty estimates. In fact, softmax will always assign a nonzero confidence to every class, even for completely irrelevant or OOD inputs such as static noise. This apparent confidence simply reflects the relative scores among the classes, not a meaningful assessment of how certain the model truly is about the predictions as it does not take into account the uncertainty in the parameters or structure of the model. As a result, the model cannot express how unsure it is when it sees something unfamiliar

or ambiguous. It doesn't take into account whether it has seen similar data before, or whether the input is very different from its training data.

However, having a reliable measure of uncertainty is essential, especially in situations where decisions depend on the model's output, or when applying the model to new tasks, *i.e.* transfer tasks. Knowing the model's confidence is therefore important, as it can be just as critical as the prediction itself.

3.2.1 Methods for uncertainty quantification

To provide context, I begin with a brief overview of commonly used methods for UQ in ANNs, highlighting their respective advantages and limitations. After this, I introduce our chosen method LLLA, which we evaluated on soil unit prediction with ANN.

Overview

Bayesian Neural Networks (BNNs) differ from standard ANNs by treating parameters like weights and biases as probability distributions instead of fixed values. Rather than learning a single best set of weights, BNNs learn a range of possible weights, capturing how uncertain the model is about them. Before training, BNNs start with prior assumptions about the weights. After seeing data, these priors are updated to posterior distributions, which better reflect the data. For predictions, BNNs average over these distributions, allowing them to express epistemic uncertainty. This makes BNNs more robust in uncertain or data-scarce situations. However, they are computationally expensive, difficult to scale, because of the complexity involved in approximating the posterior distributions, and challenging to train, especially for large models. Their performance also depends heavily on the choice of priors; poorly chosen priors can lead to overconfident predictions on OOD data [33, 131].

In contrast, Monte Carlo (MC) methods offer practical alternatives for uncertainty estimation, with two common approaches in soil science being Markov Chain Monte Carlo (MCMC) and MC Dropout [114, 120, 141, 190]. MCMC performs full Bayesian inference by sampling from the posterior over neural network weights, providing accurate uncertainty estimates [36, 58]. However, it is computationally intensive and scales poorly with model size [11, 129]. In contrast, MC Dropout approximates Bayesian inference by applying dropout during testing. This means the model makes multiple predictions with different dropout masks, which allows it to estimate uncertainty by sampling from its predictive distribution [54]. It is easy to implement and scales well, though its uncertainty estimates depend on dropout settings and may be unreliable for complex or OOD data [163].

Deep Ensembles offer a simple yet powerful approach to UQ by training multiple independent ANNs and combining their predictions. This captures epistemic uncertainty through the variability across models. Unlike BNNs or MC methods, Deep Ensembles do not rely on explicit probabilistic modeling or posterior sampling. This makes them easier to implement and scale, but it requires significantly more computation and

memory, which can be limiting in resource-constrained environments [55].

Gaussian Processes (GPs) are non-parametric models that offer a principled Bayesian approach to UQ by defining a distribution over functions. While not DL models themselves, GPs can be integrated with ANNs in hybrid frameworks [42, 110]. They naturally provide uncertainty estimates, capturing both epistemic and, with appropriate noise modeling, aleatoric uncertainty. GPs are highly interpretable and well-suited to low-data regimes, offering exact inference for small datasets [149]. However, they scale poorly with data size and input dimensionality, limiting their applicability in large-scale or high-dimensional problems. Although numerous alternative approaches and hybrid models for UQ exist, this chapter has focused on the most widely used and established methods.

Last-Layer Laplace Approximation

One such technique is the Last-Layer Laplace Approximation (*e.g.* Kristiadi et al. [104]). It is a simple and efficient method for estimating uncertainty in ANNs. Unlike full Bayesian approaches that model uncertainty across all layers, LLLA applies a Gaussian approximation only to the last layer. This reduces computational complexity while still enhancing the model's confidence calibration, particularly in OOD scenarios.

As discussed in the previous subsection, BNNs and Deep Ensembles are known for providing good uncertainty estimates, but they are often expensive to run and require a lot of computational resources. In contrast, LLLA is much more efficient. It avoids the high cost by approximating uncertainty only in the last layer of an ANN. Simpler methods like MC Dropout are fast and easy to use, but they often fail to give reliable uncertainty estimates, especially when the model sees OOD data. GPs can model uncertainty well in theory, but they do not scale to large datasets. With all that, LLLA offers a good balance in all of this. It is fast, easy to apply to pre-trained models without retraining, and gives more accurate uncertainty estimates than lightweight methods [109, 193].

Let's take a closer look at the mathematical details of the LLLA based on Kristiadi et al. [104] and Daxberger et al. [44]: As mentioned in Section 3.1, given an ANN f_w, the standard approach of training an ANN is to minimize a regularized empirical loss

$$w_{\text{MAP}} = \arg\min_{w \in \mathbb{R}^D} \underbrace{\left[r(w) + \ell(x_n, y_n; w) \right]}_{=: \mathcal{L}(\mathcal{D}; w)} . \tag{3.5}$$

As the loss function $\ell(x_n, y_n; w)$ we use the cross-entropy loss $E(w)$ from Equation (3.4) and $r(w)$ is a regularization term, such as weight decay. From a Bayesian perspective, this objective corresponds to computing the maximum a posteriori (MAP) estimate of the weights. That is, we are finding the most probable set of weights w given the observed data and a prior distribution over the weights. To go beyond the point estimate w_{MAP} and capture uncertainty in the model parameters, we approximate the

posterior distribution over the weights using the Laplace approximation (LA) [1].

LA fits a Gaussian distribution centered at the MAP estimate with covariance given by the inverse of the Hessian of the loss

$$p(w \mid \mathcal{D}) \approx \mathcal{N}(w; w_{\text{MAP}}, \Sigma), \quad \text{with} \quad \Sigma = \left(\nabla_w^2 \mathcal{L}(\mathcal{D}; w) \big|_{w=w_{\text{MAP}}} \right)^{-1}. \tag{3.6}$$

The shape of this distribution is determined by the curvature of the loss function at the MAP, specifically through the inverse of the Hessian matrix. This inverse curvature describes how much the loss function changes in different directions in parameter space. Intuitively, if the loss landscape is steep and narrow in a particular direction, the corresponding Hessian eigenvalue is large, resulting in low variance along that direction in the posterior. This implies the data strongly constrain the parameter in that dimension, indicating high confidence or low uncertainty. Conversely, if the landscape is flat, the curvature is small, the variance in the posterior is high, and the model expresses greater uncertainty in that direction. This local geometry allows LA to capture epistemic uncertainty. However, computing the inverted full Hessian $\nabla_w^2 \mathcal{L}$ is computationally infeasible for ANNs. Both time and memory requirements scale cubically with the number of parameters, *i.e.*, on the order of $\mathcal{O}(D^3)$, where D is the total number of learnable parameters in the ANN [136].

To address this, LLLA restricts the Laplace approximation to the final linear layer only, treating it probabilistically while keeping the rest of the network fixed at w_{MAP}. This results in a predictive distribution of the form:

$$p(y = i \mid x, \mathcal{D}) = \int \sigma(f_w(x))_i \, p(w \mid \mathcal{D}) \, dw, \tag{3.7}$$

where $\sigma(\cdot)_i$ denotes the softmax output for class i (Equation (3.3)). The predictive distribution captures the model's output uncertainty by integrating over the posterior distribution of the last-layer weights. After training the ANN, the weights are fixed at their MAP estimate, except for the final linear layer, for which a Gaussian distribution is fitted using the LA. The predictive distribution for a class i is then obtained by averaging the softmax output $\sigma(f_w(x))_i$ across all possible values of the final-layer weights, weighted by their posterior probability. This process incorporates model uncertainty directly into the prediction by accounting for how sensitive the loss is to changes in those weights.

One key advantage of the LLLA is the clear separation between the deterministic training phase, where the model is trained using standard optimization to obtain the MAP estimate w_{MAP} and the probabilistic modeling of uncertainty, which is applied post-hoc. This means that LLLA can be applied after training without requiring changes to the training procedure, simplifying both development and hyperparameter tuning. While the LA itself involves additional computational overhead due to the need to estimate the Hessian, prior work such as Daxberger et al. [44]

has demonstrated that approximating the Hessian using only the last layer still provides well-structured and informative uncertainty estimates, which leads us to the LLLA. Moreover, this restriction reduces both time and memory complexity from $\mathcal{O}(D^3)$ to $\mathcal{O}(D)$ or less. While there are alternative curvature approximations available like diagonal factorisation or Kronecker-factored approximate curvature (*e.g.*, Daxberger et al. [44]), this work specifically focuses on the LLLA due to its balance of simplicity, efficiency, and effectiveness.

Part II.

Spatial Uncertainty and Transferability in Digital Soil Mapping

How can we quantify, explain, and apply the uncertainty of complex soil maps predicted with neural networks? 4.

Abstract

Artificial Neural Networks (ANNs) have proven to be a useful tool for complex questions that involve large amounts of data. Our use case of predicting soil maps with ANNs is in high demand by government agencies, construction companies, or farmers, given cost and time intensive field work. However, there are two main challenges when applying ANNs. In their most common form, Deep Learning (DL) algorithms do not provide interpretable predictive uncertainty. This means that properties of an ANN such as the certainty and plausibility of the predicted variables, rely on the interpretation by experts rather than being quantified by evaluation metrics validating the ANNs. Further, these algorithms have shown a high confidence in their predictions in areas geographically distant from the training area or areas sparsely covered by training data. To tackle these challenges, we use the Bayesian DL approach Last-Layer Laplace approximation (LLLA), which is specifically designed to quantify uncertainty into deep networks, in our explorative study on soil classification. It corrects the overconfident areas without reducing the accuracy of the predictions, giving us a more realistic uncertainty expression of the model's prediction. In our study area in southern Germany, we subdivide the soils into soil regions and as a test case we explicitly exclude two soil regions in the training area but include these regions in the prediction. Our results emphasise the need for uncertainty measurement to obtain more reliable and interpretable results of ANNs, especially for regions far away from the training area. Moreover, the knowledge gained from this research addresses the problem of overconfidence of ANNs and provides valuable information on the predictability of soil units and the identification of knowledge gaps. By analysing regions where the model has limited data support and, consequently, high uncertainty, stakeholders can recognize the areas that require more data collection efforts.

4.1 Introduction

The use of ML in science has become incredibly valuable and has significantly transformed many areas of research. The number of studies in which methods from the field of ML are used is constantly increasing [192]. Soil science is one of the pioneers here, where extensive applications in the field of soil mapping were already developed at the beginning of this century [13, 121]. Today, DSM is one of the largest areas in which the methods are widely used for all kinds of climatic and geomorphometric regions of the World and in different areas of soil science, which has been demonstrated by numerous papers [126, 152, 161, 173, 192]. Methodologically, applications of ML in soil science range from linear regression to modelling soil properties and their relationships to complex DL methods [128, 179]. The increasing use of these methods is not only due to their

suitability for soil scientific and geographical questions, but also because producing soil unit maps in the conventional way with cartographers surveying the landscape is very costly and time-consuming. This effort can be reduced with ML, especially for larger or even difficult to access areas [13, 66, 83]. At the same time, ML methods and their source code are becoming more accessible due to open-source software and widely available computational resources [47], and with the publication of several large open source datasets containing digital elevation models, climate data and other remote sensing data, especially those describing the vegetation, it is getting more convenient to apply them [57, 121].
Looking at the properties and functions of soils, for example carbon and water storage and plant nutrition, the soil unit as a highly integrated prediction variable has the advantage that we can infer mechanical properties, dynamic processes and general characteristics from it with little effort [6, 73]. For example, Zhou et al. [194] showed new spatial patterns in the predicted soil unit map with their Bayesian predictive modelling approach. Grinand et al. [64] uses classification tree analysis, which also supports decision-making in soil map extrapolation using ML methods. Adhikari et al. [4] compared an existing soil map from a field survey with a predicted map calculated using a decision tree model. ANNs are currently one of the most popular ML methods [170, 172], as they are able to process large amounts of data and compute predictions comparably fast [75, 156, 164]. Brungard et al. [29] predicted soil taxonomy classes using eleven different models and found that the complex models containing ANNs were more accurate. Furthermore, Zhu [195] found that ANNs can be used to obtain high-resolution soil maps. Although Heung et al. [80] has also achieved good results with ANNs, they also have to admit that ANNs are difficult to interpret. Despite the results being rich in information, a major drawback of the predicted soil maps, and especially of the survey maps, is that they do not quantify the uncertainty of the individual soil units at a given geographical location [79]. Instead, mostly is only given an overall accuracy statement in the form of a single statistical number. This is usually calculated as a coefficient using cross-validation techniques, where a subset of the training dataset is used to quantify the uncertainty of the overall performance [183]. However, this is not sufficient, especially for regional or global tasks using unbalanced data sets, and that further analysis on uncertainty statements is needed, which was highlighted by Meyer and Pebesma [125]. Also, studies considering the uncertainty of predicted classes, like soil or vegetation classes, only looking at the probability of the predicted class or its confidence interval, have been criticized as well [183]. They reported that out of 175 papers, only 30 % included UQ, most were focused on achieving high prediction accuracy and only a handful used ML methods for the UQ. It is obvious that a better understanding and quantification of the uncertainty of soil maps modelled with ML is needed, especially when extrapolating from the training domain or when transferring the model to other more or less similar domains. In particular, working with ANNs as a black box requires such an assessment, as this model class is also known to be overconfident [25, 77, 135]. This means that ANNs can predict very reliable results, in our case soil classes, with a probability of up to 100 %, even if the input data is incorrect or uncertain. The lack of uncertainty measurement by the ANNs themselves makes it difficult to assess the reliability of the model predictions, which can lead to misinterpretations

and incorrect decisions [68]. With this study, we apply an ANN that predicts soil units inside and outside the known training domain in a trial study. We quantify the uncertainty of our model at every pixel in the area using LLLA [104]. Our aim is to add this uncertainty measurement to a soil classification problem to identify and correct the overconfidence of ANNs and to be able to spatially analyse and interpret in a following step the prediction of the ANN and its uncertainty derived from the LLLA. Further, we will discuss the transferability of the ANN to adjacent similar areas. Overall, our analyses will help to better understand and interpret results from ML models in soil science to provide new insights into soil processes and the spatial structure of the different domains.

4.2 Material & Methods

4.2.1 Study area

Our study area is located in the centre of Baden-Württemberg in Germany and covers an area of about 35 km^2. The average altitude in this region is 504 m, but if we look at the Swabian Jura (SJ) and the rest of the region separately, we have an average of 731 m and 444 m respectively, which can also be seen in Figure 4.1A. The Neckar valley with its tributaries dominates this area, which extends from the southwest to the northeast over the area of the study site. In between are predominantly agriculturally used landscapes with settlements and towns, and the Schönbuch with its extensive, characteristic forests.

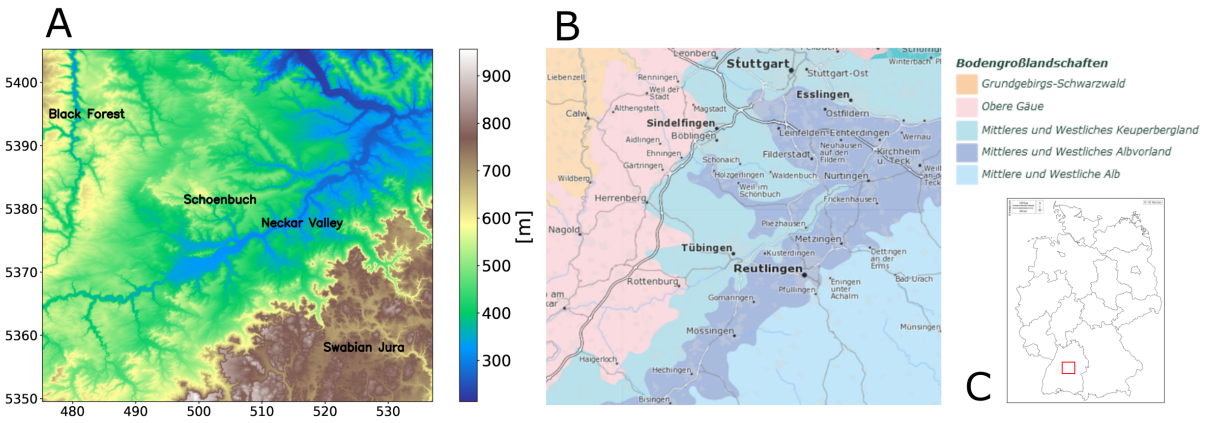

Figure 4.1: (A) Digital elevation model of the study area with its important landscapes, (B) distribution of the soil families, (C) location of the study area in Germany

In the southeast is the SJ, which is characterized by its unique maritime geologic formation with calcaric sedimentary parent material and the resulting different terrain, climate, geological substrates and thus soil units. Because of these features, it stands out from the region and in our case can be considered as an almost distinct area. In the northwest we have the Black Forest (BF), which also differs from the rest of the area in terms by the features mentioned, but at the same time has similarities due to the likewise terrestrial geologic formation including sandstone. This great difference in such a small area naturally influences the vegetation and the processes in the soil, including soil formation. In total, the area

Table 4.1: Overview of the features for the ANN

Environmental input data		Definition after
Topographic features	Altitude above channel network	[15] and [39]
	Eastness	
	Northness	
	Catchment area	
	Convergence index	
	Crest index for lowlands	
	Crest index for mountain areas	
	Diffuse radiation	
	Direct radiation	
	Elevation	
	Elevation below culmination line for lowlands	
	Elevation below culmination line for mountain areas	
	Horizontal curvature	
	Mean slope	
	Plan curvature	
	Profile curvature	
	Projected distance to stream	
	Relative elevation	
	Relative hillslope position for lowlands	
	Relative hillslope position for mountain areas	
	Steepest slope	
	Terrain classification index for lowlands	
	Topography	
Spectral features	Brightness index	[85]
	Colouration index	
	Hue index	
	Normalized difference vegetation index	
	Redness index	
	Saturation index	
Geological features	Geological map	LGRB

comprises five major soil landscapes with different characterization, shown in Figure 4.1B. These are areas in which, under similar geological, morphological and climatic conditions and under the influence of human, a landscape-typical association of soils has developed.

4.2.2 Data

Figure 4.2A shows the soil units in the study area, with each number representing a soil unit and its characterization. The soil units are determined according to the German soil classification system, which is based on the processes taking place in the soil and their properties [48]. In our area, there are 40 different soil units and the urban area, which is represented by the number 0. A detailed description can be seen in the Table 4.2, including the translation from the German into the WRB soil systematics [89].

In order to preserve the diversity that is lost in this translation, we will stick to the German classification. The soil unit map used for our prediction variable was initially provided by LGRB as a polygon map (Figure 4.2A). We converted this polygon map to a raster file using a rasterization function based on the digital elevation grid. While the original scale of the map is 1:50,000, its rasterization allowed to produce a raster with pixels of 10 x 10 meters. As features for the ANN, exemplified in Figure 4.2B, we looked for spatially dense data over the whole region to get as detailed data as possible, which is also important for the performance of the ANN. For this purpose, we use a digital elevation model (Figure 4.2B (a)), which was also provided by the LGRB with a resolution of 10m, based on which topographic indices were calculated, also with a resolution of 10m. The decision on which of the variables we use as features is based on expert geographical knowledge of the region, commonly used features in the geosciences and by using the scorpan

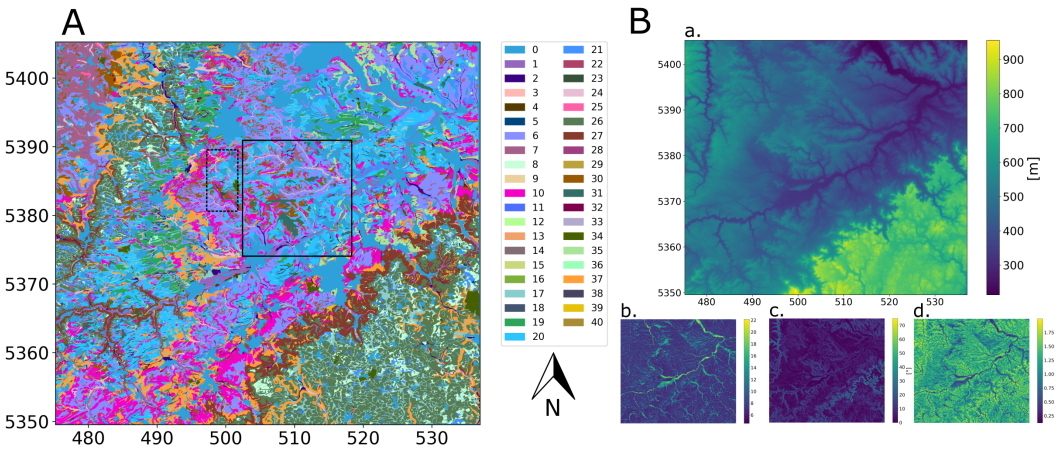

Figure 4.2: (A) soil unit map over the study area, created by the LGRB with the train and validation area (solid rectangle) and the test area (dashed rectangle) (B) Examples of four used features over the study area: a. digital elevation model, b. soil moisture, c. slope and d. NDVI

model introduced by McBratney et al. [121], which is based on Jenny [90]. To cover most of the features presented in the scorpan model, we also included satellite data. Copernicus provides the Sentinel-2 data, available from 2017 in 13 spectral bands with a 5-day repetition frequency. For us, the most important variables are the visible (R, G, B) and near-infrared bands, which have a resolution of 10 m. We use these spectral bands to calculate important indices such as the NDVI (Figure 4.2B (d)) to describe vegetation cover. Finally, we calculate the median value for each index over the time series from March to May 2019. In our analysis, we used the median as the mean over years to mitigate the influence of outliers and to ensure a more robust representation of the data. To capture the influence of geology, we add a geological map with the scale of 1:50,000, provided by the LGRB and rasterized in the same way as the soil unit map. We provide an overview of all the features used for the ANN and the corresponding references in Table 4.1.

Table 4.2: Detailed description of the soil units

Classes number	Label	German Soil Classification	WRB-Classification	Detailed information
0	None	None	None	Ablation, order, settlement
1	A1	Brauner Auenboden, Auenbraunerde	Fluvisol, Cambisol	partly with gleying in the near subsoil, of alluvial sand and alluvial loam
2	A2	Auengley-Brauner Auenboden, Auengley-Auenbraunerde	Cambisol	from alluvial sand and alluvial clay
3	A3	Auengley, Auenpseudogley-Auengley, Brauner Auenboden-Auengley	Fluvisol	from alluvial sand and alluvial clay
4	A7	Auenbraunerde, Auenparabraunerde	Cambisol	from older alluvial sediment
5	B1	Terra fusca-Braunerde, Terra fusca-Parabraunerde, Reliktbraunerde	Leptosol, Cambisol	from solifluction soils over limestone and dolomite stone
6	B2	Braunerde, Pelosol-Braunerde, Pseudogley-Braunerde	Cambisol	from solifluction soils, partly alluvial and flood loam
7	B4	Braunerde, Podsol-Braunerde	Arenosol	mostly podzolic, from sandstone, debris-rich fluvial soils and slope debris
8	CF1	Braunerde-Terra fusca, Terra fusca	Cambisol	from limestone and dolomite
9	CF2	Terra fusca, Braunerde-Terra fusca	Cambisol	from relocated river gravels
10	D1	Pelosol, Braunerde-Pelosol, Pseudogley-Pelosol	Luvisol	from solifluction soils, subordinate from alluvial debris
11	D2	Pelosol, Pseudogley-Pelosol	Luvisol	from flood and terrace sediments
12	D3	Pelosol, Braunerde, Parabraunerde, Nass- and Quellengley	Cambisol	from sliding masses
13	G1	Gley, Quellengley, Kolluvium-Gley	Gleysol	from solifluction soils and sedimentary formations, mostly alluvial deposits
14	G2	Pseudogley-Gley, Braunerde-Gley, Gley	Planosol	from flood loam, old water and alluvial sediment
15	G3	Anmoorgley, Nassgley, Humus- and Moorgley	Gleysol	from alluvial deposits, floodplain and flood sediments, and glacial deposits
16	H1	Niedermoor, Gley-Niedermoor, Hochmoor	Gleysol	from peat
17	K1	Kolluvium	Anthrosol	partly over Braunerde and Parabraunerde, from alluvial deposits over solifluction soils
18	K2	Pseudogley-Kolluvium, Gley-Kolluvium	Planosol	from alluvial deposits
19	L1	Parabraunerde	Luvisol	from loess and sand loess
20	L2	Parabraunerde, Braunerde-Parabraunerde, Pseudogley-Parabraunerde	Luvisol	of loess loam and loess-loam-rich solifluction soils
21	L3	Parabraunerde, Pelosol-Parabraunerde, Terra fusca-Parabraunerde, Pseudogley-Parabraunerde	Luvisol	from solifluction soils and slope debris
22	L5	Parabraunerde, Parabraunerde-Braunerde, Pseudogley-Parabraunerde	Luvisol	from terrace sediments, river and meltwater gravels
23	N1	Ranker und Braunerde-Ranker	Leptosol-Cambisol	from sandstone
24	P1	Podsol und Braunerde-Podsol	Podzol-Cambisol	of sandstone, sandstone and flint rubble and solifluction soils
25	Q1	Regosol, partly (Locker)Syrosem	Regosol	of slope debris, partly anthropogenically redeposited debris
26	R1	Rendzina	Leptosol	from limestone and dolomite, partly from slope or alluvial debris
27	R2	Rendzina und Pararendzina	Leptosol	from slope debris, partly from landslide debris
28	R3	Rendzina und Terra fusca-Rendzina	Leptosol	from river gravels
29	R4	Rendzina	Leptosol	from calcareous tuff and tertiary freshwater limestone
30	S1	Pseudogley, Braunerde-Pseudogley, Pelosol-Pseudogley	Planosol-Cambisol	from solifluction soils, partly Pleistocene alluvial debris
31	S2	Pseudogley, Parabraunerde-Pseudogley	Planosol-Luvisol	of loess loam and loess-loam-rich solifluction soils
32	S3	Pseudogley, Kolluvium-Pseudogley	Planosol	from alluvial deposits
33	SS1	Stagnogley, Moorstagnogley	Gleysol	from solifluction soils, basin sediments and alluvial deposits
34	X1	Disturbed terrain		original soils often heavily modified
35	Y1	Rigosol	Anthrosol	from solifluction soils, loess and various solid rocks
36	YY1	Deposit soil		from different substrates
37	Z1	Pararendzina, Pelosol-Pararendzina, Braunerde-Pararendzina	Leptosol-Vertisol	from solifluction soils and slope debris, partly from landslide masses
38	Z2	Pararendzina	Leptosol	of loess and sandy loess, partly washed away or periglacially redeposited
39	Z4	Pararendzina	Leptosol	from flood deposits, alluvial debris, river and meltwater gravels
40	Z7	Pararendzina, Braunerde-Pararendzina	Leptosol-Cambisol	from volcanic weathering, partly covered by sedimentary material

4.2.3 Model architecture

The origin of ANNs lies in the field of image recognition, especially in the area of classification [61]. These models are known for their ability to model multiple outcomes quickly and efficiently with a large amount of data, even with absence of prior knowledge about the data. Inspired by the neuronal structure, they look for dependencies and patterns in the given data that include input features and a responding output variable. ANNs are organized in layers consisting of neurons using a (non-)linear activation function to transform and forward their inputs to the next layer, allowing the ANN to learn complex patterns. The input layer receives the input data and consists of one neuron per input feature, in our case, one neuron per feature. The neurons in the hidden layers pass the weighted sum of the outputs from the previous layer to their activation function. The final layer outputs the prediction and consists of one neuron per output variable, in our case, one neuron per soil unit. During training the weights of the connections between the layers are learned via stochastic gradient descent to minimize a loss function measuring the error of the predictions. There is a wide variability of different constructs for an ANN for computation or information processing in terms of the architecture of the ANN, the number, types and dimensions of layers, or the activation function chosen. Since the focus of our study is on uncertainty of ML models in a soil context rather than on model performance, the simplicity of the model was very important to us. We choose a fully connected MLP as described in Table 4.3. As the activation function for the hidden layer, the rectified linear unit function was chosen, first used by Hahnloser et al. [70] and defined as

$$\mathrm{ReLU}(z) = \max(0, z)$$

with z as input to a neuron.

Table 4.3: Architecture of the ANN

Layer	Number of Neurons	Activation Function
Input layer	30	ReLU
Layer 1	395	ReLU
Layer 2	510	ReLU
Layer 3	489	ReLU
Output Layer	41	Softmax

4.2.4 Confidence and uncertainty measurement of Artificial Neural Networks

We use the softmax function for the output layer to transform the previous layer's outputs into a vector of probabilities, essentially a probability distribution over the input units. Mathematically, the softmax function is defined as follows [27, 61]:

$$\sigma(a_k) = \frac{\exp(a_k)}{\sum_j \exp(a_j)},$$

where a_k is the vector of raw values for all soil units. The soil unit with the highest value is often used as the ANNs' prediction. However, we are not only interested in determining the soil units for a given location and creating a soil unit map, but also in evaluating the probability distribution of these units over individual pixels. These probability values can be interpreted as a measure of confidence in the classification result. A higher maximum probability indicates that the predicted soil unit is more likely to accurately represent the soil unit at the given pixel position. In other words, the ANN has predicted this unit with a low uncertainty and is therefore very confident about the prediction. A value of 1 demonstrates that there is a very high level of confidence in the predicted unit for that particular pixel, suggesting that the model is almost certain about its prediction. On the other hand, a value of 0 reflects a very low level of confidence, indicating that the model shows a high uncertainty about the predicted unit for that pixel.

4.2.5 Last-Layer Laplace Approximation

To analyse the uncertainty of ANNs, most studies only consider the confidence measurement from the previous section. However, this does not yet take into account the uncertainty of the model itself. As mentioned earlier, ANNs are known to be miscalibrated in terms of their uncertainty [68] or even tend to be overconfident in areas that are not well covered by training points or that are far away from them [77]. To improve this overconfidence and at the same time obtain a spatially broad uncertainty expression, we use the LLLA by Kristiadi et al. [104]. This method is based on a probabilistic and Bayesian method by [1], which calculates the a posteriori uncertainty for the ANN weights. More precisely, we approximate the posterior with the Laplace approximation by estimating the posterior with

$$p(w \mid \mathcal{D}) \approx \mathcal{N}(w; w_{\text{MAP}}, \Sigma), \quad \text{with} \quad \Sigma = \left(\nabla^2_w \mathcal{L}(\mathcal{D}; w)\big|_{w=w_{\text{MAP}}}\right)^{-1}.$$

where w_{MAP} is the maximum a posteriori estimate of the last-layer parameters, obtained by minimizing the negative log posterior $\mathcal{L}(\mathcal{D}; w)$, *e.g.* standard DL with a cross-entropy loss and an isotropic Gaussian prior. In other words, this amounts to training the network as usual to find w_{MAP}, then computing the Hessian of the training loss at this point. This approach is significantly cheaper than alternative methods based on sampling weights (both in terms of compute and memory cost). It also has the benefit that the point estimate (Θ_{MAP}) is unaffected by the uncertainty estimation, which simplifies development and tuning. Nevertheless, the computation of the Hessian adds a computational overhead. Computing the full Hessian is not feasible for large networks. But prior work by other authors [104] has shown that limiting the Hessian to just the last layer, which is much less costly, already produces structured and useful uncertainty. Other approximations to curvature are also possible [44], but we limit ourselves to the LLLA in this work.

4.2.6 Training and optimization process of the model

In contrast to most studies, our train and validation area is not arbitrarily chosen, but we deliberately select a specific area, that contains all soil units typical for the four soil landscapes, developed under terrestrial conditions, shown as a solid rectangle in Figure 4.2A. The reason for the intentional selection is that we want to ensure that the soil units of the maritime soil landscape, *i.e.*, those from the SJ, are not included in order to look at how the model relates to regions about which it does not receive information on soil units. In addition, we choose the centre of our map because we want to simulate a situation that often occurs, that some parts of a study area are well sampled due to previous individual projects.

Table 4.4: Distribution of soil units in data subsets

	Ground truth	Training and validation area	Test area
classes in the area concerned	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	0, 1, 2, 3, 4, 6, 7, 10, 13, 15, 17, 18, 19, 20, 21, 23, 25, 30, 31, 32, 35, 36, 37, 38	0, 2, 3, 6, 7, 10, 13, 15, 17, 18, 20, 21, 23, 30, 31, 32, 33, 34, 36
Total number of classes	41	24	19
classes that change compared to previous column	-	5, 8, 9, 11, 12, 14, 16, 22, 24, 26, 27, 28, 29, 33, 34, 39, 40	1, 4, 19, 25, 33, 34, 35, 37, 38

The datasets contain the 30 features from Table 4.1 and the soil unit labels. The SJ and its typical soil units are intentionally not present in the training and validation dataset to avoid providing information to the model. The test dataset (dashed rectangle in Figure 4.2A) is similar in respect of the occurring soil units to the training and validation area to be able to determine the overall accuracy of the model. The detailed breakdown of soil units in the different data sets is shown in Table 4.4 and in Figure 4.3. The training and validation set comprise 1,922,946 datapoints, while the test set contains 824,120 datapoints.

We tuned architectural and training hyperparameters using sequential model-based optimization on 1% of the full dataset, to obtain results in a feasible time. Specifically, we used Bayesian optimization [56] combined with Successive Halving to allocate resources to promising settings, as implemented in the tool SMAC [113]. We tuned a total of six hyperparameters comprising the number of neurons for each of the three layers (between 32 and 512 units), the initial learning rate (between $1e-4$ and 1.0), the learning rate scheduler (cosine annealing or exponentially decaying), and weight decay (between $1e-7$ and $1e-1$). We used 500 configurations as the optimization budget and trained each network using Adam [97, 115] with a fixed batch size of 1024 on 70% of the training data and used the obtained accuracy on the 30% validation set (obtained via a stratified split from the subsampled train dataset) as the optimization objective. For the remaining analysis, the best performing network configuration,

shown in Table 4.3 with a weight decay of 0.008394464526246698 and a learning rate of 0.006203468269518196 was used.

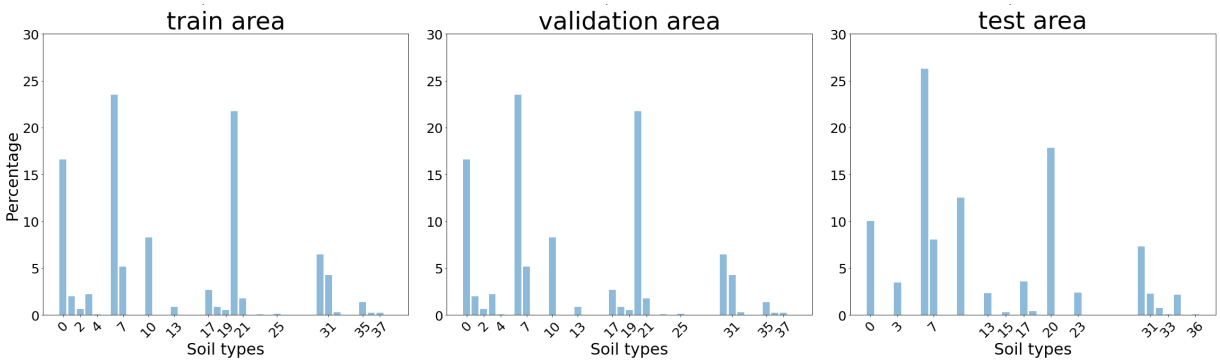

Figure 4.3: Histograms of the distribution of soil units in the training, validation and test area

After optimization, we made the decision to use the ANN with an early stop triggered when the ANN predicted the training accuracy, measured as the percentage of pixels where the model correctly predicts the outcome, to be above 90 % and the accuracy of the test data set did not increase significantly.

4.3 Results & Discussion

4.3.1 Loss and accuracy of the Artificial Neural Network

It is important to note that our research's aim is not to outperform other state-of-the-art ANNs in DSM. Instead, our goal was to provide a realistic representation of the model's capabilities and limitations, especially when dealing with predictions in new areas with different soil units. We deliberately kept the model architecture simple and performed only a brief tuning process to simulate a real-life situation where quick predictions with a prepared ANN are required. At the same time, we focused on a specific scenario where we had two well-sampled areas: one for training the model and the other for testing. For the rest of the study area, we assume that we have no information on soil units. The developed model achieved a low loss of 0.07, which can indicate its effectiveness. The training accuracy of 95.11% and the validation accuracy of 94.37% also indicate that the model is capable of performing well within the specific training area, which only includes a deliberate selection of soil units. By limiting the training data to a geographically enclosed area with similar soil characteristics, we could assess how well the model generalises to unseen similar regions. The test area, while geographically close to the training area, had some differences in soil units and their distribution, as shown in Figure 4.3. Even with these variations, the model's test accuracy was 47.95%, which was expected and still can be considered as good. Comparing our model to other ANNs used in DSM, particularly for predicting soil units, it performed at an average level, which aligns with our expectations based on previous studies [9, 13, 24, 195].

4.3.2 Prediction of the Artificial Neural Network

Predicted soil unit map

When we directly compare the map generated by the ANN prediction in Figure 4.4 (Plot A) with the ground truth map derived from the LGRB in Figure 4.2 (Plot A), several significant results emerge. Not all soil units occur in the prediction, which is expected, as certain soil units were not in the chosen training domain (see Table 4.4 column two). Consequently, the ANN was unable to predict these missing soil units because it had no knowledge of them. The decision to consider a specific area as our training data, where some soil units are missing, stems from the recognition that even within a relatively small geographical region, there can be significant variability and transitions between different soil units [186]. In such complex landscapes, it is entirely plausible that certain soil units may not have been sampled due to their close proximity or subtle variations that might have been overlooked during the sampling process [82]. By deliberately incorporating this aspect into our training dataset, this approach provides valuable insights into how the ANN responds to such common scenarios and assesses its ability to generalise and extrapolate predictions across the entire area, including regions with missing soil units [125]. In our study the absent soil units include the soil family Terra fusca (class 5, 8, and 9), a specific soil unit of the Rendzina family (class 29) and the Pararendzina family (class 39 and 40), which are typical soil units from SJ. The other part of the soil family Rendzina (classes 26, 27 and 28) is also typical for the SJ region and is also found at the border to the BF, specifically in the Oberen Gäue and along the upper course of the Neckar. In addition, some soil units are only found in specific areas. For instance, Pelosol (classes 11 and 12) and Gley (class 14) occur exclusively within the SJ region and the northern area of the Neckar valley. Conversely, classes 16, 22, 33, and 34 are sparsely distributed throughout the entire area. Lastly, the soil unit Podsol (class 24) is typical for the BF. Because these soil units are only found in the specific regions SJ (southeast) and BF (northwest), it was anticipated that the SJ and BF areas have incorrect predictions for these soil units.The results in Figure 4.4B illustrate this, where each green pixel represents a correctly predicted soil unit.

The square in the centre of the Figure 4.4B, with a high number of correctly predicted pixels, corresponds to the training area where the model performed very well. Additionally, when observing the diagonal of the image, which corresponds to areas similar to the training region, the predictions are also accurate. Notably, in the SJ area in the southeast, class 17 is the only one that was partially predicted correctly. Class 17 contains soils that are a result of human land use, including slope sediments from soil erosion (Colluvisol), soils that have been formed by mixing of natural material through human activity (Anthrosol) and alluvial deposits along creeks and rivers. All these azonal soils lack a close relation to natural soil forming factors as described by Jenny [90] and used in the scorpan model by McBratney et al. [121] except topography. They occur in any populated landscape, which is true for our full prediction area. This result demonstrates that the spatial pattern of the occurrence of azonal soils like Colluvisols and Anthrosols strongly relates to their topographic position [142]. This facilitates a transfer of the model results to other areas. In

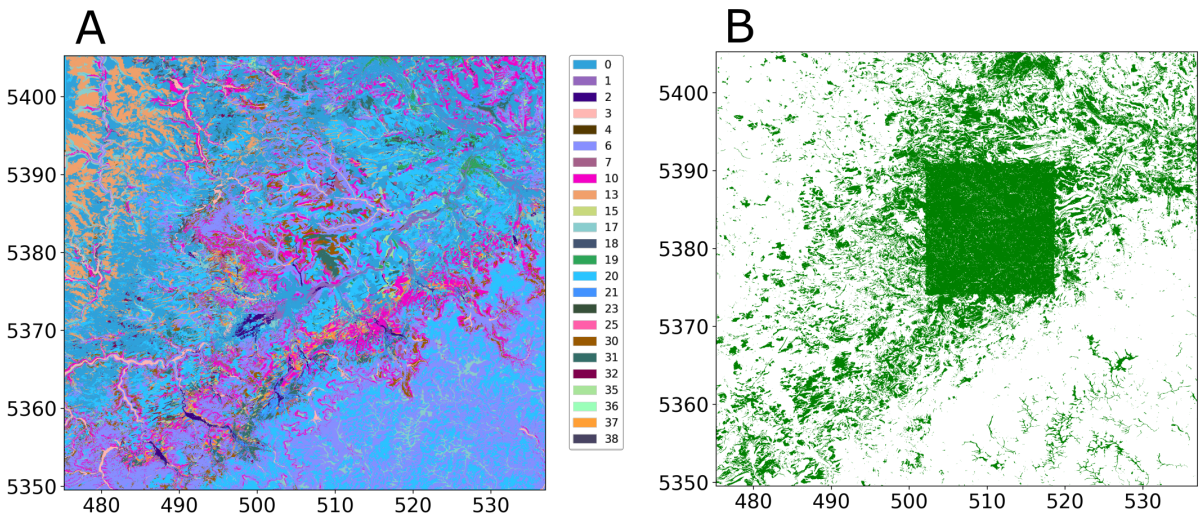

Figure 4.4: (A) Prediction map of the soil units in our study area by the ANN, (B) Comparison of the prediction with the ground truth: green means correct prediction of the soil unit

addition, it can be understood as an indicator for process proximity of an ANN, in that basic process-based rules of soil formation are recognized and reproduced by the model, as indicated before from several studies [31, 128, 137].

Consequently, as observed from the predicted map, it is evident that certain soil units occur more frequently and are more widely distributed, while others are less prevalent compared to the ground truth map. A detailed breakdown of the soil units from the soil truth map and their descriptions can be found in Table 4.2. The most substantial increase in predicted pixels compared to ground truth pixels, both in absolute numbers and relative proportions, was observed for classes 0, 6, 13, and 20. Additionally, in relative numbers, there was an increase in classes 3, 31, and 15. A more moderate increase was noted for classes 2, 10, 30, 32, and 36. On the other hand, a decrease was observed mainly in absolute numbers for classes 7, 17, 19, 21, and 37, and in relative proportions for classes 1 and 4. A minor decrease was recorded for classes 18, 35, 38, 25, and 23. Despite these decreases, it is essential to note that they are not as significant as the increases observed in some other classes.

The noticeable increase in the occurrence of classes 0, 6, and 20 can be attributed to the fact that our training data set is unbalanced. These particular classes are more frequently represented in the training set compared to other soil units. As a result, the ANN tends to predict them more often in the output [91]. During the prediction, not only are soil units that are absent in the training area assigned to certain regions, but also less frequently occurring soil units, as clearly shown in Figure 4.5. The increase in the occurrence of certain soil units in our ANN is not a random phenomenon. For instance, even though class 0 exists in the SJ region in the ground truth, it is not predicted there by our ANN. Instead, classes 6 and 20 dominate in that region. Both represent the two most commonly occurring soil units in Central Europe, making it reasonable for the ANN to classify areas about which it has no information [7]. The occurrence of class 13 in the prediction is not accounted for by the frequency of it the training data. When we closely examine its distribution in the prediction, we find that this class is predicted for the high ridges

around the BF. Previously, these areas were dominated by classes 6, 7, and 30, but now they are predicted as class 13, along with classes 17 and 0. Both class 13 and class 17 share similar characteristics to those in the ground truth soil units. For example, Gley soils can indeed occur in the BF as associated soils, although not as extensively as depicted in the prediction [20]. The overestimation of class 17 in the region is not surprising, as Colluvium is a correlated sediment of soil erosion, and it can occur independently of climate and geology, primarily in depressions and valleys, just as shown in the prediction [99]. Another significant change is observed in classes 1 to 4, which collectively represent the soils in the floodplains. The increase in class 3 is a result of soil units 1 and 2 being assigned to it. On the other hand, the decrease in class 4 is because this soil unit is now primarily predicted as class 2. Despite these shifts, it is important to note that overall, the soil units in classes 1 to 4 have remained within their respective soil family. The prediction map highlights this phenomenon, particularly in the upper course of the Neckar and its tributaries, where class 3 is predicted, and in the Nagold valley, which is located in the BF. Meanwhile, the middle and lower Neckar valley are primarily predicted as class 1. These predictions support the notion that the floodplain soils retain their general characteristics, even with some changes in specific soil unit assignments. The repetitive nature of river systems, with their well-defined channels and floodplain areas, provides a distinct and recognizable pattern that can be learned effectively by the ANN during training [189].

4.3.3 Confidence and Uncertainty of the Artificial Neural Network

Based on the previous results, especially the incorrect prediction of soil units in the SJ and BF regions, the next step is to analyse the confidence of the ANN in predicting the soil unit. In the case of an ANN, apart from cross-validation methods and other techniques, a common step is to assess the probability of the predicted class [183]. This probability can be interpreted as the model's confidence in its predictions. In Figure 4.5A, this confidence value is plotted.

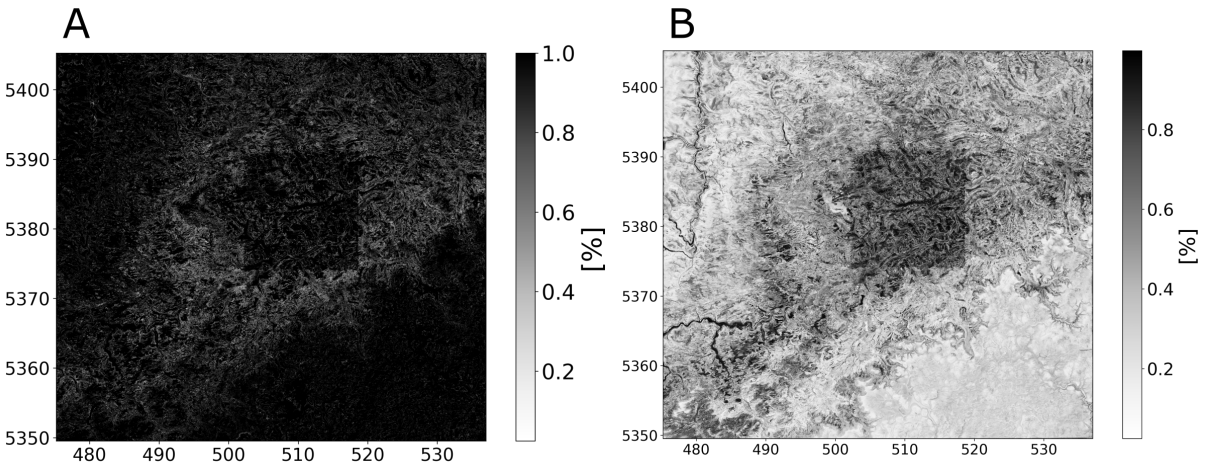

Figure 4.5: (A) probability of the class in our study area predicted by the ANN calculated with the softmax function, interpreted as confidence of the ANN (B) probability after applying the LLLA, interpreted as uncertainty of the ANN

A value of 1 indicates high confidence in the corresponding pixel's predicted class, while a value of 0 indicates low confidence in the prediction. This information provides insights into the reliability of the model's predictions for each specific pixel and the corresponding predicted class. One striking observation is the presence of three areas where pixels with high confidence are concentrated. The central area corresponds to the training region of the ANN, and it aligns with previous results, indicating that over 90% of the pixel predictions in this region were accurate. However, it is not the area with the highest overall confidence, as we would expected it. Instead, the regions of SJ in the southeast and BF in the northwest show almost uniform high confidence values approaching one, despite the ANN's poor performance in these areas. These regions are geographically and in terms of soil units distant from the training area and in addition the predicted classes often differ from the ground truth labels. As previously mentioned, ANNs tends to be overconfident in situations where it lacks sufficient training data, leading to inaccurate interpretations [77, 92]. Looking specifically at the pixels where the ANN correctly predicted soil units, we find that the mean confidence is 93%. Interestingly, even for the wrong predictions, the mean confidence remains relatively high at 92%. This indicates that the ANN assigns high confidence to both correct and incorrect predictions, further exacerbating the problem of overconfidence. Furthermore, in the surrounding regions of our training area, there are lower confidence values for the prediction of classes, particularly extending along the diagonal from southwest to northeast. This is also problematic because this area outside the training domain is where the prediction worked well previously.

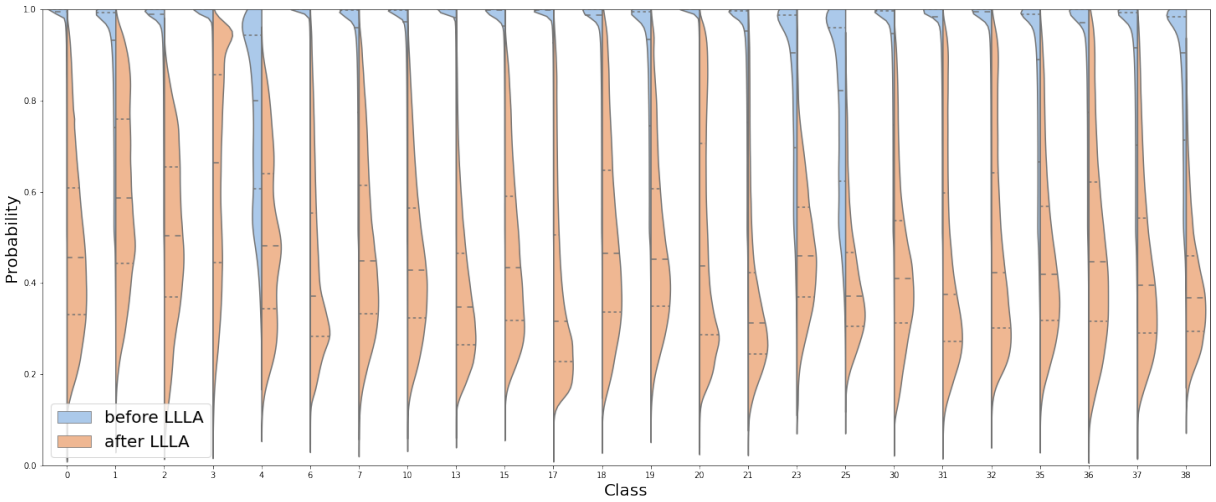

Figure 4.6: Confidence and uncertainty distribution of the soil unit classes individually, illustrated with the help of a kernel density estimate and additionally by quartiles of the distribution of each predicted class (blue shows the distribution before the application of LLLA, salmon afterwards)

Our examination of the confidence distribution of the probability of the predicted classes for each individual class reinforces a similar pattern, as visually depicted in Figure 4.6. The confidence distribution is represented by the blue curve on the left-hand side of the axis at each class. The soil units shown here correspond to those in the training set, as only these could be taken into account by the ANN in its prediction and thus in the confidence distribution. Strikingly, for each class, there is a peak near

the value of 1, indicating that most confidence values for these classes lying there. As well, as the confidence values decreases, the number of points lying around the lower values rapidly decreases, approaching zero. However, we noticed a distinct pattern for classes 4, 23, and 25, where a wider distribution of confidence values was observed. This can be attributed to the fact that these soil units appear infrequently in the predictions and are located in a diagonal region on the map, where we have a more heterogeneous representation of the model's confidence. Interestingly, classes 1, 19, 21, 35 to 38, which are also situated in this diagonal region, exhibit a weakly spread confidence distribution. Given the characteristics of the confidence distributions observed, it is evident that conducting a reliable analysis of the predictions made by the ANN is challenging. The presence of sharp peaks near 1 for most classes, suggests that the ANN is generally highly confident in its predictions. But as shown before, this high confidence may not always be well-calibrated or reliable, which is to be expected due to the absence of these soil units in the training area [54, 103]. These findings emphasize the need to address the issue of overconfidence in the ANN, particularly for distant regions from the training area, and to improve the model's confidence prediction with an uncertainty measurement in the surrounding regions where it previously showed correct predictions. These further investigations and adjustments are necessary to enhance the reliability and applicability of ANNs.

4.3.4 Uncertainty of the Artificial Neural Network due to Last-Layer Laplace Approximation

As described in Section 4.2.5, the application of the LLLA method has allowed us to obtain uncertainty statements for our model's predictions, effectively addressing the issue of overconfidence commonly associated with ANNs [104]. Following the post-processing of the model confidence with the LLLA method, we obtained Plot B in Figure 4.5, where lighter shades represent higher uncertainty. In the plot, we observe distinctive patterns among different regions. The training area stands out prominently with low uncertainty, boasting the lowest overall uncertainty levels across the map. This aligns with the predictions in the training area, as it has the highest concentration of correctly predicted soil units. Conversely, the SJ region now exhibits the highest uncertainty in the plot, indicating a higher level of uncertainty for the predicted soil units in that area. This observation is consistent with our knowledge that this area could not be predicted accurately because its main soil units were not provided to the ANN during training [54]. However, it is worth noting that the few correctly predicted pixels belonging to class 17 in the SJ area have the lowest uncertainty there. This is again consistent with the previously observed fact that the model correctly observed the soils and remains confident after applying LLLA with the correctly predicted class 17 representing the colluvisols, which also simplifies the application of the model by adding LLLA to different regions. Similar to the SJ area, we also observe the BF region with a relatively high uncertainty, except for the Nagold valley, where the uncertainty is notably lower. This indicates that the model's predictions in the BF region are generally less reliable, except for the specific area of the Nagold valley where the model exhibits

higher confidence. In the southwest region, we encounter a more diverse picture with varying levels of uncertainty. As in the previous cases, there is surprisingly little uncertainty in the Neckar Valley. Upon closer examination, we find that the predicted soil unit for both the Neckar valley and the Nagold valley corresponds to the previously described soil unit number 3. Although this prediction was incorrect, the low uncertainty in these cases can be explained by the fact that the original soil units in both valleys belong to the same soil unit family as class 3. Even though they are ultimately incorrect, this similarity in soil unit family allows the model to be more certain in its predictions [154]. The area in the diagonal of the map, which is the most similar to the training area, has now slightly higher uncertainty after the application of LLLA. Nevertheless, these uncertainty values are still higher than those observed in the SJ and BF regions. The slight increase in uncertainty after LLLA is acceptable, even if it is the most correct prediction after the training range. Considering the other results we have obtained, this increase in uncertainty is still within an acceptable range. Upon revisiting the mean uncertainty of all correctly predicted pixels, we find that it has dropped to 60.4 %, which is significantly less than before. If we instead examine the mean uncertainty of incorrectly predicted values, we observe a decrease of nearly 20%, indicating that we are able to detect the uncertainty even on average.

When we delve into more detail by examining the corresponding part in Figure 4.6, which is depicted in orange on the right side of the axis for each class, we gain a more diverse and informative understanding compared to the previous representation. The plots are wider, indicating a broader distribution of uncertainty values. Notably, class 3 stands out distinctly, as it is the only class with a peak of uncertainty surpassing 90%. This aligns with the previous knowledge that class 3 is found in the valleys, precisely representing the locations where the uncertainty is lowest. Additionally, the only other wider plot at higher values belongs to class 1, which belongs to the soil family of floodplains as class 3, and exhibits similar behaviour. It has been further observed that class 2 and class 4, also belonging to this soil family, show a notable concentration of uncertainty values around the average, unlike all the rest of the classes. These clear patterns of uncertainty within the alluvial soil family underline the ability of the LLLA uncertainty measurement to recognize subtle geographical and pedological features. This shows that the model not only recognizes that these soil units belong to the alluvial soil family, but also distinguishes them from other soil units with different characteristics.

The analysis of class 0 reveals that its uncertainty values span a wide range from nearly 0 to 1, but there is a significant concentration at low values. Similar patterns are observed for other classes, specifically class 6 and class 20. This is a positive outcome compared to the previous situation, as these soil units were previously consistently overpredicted with high confidence at the SJ, but at the same time still well predicted in the other areas. The uncertainty statement thus covers both situations well for these classes. This shows that the uncertainty measurement can better classify the spatial variability of the soil classes and is not only based on the fact that these classes occur with the highest number of data points in the training data, which is important for a reliably prediction with an imbalanced dataset [125]. Another important observation is related to classes 13 and 17, which now exhibit distributions concentrated

at the lowest confidence values, indicating higher uncertainty for these classes. This increased uncertainty now captures the fact that these classes were completely overestimated in the BF region, leading to their relative increase in the number of pixels in the prediction range. In summary, the integration of the Bayesian method LLLA has allowed us to obtain more reliable uncertainty estimates for the model's predictions, which has proven to be in line with other Bayesian methods for predicting soil variables [146]. The distinct patterns observed in the uncertainty values for different regions provide valuable insights into the model's performance and its reliability in various areas.

4.4 Implication on soil management

In the field of soil management, the integration of ANNs is promising to improve precision and efficiency [96]. However, a major challenge is to address the problem of overconfidence in ANNs, which can lead to inaccurate predictions and suboptimal decision-making. To combat this, we can leverage the LLLA method, which generates reliable uncertainty statements from ANNs. This method allows us to create uncertainty maps that play a crucial role in decision-making processes [81, 122]. Uncertainty maps allow soil mappers to target and prioritize areas to increase efficiency. By incorporating targeted sampling in regions with high uncertainty into the prediction workflow, we can quickly improve the quality of soil maps generated by ANNs [50, 168]. Thereby optimizing soil management practices by providing the dual benefit of broader spatial coverage and targeted re-sampling. In addition, exploring the potential transferability of a trained ANN to similar regions with comparable soil classes is of great interest [127]. When an ANN trained in one region is applied to another with similar soil characteristics, it is crucial to understand how well the model performs in the new context. A map indicating areas of greater uncertainty can directly show where the ANN may not function effectively. This visualization of uncertainty is crucial as it helps us understand the limitations and constraints of the model in practical applications. Such information provides essential insights for future investigations and improvements in both data collection and model development.

4.5 Conclusion

The primary objective of this research was to develop a reliable and straightforward method for quantifying uncertainty in ANNs used in DSM, particularly for soil unit predictions in order to make the predictions more reliable and interpretable. This also includes the correction of overconfidence of ANNs, a tendency of ANNs to make overconfident predictions, especially in regions with limited data. This issue is particularly concerning in soil mapping because accurate and precise soil information is essential for effective soil management. To tackle the issue of insufficient uncertainty measurements for ANNs, we introduced the Bayesian technique LLLA. The LLLA method is designed to produce more trustworthy uncertainty statements for the specified

study area predicted by ANN. By incorporating LLLA into the ANN modelling process, the study demonstrates its effectiveness in showing of the areas, where the ANN erroneously has a high confidence and thereby preserves the low uncertainty in the correctly predicted areas. This improvement in uncertainty estimation ensures that the model's predictions are more reliable and trustworthy. This is crucial, because it helps users understand the level of confidence they can place in the model's predictions and therefore make decisions. Additionally, this research provides valuable insights into identifying knowledge gaps by analyzing areas with limited data support and high uncertainty. This information guides researchers in prioritizing data collection efforts in regions where the model's predictions are less reliable. Moving forward, we plan to apply the ANN with the LLLA approach to real soil samples and compare the results with maps generated by soil mappers. This comparison will further enhance our understanding of the underlying processes. In conclusion, the incorporation of LLLA into ANN modeling offers a promising solution to improve uncertainty estimation, making DSM predictions more reliable, interpretable, and actionable for effective soil management.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645 and the Tübingen AI Center (FKZ 01IS18039A). We also thank the State Authority for Geology, Mineral Resources and Mining (LGRB), Baden-Württemberg, Freiburg, Germany, for providing soil data.

Quantifying spatial uncertainty to improve soil predictions in data-sparse regions

5.

Abstract

Artificial Neural Networks (ANNs) are valuable tools for predicting soil properties using large datasets. However, a common challenge in soil sciences is the uneven distribution of soil samples, which often results from past sampling projects that heavily sample certain areas while leaving similar yet geographically distant regions under-sampled. One potential solution to this problem is to transfer an already trained model to other similar regions. Robust spatial uncertainty quantification (UQ) is crucial for this purpose, yet often overlooked in current research. We address this issue by using a Bayesian deep learning (DL) technique, Laplace Approximations, to quantify spatial uncertainty. This produces a probability measure encoding where the model's prediction is deemed reliable, and where a lack of data should lead to a high uncertainty. We train such an ANN on a soil landscape dataset from a specific region in southern Germany and then transfer the trained model to another unseen but to some extend similar region, without any further model training. The model effectively generalised alluvial patterns, demonstrating its ability to recognize repetitive features of river systems. However, the model showed a tendency to favor overrepresented soil units, underscoring the importance of balancing training datasets to reduce overconfidence in dominant classes. Quantifying uncertainty in this way allows stakeholders to better identify regions and settings in need of further data collection, enhancing decision-making and prioritizing efforts in data collection. Our approach is computationally lightweight and can be added post-hoc to existing DL solutions for soil prediction, thus offering a practical tool to improve soil property predictions in under-sampled areas, as well as optimizing future sampling strategies, ensuring resources are allocated efficiently for maximum data coverage and accuracy.

5.1 Introduction

ML has become an indispensable tool in scientific research, leading to significant advances in many fields, including soil science [192]. Since the early 2000s, ML methods have been steadily integrated into soil mapping [13, 121, 161]. Over the past two decades, the use of ML in soil science has grown substantially, reflecting its increasing importance and effectiveness [93, 126, 152, 174, 192]. However, new research challenges have emerged as ML methods become more widely used. One key challenge is improving model interpretability in order to promote scientific knowledge [140]. Chen et al. [34] further added that future research should focus on making the most of legacy datasets, using smarter sampling strategies, improving model accuracy and interpretability, and developing advanced mapping methods to create detailed and high-quality soil maps. In line with this, Bohn and Miller [22] showed that locally enhanced, bottom-up oriented DSM approach has been shown to deliver higher accuracy compared to both conventional soil maps and global DSM products in many cases.

Using legacy datasets effectively means applying data from already sampled areas to predict conditions in similar but unsampled regions. This extrapolation process has been discussed for many years. For example, Lagacherie et al. [107] pointed out the need to develop self-learning systems that dynamically adapt predictions. Bui and Moran [30] demonstrated that existing soil maps can be combined with environmental and geological data to extend their usefulness beyond their original boundaries. Additionally, Scull et al. [162] showed with classification trees that this technique allows soil experts to focus field mapping on unique areas and efficiently extrapolate soil-landscape relationships, making it a valuable tool for soil surveys. Extrapolation approaches also address the challenges of conventional soil mapping, which relies on cartographers manually surveying landscapes, a process that is both costly and time-consuming. These methods offer a particularly cost-effective solution for predicting soil classes in regions with limited data, helping to fill the gaps in soil maps and improving the efficiency of DSM [175]. For example, decision trees demonstrated a 46.00 % overall accuracy for extrapolating soil subgroups using digital mapping methods, making them a cost-effective option for areas with limited data or challenging sampling conditions [134]. Similarly, multinomial logistic regression and classification trees have been used successfully to extrapolate soil classes [2, 64, 111]. To summarize, the increasing adoption of ML is driven not only by its relevance to soil science but also by its ability to significantly reduce the effort required for mapping, especially in large or hard-to-access areas [66, 83, 168].

More advanced methods, in particular ANNs, have proven effective for extrapolation in soil mapping. For instance, the study by Arruda et al. [8] demonstrated the potential of ANNs to produce digital soil maps, providing initial classifications for unexplored areas. Building on this, Coelho et al. [38] introduced an innovative methodology that combined georeferenced soil profile point data and ANN models for extrapolation tasks. Responding to the growing demand for high-resolution soil maps in, for instance, precision agriculture, environmental management, and land-use planning, ANNs are becoming more and more popular due to their ability to process large amounts of data and provide predictions comparably fast [75, 156, 164]. Brungard et al. [29] demonstrated the superior accuracy of complex models containing ANNs in predicting soil taxonomy classes compared to simpler models. Similarly, Zhu [195] highlighted the capability of ANNs in generating high-resolution soil maps.

Despite their advantages, one notable challenge is the lack of inherent interpretability [80]. As “black box” models, ANNs make predictions through complex internal processes that are difficult to understand and interpret. Recent studies have addressed this limitation by introducing model-agnostic interpretation techniques and game theory-based Shapley additive explanations (SHAP), which provide valuable insights into the relationships between environmental features and model predictions [138, 184]. In addition, ANNs typically lack built-in UQ, which complicates the evaluation of their predictive reliability and may lead to misinterpretations or suboptimal decision-making [68]. They often produce overly confident predictions, sometimes reaching 100.00 % certainty, even when the input data is flawed or noisy [25, 77, 135]. In the

context of DSM, this issue is compounded by the broader challenge of quantifying spatial uncertainty in soil maps [79, 151, 183]. Between 2017 and 2022, only 35.00 % of studies that addressed significant DSM tasks incorporated uncertainty in their analysis [16]. Similarly, while DSM research is expanding in countries such as India and Iran, the integration of uncertainty mapping remains limited. In India, only 34.00 % of DSM studies include uncertainty maps, while in Iran, fewer than 20.00 % address uncertainty [43, 191]. Typically, these maps then present just an overall accuracy expressed as a single statistical measure, often derived through cross-validation techniques, an iterative process that partitions the training data into multiple subsets to repeatedly train and validate the model to estimate overall performance uncertainty [183]. Although this approach offers some insight, it falls short, especially for applications involving unbalanced datasets. This gap has led to calls for more detailed uncertainty analysis [125], particularly for tasks involving extrapolation to new areas because of poor uncertainty performance in such contexts [64]. Some recent studies have made strides in incorporating UQ. For instance, Carvalho Monteiro et al. [32] and van der Westhuizen et al. [177] have demonstrated progress in quantifying uncertainties for RFs, while Saygın et al. [155] have explored the use of ANNs. However, many of these methods rely on variance estimates, which fail to adequately address critical issues such as model overconfidence. This problem has emerged in the study by Schmidinger and Heuvelink [157] that ANNs produce overly optimistic probabilistic predictions, resulting in low-reliability scores. Additionally, these approaches frequently neglect spatial uncertainty, an essential aspect of practical soil mapping [10]. The most commonly used methods for UQ in DL algorithms, particularly in ANNs, include MC Dropout, ensemble methods, and full Bayesian approaches. These methods, while effective, often require significant computational resources and memory [3]. These techniques have begun to gain traction in soil science applications, particularly for estimating uncertainty in soil moisture retrieval or soil spectral models [112]. For example, Padarian et al. [141] and Huang et al. [87] utilized these approaches to assess uncertainty in their models, demonstrating their relevance and utility despite the computational demands. These findings underscore the urgent need for methodological advancements that go beyond variance estimation to also tackle overconfidence together with spatial uncertainty while remaining computationally efficient and easy to integrate into existing workflows. Such improvements are crucial to ensure that ML models for DSM provide both accurate and reliable predictions. Our previous work Rau et al. [151] introduced for DSM the LLLA, a computationally efficient technique that addresses these challenges. Building on this methodological foundation, the current study applies an ANN model to an extrapolation task, predicting soil units non-adjacent target area outside the training area. To identify and correct the overconfidence of the ANN and perform a spatial analysis of the model's predictions and associated uncertainties, we use the LLLA, providing corrected uncertainty estimates for every pixel in the target area. Through this, we assess the transferability of the ANN by improving its interpretability and reliability for soil mapping tasks. Ultimately, our work aims to promote more robust, accurate, and insightful applications of DSM.

5.2 Material & Methods

5.2.1 Study area

This study investigates two regions in central Baden-Württemberg, Germany, near the city of Tübingen. The reference area is located northwest of the city, and the target area lies to the southwest, as shown in Figure 4.2 (A). These regions were chosen because they share similar geology, climate, and cultural development, making them suitable for comparative analysis. The reference area, named after the Goldersbach stream, covers 8.86 km^2 with an average elevation of 445.51 m above sea level, ranging from 325.31 m to 552.48 m. It represents the lower section of the upstream part of the Goldersbach River and its catchment. The main land use since the 19th century in this area is forestry and since 1972 it has been part of a nature park. The target area, named after the Bühlertalbach stream, is larger, covering 18.5 km^2 with an average elevation of 498.26 m above sea level, ranging from 388.86 m to 583.04 m. It includes the entire Bühlertalbach stream valley, from its upstream to downstream sections. Similar to the reference area, forestry is the main land use, and this area is extensively used for forest-related activities. Both areas have the same underlying geology, belonging to the Middle and Upper Keuper series, which consist of layers of sandstone, claystone, and marlstone, creating typical soil patterns of the Keuperbergland. The climate in both regions is cool temperate moist, with an average annual temperature between 8.3 °C and 8.7 °C and annual precipitation ranging from 740 mm to 770 mm. The target area was deliberately chosen to be larger, encompassing the entire catchment of the Bühlertalbach River. This strategic decision allows for the investigation of how predictions and findings extend beyond the upstream areas on which the reference area is based. By including the full catchment, this approach provides a broader and more comprehensive understanding of processes in similar but not equal landscapes.

5.2.2 Data

Figure 5.1 illustrates the distribution of soil units across the reference area (blue, subfigure (B)) and target area (red, subfigure (C)), with each number corresponding to a specific soil unit and its associated characterization. The classification of the soil types in these units follows the LGRB soil classification system, a local variant of the German soil classification KA5, which is structured around soil formation processes and properties [48]. Within the reference area, there are eight distinct soil units, alongside an urban zone represented as unit 0. In contrast, the target area exhibits greater diversity, comprising 14 unique soil units. A comprehensive description of all these units is provided in Table 5.1, including their correspondence with the WRB soil classification system [89].

For international understanding, we use the WRB classification system. The soil unit maps, initially sourced from the LGRB, were provided in vector format. To facilitate our analysis, these polygons were converted into raster files using a rasterization process based on digital elevation grids. The original map scale of 1:50 000 was rasterized to produce a 10 m $\times$ 10 m resolution. It should be noted that this study is based entirely

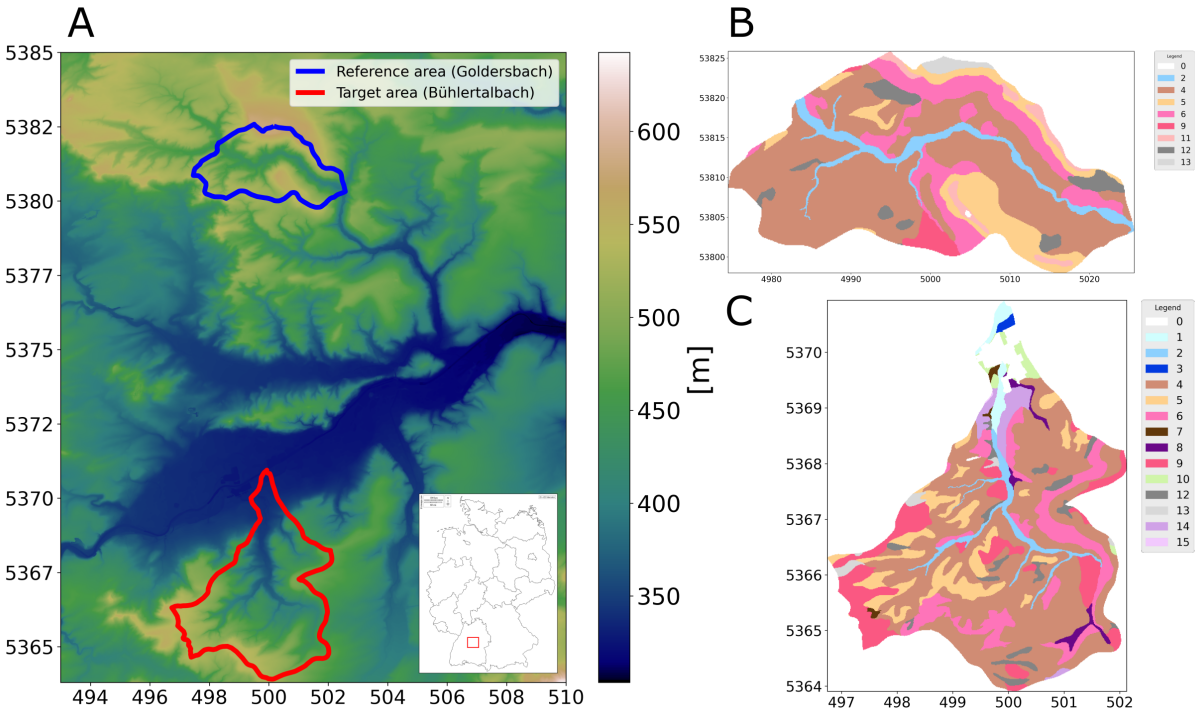

Figure 5.1: (A) Digital elevation model of the study area with the location of the study area in Germany, reference area in blue, target area in red, (B) and (C) Soil unit maps over the reference and target areas, created by the LGRB

on pixel-based soil unit prediction using these rasterized soil maps as training and validation labels rather than direct field observations. To enhance the performance of the ANN and ensure detailed analysis, spatially dense variable data were required for the entire region. For this purpose, digital elevation models (Figure 5.1 (A)) were used for both areas. These models offer a 10 m resolution and serve as the basis for calculating topographic indices, also at 10 m resolution. The variable selection was informed by local expert geographical knowledge, guided by commonly used proxies representing the scorpan model introduced by McBratney et al. [121], which draws upon Jenny [90]. In addition, we included spectral indices based on satellite data from the *Copernicus Sentinel-2* program. Since 2017, *Sentinel-2* provides data in 13 spectral bands with a 5-day revisit time. For this study, we focused on the visible (R, G, B) and near-infrared bands, which have a 10 m resolution. Using these bands, we calculated indices such as the NDVI to measure vegetation cover. To ensure robust data representation and reduce the impact of outliers, we computed the median values of these indices for the time series of cloud-free images from March to May 2019. We also included geological maps, scaled at 1:50 000, provided by the *LGRB*. These maps were rasterized in the same way as the soil unit maps.

Table 5.1: Detailed description of the soil units

Class no.	Label	German Soil Classification	WRB-Classification	Detailed information
0	None	None	None	Ablation, order, settlement
1	A1	Brauner Auenboden, Auenbraunerde	Fluvisol, Cambisol	partly with gleying in the near subsoil, of alluvial sand and alluvial loam
2	A3	Auengley, Auenpseudogley-Auengley, Brauner Auenboden-Auengley	Fluvisol	from alluvial sand and alluvial clay
3	A7	Auenbraunerde, Auenparabraunerde	Cambisol	from older alluvial sediment
4	B2	Braunerde, Pelosol-Braunerde, Pseudogley-Braunerde	Cambisol	from solifluction soils, partly alluvial and flood loam
5	B4	Braunerde, Podsol-Braunerde	Arenosol	mostly podzolic, from sandstone, debris-rich fluvial soils and slope debris
6	D1	Pelosol, Braunerde-Pelosol, Pseudogley-Pelosol	Luvisol-Vertisol	from solifluction soils, subordinate from alluvial debris
7	K1	Kolluvium	Anthrosol	partly over Braunerde and Parabraunerde, from alluvial deposits over solifluction soils
8	K2	Pseudogley-Kolluvium, Gley-Kolluvium	Gleyic Anthrosol	from alluvial deposits
9	L2	Parabraunerde, Braunerde-Parabraunerde, Pseudogley-Parabraunerde	Luvisol	of loess loam and loess-loam-rich solifluction soils
10	L3	Parabraunerde, Pelosol-Parabraunerde, Terra fusca-Parabraunerde, Pseudogley-Parabraunerde	Luvisol	from solifluction soils and slope debris
11	N1	Ranker und Braunerde-Ranker	Leptosol-Cambisol	from sandstone
12	S1	Pseudogley, Braunerde-Pseudogley, Pelosol-Pseudogley	Planosol-Cambisol	from solifluction soils, partly Pleistocene alluvial debris
13	S2	Pseudogley, Parabraunerde-Pseudogley	Planosol-Luvisol	of loess loam and loess-loam-rich solifluction soils
14	Z1	Pararendzina, Pelosol-Pararendzina, Braunerde-Pararendzina	Leptosol-Vertisol	from solifluction soils and slope debris, partly from landslide masses
15	R3	Rendzina und Terra fusca-Rendzina	Leptosol	from river gravels

Table 5.2: Overview of the features for the ANN

Environmental input data		Definition after
Topographic indices	Eastness, Elevation, Northness, Slope	[12]
	Diffuse radiation, Direct radiation, Slope discontinuities, Terrain classification index for lowlands	[21]
	Relative height above the depth line, Soil moisture	[23]
	Catchment area	[51]
	Plan curvature, Profile curvature	[76]
	Convergence divergence index, Crest index for lowlands, Crest index for mountain areas, Culmination line for lowlands, Culmination line for mountain areas, Elevation below the culmination line for lowlands, Elevation below the culmination line for mountain areas, Horizontal distance to the depth line, Relative hill slope position for lowlands, Relative hill slope position for mountain areas, Relative altitude, Relief	[100]
	Depth of closed surface depressions	[185]
Spectral indices	Brightness index, Colouration index, Hue index, Normalized difference vegetation index, Redness index, Saturation index	[85]
Geological variable	Geological map	*LGRB*

Table 5.2 summarizes the indices and variables used for the ANN as features and their respective references. To compare the features in the reference and target areas, we applied the cosine similarity index, as outlined by Schütze et al. [160]. This method, which measures similarity on a scale from −1 (completely opposite) to 1 (identical), resulted in a mean value of 0.85. Using this score confirmed a strong similarity between the areas. In addition, we collaborated with experts from the *LGRB*, whose extensive regional knowledge ensured the appropriate selection of study areas. Both the similarity assessment and the expert consultation were carried out in recognition of the fact that, even at the local scale, it is crucial to apply models only where they are valid, a principle already established in global-scale research [116].

5.2.3 Model design

ANNs originated in the field of image recognition, particularly for classification tasks [61]. These models are highly effective at identifying patterns and relationships in data, even without prior domain knowledge, and excel in handling large datasets. ANNs are composed of layers of neurons leveraging activation functions to learn complex patterns. The structure of an ANN can vary widely in terms of its architecture, the number and type of layers, their dimensions, and the activation functions used. Since our research prioritizes understanding uncertainty in ML models applied to soil data rather than optimizing model performance, we opted for a straightforward design: a fully connected MLP, as described in Table 5.3. For the hidden layers, we employed the rectified linear unit (ReLU) activation function, which is defined as:

$$\mathrm{ReLU}(z) = \max(0, z)$$

with z as input to a neuron [53, 59, 130]. For training and validation, we used data from the reference area, which includes 33 features (listed in Table 5.2) and soil unit labels. The reference area comprises eight distinct soil units, which the model aims to predict. To evaluate the model's performance, we tested it on the ground truth map of soil units from the

target area (Figure 5.1 (C)). The training and validation dataset consisted of 142 569 data points, *i.e.* the number of raster cells, which was separated through random sampling to a 70%-30% split, while the test dataset contained 378 214 data points. We used the architecture mentioned above. A detailed description of the model tuning protocol is provided in Rau et al. [151], where the method was first tested in a simplified, controlled soil classification setup. The optimized hyperparameters derived from this process were successfully transferred and applied to the reference area, yielding excellent results. To enhance the model's robustness and prevent overfitting, we implemented an early stopping criterion. The training process was halted when the model's training accuracy, defined as the percentage of correctly predicted pixels, exceeded 95.00 %, and no significant improvement in test dataset accuracy was observed.

Table 5.3: Architecture of the ANN

Layer	Number of Neurons	Activation Function
Input layer	33	ReLU
Layer 1	395	ReLU
Layer 2	510	ReLU
Layer 3	489	ReLU
Output Layer	9	Softmax

5.2.4 Uncertainty measurement of ANNs with Last-Layer Laplace Approximation

For ANNs commonly the softmax function in the output layer is used to convert raw scores into a probability distribution over the predicted classes. The softmax function transforms the output of the previous layer into a vector of probabilities, essentially forming a distribution across input classes. The softmax function is defined as follows:

$$\sigma(a_k) = \frac{\exp(a_k)}{\sum_j \exp(a_j)},$$

where a_k is the vector of raw score for all soil units [27, 61]. These probability values can be interpreted as uncertainty about the classification output. A higher probability indicates greater certainty, while a lower value signifies uncertainty. In other words, the ANN has predicted a class with low uncertainty and is therefore very confident about the prediction. Nevertheless, relying solely on softmax-derived uncertainty measures has limitations, particularly in regions where the ANN encounters data points far from its training distribution, as they do not account for the uncertainty in the model's parameters or structure [68, 77]. To address these limitations and quantify model uncertainty, *i.e.* epistemic uncertainty, we employ the LLLA following Daxberger et al. [44]. This method is based on Bayesian principles and provides a computationally efficient approach to estimate posterior uncertainties for neural network parameters.

The LLLA approximates the posterior distribution of the weights as

$$p(w \mid \mathcal{D}) \approx \mathcal{N}(w; w_{\text{MAP}}, \Sigma), \quad \text{with} \quad \Sigma = \left(\nabla^2_w \mathcal{L}(\mathcal{D}; w)\big|_{w=w_{\text{MAP}}}\right)^{-1}.$$

Here, w_{MAP} represents the maximum a posteriori estimate of the last-layer parameters, obtained by minimizing the negative log posterior $\mathcal{L}(\mathcal{D}; w)$, typically the cross-entropy loss with an isotropic Gaussian prior. Epistemic uncertainty is captured by the LLLA through the local curvature (Hessian) of the loss: flat directions in this geometry indicate parameters that are weakly constrained by the data and thus remain uncertain. Such flatness arises in regions of limited data, structural uncertainty, or poor transferability, all of which reflect ambiguity in the posterior distribution over model parameters. The method is accessible through the open-source `laplace.torch` package, which facilitates easy integration into PyTorch-based workflows. The LLLA method offers key advantages: it is computationally efficient by focusing on the last layer [104] and has the benefit that the point estimate (w_{MAP}) is unaffected by the uncertainty estimation, which simplifies development and tuning. We already have shown that LLLA effectively identifies areas of high uncertainty in soil classification tasks [151], making it crucial for generating uncertainty maps with uneven training data coverage.

5.3 Results & Discussion

5.3.1 Loss and accuracy of the ANN

In our study, we used a simple neural network architecture rather than a highly specialized one tailored to the soil classification extrapolation task. This decision reflects common scenarios where pre-built models are preferred due to their ease of use and quick deployment. Our goal was to assess and enhance the ANN's ability to extrapolate, and not to achieve the highest possible overall accuracy outperforming other state-of-the-art ANNs, an objective that could be pursued through deliberate and targeted hyperparameter optimization [147]. For this reason, we focus on the spatial uncertainty at pixel level rather than the total uncertainty of the soil map, which is used in DSM [183]. The study was therefore based on two different areas: a well-sampled reference area and a completely unsampled target area (Figure 5.1). This setup simulated a realistic challenge, where models are often required to make predictions in areas with limited or no prior information [82]. The results indicated that the model learned the training data effectively, achieving a low loss value of 0.01, a high training accuracy of 98.57 %, and a validation accuracy of 96.73 %. However, when applied to the target area, the test accuracy dropped significantly to 47.38 %. This decrease was expected since there was no area-specific tuning for the ANN and our target area is substantially larger than the reference area including the entire course of the river. The reference area can thus not fully represent the target area [186]. Compared with other studies in DSM that use ANNs for soil classification (*e.g.*, [9, 13, 24, 134, 195]), our model performed at an average level, consistent with our expectations due to similar parent material and climate as well as cultural development of the areas over the last centuries. These initial findings emphasize the trade-offs between

simplicity and predictive performance when using simple ANNs in soil mapping applications. While these are convenient and easy to deploy, their performance is often limited in data-sparse regions. This highlights the importance of complementing ANN predictions with UQ to effectively identify data gaps.

5.3.2 Prediction of the ANN

The prediction of soil units across the target area reveals several notable patterns and challenges when we compare the ANN-predicted map in Figure 5.2 (A) with our ground truth derived from the *LGRB* in Figure 5.1 (C). Not all soil units were predicted, which is expected as certain soil units (in that case soil units 1, 3, 7, 8, 10, 14, and 15) were absent from the reference area. The ANN could not predict these soil units due to its lack of training data. This phenomenon is not uncommon in practice, as soil units in complex areas often remain untested in reality [82]. For example, units 1 and 3 belong to floodplain soils, more precisely to Fluvisols after WRB, and are not included in the reference area. Similarly, soil unit 14, representing soils over the Gipskeuper formation, is also missing. The absence of certain soil units in the training dataset reveals how the ANN handles such cases and provides insight into the associated uncertainty. A detailed breakdown of the soil units is provided in Table 5.1.

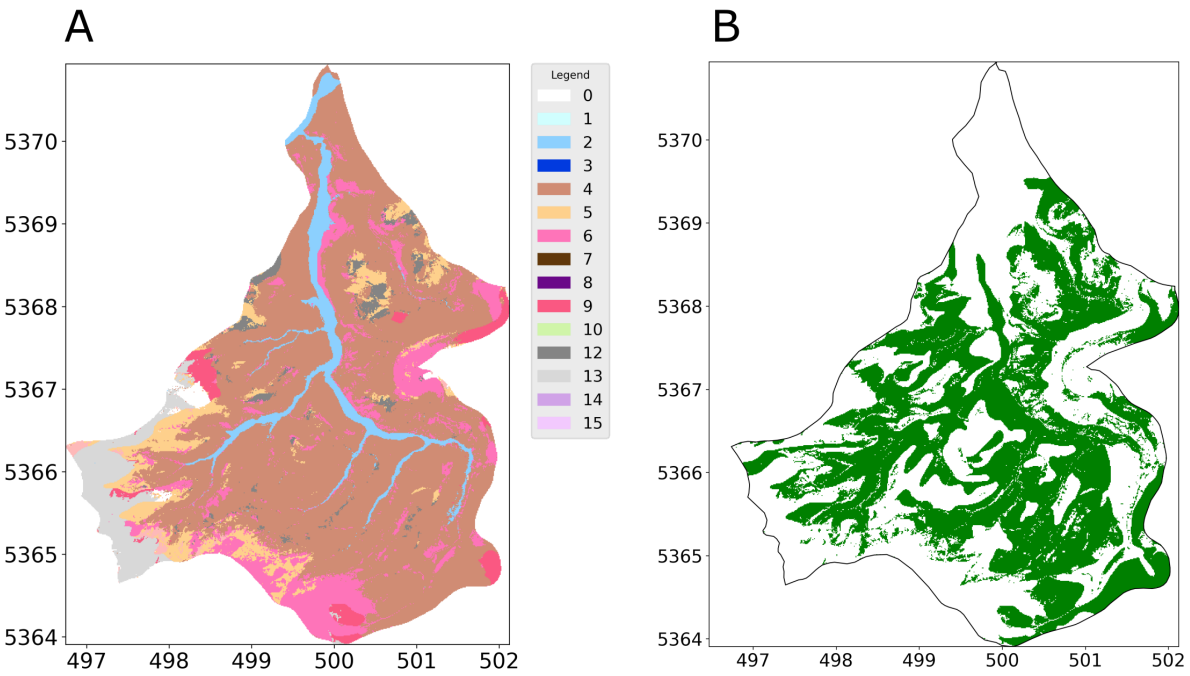

Figure 5.2: (A) Prediction map of the soil unit in our target area by the ANN, (B) Comparison of the prediction with the ground truth: green means a correct prediction of the soil unit

An overall evaluation of the predictions reveals that certain regions, particularly the northern and southwestern parts of the target area, were predicted wrong (Figure 5.2 (B)). Interestingly, these areas correspond to the upstream and downstream sections of the river, which were areas not well represented in the reference area. We can see that predictions were more accurate in the middle stream of the river, which closely aligns with

the reference area, compared to the downstream and upstream regions. This is likely due to the increasing distance from the training data. A brief and selected breakdown of the prediction of the specific soil units is now given and can be followed with the comparison of the ground truth map in Figure 5.1 (C) with the prediction of the ANN in Figure 5.2 (A) and with the confusion matrix in Figure 5.3.

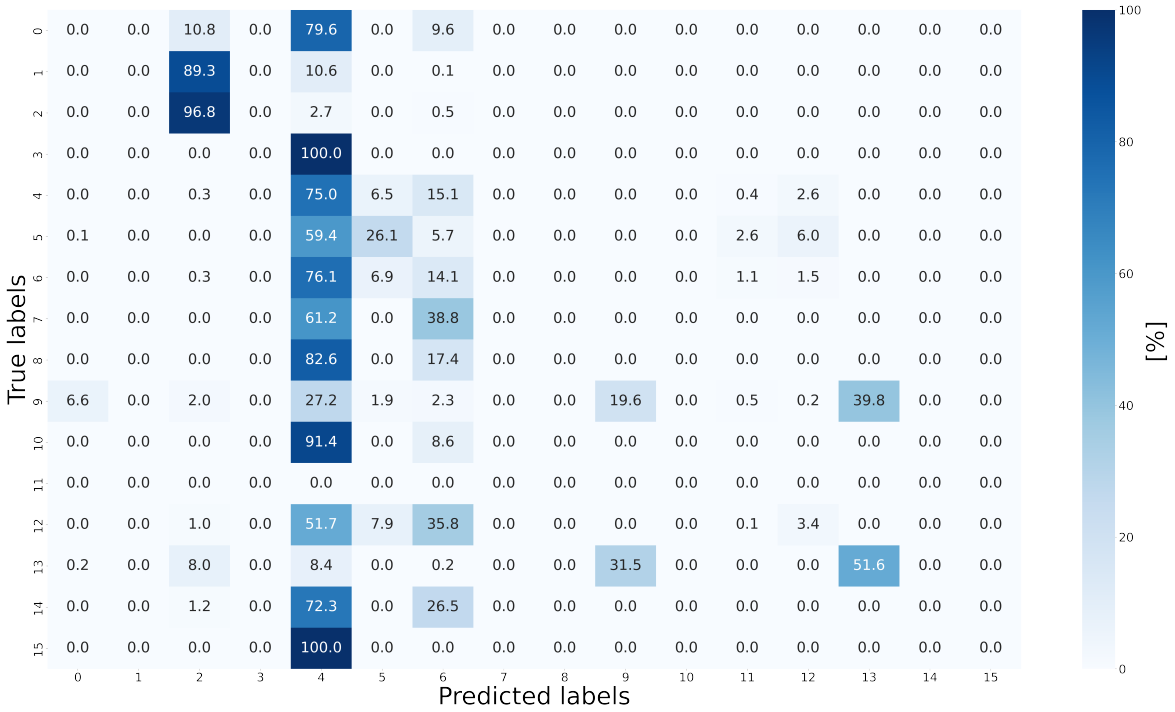

Figure 5.3: The confusion matrix of the ANN displays true vs. predicted classifications, with diagonal values indicating correct predictions and off-diagonal values showing misclassifications. The percentages reflect how much of each actual class was predicted either correctly (diagonal) or incorrectly (off-diagonal).

Unit 0, which represents the settlements, was completely misclassified and often predicted as unit 2 or more often as units 4 and 6. Because unit 0 is found in the river valley and extends towards the receiving stream, these predictions are not surprising. It is interesting to see what happens to soil units 1 and 3, both alluvial soil units, neither of which occur in the reference area. Soil unit 1, a Fluvisol from alluvial sand and loam, was predicted as soil unit 2, which is the only Fluvisol in the reference area, also from alluvial sand and clay. Meanwhile, unit 3, classified as Cambisol, is often predicted as soil unit 4, representing Cambisol formed from alluvial deposits, which aligns with unit 3's characteristics as a floodplain Cambisol from older sediments. In conclusion, the model's predictions for units 1 and 3 demonstrate its ability to recognize and generalise alluvial patterns, even when specific soil units are missing from the reference data. These results suggest that floodplain soils maintain recognizable characteristics across regions, and the model effectively learns the repetitive patterns of river systems, such as floodplains and channel deposits, during training. Unit 4 is dominant in both the reference and target areas, as confirmed by the ground truth map. Its proportion increased from 51.09 % in the ground truth from the target area to 64.21 % in the prediction by the ANN. However, this soil unit was overestimated in the central region and underestimated in the south, demonstrating the

importance of spatial analysis. This is also the case for soil unit 5. This soil unit was underestimated in the central region but correctly identified in the southwest and north. Overestimations occurred in the south, and misclassifications primarily involved unit 4. This can be explained by the fact that both represent the most common soil units in Germany and the domination of 4 in the training data [7, 189]. As a result, the ANN tends to predict them more often in the output [91]. Soil unit 6 was also underestimated overall but maintained its proportional representation due to false predictions along the southern margins. The colluvial soil units 7 and 8, were not part of the training set, so they were misclassified as 4 and 6, soil unit 7 specifically as unit 6 in the southwest and unit 4 in the north. Unit 9 was extremely accurately detected in four areas in the south, east, and west but was underrepresented overall. In the southwest, areas belonging to unit 9 were often predicted as unit 13, which is also a Luvisol from loess loam.

The model effectively recognizes familiar soil units, indicating that it successfully learns and applies process-based rules of soil formation, as demonstrated by the predictions for soil units 1 and 3. However, it shows a tendency to generalise soil units based on shared properties, as seen in the misclassification of soil unit 13 as unit 9 due to their similar origin as Luvisols from loess. This suggests the model is adept at identifying broad patterns but lacks sensitivity to regional nuances and finer distinctions between similar units. The substantial regional variability in the predictions highlights the need for spatial uncertainty analyses to improve accuracy and address the model's limitations in handling less common or unfamiliar soil units.

5.3.3 Confidence of the ANN

Based on the previous results, especially the large distribution of correctly and incorrectly predicted classes in an unit and a non-spatial accuracy of 47.38 %, we now analyze the uncertainty of the ANN prediction of every single soil unit before applying the LLLA. In the case of an ANN, besides cross-validation methods and other techniques, a common step is to evaluate the probability of the predicted class [183]. This probability can be interpreted as the confidence of the model in its predictions, thus the degree of uncertainty of the model regarding the predictions per pixel (Figure 5.4 (A)).

Notably, the highest confidence values, often reaching 100 %, are observed at the borders in the south, west, and north, as well as in the central region near the river. In contrast, the intermediate regions display a more diverse confidence distribution, though the values remain generally high. This trend is reflected in the mean confidence value, which stands at 96.22 %. When examining the relationship between confidence and prediction accuracy, it can be seen that in areas where the ANN performs poorly (Figure 5.2 (B)), the confidence values paradoxically remain high (Figure 5.4 (A)). This indicates overconfidence in regions, which happens with ANNs when the training data does not represent the target area. For example, pixels where soil units are correctly predicted exhibit a mean confidence of 97.17 %, while incorrectly predicted pixels also demonstrate a high mean confidence of 95.36 %. This pattern underscores the ANN's tendency to assign high confidence to both correct and incorrect

predictions, exacerbating the issue of overconfidence. Such behavior aligns with findings from previous studies, which have highlighted the tendency of ANNs to exhibit overconfidence in data-scarce regions [77, 92, 151].

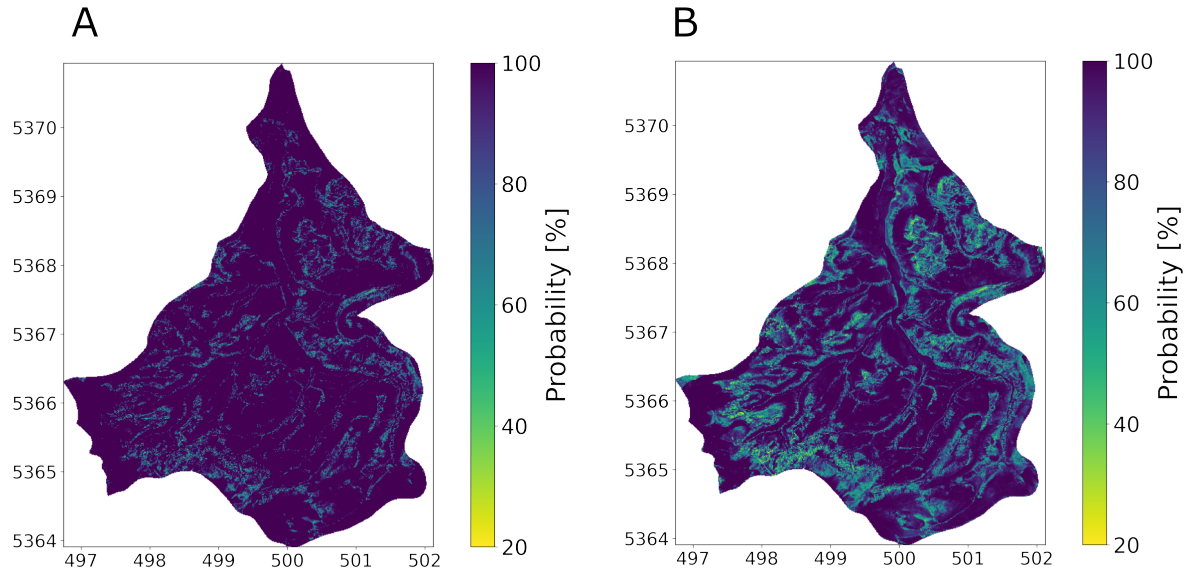

Figure 5.4: (A) probability of the soil unit in our target area predicted by the ANN calculated with the softmax function, interpreted as the confidence of the ANN (B) probability after applying the LLLA, interpreted as the uncertainty of the ANN

Further analysis of the confidence distribution for each soil unit is presented using a violin plot in Figure 5.5. The blue curves represent the distribution of confidence values for each unit, focusing only on soil units present in the reference area, as these are the only ones the ANN can predict. The width of the plot indicates where confidence values are more frequent, and the shape shows the range of these values. For most soil units, there are sharp peaks around 100 %, which means that the ANN is overly confident in all units. However, soil units 11 and 12 stand out, as the ANN also shows high confidence here, but the shape is in a wider range. This analysis highlights the issue of overconfidence of ANNs: here the ANN is too confident, even in areas where it performs poorly. This overconfidence is especially apparent in regions far from the training data or underrepresented areas. To improve the ANN's reliability, its ability to estimate its uncertainty needs to be enhanced. Further detailed analysis by soil units will be provided when comparing the ANN's predictions with those after applying the LLLA method as described in Section 5.2.4.

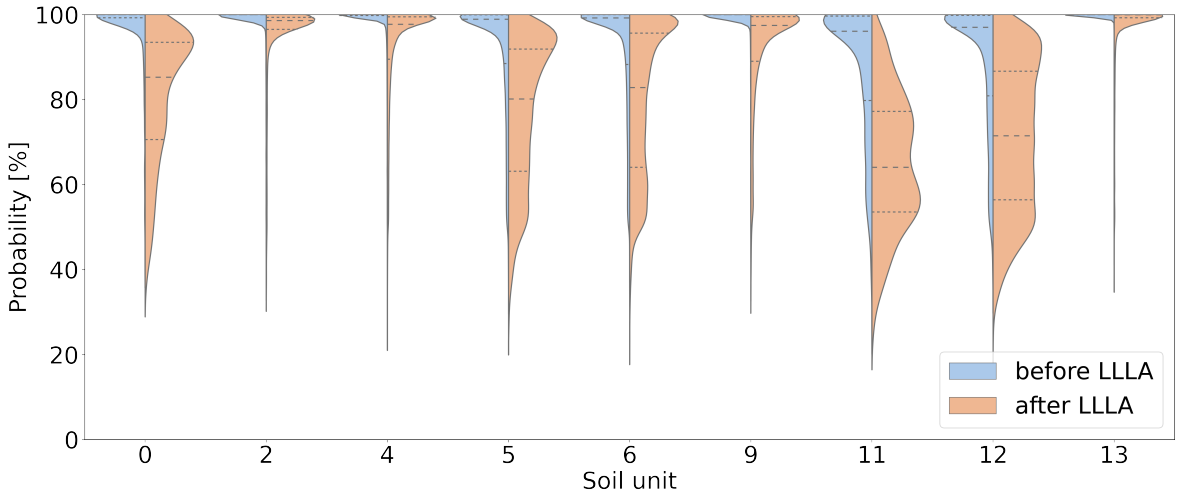

Figure 5.5: Distributions of the probability of predictions across the soil units (Blue: before LLLA; orange: after LLLA).

5.3.4 Uncertainty of the ANN from Last-Layer Laplace Approximation

As discussed in Section 5.2.4, the application of the LLLA method enabled us to generate uncertainty estimates for the model predictions, addressing the overconfidence issue typically associated with ANNs [104]. It is important to note that LLLA captures only epistemic uncertainty. Aleatoric uncertainty remains, as predictions are inherently constrained by the data on which they are based. After applying LLLA, the model's adjusted confidence values are shown in Figure 5.4 (B), where lighter colors indicate higher uncertainty. Some areas, like the western edge and the river region, showed almost no change, but overall, the average confidence dropped from 96.22 % to 88.66 %. This decrease shows that the LLLA method helped adjust the ANN's confidence to be more realistic. When looking at areas where the ANN made correct predictions, the mean confidence decreased by 5.97 %, so just minor adjustments are needed. A larger reduction can be observed in areas where the predictions are wrong. The mean confidence decreased more, by 9.00 %. This shows that LLLA was effective in reducing the model's overconfidence, especially where it previously made incorrect predictions.

Considering spatial differences inside correctly predicted areas, confidence reductions mainly occur along the edges. The opposite is true for wrongly predicted areas, where larger reductions occur more in the center, which is particularly apparent in the northern and southern regions. This indicates that it is important to look at the spatial variability of the single soil units. A first insight is provided by the violin plot (Figure 5.5), where the orange-colored part shows the confidence after LLLA. It shows that the spread of confidence values has increased for some soil units, like units 0, 5, 6, 11, and 12. This suggests that LLLA made the model uncertainty more precise for these units. For other units, like 2, 4, 9, and 13, the confidence distribution stayed mostly the same. The highest points of confidence, called peaks, shifted for some units. For example, units 0, 5, and 6 still had high peaks, but units 11 and 12

showed much lower peaks after LLLA. For most other units, the peaks remained high, meaning the ANN stayed confident in its predictions for those units.

Examining the behavior of individual soil units gives further insights. For soil unit 0, which was completely misclassified, the confidence in the areas, where it was predicted, dropped significantly, and the same happened for areas where this soil unit was wrongly predicted. In contrast, soil unit 1, which was mostly misclassified as soil unit 2, maintained high confidence in the correct areas, except for one point in the north, where it was misclassified as soil unit 4. This pattern indicates that when the model predicted soil unit 2, a familiar and similar soil unit, it remained confident, whereas the misclassification to soil unit 4, a less related unit, triggered a higher uncertainty adjustment. This finding indicates that the model is capable of differentiating between plausible misclassifications and more significant errors. A similar pattern is observed with soil unit 3, which was often misclassified as soil unit 4, a closely related soil unit, where the confidence remained high after LLLA. The most extreme example of a plausible misclassification is soil unit 9, which was misclassified as unit 13 for a large area in the west. Notably, there was no reduction in confidence despite the large spatial error, suggesting that the LLLA method failed to detect the misclassification, likely because soil units 9 and 13 are very similar. This highlights a limitation of the LLLA adjustment when soil units have closely overlapping characteristics, making it difficult for the model to recognize the need for uncertainty in such cases. Soil unit 2, which was well-predicted overall, maintained high confidence in both the correct areas and the false positive areas, where it was misclassified as soil unit 1. In the correct areas, it retained the highest confidence levels of all soil units, indicating that the ANN remained highly confident after the LLLA in its accurate predictions. The behavior of the LLLA for soil unit 4 is more complex. Recall that it was overestimated in the center and north and underestimated in the south. After applying LLLA (Figure 5.4 (B)), the confidence decreased significantly in the central areas, where it was wrongly predicted over soil units 5, 9, and 12, indicating that the model recognized uncertainty in those regions. However, in most of the other misclassified areas, with a small exception in the north, the model's confidence remained high, suggesting that it did not adjust sufficiently for those errors. This lack of uncertainty adjustment could be explained by the large proportion of soil unit 4 in the training data, leading the model to overtrust its predictions for this unit. The model seems to favor overrepresented units, even when faced with evidence of misclassification, which highlights the importance of balancing the training dataset to avoid overconfidence in dominant soil units [101]. The soil units 7 and 8, both absent from the training data, were misclassified as units 4 and 6. However, these misclassifications were detected extremely well by the LLLA adjustment. Initially, the model assigned high confidence to these areas, but after LLLA, the regions corresponding to soil unit 8 showed some of the lowest confidence values. This indicates that LLLA effectively identified areas of high uncertainty, particularly where the model faced unknown soil units, suggesting that the method is highly effective in detecting errors related to unfamiliar inputs. Soil unit 6 was underestimated overall, with false predictions along the southern margins. In the correct areas, where this unit should have been identified, LLLA significantly lowered the confidence, indicating

that the model recognized the initial overconfidence. However, in the wrongly predicted areas, the confidence decrease varied spatially. In the south and east, the model remained highly confident, even where soil unit 6 was misclassified as unit 4, showing that LLLA had difficulties detecting the uncertainty of soil unit 4. This is another example of overtrusted predictions of the model in regions where dominant units like soil unit 4 were prevalent in the training set. Conversely, in the center and north, LLLA effectively detected the misclassification, leading to a clear reduction in confidence. In conclusion, the application of LLLA effectively addressed the overconfidence issue of the ANN by providing uncertainty estimates and adjusting confidence levels in both correct and incorrect predictions. The method successfully reduced confidence in misclassified areas, particularly for unknown soil units like 7 and 8, indicating its effectiveness in detecting unfamiliar inputs. However, the results also show regional variability in the uncertainty adjustments. For some soil units, such as unit 4, the model remained overconfident in dominant units, especially for units that were prevalent in the training data. This highlights a limitation of LLLA in handling closely related or overrepresented soil units, emphasizing the need for balanced training data to improve the model's uncertainty calibration and overall robustness. Nevertheless, compared to broader global approaches like Homosoils, where even the study by Nenkam et al. [132] acknowledged that model accuracy improved significantly when incorporating local data, LLLA provides a key advantage by offering spatially resolved uncertainty estimates. This allows for more localized and detailed insights into the reliability of predictions, making it a valuable tool for identifying regional variations in model performance and improving uncertainty calibration at finer scales. In addition to established UQ methods for ANNs, such as MC Dropout, ensembles and full BNNs, the application of LLLA presents a practical and computationally efficient alternative. As a post hoc method, LLLA enables uncertainty estimation to be incorporated after model training without requiring any modifications to the architecture or learning process. This simplicity made it especially attractive for our soil prediction task, where retraining the model or restructuring the network would have been costly and unnecessary. LLLA operates by approximating the posterior distribution of the final layer weights, capturing model uncertainty with a single forward pass at inference time. Though the computation of the Hessian or its approximation introduces a one-time cost, it does not impact the efficiency of prediction, unlike MC Dropout which multiplies inference costs with repeated forward passes [44, 104]. In our case, LLLA was highly effective at mitigating overconfidence and highlighting spatial uncertainty in the extrapolation domain, especially in under-sampled areas, confirming its value as a robust, scalable, and lightweight UQ tool for DSM applications.

5.4 Conclusion

This study explored the use of ANNs for extrapolation tasks in DSM in under-sampled regions and proposed a novel UQ approach using LLLA. The uneven distribution of soil samples limits the reliability of models when extrapolating to new areas. Our research addressed this issue by training an ANN on soil data from a reference area and applying it to a similar but unsampled target area. The results showed that while the ANN could recognize familiar soil patterns, it often produced overconfident predictions, particularly in regions outside the training domain. By applying the LLLA method, we successfully reduced the overconfidence of the ANN and generated spatial uncertainty estimates. This approach provided more realistic confidence values and identified regions where the model's predictions were less reliable. Importantly, LLLA was particularly effective in detecting areas with unfamiliar soil units, reducing confidence in those regions and highlighting the need for further data collection. Our findings underline the importance of UQ in DSM, particularly when using ML models in spatially diverse landscapes. While ANNs excel in recognizing patterns and extrapolating soil units, their inherent "black-box" nature and tendency for overconfidence pose significant risks when models are deployed in new areas. The LLLA method offers a practical, computationally lightweight solution to address these issues, making it a valuable tool for improving the reliability of soil predictions. Future work should focus on improving the balance and representativeness of training datasets to enhance the accuracy of uncertainty estimates. Integrating spatial uncertainty maps into sampling strategies can further optimize data collection by directing limited resources to regions of high model uncertainty. Additionally, research should examine how the LLLA responds to established strategies for improving model transferability through the targeted addition of samples, as shown for example by Broeg et al. [28]. In particular, evaluating how LLLA-based uncertainty estimates evolve with such sample augmentation under transfer learning conditions is essential. To further assess the generalisability of LLLA, systematic benchmarking on diverse datasets is necessary. A valuable foundation for this purpose offers for example the LimeSoDa dataset collection, with its broad range of environmental conditions and standardized DSM features [157]. In conclusion, our research demonstrates that combining ANNs with post-hoc Bayesian UQ techniques can significantly enhance the interpretability, reliability, and transferability of DSM models. This advancement is essential for making ML models more robust and trustworthy in practical applications, particularly in regions with sparse or uneven soil data.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645 and the Tübingen AI Center (FKZ 01IS18039A). We also thank the Department 9: State Authority for Geology, Mineral Resources and Mining (LGRB), Freiburg, Germany, for providing soil data.

Part III.

Conclusion & Future Directions

6. Conclusion & Future Directions

This thesis evaluated the capability of Artificial Neural Networks (ANNs) for Digital Soil Mapping (DSM) in terms of three core aspects: uncertainty quantification (UQ), spatial generalisation, and the practical use of spatial uncertainty in spatial soil predictions. Central to this research was the application of the Last-Layer Laplace Approximation (LLLA), a Bayesian approach that allows post-hoc computationally efficient uncertainty estimation in ANNs. Using two case studies based on heterogeneous soil landscapes in southern Germany, this research has resulted in a number of important findings that are promising and have implications for DSM, environmental modelling and spatial data analysis.

(RQ1) *How effectively can the LLLA quantify and correct the uncertainty in ANNs predictions for soil mapping?*

The LLLA proves to be an effective and computationally efficient method for estimating uncertainty in ANN-based soil predictions in both studies. While standard ANNs are powerful, they tend to be overconfident, especially in regions with little or no training data. They can predict soil types with 100% confidence in certain locations, even if the true soil unit is not included in the input data or is contaminated by noisy data. This overconfidence can mislead users and decision-makers and lead to unreliable soil maps. LLLA solves this problem by using a Bayesian approach applied only to the last layer of the ANN, which makes the computation very fast and efficient. It estimates a probability distribution over the model's final weights, which allows the model to express how uncertain it is about each prediction. This leads to uncertainty maps that show, pixel by pixel, where the model is confident and where it's not, including the epistemic uncertainty due to sparse data or extrapolation. Furthermore, LLLA is a post-hoc method, meaning it can be applied to already trained ANNs without requiring retraining. In the two contributions, we showed that using LLLA adds uncertainty to the softmax outputs and thus avoids false confidence in the predictions. It reveals high uncertainties in regions that were not covered during training. At the same time, the overall accuracy of the model is maintained. Thus, LLLA is effective at making ANN outputs more trustworthy without significantly increasing computational cost in the case of soil unit prediction. However, there are limitations. LLLA assumes a Gaussian posterior which may not accurately reflect the true distribution of weights, especially in highly non-convex landscapes like those seen in deep learning. Finally, while LLLA improves uncertainty estimation, it does not take into account any uncertainty in the input data, which remains unmodeled in this approach. Overall, though, LLLA represents a highly promising method for enhancing the trustworthiness and communication of ANN-based soil predictions.

(RQ2) *To what extent can Artificial Neural Networks (ANNs) trained on well-sampled regions generalise to soil predictions on geographically similar but distant, data-sparse regions via model transfer?*

The studies show that ANNs can generalise to new, similar regions, but with limitations. An important strength was that ANNs were able to recognise certain patterns, especially in the context of colluvisols soils and river systems and their floodplain soils, even when the specific soil units were not included in the training set, the prediction remained in the soil family. As expected, the predictions were more accurate in areas that were very similar to the training area in terms of environmental characteristics. Nevertheless, the results also revealed significant limitations. The ANN often selected overrepresented soil units, indicating a bias towards predictions in favour of dominant soil units. This highlights a fundamental weakness of ANNs in dealing with imbalanced classes and limited representation. This could be recognised by the uncertainty statement developed with the LLLA, but also shows the need for more balanced training datasets, better domain adaptation techniques and hybrid modelling strategies. In summary, the transfer of ANN models works moderately well for similar regions, but additional quantification of uncertainty is crucial to make these predictions useful and trustworthy.

(RQ3) *What role does spatial uncertainty play in the reliability and practical application of digital soil maps?*

Spatial uncertainty is very important for making digital soil maps more reliable and useful in practice, as shown in our research. When we clearly show where the model is less certain, users can better understand where predictions might not be accurate. This is especially helpful in real-world applications like farming, building projects, or environmental planning, where using uncertain data without caution could lead to poor decisions. Knowing where uncertainty is high also helps make future soil sampling more effective [168]. Instead of collecting new data randomly, we can focus on the areas where the model is unsure. This saves time and money, and improves the quality of the map. In addition, spatial uncertainty helps us understand the limits of the model. For example, it can reveal if the model is biased toward more common soil types, if it struggles to predict soils in certain landscapes, or if it cannot be reliably transferred to areas outside the training region. This information is useful for improving both the training data and the model itself. Overall, including spatial uncertainty is essential for ML in soil mapping. It helps make predictions more transparent, supports better decision-making, and guides improvements in data collection and model design, especially when the model is used in new or different areas.

6.1 Impact & Implications

Advancing Digital Soil Mapping Through Uncertainty Quantification

The integration of UQ into DSM, already presented as one of the challenges in Wadoux et al. [182], is a significant advancement in the field of geospatial environmental modeling. Conventional soil maps, whether derived from field surveys or DSM maps, often do not provide accurate confidence in their results. This gap can lead to overconfidence in decision-making processes, especially in under-sampled or extrapolated regions. By embedding spatially explicit uncertainty metrics into ANN-derived soil maps, this research enhances the interpretability, reliability, and ultimately the usability of DSM products.

Practical Relevance for Stakeholders

The studies emphasize the real-world applicability of uncertainty-aware models for a wide range of stakeholders, including land-use planners, agricultural managers, geologists, and policy makers. For example, construction companies and environmental agencies require high-confidence assessments of subsoil conditions, while precision agriculture depends on accurate soil data to optimize inputs and yields. By identifying areas of high model uncertainty, these methods allow stakeholders to prioritize targeted field sampling, especially in poorly characterized or extrapolated zones. That could lead to allocate resources more efficiently, improving the cost-benefit ratio of field campaigns. With uncertainty maps they can interpret predictions with appropriate caution.

Scientific Contributions and Methodological Advancements

The combination of ANNs with LLLA addresses several challenges in DSM. One is the overconfidence mitigation. ANNs are known for making predictions with unwarranted certainty, particularly outside their training domain. This research effectively diagnoses and corrects this error and promotes more robust model behaviour. By deliberately testing model performance in areas absent from the training data, the studies contribute critical insight into the limits of model generalisability, a common yet under-addressed challenge in spatial modeling. The most practical in terms of programming is the post-hoc uncertainty integration, which only requires access to the weights of the final linear layer. The proposed framework enables retrofitting existing ANNs with uncertainty estimates, making it practical and scalable for legacy systems or operational models where retraining is infeasible. Additionally the interpretability comes without sacrificing performance. Importantly, these methods enhance model interpretability without degrading accuracy. In fact, the results highlight that even average-performing models benefit meaningfully from the added uncertainty layer, particularly when accuracy alone is insufficient to guide decisions.

Broader Implications for Environmental Modeling

Beyond the specific domain of soil science, the implications of this research could be extended to other geospatial modeling applications. The methodological framework laid out here, based on a simple ANN and classification task, combining ML with probabilistic reasoning, aligns with broader trends toward interpretable, trustworthy artificial intelligence in environmental science. Moreover, these findings underline the ethical and practical importance of quantifying uncertainty in spatial models. In an era increasingly reliant on automated predictions to inform land management and policy decisions, acknowledging what we do not know is as critical as what we claim to know.

6.2 Future Work

Extension to Continuous Soil Properties

So far, our work has focused on classifying soil units. However, in agriculture and environmental science, stakehoders are often more interested in continuous soil properties, like how much soil organic carbon is present, the soil's acidity, or its cation exchange capacity. These values are very important for understanding soil health and deciding how to manage land or grow crops. Future work should extend the UQ with LLLA to ANNs predicting these continuous values instead of just classifying categories. To do this, we will need to adjust the ANN, so that it can handle continuous outputs. The LLLA method we used is also well-suited for continuous data, so we can still use it without major changes.

Integration of Real Sampled Soil Data

So far, our models rely on existing soil maps and raster data from official sources. While this provides a useful starting point, future research should also include new soil data collected directly from the field. This would allow us to compare modern DSM methods together with statistical sampling strategies like conditioned Latin Hypercube Sampling (cLHS) with conventional soil sampling and mapping techniques. By doing so, we can better understand the strengths and weaknesses of each approach, and identify which methods work best in different situations.

Closed-Loop Feedback Between Model and Field Data

To improve the accuracy and usefulness of soil predictions, future work should explore a closed-loop feedback system between model outputs and field data. In this approach, predictions made by the model would guide where to collect new soil samples in the field. Then, the new data would be fed back into the model to improve its accuracy. This feedback loop can help identify areas where the model is uncertain or likely to be wrong, allowing for smarter, more efficient sampling. It also helps the model learn from its mistakes, leading to continuous improvement over time. By combining field observations and model predictions in an ongoing cycle, we can create a more adaptive and reliable system for soil mapping and monitoring.

Benchmarking the LLLA

To better understand the performance and potential of the LLLA, future work should include a detailed benchmarking study. This involves comparing LLLA with other commonly used uncertainty methods in DSM, such as MC Dropout, Deep Ensembles, and other Bayesian approaches, using the same tasks and datasets. To carry out such comparisons, well-curated benchmarking datasets are essential. Efforts like the LimeSoda project by Schmidinger et al. [158] are already underway to provide suitable datasets for this purpose. By evaluating performance in terms of accuracy, uncertainty estimation, computation time, and interpretability, we can assess deeper insights in the strengths and limitations of LLLA beyond my contributions. Benchmarking will also help determine where LLLA offers practical advantages. These comparisons are crucial for building confidence in LLLA and for informing its effective use in DSM applications.

Bibliography

[1] D. J. C. M. "Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks". *Network: Computation in Neural Systems* 6.3 (1995), page 469. eprint: https://dx.doi.org/10.1088/0954-898X/6/3/011.

[2] F. Abbaszadeh Afshar, S. Ayoubi, and A. Jafari. "The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran". *Geoderma* 315 (2018), pages 36–48. eprint: https://doi.org/10.1016/j.geoderma.2017.11.030.

[3] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". *Information Fusion* 76 (2021), pages 243–297. eprint: https://doi.org/10.1016/j.inffus.2021.05.008.

[4] K. Adhikari, B. Minasny, M. B. Greve, and M. H. Greve. "Constructing a soil class map of Denmark based on the FAO legend using digital techniques". *Geoderma* 214 (2014), pages 101–113. eprint: https://doi.org/10.1016/j.geoderma.2013.09.023.

[5] M. Aitkenhead and M. Coull. "Mapping soil carbon stocks across Scotland using a neural network model". *Geoderma* 262 (2016), pages 187–198. eprint: https://doi.org/10.1016/j.geoderma.2015.08.034.

[6] C. Albrecht, R. Jahn, and B. Huwe. "Bodensystematik und Bodenklassifikation Teil I: Grundbegriffe". *Journal of Plant Nutrition and Soil Science* 168.1 (2005), pages 7–20. eprint: https://doi.org/10.1002/jpln.200421475.

[7] W. Amelung, H.-P. Blume, H. Fleige, R. Horn, E. Kandeler, I. Kögel-Knabner, R. Kretzschmar, K. Stahr, and B.-M. Wilke. "Scheffer/Schachtschabel Lehrbuch der Bodenkunde". *Springer-Verlag* (2018). eprint: https://doi.org/10.1007/978-3-662-55871-3.

[8] G. P. de Arruda, J. A. M. Dematte, C. d. S. Chagas, P. R. Fiorio, A. B. e. Souza, and C. T. Fongaro. "Digital soil mapping using reference area and artificial neural networks". *SCIENTIA AGRICOLA* 73.3 (2016), pages 266–273. eprint: http://dx.doi.org/10.1590/0103-9016-2015-0131.

[9] M. Bagheri Bodaghabadi, J. Martínez-Casasnovas, M. H. Salehi, J. Mohammadi, I. Esfandiarpoor Borujeni, N. Toomanian, and A. Gandomkar. "Digital Soil Mapping Using Artificial Neural Networks and Terrain-Related Attributes". *Pedosphere* 25.4 (2015), pages 580–591. eprint: https://doi.org/10.1016/S1002-0160(15)30038-2.

[10] Y. Bao, F. Yao, X. Meng, J. Wang, H. Liu, Y. Wang, Q. Liu, J. Zhang, and A. M. Mouazen. "A fine digital soil mapping by integrating remote sensing-based process model and deep learning method in Northeast China". *Soil and Tillage Research* 238 (2024), page 106010. eprint: https://doi.org/10.1016/j.still.2024.106010.

[11] R. Bardenet, A. Doucet, and C. Holmes. "On Markov chain Monte Carlo methods for tall data". *Journal of Machine Learning Research* 18.47 (2017), pages 1–43. eprint: http://jmlr.org/papers/v18/15-205.html.

[12] J. Bauer, H. Rohdenburg, and H. Bork. "Ein digitales Reliefmodell als Vorraussetzung für ein deterministisches Modell der Wasser-und Stoff-Flüsse". *Landschaftsgenese und Landschaftsökologie* 10.1 (1985), page 15.

[13] T. Behrens, H. Förster, T. Scholten, U. Steinrücken, E.-D. Spies, and M. Goldschmitt. "Digital soil mapping using artificial neural networks". *Journal of plant nutrition and soil science* 168.1 (2005), pages 21–33. eprint: https://doi.org/10.1002/jpln.200421414.

[14] T. Behrens and T. Scholten. "Digital soil mapping in Germany—a review". *Journal of Plant Nutrition and Soil Science* 169.3 (2006), pages 434–443. eprint: https://doi.org/10.1002/jpln.200521962.

[15] T. Behrens, A.-X. Zhu, K. Schmidt, and T. Scholten. “Multi-scale digital terrain analysis and feature selection for digital soil mapping”. *Geoderma* 155.3-4 (2010), pages 175–185. eprint: https://doi.org/10.1016/j.geoderma.2009.07.010.

[16] W. H. Belkadi and Y. Drias. “Advancements in Digital Soil Mapping: From Data Acquisition to Uncertainty Estimation - A Comprehensive Review”. *Artificial Intelligence Doctoral Symposium*. Edited by H. Drias, F. Yalaoui, and A. Hadjali. Singapore: *Springer Nature Singapore*, 2023, pages 162–177. eprint: https://doi.org/10.1007/978-981-99-4484-2_13.

[17] Y. Bengio. “Learning Deep Architectures for AI”. *Foundations and Trends in Machine Learning* 2.1 (2009), pages 1–127. eprint: http://dx.doi.org/10.1561/2200000006.

[18] J. Bergstra and Y. Bengio. “Random Search for Hyper-Parameter Optimization”. *Journal of Machine Learning Research* 13.10 (2012), pages 281–305. eprint: http://jmlr.org/papers/v13/bergstra12a.html.

[19] C. M. Bishop. “Pattern Recognition and Machine Learning (Information Science and Statistics)”. *Springer-Verlag* (2006). ISBN/EAN: 9781493938438.

[20] K. Bleich, F. Hädrich, P. Hummel, S. Müller, D. Ortlam, and J. Werner. “Paläoböden in Baden-Württemberg”. *Geologisches Jahrbuch. Reihe F, Bodenkunde* 14 (1982). ISBN: 9783510960170, pages 61–99.

[21] M. Bock, J. Böhner, O. Conrad, R. Köthe, and A. Ringeler. “XV. Methods for creating functional soil databases and applying digital soil mapping with SAGA GIS”. *JRC Scientific and technical Reports, Office for Official Publications of the European Communities, Luxemburg* (2007).

[22] M. P. Bohn and B. A. Miller. “Locally enhanced digital soil mapping in support of a bottom-up approach is more accurate than conventional soil mapping and top-down digital soil mapping”. *Geoderma* 442 (2024), page 116781. eprint: https://doi.org/10.1016/j.geoderma.2024.116781.

[23] J. Böhner and R. Köthe. “Bodenregionalisierung und Prozessmodellierung: Instrumente für den Bodenschutz”. *Petermanns Geographische Mitteilungen* 147.3 (2003). ISSN: 0031-6229, pages 72–82.

[24] L. Boruvka and V. Penizek. “Chapter 30 A Test of an Artificial Neural Network Allocation Procedure using the Czech Soil Survey of Agricultural Land Data”. *Digital Soil Mapping*. Edited by P. Lagacherie, A. McBratney, and M. Voltz. Volume 31. Developments in Soil Science. *Elsevier*, 2006, pages 415–424. eprint: https://doi.org/10.1016/S0166-2481(06)31030-6.

[25] L. Breiman. “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. *Statistical science* 16.3 (2001), pages 199–231. eprint: http://dx.doi.org/10.1214/ss/1009213726.

[26] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. “Classification and regression trees”. *Routledge* (2017). eprint: https://doi.org/10.1201/9781315139470.

[27] J. S. Bridle. “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition”. *Neurocomputing*. Edited by F. F. Soulié and J. Hérault. Berlin, Heidelberg: *Springer Berlin Heidelberg*, 1990, pages 227–236. eprint: https://doi.org/10.1007/978-3-642-76153-9_28.

[28] T. Broeg, M. Blaschek, S. Seitz, R. Taghizadeh-Mehrjardi, S. Zepp, and T. Scholten. “Transferability of Covariates to Predict Soil Organic Carbon in Cropland Soils”. *Remote Sensing* 15.4 (2023). eprint: https://doi.org/10.3390/rs15040876.

[29] C. W. Brungard, J. L. Boettinger, M. C. Duniway, S. A. Wills, and T. C. Edwards. “Machine learning for predicting soil classes in three semi-arid landscapes”. *Geoderma* 239-240 (2015), pages 68–83. eprint: https://doi.org/10.1016/j.geoderma.2014.09.019.

[30] E. N. Bui and C. J. Moran. “A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia”. *Geoderma* 111.1-2 (2003), pages 21–44. eprint: https://doi.org/10.1016/S0016-7061(02)00238-0.

[31] B. J. Carter and E. J. Ciolkosz. “Slope gradient and aspect effects on soils developed from sandstone in Pennsylvania”. *Geoderma* 49.3 (1991), pages 199–213. eprint: https://doi.org/10.1016/0016-7061(91)90076-6.

[32] M. E. Carvalho Monteiro, F. P. Avalos, M. H. Procópio Pelegrino, R. Brito Vilela, F. Weimar Acerbi Júnior, I. T. Bueno, N. Li, S. H. Godinho Silva, E. Giasson, N. Curi, and M. Duarte de Menezes. "Digital mapping of soil classes in Southeast Brazil: environmental covariate selection, accuracy, and uncertainty". *Journal of South American Earth Sciences* 132 (2023), page 104640. eprint: https://doi.org/10.1016/j.jsames.2023.104640.

[33] T. Charnock, L. Perreault-Levasseur, and F. Lanusse. "Bayesian Neural Networks". *Artificial Intelligence for High Energy Physics* (2022), pages 663–713. eprint: https://doi.org/10.1142/9789811234033_0018.

[34] S. Chen, D. Arrouays, V. Leatitia Mulder, L. Poggio, B. Minasny, P. Roudier, Z. Libohova, P. Lagacherie, Z. Shi, J. Hannam, J. Meersmans, A. C. R. de Forges, and C. Walter. "Digital mapping of GlobalSoilMap soil properties at a broad scale: A review". *Geoderma* 409 (2022), page 115567. eprint: https://doi.org/10.1016/j.geoderma.2021.115567.

[35] S. Chen, V. L. Mulder, M. P. Martin, C. Walter, M. Lacoste, A. C. R. de Forges, N. P. Saby, T. Loiseau, B. Hu, and D. Arrouays. "Probability mapping of soil thickness by random survival forest at a national scale". *Geoderma* 344 (2019), pages 184–194. eprint: https://doi.org/10.1016/j.geoderma.2019.03.016.

[36] S. Chib. "Chapter 57 - Markov Chain Monte Carlo Methods: Computation and Inference". *Elsevier*. Handbook of Econometrics 5 (2001). Edited by J. J. Heckman and E. Leamer, pages 3569–3649. eprint: https://doi.org/10.1016/S1573-4412(01)05010-3.

[37] A. Cialella, R. Dubayah, W Lawrence, and E Levine. "Predicting soil drainage class using remotely sensed and digital elevation data". *Photogrammetric Engineering and Remote Sensing* 63.2 (1997), pages 171–177. eprint: https://www.asprs.org/wp-content/uploads/pers/1997journal/feb/1997_feb_171-178.pdf.

[38] F. F. Coelho, E. Giasson, A. R. Campos, R. G. P. de Oliveira e Silva, and J. J. F. Costa. "Geographic object-based image analysis and artificial neural networks for digital soil mapping". *CATENA* 206 (2021), page 105568. eprint: https://doi.org/10.1016/j.catena.2021.105568.

[39] O. Conrad, B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, J. Wehberg, V. Wichmann, and J. Böhner. "System for automated geoscientific analyses (SAGA) v. 2.1. 4". *Geoscientific Model Development* 8.7 (2015), pages 1991–2007. eprint: https://doi.org/10.5194/gmd-8-1991-2015.

[40] C. Cortes and V. Vapnik. "Support-vector networks". *Machine learning* 20.3 (1995), pages 273–297. eprint: https://doi.org/10.1007/BF00994018.

[41] T. Cover and P. Hart. "Nearest neighbor pattern classification". *IEEE Transactions on Information Theory* 13.1 (1967), pages 21–27. eprint: 10.1109/TIT.1967.1053964.

[42] A. Damianou and N. D. Lawrence. "Deep gaussian processes" (2013), pages 207–215. eprint: https://doi.org/10.48550/arXiv.1211.0358.

[43] P. K. Dash, N. Panigrahi, and A. Mishra. "Identifying opportunities to improve digital soil mapping in India: A systematic review". *Geoderma Regional* 28 (2022). eprint: https://doi.org/10.1016/j.geodrs.2021.e00478.

[44] E. A. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. "Laplace Redux - Effortless Bayesian Deep Learning". *CoRR* abs/2106.14806 (2021). eprint: https://doi.org/10.48550/arXiv.2106.14806.

[45] S. Dharumarajan, R. Hegde, N. Janani, and S. Singh. "The need for digital soil mapping in India". *Geoderma Regional* 16 (2019), e00204. eprint: https://doi.org/10.1016/j.geodrs.2019.e00204.

[46] V. V. Dokuchaev. "The Russian Chernozem: Report to the Free Economic Society". *Free Economic Society* (1883). Original publication in Russian.

[47] J. S. Dramsch. "Chapter One - 70 years of machine learning in geoscience in review". *Machine Learning in Geosciences*. Advances in Geophysics 61 (2020). Edited by B. Moseley and L. Krischer, pages 1–55. eprint: https://doi.org/10.1016/bs.agph.2020.08.002.

[48] W. Eckelmann, H Sponagel, W Grottenthaler, et al. "Bodenkundliche Kartieranleitung.-5. verbesserte und erweiterte-Auflage". *Schweizerbart'sche Verlagsbuchhandlung* (2005). ISBN: 978-3-510-95920-4.

[49] E. Fix and J. L. Hodges. "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties". *International Statistical Review / Revue Internationale de Statistique* 57.3 (1989), pages 238–247. eprint: https://doi.org/10.2307/1403797..

[50] A. C. Richer-de Forges, D. Arrouays, M. Bardy, A. Bispo, P. Lagacherie, B. Laroche, B. Lemercier, J. Sauter, and M. Voltz. "Mapping of Soils and Land-Related Environmental Attributes in France: Analysis of End-Users' Needs". *Sustainability* 11.10 (2019). eprint: https://doi.org/10.3390/su11102940.

[51] T. G. Freeman. "Calculating catchment area with divergent flow based on a regular grid". *Computers & geosciences* 17.3 (1991), pages 413–422. eprint: https://doi.org/10.1016/0098-3004(91)90048-I.

[52] J. H. Friedman. "Greedy function approximation: a gradient boosting machine". *Annals of statistics* (2001), pages 1189–1232. eprint: http://dx.doi.org/10.1214/aos/1013203451.

[53] K. Fukushima. "Visual feature extraction by a multilayered network of analog threshold elements". *IEEE Transactions on Systems Science and Cybernetics* 5.4 (1969), pages 322–333. eprint: https://doi.org/10.1109/TSSC.1969.300225.

[54] Y. Gal et al. "Uncertainty in deep learning". *University of Cambridge* (2016). eprint: https://www.cs.ox.ac.uk/people/yarin.gal/website/blog_2248.html.

[55] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan. "Ensemble deep learning: A review". *Engineering Applications of Artificial Intelligence* 115 (2022), page 105151. eprint: https://doi.org/10.1016/j.engappai.2022.105151.

[56] R. Garnett. "Bayesian Optimization". *Cambridge University Press* (2023). ISBN: 978-1108425780.

[57] F. Gascon, C. Bouzinac, O. Thépaut, M. Jung, B. Francesconi, J. Louis, V. Lonjou, B. Lafrance, S. Massera, A. Gaudel-Vacaresse, et al. "Copernicus Sentinel-2A calibration and products validation status". *Remote Sensing* 9.6 (2017), page 584. eprint: https://doi.org/10.3390/rs9060584.

[58] W. R. Gilks, S. Richardson, and D. Spiegelhalter. "Markov chain Monte Carlo in practice". *CRC press* (1995). eprint: https://doi.org/10.1201/b14835.

[59] X. Glorot, A. Bordes, and Y. Bengio. "Deep sparse rectifier neural networks". *JMLR Workshop and Conference Proceedings* (2011), pages 315–323. eprint: https://proceedings.mlr.press/v15/glorot11a.html.

[60] L. C. Gomes, R. M. Faria, E. de Souza, G. V. Veloso, C. E. G. Schaefer, and E. I. F. Filho. "Modelling and mapping soil organic carbon stocks in Brazil". *Geoderma* 340 (2019), pages 337–350. eprint: https://doi.org/10.1016/j.geoderma.2019.01.007.

[61] I. Goodfellow, Y. Bengio, and A. Courville. "Deep Learning". *MIT Press* (2016). eprint: http://www.deeplearningbook.org.

[62] P. Goovaerts. "Geostatistics in soil science: state-of-the-art and perspectives". *Geoderma* 89.1 (1999), pages 1–45. eprint: https://doi.org/10.1016/S0016-7061(98)00078-0.

[63] R. Grimm, T. Behrens, M. Märker, and H. Elsenbeer. "Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis". *Geoderma* 146.1-2 (2008), pages 102–113. eprint: https://doi.org/10.1016/j.geoderma.2008.05.008.

[64] C. Grinand, D. Arrouays, B. Laroche, and M. P. Martin. "Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context". *Geoderma* 143.1-2 (2008), pages 180–190. eprint: http://dx.doi.org/10.1016/j.geoderma.2007.11.004.

[65] S. Grunwald. "Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils". *Carbon Footprints* 1.1 (2022). eprint: https://doi.org/10.20517/cf.2022.03.

[66] S. Grunwald, J. Thompson, and J. Boettinger. "Digital soil mapping and modeling at continental scales: Finding solutions for global issues". *Soil Science Society of America Journal* 75.4 (2011), pages 1201–1213. eprint: http://dx.doi.org/10.2136/sssaj2011.0025.

[67] M. Guevara, G. F. Olmedo, E. Stell, Y. Yigini, Y. Aguilar Duarte, C. Arellano Hernández, G. E. Arévalo, C. E. Arroyo-Cruz, A. Bolivar, S. Bunning, N. Bustamante Cañas, C. O. Cruz-Gaistardo, F. Davila, M. Dell Acqua, A. Encina, H. Figueredo Tacona, F. Fontes, J. A. Hernández Herrera, A. R. Ibelles Navarro, V. Loayza, A. M. Manueles, F. Mendoza Jara, C. Olivera, R. Osorio Hermosilla, G. Pereira, P. Prieto, I. A. Ramos, J. C. Rey Brina, R. Rivera, J. Rodríguez-Rodríguez, R. Roopnarine, A. Rosales Ibarra, K. A. Rosales Riveiro, G. A. Schulz, A. Spence, G. M. Vasques, R. R. Vargas, and R. Vargas. "No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America". *SOIL* 4.3 (2018), pages 173–193. eprint: https://doi.org/10.5194/soil-4-173-2018.

[68] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On Calibration of Modern Neural Networks". *CoRR* abs/1706.04599 (2017). eprint: https://doi.org/10.48550/arXiv.1706.04599.

[69] P. Guo, M. Lyu, and C. Chen. "Regularization parameter estimation for feedforward neural networks". *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 33.1 (2003), pages 35–44. eprint: https://doi.org/10.1109/TSMCB.2003.808176.

[70] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung. "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit". *nature* 405.6789 (2000), pages 947–951. eprint: https://doi.org/10.1038/35016072.

[71] N. Hamzehpour, H. Shafizadeh-Moghadam, and R. Valavi. "Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture". *CATENA* 182 (2019), page 104141. eprint: https://doi.org/10.1016/j.catena.2019.104141.

[72] J. Han, K. Mao, T. Xu, J. Guo, Z. Zuo, and C. Gao. "A soil moisture estimation framework based on the CART algorithm and its application in China". *Journal of hydrology* 563 (2018), pages 65–75. eprint: https://doi.org/10.1016/j.jhydrol.2018.05.051.

[73] A. E. Hartemink and J. Bockheim. "Soil genesis and classification". *CATENA* 104 (2013), pages 251–256. eprint: https://doi.org/10.1016/j.catena.2012.12.001.

[74] A. E. Hartemink, J. Hempel, P. Lagacherie, A. McBratney, N. McKenzie, R. A. MacMillan, B. Minasny, L. Montanarella, M. L. de Mendonça Santos, P. Sanchez, M. Walsh, and G.-L. Zhang. "GlobalSoilMap.net – A New Digital Soil Map of the World". *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. Edited by J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, and S. Kienast-Brown. Dordrecht: *Springer Netherlands*, 2010, pages 423–428. eprint: https://doi.org/10.1007/978-90-481-8863-5_33.

[75] S. Haykin. "Neural Networks: A Comprehensive Foundation". *Prentice Hall PTR* (1994). ISBN: 978-0023527616.

[76] R. G. Heerdegen and M. A. Beran. "Quantifying source areas through land surface curvature and shape". *Journal of Hydrology* 57.3-4 (1982), pages 359–373. eprint: https://doi.org/10.1016/0022-1694(82)90155-X.

[77] M. Hein, M. Andriushchenko, and J. Bitterwolf. "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pages 41–50. eprint: https://doi.org/10.1109/CVPR.2019.00013.

[78] B. Henderson, E. Bui, C. Moran, and D. Simon. "Australia-wide predictions of soil properties using decision trees". *Geoderma* 124.3 (2005), pages 383–398. eprint: https://doi.org/10.1016/j.geoderma.2004.06.007.

[79] T. Hengl, J. Mendes de Jesus, G. B. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, et al. "SoilGrids250m: Global gridded soil information based on machine learning". *PLoS one* 12.2 (2017), e0169748. eprint: https://doi.org/10.1371/journal.pone.0169748.

[80] B. Heung, H. C. Ho, J. Zhang, A. Knudby, C. E. Bulmer, and M. G. Schmidt. "An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping". *Geoderma* 265 (2016), pages 62–77. eprint: https://doi.org/10.1016/j.geoderma.2015.11.014.

[81] G. Heuvelink and R. Webster. "Uncertainty assessment of spatial soil information". *Encyclopedia of Soils in the Environment* 4 (2023), pages 671–683. eprint: https://doi.org/10.1016/B978-0-12-822974-3.00174-9.

[82] G. Heuvelink and R Webster. "Modelling soil variation: past, present, and future". *Geoderma* 100.3 (2001). Developments and Trends in Soil Science, pages 269–301. eprint: https://doi.org/10.1016/S0016-7061(01)00025-8.

[83] A. Hewitt. "Predictive modelling in soil survey". *Soils and Fertilizers* 56.3 (1993), pages 305–314.

[84] S. Hochreiter and J. Schmidhuber. "Long short-term memory". *Neural computation* 9.8 (1997), pages 1735–1780. eprint: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[85] K. O. Hounkpatin, K. Schmidt, F. Stumpf, G. Forkuor, T. Behrens, T. Scholten, W. Amelung, and G. Welp. "Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso)". *Sci rep* 8.1 (2018), pages 1–16. eprint: https://doi.org/10.1038/s41598-018-28244-w.

[86] A. S. Householder. "A theory of steady-state activity in nerve-fiber networks: I. Definitions and preliminary lemmas". *The bulletin of mathematical biophysics* 3 (1941), pages 63–69. eprint: https://doi.org/10.1007/BF02478220.

[87] Y.-C. Huang, J. Padarian, B. Minasny, and A. B. McBratney. "Using Monte Carlo conformal prediction to evaluate the uncertainty of deep learning soil spectral models". *EGUsphere* 2025 (2025), pages 1–15. eprint: https://doi.org/10.5194/egusphere-2024-3703.

[88] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". *CoRR* abs/1502.03167 (2015). eprint: https://doi.org/10.48550/arXiv.1502.03167.

[89] IUSS Working Group. "World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps, 4th edition." English. Edited by P. Schad, L. Anjos, J. Boixadera Llobet, S. Deckers, S. Dondeyne, E. Eberhardt, M. Gerasimova, B. Harms, C. Kabala, S. Mantel, K. Stahr, and C. van Huyssteen. *International Union of Soil Sciences (IUSS)*, (2022). eprint: https://www.isric.org/sites/default/files/WRB_fourth_edition_2022-12-18.pdf.

[90] H. Jenny. "Factors of soil formation. 281 pp". *New York* 801 (1941).

[91] J. M. Johnson and T. M. Khoshgoftaar. "Survey on deep learning with class imbalance". *Journal of Big Data* 6.1 (2019), pages 1–54. eprint: https://doi.org/10.1186/s40537-019-0192-5.

[92] K. Kasiviswanathan, K. Sudheer, and J. He. "Probabilistic and ensemble simulation approaches for input uncertainty quantification of artificial neural network hydrological models". *Hydrological Sciences Journal* 63.1 (2018), pages 101–113. eprint: https://doi.org/10.1080/02626667.2017.1393686.

[93] N. M. Kebonye, P. C. Agyeman, and J. K. Biney. "Optimized modelling of countrywide soil organic carbon levels via an interpretable decision tree". *Smart Agricultural Technology* 3 (2023), page 100106. eprint: https://doi.org/10.1016/j.atech.2022.100106.

[94] 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). eprint: https://doi.org/10.48550/arXiv.1703.04977.

[95] H. Keskin, S. Grunwald, and W. G. Harris. "Digital mapping of soil carbon fractions with machine learning". *Geoderma* 339 (2019), pages 40–58. eprint: https://doi.org/10.1016/j.geoderma.2018.12.037.

[96] Y. Khaledian and B. A. Miller. "Selecting appropriate machine learning methods for digital soil mapping". *Applied Mathematical Modelling* 81 (2020), pages 401–418. eprint: https://doi.org/10.1016/j.apm.2019.12.016.

[97] D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". *Proceedings of the International Conference on Learning Representations (ICLR)* (2015), pages 1–15. eprint: https://doi.org/10.48550/arXiv.1412.6980.

[98] R. Koenker and G. Bassett Jr. "Regression quantiles". *Econometrica: Journal of the Econometric Society* (1978), pages 33–50. eprint: https://doi.org/10.2307/1913643.

[99] B. Kopecky-Hermanns, R. Vogt, and S. Berg. "Kolluvien". *Geoarchäologie* (2022). Edited by C. Stolz and C. E. Miller, pages 207–216. eprint: https://doi.org/10.1007/978-3-662-62774-7_11.

[100] R. Köthe and M. Bock. "Development and use in practice of SAGA modules for high quality analysis of geodata". *FREE AND OPEN GIS-SAGA-GIS* 115 (2006). ISSN: 0341-3780, pages 85–96.

[101] J. Kotzé and J. van Tol. "Extrapolation of Digital Soil Mapping Approaches for Soil Organic Carbon Stock Predictions in an Afromontane Environment". *Land* 12.3 (2023). eprint: https://doi.org/10.3390/land12030520.

[102] M. Kovačević, B. Bajat, and B. Gajić. "Soil type classification and estimation of soil properties using support vector machines". *Geoderma* 154.3-4 (2010), pages 340–347. eprint: http://dx.doi.org/10.1016/j.geoderma.2009.11.005.

[103] M. Kramer and J. Leonard. "Diagnosis using backpropagation neural networks—analysis and criticism". *Computers and Chemical Engineering* 14.12 (1990), pages 1323–1338. eprint: https://doi.org/10.1016/0098-1354(90)80015-4.

[104] A. Kristiadi, M. Hein, and P. Hennig. "Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks". *Proceedings of Machine Learning Research* 119 (2020). Edited by H. D. III and A. Singh, pages 5436–5446. eprint: https://proceedings.mlr.press/v119/kristiadi20a.html.

[105] A. Krizhevsky, G. Hinton, et al. "Learning multiple layers of features from tiny images". *Toronto, ON, Canada* (2009). eprint: https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

[106] P. Lagacherie. "Digital Soil Mapping: A State of the Art". *Springer Netherlands* (2008). Edited by A. E. Hartemink, A. McBratney, and M. d. L. Mendonça-Santos, pages 3–14. eprint: https://doi.org/10.1007/978-1-4020-8592-5_1.

[107] P. Lagacherie, J. Legros, and P. Burfough. "A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area". *Geoderma* 65.3 (1995), pages 283–301. eprint: https://doi.org/10.1016/0016-7061(94)00040-H.

[108] P. Lagarcherie and S. Holmes. "Addressing geographical data errors in a classification tree for soil unit prediction". *International Journal of Geographical Information Science* 11.2 (1997), pages 183–198. eprint: https://doi.org/10.1080/136588197242455.

[109] M.-H. Laves, S. Ihler, K.-P. Kortmann, and T. Ortmaier. "Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference". *arXiv e-prints* (2019). eprint: https://doi.org/10.48550/arXiv.1909.13550.

[110] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. "Deep Neural Networks as Gaussian Processes". *arXiv eprints* (2018). eprint: https://doi.org/10.48550/arXiv.1711.00165.

[111] B. Lemercier, M. Lacoste, M. Loum, and C. Walter. "Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach". *Geoderma* 171-172 (2012), pages 75–84. eprint: https://doi.org/10.1016/j.geoderma.2011.03.010.

[112] Y. Li, S. Yan, and J. Gong. "Quantifying uncertainty in soil moisture retrieval using a Bayesian neural network framework". *Computers and Electronics in Agriculture* 215 (2023), page 108414. eprint: https://doi.org/10.1016/j.compag.2023.108414.

[113] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass, and F. Hutter. "SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization". *Journal of Machine Learning Research* 23.54 (2022), pages 1–9. eprint: http://jmlr.org/papers/v23/21-0888.html.

[114] W. Liu, X. Luo, F. Huang, and M. Fu. "Uncertainty of the Soil–Water Characteristic Curve and Its Effects on Slope Seepage and Stability Analysis under Conditions of Rainfall Using the Markov Chain Monte Carlo Method". *Water* 9.10 (2017). eprint: https://doi.org/10.3390/w9100758.

[115] I. Loshchilov and F. Hutter. "Fixing weight decay regularization in Adam". *Proceedings of the International Conference on Learning Representations (ICLR)* (2018), pages 1–13. eprint: https://arxiv.org/pdf/1711.05101v2/1000.

[116] M. Ludwig, A. Moreno-Martinez, N. Hölzel, E. Pebesma, and H. Meyer. "Assessing and improving the transferability of current global spatial prediction models". *Global Ecology and Biogeography* 32.3 (2023), pages 356–368. eprint: https://doi.org/10.1111/geb.13635.

[117] Y. Ma, B. Minasny, B. P. Malone, and A. B. Mcbratney. "Pedology and digital soil mapping (DSM)". *European Journal of Soil Science* 70.2 (2019), pages 216–235. eprint: https://doi.org/10.1111/ejss.12790.

[118] B. P. Malone, B. Minasny, N. P. Odgers, and A. B. McBratney. "Using model averaging to combine soil property rasters from legacy soil maps and from point data". *Geoderma* 232 (2014), pages 34–44. eprint: https://doi.org/10.1016/j.geoderma.2014.04.033.

[119] N. Mansuy, E. Thiffault, D. Paré, P. Bernier, L. Guindon, P. Villemaire, V. Poirier, and A. Beaudoin. "Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method". *Geoderma* 235-236 (2014), pages 59–73. eprint: https://doi.org/10.1016/j.geoderma.2014.06.032.

[120] A. K. Matazi, E. E. Gognet, and R. G. Kakaï. "Digital soil mapping: a predictive performance assessment of spatial linear regression, Bayesian and ML-based models". *Modeling Earth Systems and Environment* 10.1 (2024), pages 595–618. eprint: https://doi.org/10.1007/s40808-023-01788-1.

[121] A. McBratney, M. Mendonça Santos, and B. Minasny. "On Digital Soil Mapping". *Geoderma* 117 (2003), pages 3–52. eprint: https://doi.org/10.1016/S0016-7061(03)00223-4.

[122] A. B. McBratney. "On variation, uncertainty and informatics in environmental soil management". *Soil Research* 30.6 (1992), pages 913–935. eprint: https://doi.org/10.1071/SR9920913.

[123] W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". *The bulletin of mathematical biophysics* 5 (1943), pages 115–133. eprint: https://doi.org/10.1007/BF02478259.

[124] N. McKenzie and M. Austin. "A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation". *Geoderma* 57.4 (1993), pages 329–355. eprint: https://doi.org/10.1016/0016-7061(93)90049-Q.

[125] H. Meyer and E. Pebesma. "Machine learning-based global maps of ecological variables and the challenge of assessing them". *Nature Communications* 13.1 (2022), page 2208. eprint: https://doi.org/10.1038/s41467-022-29838-9.

[126] B. Minasny and A. McBratney. "Digital soil mapping: A brief history and some lessons". *Geoderma* 264 (2016). Soil mapping, classification, and modelling: history and future directions, pages 301–311. eprint: https://doi.org/10.1016/j.geoderma.2015.07.017.

[127] R. Mirzaeitalarposhti, H. Shafizadeh-Moghadam, R. Taghizadeh-Mehrjardi, and M. S. Demyan. "Digital Soil Texture Mapping and Spatial Transferability of Machine Learning Models Using Sentinel-1, Sentinel-2, and Terrain-Derived Covariates". *Remote Sensing* 14.23 (2022). eprint: https://doi.org/10.3390/rs14235909.

[128] I. D. Moore, P. E. Gessler, G. A. Nielsen, and G. A. Peterson. "Soil Attribute Prediction Using Terrain Analysis". *Soil Science Society of America Journal* 57.2 (1993), pages 443–452. eprint: https://doi.org/10.2136/sssaj1993.03615995005700020026x.

[129] P. Moraga, C. Dean, J. Inoue, P. Morawiecki, S. R. Noureen, and F. Wang. "Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: A case study predicting malaria risk in Mozambique". *Spatial and Spatio-temporal Epidemiology* 39 (2021), page 100440. eprint: https://doi.org/10.1016/j.sste.2021.100440.

[130] V. Nair and G. E. Hinton. "Rectified linear units improve restricted boltzmann machines". *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*. ICML'10 (2010), 807–814. eprint: https://icml.cc/2010/papers/432.pdf.

[131] R. M. Neal. "Bayesian learning for neural networks". *Springer Science & Business Media* 118 (2012). eprint: https://doi.org/10.1007/978-1-4612-0745-0.

[132] A. M. Nenkam, A. M. J.-C. Wadoux, B. Minasny, A. B. McBratney, P. C. S. Traore, G. N. Falconnier, and A. M. Whitbread. "Using homosoils for quantitative extrapolation of soil mapping models". *European Journal of Soil Science* 73.5 (2022), e13285. eprint: https://doi.org/10.1111/ejss.13285.

[133] A. M. Nenkam, A. M.-C. Wadoux, B. Minasny, F. B. Silatsa, M. Yemefack, S. U. Ugbaje, S. Akpa, G. V. Zijl, A. Bouasria, Y. Bouslihim, L. M. Chabala, A. Ali, and A. B. McBratney. "Applications and challenges of digital soil mapping in Africa". *Geoderma* 449 (2024), page 117007. eprint: https://doi.org/10.1016/j.geoderma.2024.117007.

[134] M. Neyestani, F. Sarmadian, A. Jafari, A. Keshavarzi, and A. Sharififar. "Digital mapping of soil classes using spatial extrapolation with imbalanced data". *Geoderma Regional* 26 (2021), e00422. eprint: https://doi.org/10.1016/j.geodrs.2021.e00422.

[135] A. Nguyen, J. Yosinski, and J. Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images". *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pages 427–436. eprint: https://doi.org/10.48550/arXiv.1412.1897.

[136] J. Nocedal and S. J. Wright. "Numerical optimization". *Springer* (1999). eprint: https://doi.org/10.1007/978-0-387-40065-5.

[137] M. Osat, A. Heidari, M. Karimian Eghbal, and S. Mahmoodi. "Impacts of topographic attributes on Soil Taxonomic Classes and weathering indices in a hilly landscape in Northern Iran". *Geoderma* 281 (2016), pages 90–101. eprint: https://doi.org/10.1016/j.geoderma.2016.06.020.

[138] J. Padarian, A. B. McBratney, and B. Minasny. "Game theory interpretation of digital soil mapping convolutional neural networks". *SOIL* 6.2 (2020), pages 389–397. eprint: https://doi.org/10.5194/soil-6-389-2020.

[139] J. Padarian, B. Minasny, and A. B. McBratney. "Using deep learning for digital soil mapping". *SOIL* 5.1 (2019), pages 79–89. eprint: https://doi.org/10.5194/soil-5-79-2019.

[140] J. Padarian, B. Minasny, and A. B. McBratney. "Machine learning and soil sciences: a review aided by machine learning tools". *SOIL* 6.1 (2020), pages 35–52. eprint: https://doi.org/10.5194/soil-6-35-2020.

[141] J. Padarian, B. Minasny, and A. McBratney. "Assessing the uncertainty of deep learning soil spectral models using Monte Carlo dropout". *Geoderma* 425 (2022), page 116063. eprint: https://doi.org/10.1016/j.geoderma.2022.116063.

[142] V. Penizek and L. Boruvka. "The Digital Terrain Model as a Tool for Improved Delineation of Alluvial Soils". *Digital Soil Mapping with Limited Data* (2008). Edited by A. E. Hartemink, A. McBratney, and M. d. L. Mendonça-Santos, pages 319–326. eprint: https://doi.org/10.1007/978-1-4020-8592-5_28.

[143] K. Piikki, J. Wetterlind, M. Söderström, and B. Stenberg. "Perspectives on validation in digital soil mapping of continuous attributes—A review". *Soil Use and Management* 37.1 (2021), pages 7–21. eprint: https://doi.org/10.1111/sum.12694.

[144] L. Poggio, L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter. "SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty". *SOIL* 7.1 (2021), pages 217–240. eprint: https://doi.org/10.5194/soil-7-217-2021.

[145] L. Poggio, L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter. "SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty". *SOIL* 7.1 (2021), pages 217–240. eprint: https://doi.org/10.5194/soil-7-217-2021.

[146] L. Poggio, A. Gimona, L. Spezia, and M. J. Brewer. "Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA". *Geoderma* 277 (2016), pages 69–82. eprint: https://doi.org/10.1016/j.geoderma.2016.04.026.

[147] P. Probst, A.-L. Boulesteix, and B. Bischl. "Tunability: Importance of Hyperparameters of Machine Learning Algorithms". *Journal of Machine Learning Research* 20.53 (2019), pages 1–32. eprint: http://jmlr.org/papers/v20/18-444.html.

[148] J. R. Quinlan et al. “Learning with continuous classes”. *5th Australian joint conference on artificial intelligence* 92 (1992), pages 343–348. eprint: https://doi.org/10.1142/1897.

[149] C. E. Rasmussen and C. K. I. Williams. “Gaussian Processes for Machine Learning”. *The MIT Press*, (2005). eprint: https://doi.org/10.7551/mitpress/3206.001.0001.

[150] K. Rau, K. Eggensperger, F. Schneider, M. Blaschek, P. Hennig, and T. Scholten. “Quantifying spatial uncertainty to improve soil predictions in data-sparse regions”. *SOIL* 11.2 (2025), pages 833–847. eprint: https://doi.org/10.5194/soil-11-833-2025.

[151] K. Rau, K. Eggensperger, F. Schneider, P. Hennig, and T. Scholten. “How can we quantify, explain, and apply the uncertainty of complex soil maps predicted with neural networks?” *Science of The Total Environment* 944 (2024), page 173720. eprint: https://doi.org/10.1016/j.scitotenv.2024.173720.

[152] T. Rentschler, M. Bartelheim, T. Behrens, T. Behrens, M. Díaz-Zorita Bonilla, S. Teuber, T. Scholten, and K. Schmidt. “Contextual spatial modelling in the horizontal and vertical domains”. *Scientific reports* 12.1 (2022), page 9496. eprint: https://doi.org/10.1038/s41598-022-13514-5.

[153] F. Rosenblatt. “Principles of neurodynamics: Perceptrons and the theory of brain mechanisms”. *Spartan Press* (1962).

[154] D. G. Rossiter, R. Zeng, and G.-L. Zhang. “Accounting for taxonomic distance in accuracy assessment of soil class predictions”. *Geoderma* 292 (2017), pages 118–127. eprint: https://doi.org/10.1016/j.geoderma.2017.01.012.

[155] F. Saygın, H. Aksoy, P. Alaboz, and O. Dengiz. “Different approaches to estimating soil properties for digital soil map integrated with machine learning and remote sensing techniques in a sub-humid ecosystem”. *Environmental Monitoring and Assessment* 195.9 (2023), page 1061. eprint: http://dx.doi.org/10.1007/s10661-023-11681-0.

[156] J. Schmidhuber. “Deep learning in neural networks: An overview”. *Neural Networks* 61 (2015), pages 85–117. eprint: https://doi.org/10.1016/j.neunet.2014.09.003.

[157] J. Schmidinger and G. B. Heuvelink. “Validation of uncertainty predictions in digital soil mapping”. *Geoderma* 437 (2023), page 116585. eprint: https://doi.org/10.1016/j.geoderma.2023.116585.

[158] J. Schmidinger, S. Vogel, V. Barkov, A.-D. Pham, R. Gebbers, H. Tavakoli, J. Correa, T. R. Tavares, P. Filippi, E. J. Jones, V. Lukas, E. Boenecke, J. Ruehlmann, I. Schroeter, E. Kramer, S. Paetzold, M. Kodaira, A. M.-C. Wadoux, L. Bragazza, K. Metzger, J. Huang, D. S. Valente, J. L. Safanelli, E. L. Bottega, R. S. Dalmolin, C. Farkas, A. Steiger, T. Z. Horst, L. Ramirez-Lopez, T. Scholten, F. Stumpf, P. Rosso, M. M. Costa, R. S. Zandonadi, J. Wetterlind, and M. Atzmueller. “LimeSoDa: A dataset collection for benchmarking of machine learning regressors in digital soil mapping”. *Geoderma* 459 (2025), page 117337. eprint: https://doi.org/10.1016/j.geoderma.2025.117337.

[159] K. Schmidt, T. Behrens, and T. Scholten. “Instance selection and classification tree analysis for large spatial datasets in digital soil mapping”. *Geoderma* 146.1 (2008), pages 138–146. eprint: https://doi.org/10.1016/j.geoderma.2008.05.010.

[160] H. Schütze, C. D. Manning, and P. Raghavan. “Introduction to information retrieval”. *Cambridge University Press* 39 (2008).

[161] P. Scull, J. Franklin, O. A. Chadwick, and D. McArthur. “Predictive soil mapping: a review”. *Progress in Physical Geography: Earth and Environment* 27.2 (2003), pages 171–197. eprint: https://doi.org/10.1191/0309133303pp366ra.

[162] P. Scull, J. Franklin, and O. Chadwick. “The application of classification tree analysis to soil type prediction in a desert landscape”. *Ecological Modelling* 181.1 (2005), pages 1–15. eprint: https://doi.org/10.1016/j.ecolmodel.2004.06.036.

[163] R. Seoh. “Qualitative Analysis of Monte Carlo Dropout”. *arXiv e-prints* (2020). eprint: https://doi.org/10.48550/arXiv.2007.01720.

[164] C. T. Silveira, C. Oka-Fiori, L. J. C. Santos, A. E. Sirtoli, C. R. Silva, and M. F. Botelho. “Soil prediction using artificial neural networks and topographic attributes”. *Geoderma* 195-196 (2013), pages 165–172. eprint: https://doi.org/10.1016/j.geoderma.2012.11.016.

[165] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. *arXiv eprint* (2014). eprint: https://doi.org/10.48550/arXiv.1409.1556.

[166] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”. *J. Mach. Learn. Res.* 15.1 (2014), 1929–1958. eprint: http://jmlr.org/papers/v15/srivastava14a.html.

[167] R. A. Stine. “Bootstrap Prediction Intervals for Regression”. *Journal of the American Statistical Association* 80.392 (1985), pages 1026–1031. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1985.10478220.

[168] F. Stumpf, K. Schmidt, P. Goebes, T. Behrens, S. Schönbrodt-Stitt, A. Wadoux, W. Xiang, and T. Scholten. “Uncertainty-guided sampling to improve digital soil maps”. *CATENA* 153 (2017), pages 30–38. eprint: https://doi.org/10.1016/j.catena.2017.01.033.

[169] T. Szandała. “Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks”. *Bio-inspired Neurocomputing* (2021). Edited by A. K. Bhoi, P. K. Mallick, C.-M. Liu, and V. E. Balas, pages 203–224. eprint: https://doi.org/10.1007/978-981-15-5495-7_11.

[170] R. Taghizadeh-Mehrjardi, M. Mahdianpari, F. Mohammadimanesh, T. Behrens, N. Toomanian, T. Scholten, and K. Schmidt. “Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran”. *Geoderma* 376 (2020), page 114552. eprint: https://doi.org/10.1016/j.geoderma.2020.114552.

[171] R. Taghizadeh-mehrjardi, N. Toomanian, A. R. Khavaninzadeh, A. Jafari, and J. Triantafilis. “Predicting and mapping of soil particle-size fractions with adaptive neuro-fuzzy inference and ant colony optimization in central Iran”. *European Journal of Soil Science* 67.6 (2016), pages 707–725. eprint: https://doi.org/10.1111/ejss.12382.

[172] R. Taghizadeh-Mehrjardi, M. Emadi, A. Cherati, B. Heung, A. Mosavi, and T. Scholten. “Bio-Inspired Hybridization of Artificial Neural Networks: An Application for Mapping the Spatial Distribution of Soil Texture Fractions”. *Remote Sensing* 13.5 (2021). eprint: https://doi.org/10.3390/rs13051025.

[173] R. Taghizadeh-Mehrjardi, N. Hamzehpour, M. Hassanzadeh, B. Heung, M. Ghebleh Goydaragh, K. Schmidt, and T. Scholten. “Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping”. *Geoderma* 399 (2021), page 115108. eprint: https://doi.org/10.1016/j.geoderma.2021.115108.

[174] R. Taghizadeh-Mehrjardi, K. Nabiollahi, N. M. Kebonye, N. Kakhani, M. Ghebleh-Goydaragh, B. Heung, A. Amirian-Chakan, S. M. T. Hossaini, and T. Scholten. “High-performance soil class delineation via UMAP coupled with machine learning in Kurdistan Province, Iran”. *Geoderma Regional* 36 (2024), e00754. eprint: https://doi.org/10.1016/j.geodrs.2024.e00754.

[175] R. Taghizadeh-Mehrjardi, R. Sheikhpour, M. Zeraatpisheh, A. Amirian-Chakan, N. Toomanian, R. Kerry, and T. Scholten. “Semi-supervised learning for the spatial extrapolation of soil information”. *Geoderma* 426 (2022), page 116094. eprint: https://doi.org/10.1016/j.geoderma.2022.116094.

[176] R. Taghizadeh-Mehrjardi, M. Zeraatpisheh, A. Amirian-Chakan, and T. Scholten. “Chapter 12 - A brief review of digital soil mapping in Iran”. *Remote Sensing of Soil and Land Surface Processes*. Edited by A. M. Melesse, O. Rahmati, K. Khosravi, and G. P. Petropoulos. Earth Observation. *Elsevier*, 2024, pages 217–228. eprint: https://doi.org/10.1016/B978-0-443-15341-9.00027-7.

[177] S. van der Westhuizen, G. B. Heuvelink, and D. P. Hofmeyr. “Multivariate random forest for digital soil mapping”. *Geoderma* 431 (2023), page 116365. eprint: https://doi.org/10.1016/j.geoderma.2023.116365.

[178] K. Vaysse and P. Lagacherie. “Using quantile regression forest to estimate uncertainty of digital soil mapping products”. *Geoderma* 291 (2017), pages 55–64. eprint: https://doi.org/10.1016/j.geoderma.2016.12.017.

[179] M. Veres, G. Lacey, and G. W. Taylor. “Deep Learning Architectures for Soil Property Prediction”. *2015 12th Conference on Computer and Robot Vision* (2015), pages 8–15. eprint: https://doi.org/10.1109/CRV.2015.15.

[180] A. M. J.-C. Wadoux. "Artificial intelligence in soil science". *European Journal of Soil Science* 76.2 (2025). e70080 EJSS-594-24.R2, e70080. eprint: https://doi.org/10.1111/ejss.70080.

[181] A. M.-C. Wadoux. "Using deep learning for multivariate mapping of soil with quantified uncertainty". *Geoderma* 351 (2019), pages 59–70. eprint: https://doi.org/10.1016/j.geoderma.2019.05.012.

[182] A. M.-C. Wadoux, G. B. Heuvelink, R. M. Lark, P. Lagacherie, J. Bouma, V. L. Mulder, Z. Libohova, L. Yang, and A. B. McBratney. "Ten challenges for the future of pedometrics". *Geoderma* 401 (2021), page 115155. eprint: https://doi.org/10.1016/j.geoderma.2021.115155.

[183] A. M.-C. Wadoux, B. Minasny, and A. B. McBratney. "Machine learning for digital soil mapping: Applications, challenges and suggested solutions". *Earth-Science Reviews* 210 (2020), page 103359. eprint: https://doi.org/10.1016/j.earscirev.2020.103359.

[184] A. M.-C. Wadoux and C. Molnar. "Beyond prediction: methods for interpreting complex models of soil variation". *Geoderma* 422 (2022), page 115953. eprint: https://doi.org/10.1016/j.geoderma.2022.115953.

[185] L. Wang and H. Liu. "An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling". *International Journal of Geographical Information Science* 20.2 (2006), pages 193–213. eprint: https://doi.org/10.1080/13658810500433453.

[186] A. W. Warrick. "Soil physics companion". *CRC press* (2001). eprint: https://doi.org/10.1201/9781420041651.

[187] R. Webster and P. A. Burrough. "Computer-based soil mapping of small areas from sample data". *Journal of Soil Science* 23.2 (1972), pages 222–234. eprint: https://doi.org/10.1111/j.1365-2389.1972.tb01655.x.

[188] B. Widrow and M. E. Hoff. "Adaptive switching circuits". *Neurocomputing, Volume 1: Foundations of Research*. *The MIT Press*, (1988), pages 96–104. eprint: https://doi.org/10.7551/mitpress/4943.003.0012.

[189] H. Wiechmann. "Die bodensystematische Kennzeichnung von Auenböden". *Stoffhaushalt von Auenökosystemen: Böden und Hydrologie, Schadstoffe, Bewertungen*. Edited by K. Friese, B. Witter, M. Rode, and G. Miehlich. Berlin, Heidelberg: *Springer Berlin Heidelberg*, (2000), pages 19–25. eprint: https://doi.org/10.1007/978-3-642-59744-2_2.

[190] H. Yan, C. M. DeChant, and H. Moradkhani. "Improving Soil Moisture Profile Prediction With the Particle Filter-Markov Chain Monte Carlo Method". *IEEE Transactions on Geoscience and Remote Sensing* 53.11 (2015), pages 6134–6147. eprint: https://doi.org/10.1109/TGRS.2015.2432067.

[191] M. Zeraatpisheh, A. Jafari, M. Bagheri Bodaghabadi, S. Ayoubi, R. Taghizadeh-Mehrjardi, N. Toomanian, R. Kerry, and M. Xu. "Conventional and digital soil mapping in Iran: Past, present, and future". *CATENA* 188 (2020), page 104424. eprint: https://doi.org/10.1016/j.catena.2019.104424.

[192] W. Zhang, X. Gu, L. Tang, Y. Yin, D. Liu, and Y. Zhang. "Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge". *Gondwana Research* 109 (2022), pages 1–17. eprint: https://doi.org/10.1016/j.gr.2022.03.015.

[193] Y. Zhang and C. Sutton. "Quasi-Newton Methods for Markov Chain Monte Carlo". *Advances in Neural Information Processing Systems*. Edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger. Volume 24. *Curran Associates, Inc.*, (2011). eprint: https://proceedings.neurips.cc/paper_files/paper/2011/file/e702e51da2c0f5be4dd354bb3e295d37-Paper.pdf.

[194] B. Zhou, X.-g. Zhang, and R.-c. Wang. "Automated soil resources mapping based on decision tree and Bayesian predictive modeling". *Journal of Zhejiang University-Science A* 5 (2004), pages 782–795. eprint: https://doi.org/10.1631/jzus.2004.0782.

[195] A.-X. Zhu. "Mapping soil landscape as spatial continua: The Neural Network Approach". *Water Resources Research* 36.3 (2000), pages 663–677. eprint: https://doi.org/10.1029/1999WR900315.