UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Karl Hendrik Tamkivi

Enhancing Mowing Event Detection by Mitigating Semi-Transparent Cloud Anomalies in Optical Satellite Image Time Series

Master's Thesis (15 ECTS)

Supervisors: Viacheslav Komisarenko, MSc
Tetiana Shtym, MSc

Tartu 2025

Enhancing Mowing Event Detection by Mitigating Semi-Transparent Cloud Anomalies in Optical Satellite Image Time Series

Abstract: Cloud contamination in optical satellite imagery poses a major challenge in remote sensing, particularly for applications that rely on high-quality temporal data and where infrequent satellite revisits make each observation valuable. Traditional pixel-based cloud detection methods often struggle with semi-transparent clouds, which can be difficult to distinguish from atmospheric effects or land surface variations. This thesis introduces a time series-based approach for detecting semi-transparent cloud contamination in Sentinel-2 optical time series for Danish grasslands. A supervised anomaly detection model was trained to estimate cloud anomaly probabilities, which were then integrated into an existing mowing event detection framework through loss function modifications, custom network layers, or post-processing techniques. The results demonstrate that incorporating cloud anomaly probabilities improved model reliability by reducing false positives caused by cloud contamination. The findings highlight the potential of uncertainty-aware learning for enhancing event detection and other remote sensing applications affected by optical data contamination.

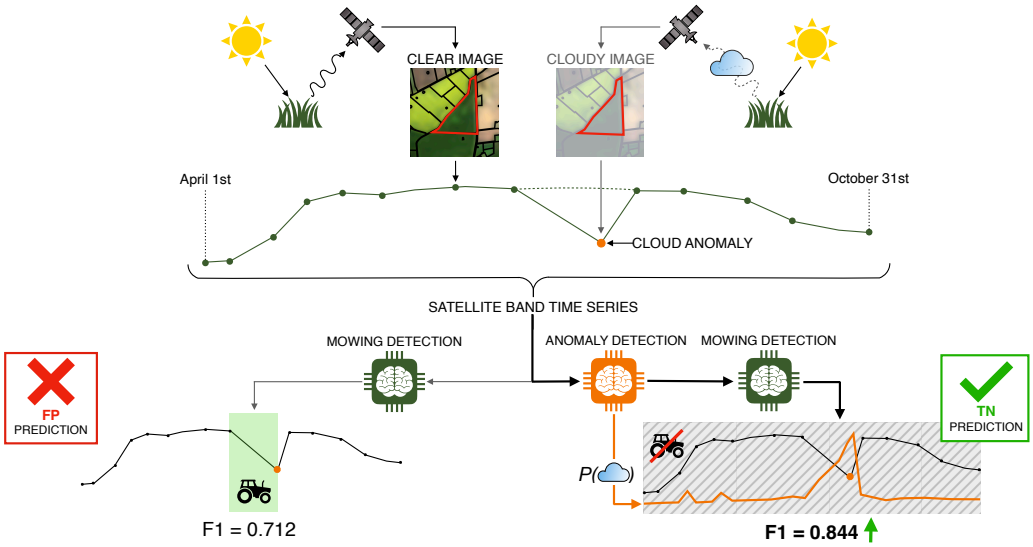
Keywords:
Remote sensing, Cloud contamination, Time series analysis, Anomaly detection, Agricultural monitoring

CERCS: P176 - Artificial intelligence, T181 - Remote sensing

Enhancing Mowing Event Detection by Mitigating Semi-Transparent Cloud Anomalies in Optical Satellite Image Time Series

Data Science (MSc), 2025

Author: Karl Hendrik Tamkivi; **Supervisors**: Viacheslav Komisarenko (MSc) & Tetiana Shtym (MSc)



Semi-transparent cloud anomaly detection model results enhance the reliability and robustness of the downstream mowing detection model

Niitmissündmuste tuvastusmudeli täiustamine optiliste satelliidipiltide aegridades olevate poolläbipaistvate pilveanomaaliate tuvastamise kaudu

Lühikokkuvõte: Atmosfääri pilvkattest põhjustatud vead optilistel satelliidipiltidel on maapinna kaugseires oluliseks väljakutseks, eriti aegridadepõhistes analüüsides, kus satelliitide harvad kordustsüklid tõstavad iga andmepunkti olulisust. Traditsioonilised pikslipõhised pilvetuvastusmeetodid ei suuda piltidelt sageli eristada poolläbipaistvaid pilvi, mistõttu ei ole ainuüksi nende meetoditega filtreeritud andmed tänapäevastes kaugseirerakendustes nõutava kvaliteedi tagamiseks piisavad. Seda eriti valdkondades, nagu näiteks põllumaade seire ja maakattemuutuste tuvastus, mis baseeruvad piltidest agregeeritud aegridade analüüsile. Magistritöös arendatakse välja uudne aegridadel põhinev lähenemisviis poolläbipaistvate pilvevigade tuvastamiseks Sentinel-2 satelliidi aegridadel Taani rohumaade näitel. Kasutades erinevaid masinõppe meetodeid treenitakse anomaaliatuvastusmudel, hindamaks pilveanomaaliate esinemise tõenäosusi aegridades. Saadud tõenäosused integreeritakse olemasolevasse niitmissündmuste tuvastamise mudelisse, kasutades mitmeid erinevaid strateegiaid - mudeli kaofunktsiooni modifikatsioone, spetsiaalselt kohandatud tehisnärvivõrgu kihte ja erinevaid anomaaliaskooridel baseeruvaid tulemuste järeltöötlustehnikaid. Tulemustest ilmneb, et pilveanomaalia tõenäosuste kaasamine vähendab pilvevigadest põhjustatud valepositiivsete niitmissündmuste tuvastuste arvu, parandades seeläbi mudeli töökindlust ja täpsust. Töö tulemused näitlikustavad edukalt sisendandmete kvaliteeti teadvustavate treeningmetoodikate potentsiaali parandadamaks optiliste andmete häiringutest mõjutatud aegridadepõhiste kaugseirerakenduste tulemuste kvaliteeti ja usaldusväärsust.

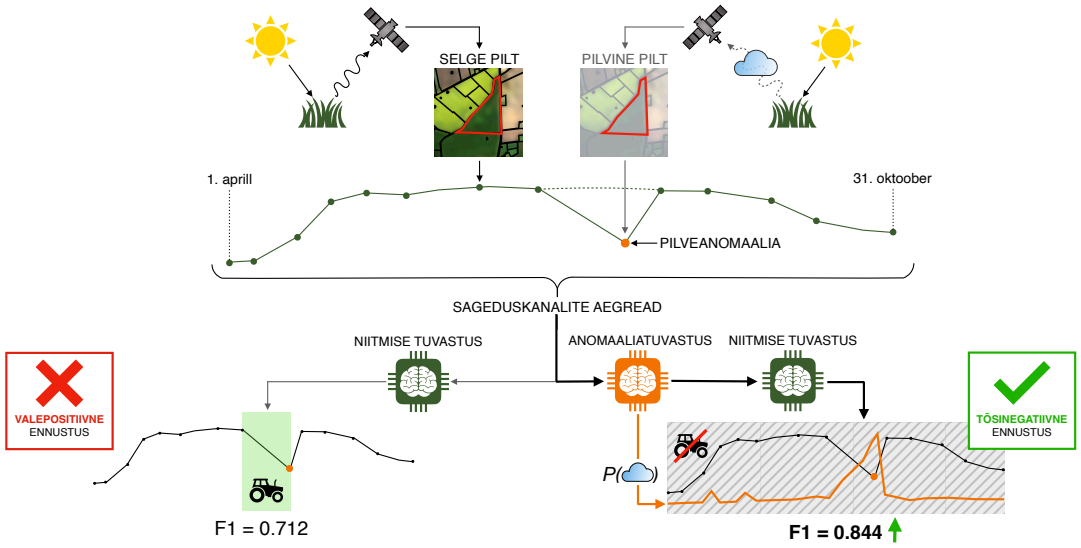
Võtmesõnad:
Kaugseire, pilvkate, aegridade analüüs, anomaaliatuvastus, põllumaade seire

CERCS: P176 - Tehisintellekt, T181 - Kaugseire

Niitmissündmuste tuvastusmudeli täiustamine optiliste satelliidipiltide aegridades olevate poolläbipaistvate pilveanomaaliate tuvastamise kaudu

Andmeteadus (MSc), 2025

Autor: Karl Hendrik Tamkivi; **Juhendajad**: Viacheslav Komisarenko (MSc) & Tetiana Shtym (MSc)



Poolläbipaistvate pilveanomaaliate tuvastamise mudeli tulemused parandavad niitmissündmuste tuvastamise mudeli robustsust ja usaldusväärsust

Contents

1 Introduction

1.1 Cloud Contamination in Satellite Remote Sensing

Satellite remote sensing, the science of acquiring information about Earth's surface and atmosphere without direct contact, has become a cornerstone of environmental monitoring and resource management in the past couple of decades (Cunningham et al., 2013, Garg, 2024, Yang et al., 2013). Using sensors mounted on satellites, remote sensing collects data such as spectral reflectance, thermal emissions, and radar backscatter, providing critical insights into phenomena like land cover changes, vegetation health, and atmospheric composition among many others (Garg, 2024, Li et al., 2020, Lillesand et al., 2015). Over recent decades, the increased availability and accessibility of remote sensing data, driven by advancements in sensor technology and open data policies, like for Sentinel (Berger et al., 2012) and Landsat (Wulder et al., 2022) missions, have led to its widespread use in diverse fields such as agriculture, water management, and climate studies (Cui et al., 2019, Khanal et al., 2020). For example, different optical vegetation indices, such as the Normalised Difference Vegetation Index (NDVI), are widely utilised in agriculture to monitor crop health, detect early signs of drought stress, and optimise irrigation practices, providing critical tools for food security in the face of climate variability (Garg, 2024, Khanal et al., 2020).

Similarly, radar-based remote sensing technologies, such as those employed by Sentinel-1, offer robust capabilities to monitor soil moisture, map flood extents, and quickly detect land cover change, even under cloud cover or during adverse weather conditions (Liu et al., 2019, Ma, 2024). The growing reliance on remote sensing applications across diverse fields underscores its transformative role in addressing global environmental challenges while also driving significant economic impacts (Benami et al., 2021, De Vroey et al., 2022, Reichstein et al., 2019). By enabling multi-temporal and large-scale analysis, remote sensing supports evidence-based decision-making and fosters sustainable development, with its economic influence extending even into sectors like insurance, logistics, and carbon credit marketing, catalysing innovation and new business models (Cunningham et al., 2013, Kenduiywo et al., 2021, Tsiropoulos et al., 2017).

The quality of remote sensing data is a critical area of research, with scientists and space agencies worldwide dedicating significant efforts to addressing the pervasive challenge of cloud contamination in optical satellite imagery. This is because clouds can obscure land features, dramatically reducing the usability of optical data, while their diverse forms and interaction with different surface types present additional challenges in downstream applications of remote sensing data (De Vroey et al., 2022, Gawlikowski et al., 2022, Li et al., 2022). For instance, bright surfaces such as snow, urban structures or recently harvested fields are often misclassified as clouds, while cloud shadows are frequently mistaken for water bodies (Gawlikowski et al., 2022, Zhai et al., 2018). Semi-transparent clouds are particularly challenging, as their optical signature can closely resemble the surface beneath, while simultaneously introducing subtle yet impactful distortions that degrade the accuracy of remote sensing analyses relying on this contaminated data (De Vroey et al., 2022, Domnich et al., 2021, Shtym et al., 2025). Various studies estimate the global average annual cloud coverage to range between 65 and 69% (Dutta et al., 2023, Stubenrauch et al., 2013), with even higher values observed at higher latitudes, and therefore the significant impact of cloud cover can not be overlooked.

Although synthetic aperture radar (SAR) data provides cloud-penetrating capabilities, making it resilient to weather conditions, it cannot fully replace optical data. Main reason being that SAR and optical sensors capture distinct physical properties, such as surface roughness and vegetation greenness, which are often complementary rather than interchangeable in real-world use cases (De Vroey et al., 2022, Komisarenko et al., 2022, Liu et al., 2019). Therefore, the detection and masking of clouds and their accompanying shadows are indispensable for ensuring the reliability and continuity of optical remote sensing data. Progress in designing reliable solutions to mitigate cloud contamination is crucial for applications such as land monitoring and emergency management, where consistent data accuracy is essential (Benami et al., 2021, Li et al., 2020, Lind and Pedersen, 2017). The ability to effectively handle cloud interference ensures that remote sensing analyses remain dependable, directly influencing the quality of decisions in high-stakes scenarios.

1.2 Advances in Cloud Detection for Optical Remote Sensing

The earliest scientific studies addressing cloud detection and removal from optical satellite image data emerged in the 1980s, and since then research in this field has grown almost exponentially. Over the decades multitude of algorithmic approaches have been proposed and developed, but most can be broadly categorized into two primary types: physical-rule-based methods or machine learning-based methods (Gupta and Nanda, 2022). Among these, physical-rule-based approaches have been used since the inception of cloud detection research and continue to be a significant focus of publications even today (Li et al., 2022). These methods rely on the spectral characteristics of clouds to identify their presence in satellite images. Clouds typically exhibit high reflectance in the visible and near-infrared spectral regions and low reflectance in the shortwave-infrared bands (Zekoll et al., 2021, Zhai et al., 2018). These properties form the foundation of widely adopted algorithms such as the Fmask algorithm (Qiu et al., 2019, Zhu et al., 2015), which combines predefined spectral thresholds with geometric principles to classify clouds and their shadows. The strength of these methods lies in their computational efficiency and interpretability, which make them well-suited for large-scale operational use (Li et al., 2022). Since rule-based approaches are grounded in explicitly defined rules, they are transparent, allowing practitioners to rather easily understand their outputs.

Even though these methods are still widely used, they also have notable limitations. One major drawback is their reliance on static thresholds, which often fail to generalize across diverse atmospheric and surface conditions (Qiu et al., 2019, Zekoll et al., 2021). This can lead to significant geographic biases and misclassifications, particularly in heterogeneous environments. Additionally, the selection of rules and thresholds is often based on manual empirical judgment and parameter sensitivity analysis, which can introduce local biases and reduce the robustness of the method (Sanchez et al., 2020, Shtym et al., 2025). Furthermore, the extreme variability in the spectral characteristics of different cloud types, ranging from opaque cumulonimbus to thin cirrus and semi-transparent clouds, further restricts the applicability and effectiveness of rule-based approaches (Gupta and Nanda, 2022, Kristollari and Karathanassi, 2020).

Machine learning-based methods, the second major category of cloud detection ap-

proaches, have experienced a remarkable rise in prominence since the 2010s. This surge coincides with rapid advancements in computational processing power and the transformative influence of artificial intelligence across numerous domains, including remote sensing (Li et al., 2022, Yuan et al., 2020). Early machine learning applications in cloud detection utilized algorithms such as fuzzy clustering, random forests and support vector machines (SVMs), which relied on manually crafted features (Joshi et al., 2019, Maxwell et al., 2018). While these methods marked a significant improvement over traditional rule-based approaches in many applications, they were limited by their dependence on feature engineering and their inability to adapt to complex patterns in the data. Recent advancements have introduced neural networks and sophisticated deep learning models, particularly convolutional neural networks (CNNs), which have revolutionized the field of remote sensing (Choi et al., 2024, Domnich et al., 2021, Yuan et al., 2020, Zhang and Zhang, 2022). CNNs can hierarchically and automatically extract features from raw satellite data, eliminating the need for manual feature selection, while also significantly improving the accuracy of cloud detection and masking (Kristollari and Karathanassi, 2020, Mohajerani and Saeedi, 2019, Zhang and Zhang, 2022). These models have demonstrated success in handling diverse cloud types, capturing subtle spectral and spatial variations that static threshold-based methods often fail to detect. In particular, deep architectures such as U-Net, DeepLabv3+, and SegNet have achieved promising results in cloud segmentation tasks, generating pixel-level cloud masks across various datasets and conditions (Choi et al., 2024, Domnich et al., 2021, Shtym et al., 2025).

Although the resolution of satellite imagery and the associated data demands are steadily increasing, the argument that computational intensity is the primary limitation for deep-learning-based cloud detection is becoming less valid in light of advancements in computational resources. Modern GPUs, cloud computing platforms, and distributed systems have significantly enhanced storage and processing capabilities, making it feasible to train and deploy complex models at scale (Choi et al., 2024, Garg, 2024, Yuan et al., 2020). Instead, the more pressing challenges lie in addressing the quality and diversity of training data. Geographical, categorical, and labeling biases in existing datasets impair the ability of models to generalize to new landscapes, atmospheric conditions and cloud types, often degrading performance in downstream applications such as land monitoring

and agricultural assessments (Domnich et al., 2021, Shtym et al., 2025, Stubenrauch et al., 2013).

Additionally, the inherent limitations of data-driven approaches, such as their potential to violate physical constraints and their reliance on data availability, highlight the need for hybrid model architectures that combine physical and machine learning-based methods (Zhang and Zhang, 2022). These models offer the interpretability and extrapolation capabilities of physics-based approaches while leveraging the flexibility and pattern recognition strengths of AI-driven systems, providing a more balanced and robust framework for cloud detection in remote sensing (Reichstein et al., 2019). Addressing these data-related challenges and adopting hybrid methodologies are crucial not only for advancing the field of cloud detection but also for ensuring the quality of the output data. High-quality cloud detection is essential, as the reliability of downstream applications, such as land monitoring and agricultural assessments, depends heavily on the accuracy of the cloud-masked datasets.

1.3 Time Series-Based Approaches to Cloud Detection

In recent years, several researchers have highlighted the significant potential of image time series-based models for cloud detection, emphasizing the importance of incorporating temporal context into the analysis (Alonso-Sarria et al., 2024, Campos-Taberner et al., 2020, Oehmcke et al., 2020, Zhang et al., 2021). These approaches leverage the temporality of satellite imagery to identify cloud-contaminated pixels, either by removing them from further analysis or employing advanced methods such as generative architectures like GANs or recurrent networks like LSTMs to predict missing information using historical cloud-free observations. Also numerous optical data indices, texture features and statistical measures have been proposed to enhance the distinction of cloudy pixels in satellite images algorithmically (Gupta and Nanda, 2022, Mahajan and Fataniya, 2020). While these approaches have demonstrated promising results, particularly in smaller-scale test cases, they share a fundamental limitation: most rely on pixel-based analysis at some stage of the workflow. This pixel-based focus is rooted in the traditional structure of optical remote sensing workflows, where image processing and information

extraction are typically performed before downstream applications, such as agricultural monitoring or land cover change detection (Garg, 2024, Lillesand et al., 2015, Zhang et al., 2024).

However, the structure of remote sensing data pipelines is rapidly evolving, with product chains increasingly adopting a modular approach where specialised counterparts handle distinct stages of the process (Di and Yu, 2023, Xu et al., 2022). For example, one entity may handle tasks such as image storage, preprocessing (e.g., applying standardised cloud-masking algorithms) and extracting information by aggregating pixel-level data over predefined geographical parcels. Another entity may then utilise these preprocessed outputs to develop downstream applications such as time series-based agricultural event detection, relying on simplified statistical summaries rather than raw pixel-level data. In such workflows, pixel-based cloud detection methods often turn out to be impractical for teams working on, for example, downstream machine learning tasks, as they lack direct access to raw pixel-level data. Instead, cloud anomalies must be addressed at the time series level, where only aggregated optical indices and their statistical measures can be analysed.

Despite the growing relevance of these scenarios, the development of cloud anomaly detection methods specifically tailored to time series data has received limited attention from the scientific community. This gap is increasingly noticeable as use cases requiring robust time series-based approaches - such as event detection and trend analysis - become more prevalent (Alonso-Sarria et al., 2024, De Vroey et al., 2022, Liu et al., 2022, Prudente et al., 2020). Especially since, unlike pixel-based approaches that rely on single-date observations and can be unreliable for smaller agricultural parcels, time series-based methods integrate information across multiple time points, enhancing resilience to data gaps and transient noise while theoretically also enabling more stable event detection (De Vroey et al., 2022, Liu et al., 2022). Exploring and advancing these methods could provide substantial added value, particularly in modern remote sensing workflows where traditional pixel-based solutions might no longer be feasible.

An anomaly in time series data according to Schmidl et al. is typically defined as a point or series of points that deviate significantly from the expected patterns, based on a chosen measure, model or embedding (Schmidl et al., 2022). In the context of remote

sensing time series, cloud-contaminated points can similarly be regarded as anomalies, as their values disrupt the regular temporal patterns of the data. This perspective opens the door to applying established methods from the field of anomaly detection, which include approaches spanning outlier detection, statistical analysis, signal processing, data mining and deep learning, to address the challenge of identifying and managing cloud-induced errors in remote sensing datasets (Li et al., 2022, Schmidl et al., 2022, Shaukat et al., 2021). Each of these approaches has distinct strengths and limitations, making the selection of an appropriate algorithm for a specific anomaly detection task highly challenging.

This complexity is further exacerbated by the expectation in many comparative assessments that the time series data points are equidistant - a condition that publicly available, preprocessed optical satellite datasets often fail to meet due to irregular observation intervals caused by cloud cover or other factors (Alonso-Sarria et al., 2024, Liang et al., 2024, Sanchez et al., 2020). Additionally, many advanced anomaly detection methods are tailored for time series in domains such as finance, transportation or medicine, where the data often exhibits specific characteristics, such as strong seasonality or periodic trends (Shaukat et al., 2021). In contrast, optical satellite data, particularly when focused on specific geographical parcels in the context of active agricultural monitoring over limited time spans, often lack such periodic characteristics (Pipia et al., 2022). This absence of seasonality introduces further challenges in adapting these methods to remote sensing applications.

Despite these difficulties, there remains significant potential for employing anomaly detection techniques in time series-based cloud detection. Comparative studies of anomaly detection methods frequently demonstrate that simpler approaches can perform nearly as well as more sophisticated ones, with the primary challenge often being parameter tuning during model development (Li et al., 2022, Schmidl et al., 2022). This highlights the importance of focusing on robust and universal anomaly detection methodologies tailored to the unique characteristics of remote sensing data, which could provide valuable contributions to the field of cloud contamination detection.

Another aspect that sets optical remote sensing time series apart from those in other domains is the high value of each data instance. In many fields, anomalous data points

can often be excluded from further analysis due to the abundance of data (Samariya and Thakkar, 2023, Shaukat et al., 2021). However, in remote sensing, the relatively low revisit frequency of satellites over specific geographical locations makes every observation particularly valuable (De Vroey et al., 2022, Sanchez et al., 2020). This highlights the potential of semi-transparent cloud-contaminated images, which, despite their contamination, could theoretically still contribute to time series analysis. For instance, in land cover change detection, such points might still be included to minimize temporal gaps, provided the contamination does not lead to false positive predictions driven by shifts in index values caused by cloud interference (Zekoll, 2023, Zhang et al., 2023). The challenge lies in leveraging these contaminated data points effectively without compromising the reliability of the analysis. Thus, ensuring that cloud-contaminated points are interpreted correctly without generating erroneous event predictions, is critical.

While incorporating anomaly detection outputs as inputs for downstream machine learning models has shown promise in other fields, such as finance or healthcare (Bakumenko and Elragal, 2022, Tabassum et al., 2024), this approach has seen little exploration in remote sensing. Moreover, no existing studies specifically address how time series-based cloud detection might be used to enhance event detection models in remote sensing applications. Thus the prospect of integrating these anomaly detection techniques to improve downstream model performance is theoretically achievable, but has remained largely unexplored.

2 Goals

Cloud contamination in optical remote sensing data presents a persistent challenge, particularly for time series-based monitoring analysis, where each data point holds significant value due to the relatively low revisit frequency of satellites. The ability to detect and manage semi-transparent cloud contamination is critical, as such points can still contribute to analyses like land cover change detection and agricultural monitoring, provided they are accurately interpreted. Traditional pixel-based cloud detection methods, while effective in certain scenarios, often fall short in modern workflows that rely on aggregated time series data for downstream applications. Thus this thesis aims to address these challenges by focusing on a fully time series-based approach for cloud contamination detection that would enhance the performance of a downstream event monitoring model - an area that has remained largely underexplored despite its growing relevance in real-world applications. To tackle these issues and provide robust solutions, this work is structured around three primary goals:

- Develop a labeled time series training dataset focusing on capturing semi-transparent cloud contamination errors in optical Sentinel-2 data for Danish grasslands;
- Train a robust anomaly detection model to identify cloud-induced anomalies in optical time series and output calibrated anomaly scores for each observation;
- Integrate the anomaly scores into an existing mowing event detection model through various integration strategies. This aims to produce a mowing detection model that is more resilient against false positive predictions caused by semi-transparent clouds, encouraging caution in predictions near cloud-affected regions, while maintaining the ability to accurately detect true positive events.

3 Data & Methods

This chapter outlines the methodology for detecting semi-transparent cloud contamination in optical satellite image time series and incorporating this information into downstream mowing event detection framework. Section 3.1 details the data sources, preprocessing steps, and manual labeling procedures used to construct the cloud anomaly training dataset. Section 3.2 describes the selection and engineering of features designed to capture cloud-induced distortions in Sentinel-2 time series. Section 3.3 presents the modeling approaches used to detect cloud contamination and generate anomaly scores, which indicate the likelihood of cloud presence at each time step. Finally, section 3.4 introduces a set of anomaly score integration strategies applied to the mowing event detection model, including loss function modifications, architectural adjustments, and post-processing techniques, followed by a performance evaluation using a specifically gathered heavily cloud-affected test case dataset.

3.1 Data Characteristics & Labeling

The training and testing data for semi-transparent cloud anomaly detection model were collected and processed in multiple steps, as shown in Figure 1. The data preparation was guided by its intended final application - improving the accuracy of mowing event detection models. Since the monitoring season for mowing events spans from April to October, this timeframe was used as the temporal filter for data collection within each year. The Sentinel-2 time series data were sourced from the KappaZeta historical time series and events database for Danish agricultural parcels. KappaZeta is an Estonian radar remote sensing company specializing in transforming Sentinel-1 data into actionable insights for agriculture, forestry and defense domains, enabling better decision-making through advanced data processing and machine learning.

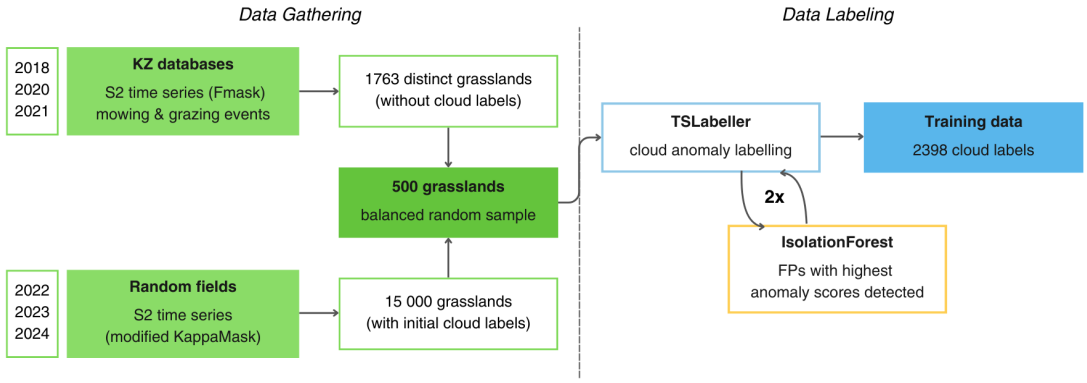


Figure 1. Training data preparation schema for the cloud anomaly detection model. The process starts with gathering data about Sentinel-2 time series linked with historical mowing or grazing events stored in KappaZeta databases (2018–2021) and extracting new random field data using a modified KappaMask workflow (2022–2024). From these two sources, a total of 500 grasslands were selected using balanced random sampling across six monitoring seasons. Manual labeling was carried out using TSlabeller software, where cloud-contaminated points were identified. This was followed by two iterations of active labeling based on IsolationForest model predictions to improve label quality.

The first set of fields for the model training were initially selected based on two criteria: they had no recorded cloud events already attached in the previously existing database, and were associated with either a mowing or grazing event. The time series data for these fields were already stored in the database and had undergone initial preprocessing. This preprocessing included cloud segmentation using the Fmask algorithm, followed by timespan filtering and a sequential point-based triplet filter, to remove larger cloud anomalies from the time series. The dataset included events and time series data from the years 2018, 2020 and 2021.

The second half of the training data was gathered using KappaMask as the cloud segmentation tool for satellite image preprocessing, with some minor modifications to its workflow. KappaMask, developed by KappaZeta, is a cloud segmentation tool specifically designed for Sentinel-2 imagery (Domnich et al., 2021, Shtym et al., 2025). It classifies satellite image pixels into five categories: clear, cloud, cloud shadow, semi-transparent cloud and missing. Compared to other widely used tools such as Fmask and Sen2Cor, KappaMask outperforms them particularly in detecting semi-transparent clouds - a class that many other approaches either neglect or handle poorly. This makes KappaMask the

most accurate tool currently available for semi-transparent cloud detection.

For this study, a total of 15 000 random Danish grassland geometries from 2022, 2023 and 2024 were selected to extract optical data time series using KappaMask instead of Fmask. Typically, KappaMask workflow extracts time series data points from 512 x 512 subtile images only if all of the pixels on a specific geographical parcel are classified as cloud-free. However, for this experiment, the criteria were modified so that fields where at least 50% of the pixels were classified as semi-transparent were also included in the time series. These cloudy points were then given an initial cloud label in the event table to accelerate the subsequent labeling process. It should be noted that the time series of the 2024 season ended in September, as the data for the latter part of the monitoring period was not available at the start of the experiment.

The next step in data preparation involved manual, point-based labeling of the time series. To achieve this, 500 random grasslands from the initial selection were chosen in a balanced manner across the six monitoring seasons, ensuring an equal number of fields from each year. The time series points were then labeled using TSlabeller (Figure 2), a time series labeling software tool developed by KappaZeta. Cloud-contaminated points were assigned the label 'cloud,' with additional comments added, specifying whether the instance was a dense cloud cover, cloud shadow or semi-transparent cloud. To improve labeling accuracy, an active learning approach was employed, leveraging initial anomaly detection results from the unsupervised IsolationForest model (detailed further in the following chapter). The model's most prevalent false positive predictions with the highest anomaly scores (upper quantile) were collected for a second round of labeling. These points were re-checked and cloud labels were attached if necessary, addressing potential inaccuracies. This process was repeated twice to enhance label quality and minimize labeler bias. By the end of this iterative process, a total of 2 398 cloud labels were assigned to a total of 500 time series, covering six different years.

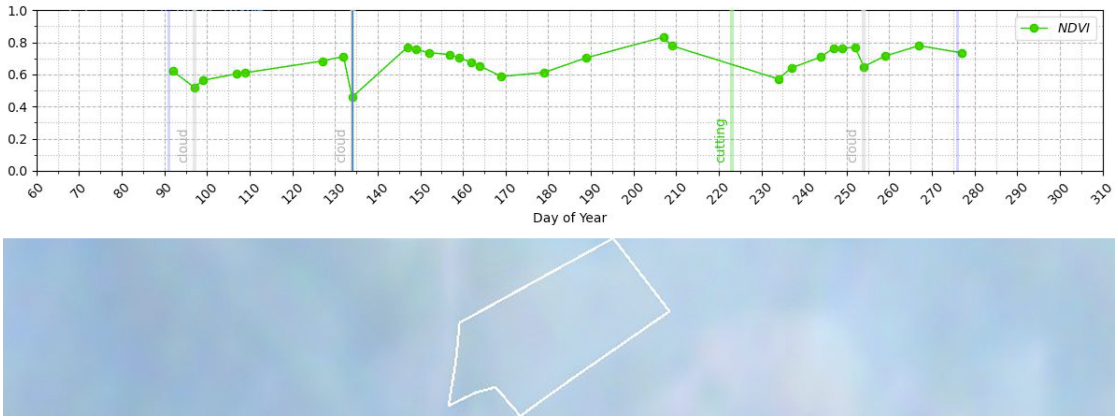


Figure 2. Example of a grassland time series with event labels attached in TSlabeller. Image below corresponds to the cloud-contaminated Sentinel-2 image at a point marked with a vertical blue line on the plot above.

3.2 Feature Selection & Engineering

To enhance the detection of semi-transparent cloud anomalies in Sentinel-2 time series data, input features were selected to balance robustness, simplicity, and predictive reliability. The focus was on capturing cloud contamination patterns without overcomplicating the model design. All features were derived from Sentinel-2 Level-2A surface reflectance data, with four band-based features ultimately selected for model training. First of them being Normalised Difference Vegetation Index e.g NDVI, which is often one of the most influential input features in many machine learning based remote sensing applications relying on vegetation condition monitoring such as land cover change detection and agricultural parcel monitoring (Alonso-Sarria et al., 2024, Campos-Taberner et al., 2020). Cloud contamination significantly degrades the quality of NDVI time series, with semi-transparent clouds often causing abrupt drops in the patterns, making it a critical feature for identifying anomalies. These anomalies can affect the performance of downstream applications, including crop classification and mowing detection. NDVI for Sentinel-2 bands is calculated using red (B04) and near-infrared (B8A) reflectance values, as follows:

$$NDVI = \frac{\text{B8A} - \text{B04}}{\text{B8A} + \text{B04}}$$

Where values close to 0 indicate sparse or dry vegetation, while values approaching 1 signify dense, healthy vegetation (Campbell and Wynne, 2011).

The second input feature for cloud anomaly detection is the Sentinel-2 blue band (B02). This band often shows sudden increases in reflectance values in the presence of clouds, particularly semi-transparent ones, due to the scattering of shorter wavelengths (Liang et al., 2024). Its sensitivity to cloud contamination makes it a commonly utilised feature in rule-based cloud detection methods within the field of remote sensing, providing complementary information to vegetation-based indices like NDVI (Zekoll et al., 2021, Zhai et al., 2018).

The third input feature in this thesis is the Tasseled Cap Brightness (TCB), one of the four subcomponents commonly derived from Tasseled Cap Transformations (TCT). TCT is a widely used method in remote sensing, valued for its ability to effectively separate land cover types and estimate biophysical characteristics (Joshi et al., 2019, Shi and Xu, 2019). Due to its relevance in estimating surface albedo (i.e., "whiteness"), TCB is frequently used in cloud detection algorithms as an indicator of potential cloud contamination (Alonso-Sarria et al., 2024). TCT coefficients vary depending on the optical satellite sensor and in this thesis, the Sentinel-2-specific coefficients proposed by Shi and Xu in 2019 are applied (Shi and Xu, 2019). These coefficients were tailored to spectral bands of Sentinel-2 to enhance their accuracy in brightness, greenness and wetness analyses. TCB in this thesis is defined as:

$$\begin{aligned} TCB = &(0.3510 \cdot B02) + (0.3813 \cdot B03) + (0.3437 \cdot B04) \\ &(0.7196 \cdot B08) + (0.2396 \cdot B11) + (0.1949 \cdot B12) \end{aligned}$$

The fourth and final input feature is a modified version of the Normalised Difference Snow Index (NDSI), which is typically used to differentiate snowy surfaces from satellite images. While NDSI is effective for distinguishing snow-covered areas, it often requires localised threshold tuning to separate snow from clouds in satellite imagery (Qiu et al., 2019, Zhu et al., 2015). This tuning process, however, can be labor-intensive and prone

to errors due to varying atmospheric conditions. However, in this thesis the cloud anomaly detection algorithm does not need to account for time series data extracted from snow-covered images. Thus the reversed NDSI (rNDSI) is utilised, as it can in theory effectively capture cloud contamination in images. Drops in rNDSI values are potentially particularly indicative of semi-transparent clouds. For Sentinel-2 data, the rNDSI is calculated using the green (B03) and shortwave infrared (B11) reflectance bands, as follows:

$$rNDSI = \frac{\text{B11} - \text{B03}}{\text{B11} + \text{B03}}$$

The time series anomaly detection model training in this thesis relied on a structured approach to feature engineering and data preparation, tailored to enhance the detection of cloud-related anomalies in time series data. The initial Sentinel-2 time series data contains irregular temporal gaps due to factors such as cloud contamination and satellite revisit intervals. To address this, the data was interpolated using linear resampling to a uniform 1-day grid, resulting in a standardized sequence length of 214 days, corresponding to the monitoring period from April 1 to October 31. This preprocessing ensured temporal consistency across all samples, enabling the model to effectively capture seasonal variations and abrupt deviations caused by cloud contamination.

A comprehensive feature engineering process was applied to enrich the dataset with temporal, statistical and feature interaction based insights, besides using just the four original input channels. Lagged and future features were created for the input variables to capture temporal dependencies and trends over preceding and subsequent time steps. Rolling statistics, such as moving averages, were calculated to normalise the data and highlight deviations from expected patterns. Peaks, including local minima for NDVI and rNDSI, and maxima for B02 and TCB were detected to identify abrupt changes in values, which can theoretically signify anomalies. Interaction terms, such as NDVI-to-B02 and NDVI-to-TCB ratios together with squared transformations of the four original input features were introduced to uncover more complex relationships and amplify patterns indicative of cloud contamination. The feature engineering techniques are summarised in Table 1.

Category	Feature Type	Description
Original Features	Sentinel-2 bands and derived indices	Core features directly derived from Sentinel-2 data.
Temporal Features	Lagged and future features	Captures past and future temporal dependencies by introducing lagged (1, 3 & 5-day) and future (1 & 3-day) versions of original features.
Rolling Statistics	Rolling averages	Highlights deviations from expected patterns by smoothing data over a 7-day window.
Peaks	Local minima and maxima	Identifies abrupt changes in values that may signify anomalies.
Interaction Terms	Ratios between features	Explores relationships between features to uncover anomalies.
Squared Features	Non-linear transformations	Highlights complex patterns by amplifying feature values through squaring.

Table 1. Summary of feature engineering categories and derived features for the cloud anomaly detection model.

3.3 Cloud Anomaly Detection Modeling

In this thesis, cloud-contaminated time series points were initially addressed and labeled using an anomaly detection approach based on IsolationForest (Figure 1), a method proposed by Liu and colleagues in 2012 (Liu et al., 2012). IsolationForest is a highly effective unsupervised anomaly detection method that identifies anomalies by isolating them based on their distinct properties of being "few and different" and thus having short average path lengths, without relying on traditional distance or density based measures like many other tree-based methods do. This approach enables efficient detection with low linear time complexity, minimal memory requirements, and robustness against challenging instances such as non-equidistant and non-periodic time series. Moreover, IsolationForest often tends to outperform widely used anomaly detection methods, such as one-class Support Vector Machines (SVM) and Local Outlier Factor (LOF) in terms of runtime, detection accuracy and scalability, particularly in large, high-dimensional

datasets (Emmott et al., 2013). These attributes make IsolationForest a promising candidate for the challenging task of detecting semi-transparent cloud anomalies in Sentinel-2 time series.

After two iterations of labeling with the help of IsolationForest results, the same model was tested against four other possible alternatives for cloud anomaly detection. First of them being logistic regression, a supervised learning method, which was picked mainly due to its efficiency, simplicity and interpretability (Jr et al., 2013). Its capacity to provide probabilistic outputs is particularly valuable for remote sensing applications, where transparent anomaly scores could support downstream tasks like agricultural event detection. Logistic regression's low computational requirements and straightforward implementation make it a strong candidate for large-scale data processing (Hand and Yu, 2007). However, the model's assumption of a linear relationship between features and outcomes can limit its ability to detect complex cloud-induced anomalies in time series influenced by interactions between atmospheric and surface conditions. This may reduce its effectiveness in capturing subtle deviations caused by semi-transparent clouds. Despite these limitations, logistic regression could still achieve competitive performance when paired with well-crafted features. This makes it particularly suitable in scenarios where interpretability, transparency and computational efficiency are essential, even if it does not fully capture complex non-linear interactions present in the data.

The second alternative method evaluated for cloud anomaly detection was the support vector machine (SVM), which offers greater flexibility through the use of non-linear kernel functions (Cortes and Vapnik, 1995). This flexibility allows SVMs to effectively capture complex relationships between features, making them potentially well-suited for identifying subtle anomalies such as semi-transparent clouds. By utilizing kernels like the radial basis function (RBF), SVMs can map input data to higher-dimensional spaces where separability of normal and anomalous points is enhanced. This improves the model's ability to detect nuanced variations that simpler models may overlook. However, this capability comes with significant drawbacks, particularly in large-scale applications - SVMs have a high computational cost, both in training and prediction, which grows with the size of the dataset and number of features (Hsu et al., 2010). This can hinder scalability when processing extensive satellite time series data. Moreover,

SVM performance is highly sensitive to hyperparameter selection, including kernel type, regularization strength and gamma parameters, requiring extensive tuning to strike a balance between model complexity, overfitting and generalization to unseen data.

Kernel Density Estimation (KDE) was the third model type evaluated as an alternative to the initial IsolationForest model. This non-parametric unsupervised method estimates the underlying probability distribution of normal points and identifies anomalies as those occurring in low-density regions (Silverman, 2018). KDE is particularly advantageous when no strong assumptions about feature relationships can be made, as it adapts to the data distribution without predefined models. Despite its adaptability, KDE's performance can degrade in high-dimensional settings due to the curse of dimensionality (Scott, 1991). Furthermore, its computational demands, especially during inference, pose challenges for scalability in large-scale or real-time remote sensing applications. However, with smaller datasets and lower-dimensional feature sets, KDE remains a promising option for capturing subtle variations in data.

A fourth model evaluated was the Random Forest (RF) classifier, a supervised ensemble learning method that constructs multiple decision trees and aggregates their outputs for improved predictive performance (Breiman, 2001). RF is well-suited for remote sensing applications due to its ability to capture complex feature interactions, handle high-dimensional data, and provide feature importance rankings for interpretability (Belgiu and Drăguţ, 2016, Pal, 2005). Unlike SVM, it does not require kernel transformations to model non-linear relationships, and it is generally more computationally efficient. However, RF tends to produce overconfident probability estimates, often necessitating post-hoc calibration methods like isotonic regression to ensure more reliable anomaly probabilities (Niculescu-Mizil and Caruana, 2005). Additionally, while its training time scales better than SVM, large-scale datasets with deep trees can still lead to increased computational demands (Pal, 2005). Despite these challenges, RF remains a strong candidate for cloud anomaly detection due to its generalization capabilities and relatively lower sensitivity to hyperparameter tuning compared to SVM or KDE.

For training and evaluation, the dataset was split into training, validation, and test sets using a 60/20/20 ratio. To ensure temporal and spatial coherence, all time series points from the same field were assigned to the same subset. This avoided spatial data

leakage and allowed for a more realistic evaluation of the model's generalisability to new locations and conditions. Hyperparameter optimization was conducted using a grid search with GroupKFold cross-validation for each model type, with relevant parameters tested across a range of values. The specific settings tested for each model are detailed in Appendix 1. The best-performing model was selected based on its performance on the test set, evaluated using metrics such as F1-score and Area Under the Receiver Operating Characteristic Curve (AUC).

The F1-score, which represents the harmonic mean of precision (PR) and recall (RC), is defined as:

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{PR} \cdot \mathrm{RC}}{\mathrm{PR} + \mathrm{RC}}$$

This metric is well-suited for imbalanced classification tasks, as it provides a balanced measure of performance by accounting for both false positives and false negatives. The selected model was used in subsequent stages to predict anomaly probabilities, referred to as anomaly scores.

To improve the reliability and calibration of the predicted anomaly probabilities i.e. anomaly scores, the selected model was refined using the `CalibratedClassifierCV` method from the `sklearn` library, employing isotonic regression. Isotonic regression was chosen for its ability to fit a non-parametric, non-decreasing function to the data, allowing flexible adjustment of predicted probabilities without imposing strict assumptions about their distribution (Zadrozny and Elkan, 2002). This calibration process aimed for a reduction of the Expected Calibration Error (ECE) metric, ensuring that the predicted probabilities more closely correspond to the actual likelihood of semi-tranparent cloud anomalies. Accurate calibration improves the interpretability of the scores by mitigating the impact of model biases that could otherwise distort probability estimates. In this thesis, this alignment is particularly valuable for downstream use in mowing event detection, where the anomaly scores are used to optimize the model performance. By enhancing the precision of probability estimates, the calibrated model in theory supports more reliable decision-making across various time series scenarios, especially in the presence of semi-transparent cloud contamination.

3.4 Mowing Event Detection Modeling

3.4.1 Original Model Training and Evaluation

The previously discussed semi-transparent cloud anomaly detection algorithms were specifically designed to enhance the performance of a grassland mowing event prediction model developed by KappaZeta. This mowing model operates within a configurable pipeline capable of supporting both training and inference runs, utilising preselected Sentinel-1, Sentinel-2 and Landsat time series channels as inputs. Built on a 1D-CNN architecture (Figure 3), the model is particularly well-suited for time series prediction, effectively capturing local temporal patterns and dependencies, such as trends and periodic behaviors, through convolutional layers (Tang et al., 2020). By stacking multiple layers, it learns hierarchical features ranging from simple fluctuations to more complex long-term patterns, all while maintaining computational efficiency, making it lighter than more resource-intensive models like RNNs (Kiranyaz et al., 2021). The model produces a probability curve for each field's time series, with event detection configured through adjustable thresholds, duration limits and customisable post-processing filters to ensure accuracy and flexibility in real-world application. Finer details about the model's architecture, training process and implementation are not discussed in this thesis, as they contain information with strategic business value that is proprietary to KappaZeta.

The original 1D-CNN mowing detection model was trained using KappaZeta's labeled event data from Danish grasslands spanning from 2018 to 2023. The training data included time series labeled as 'mowing' for positive samples and 'not cut' for negative samples, comprising 3577 labeled events across 1832 unique time series collected over multiple years. Field samples were divided into training, validation and test sets with a 70/15/15 split, ensuring that event distributions remained consistent across subsets. To handle missing observations and varying intervals between data streams, the time series were linearly interpolated to a fixed length of 214 days, matching the time span used in the anomaly detection model setup. The mowing event detection model employed a detection probability threshold of 0.6, with no restrictions on the number of events detected per time series. Model performance was evaluated at the field level, assessing whether a given field had experienced a mowing event during the monitoring season,

rather than on individual time series points. This setup aligned the evaluation process with real-world applications, where field-level event detection is often more relevant.

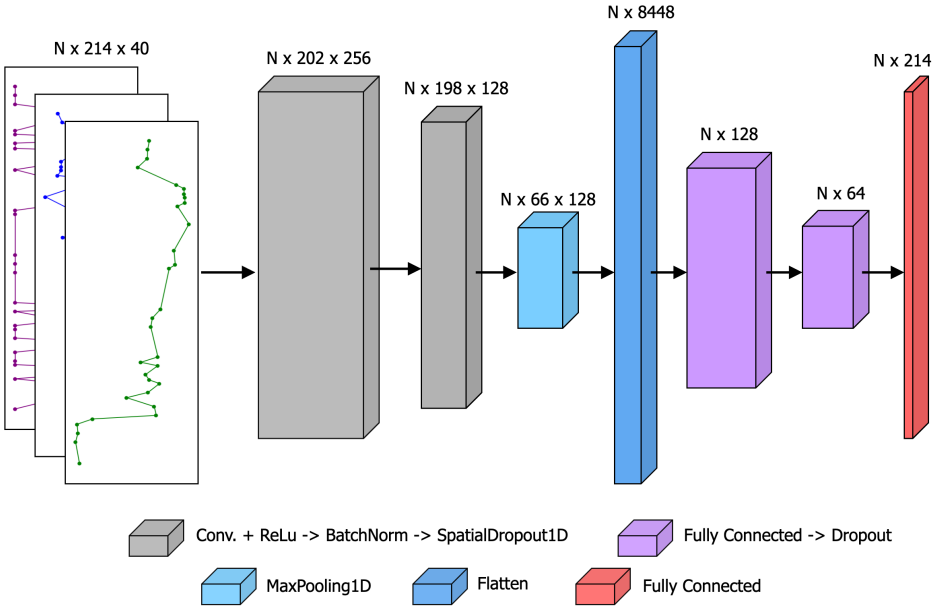


Figure 3. Diagram of the 1D-CNN mowing event detection model architecture. N refers to the batch size during the training process.

3.4.2 Anomaly Score-Based Model Modifications

Although the original 1D-CNN mowing detection model effectively identified mowing events under clear weather conditions, it lacked a mechanism to account for spectral distortions introduced by semi-transparent cloud contamination. This limitation led to increased false positive rates in cloud-affected regions, reducing the model's reliability in real-world applications. To mitigate this issue, multiple integration strategies were tested to incorporate anomaly scores produced by the upstream cloud anomaly detection model into the mowing detection framework. These strategies aimed to improve the model's ability to handle cloud-contaminated time series by modifying different stages of the workflow. By leveraging anomaly scores, the model could theoretically learn to either discount unreliable observations, dynamically adjust its learning process to reduce the impact of cloud-induced distortions, or refine its final probability estimates.

The first method to integrate information about cloud anomalies was to add anomaly scores as an additional input feature to the original detection model. This approach allows the network to incorporate explicit information about cloud contamination directly into the learning process, making it theoretically possible to model the relationship between cloud presence and event probability. However, despite its conceptual simplicity, this approach does not directly regulate how the model should treat the cloud-affected samples, meaning that the network could still overfit to cloud-induced spectral patterns afterwards. Furthermore, the limited variability and frequency of cloud-contaminated examples in the training data may hinder the model's ability to generalize this relationship effectively. Despite these limitations, similar feature augmentation approaches have been used successfully in machine learning to enhance robustness against domain-specific noise and artifacts in other fields (Ho and Wookey, 2020, Hwang et al., 2024).

Loss Function Modifications. To more explicitly control the model's learning process in cloud-affected conditions, the second approach involved modifying the loss function by introducing a weighted binary cross-entropy (WCE) term. While traditional binary cross-entropy (BCE) loss treats all time steps equally, this approach adjusts the contribution of each sample based on its cloud anomaly score. Specifically, higher weights were assigned to cloud-contaminated samples, increasing their influence during training. Rather than suppressing these uncertain regions, the goal was to penalize the model more heavily for making overconfident predictions in ambiguous, potentially distorted contexts. This strategy encourages the model to learn more cautious decision boundaries in cloud-affected segments, promoting robustness and reducing the rate of false positives. Loss weighting techniques similar to this have been studied in the context of handling noisy labels, imbalanced training distributions, and uncertainty-aware learning, particularly in applications where certain samples are inherently more informative or reliable than others (Lin et al., 2017, Zhang and Sabuncu, 2018).

The custom WCE loss for a sample is calculated as a weighted BCE, where the anomaly scores are used to adjust the weight of each time step i:

$$L_{\text{WCE}} = \frac{1}{N} \sum_{i=1}^{N} [w_{\text{anomaly},i} \cdot \text{BCE}(y_{\text{true},i}, y_{\text{pred},i})]$$

Where:

y_{true}: Ground truth labels (1 for mowing, 0 otherwise).
y_{pred}: Predicted probabilities for each time step.
w_{anomaly}: A weight vector defined as $w_{\text{anomaly}} = \mathbf{A}_{\text{cloud}} + 1.0$,
where $\mathbf{A}_{\text{cloud}}$ represents the cloud anomaly score vector for the sample.
This ensures a minimum weight of 1.0 for all points.

The BCE is calculated as:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

Where:

y_i: The true label for time step i.
p_i: The predicted event probability for time step i.

Alternative loss function modification approach tested was applying an anomaly-weighted regularization term. Regularization techniques such as L1 (Lasso) and L2 (Ridge) penalties have been widely used to stabilize deep learning models by discouraging overly complex representations that may overfit to noisy patterns in the data (Hoerl and Kennard, 1970, Tibshirani, 1996). In this study, regularization strength was dynamically scaled based on cloud anomaly scores, allowing stronger penalties in regions where cloud contamination was higher. The cloud anomaly score scaled the regularization effect per sample, ensuring that the model learned more conservatively when processing more cloud-affected time series. This adaptive regularization approach has been explored in other domains, including adversarial learning and uncertainty-based deep learning, where structured noise in the training data can otherwise lead to overfitting (Kendall and Gal, 2017, Miyato et al., 2021). The modified loss function incorporating anomaly-weighted regularization term was formulated as:

$$L_{\text{Reg}} = \frac{1}{N}\sum_{i=1}^{N}\text{BCE}(y_{\text{true},i}, y_{\text{pred},i}) + \lambda \cdot \frac{1}{N}\sum_{j=1}^{N}\mathbf{A}_{\text{cloud},i} \cdot \frac{1}{M}\sum_{j=1}^{M}||\theta_j||^p$$

Where:

λ: Strength of the regularization term.

M: Total number of trainable parameters in the model.

$||\theta_j||^p$: Penalty term denoting either *L1 (p=1)* or *L2 (p=2)* regularization.

Model Architecture Modifications. The second direction of tested model modifications focused on refining the 1D-CNN model architecture through custom layer designs aimed at mitigating cloud-induced distortions. The first of these architectural modifications introduced an anomaly-based scaling mechanism that dynamically adjusted Sentinel-2 optical feature values before they were processed by the convolutional layers. This approach aimed to suppress spectral distortions in regions with high anomaly scores, preventing the model from learning misleading patterns associated with semi-transparent cloud contamination. The transformation was defined as follows:

$$\mathbf{X}_{\text{S2 scaled},i,j} = \mathbf{X}_{\text{S2},i,j} \cdot S_j \cdot \sigma(\mathbf{A}_{\text{cloud},i}, \beta)$$

Where:

$\mathbf{X}_{\text{S2 scaled},i,j}$: Adjusted values for j-th Sentinel-2 feature at time step i.

$\mathbf{X}_{\text{S2},i,j}$: Sentinel-2 feature value for feature j at time step i.

S_j: Trainable scaling parameter for the j-th Sentinel-2 feature.

$\mathbf{A}_{\text{cloud},i}$: Cloud anomaly score at time step *i*.

$\sigma(\mathbf{A}_{\text{cloud},i}, \beta)$: Anomaly-dependent sigmoid scaling function, defined as:

$$\sigma(\mathbf{A}_{\text{cloud},i}, \beta) = \frac{1}{1 + e^{\beta(\mathbf{A}_{\text{cloud},i} - T)}}$$

Where:

T: Anomaly score threshold (pre-determined during anomaly detection model training).

β: Sigmoid scaling steepness parameter.
Higher β results in a more abrupt suppression of feature values when the anomaly score exceeds the threshold T.

This approach theoretically enhances the model's adaptability by allowing feature-specific attenuation of cloud-induced distortions through the trainable scaling factor S_j. Unlike uniform suppression methods, it enables the network to learn optimal adjustments per spectral band, preserving useful information while mitigating contamination. Inspired by feature-wise transformation techniques and adaptive gating mechanisms, that have seen notable attention in image-related tasks, this method improves robustness by dynamically regulating feature scaling based on cloud anomaly scores (Ehteshami Bejnordi and Krestel, 2020, Perez et al., 2018). This fine-grained modulation thus helps prevent overfitting to cloud artefacts while aiming to maintain critical spectral signals for accurate mowing event detection.

Second architectural modification tested in this experiment aimed at improving the model's robustness to cloud-induced distortions involved the introduction of a bottleneck layer before the convolutional layers. This layer compressed the high-dimensional feature space into a lower-dimensional representation, forcing the network to learn a compact and more generalizable encoding of the input time series. Unlike the anomaly-based scaling mechanism, which directly attenuated affected features, this approach sought to reduce redundancy and suppress spurious correlations introduced by cloud contamination, thereby encouraging the model to focus on more stable temporal patterns. The bottleneck transformation was implemented using a dense layer with a reduced number of output units, applied to all features except the anomaly score itself (Figure 4). To preserve explicit information about cloud contamination, the anomaly score feature was excluded from the dimensionality reduction process and reintroduced after the bottleneck transformation. This ensured that while the primary features were compressed, the model still retained explicit access to the anomaly score when making predictions. Similar dimensionality reduction techniques have been widely explored in deep learning, particularly in autoencoders and latent space representations, where they help to remove noise and improve generalization in various applications (Bengio et al., 2013, Hinton and Salakhutdinov, 2006).

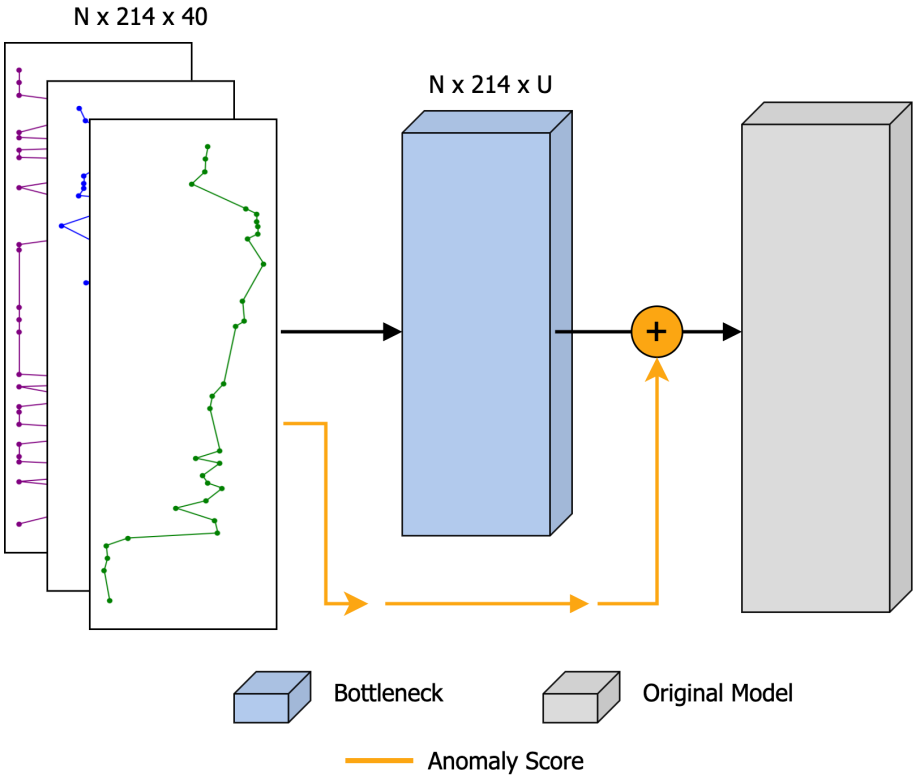


Figure 4. Diagram of the bottleneck layer modification to the original 1D-CNN mowing event detection model architecture. N refers to the batch size during the training process, while U refers to the number of bottleneck output units.

Post-Processing Techniques. Besides the previously discussed model loss function and architecture modifications, an alternative set of model improvement strategies was explored by applying post-processing techniques to the original, unmodified model predictions, rather than adjusting the network itself. These post-processing methods aimed to refine the final probability outputs by incorporating cloud anomaly scores as an external correction factor, ensuring that uncertainty associated with cloud-contaminated observations was explicitly accounted for. The first post-processing approach employed a linear scaling rule to systematically adjust the mowing event probabilities based on the anomaly scores at each time step. The transformation was formulated as follows:

$$p_i' = p_i \cdot (1 - \alpha \cdot \mathbf{A}_{\text{cloud},i}), \quad \text{for } \mathbf{A}_{\text{cloud},i} > T$$

Where:

p_i': Adjusted mowing event probability at time step i.
p_i: Original predicted probability for mowing event at time step i.
$\mathbf{A}_{\text{cloud},i}$: Cloud anomaly score at time step i.
α: Scaling factor controlling the magnitude of the probability adjustment.
T: Anomaly score threshold (pre-determined during anomaly detection model training).

This approach systematically reduced the confidence of predictions in time steps where cloud contamination was detected, thus in theory mitigating the likelihood of false positive event detections driven by transient spectral distortions. While computationally efficient and easy to implement, this method applied a uniform penalty across all affected samples, limiting its flexibility in adapting to varying degrees of cloud contamination across different scenarios and conditions. Despite that, linear correction strategies similar to this have been previously employed in other domains to account for uncertainty in model outputs, particularly in applications where environmental artifacts introduce systematic biases (Jiang et al., 2021).

The second postprocessing technique applied conditional isotonic calibration, a non-parametric approach that adjusted model predictions based on cloud anomaly scores. Unlike the linear scaling method, which applied a uniform correction, isotonic calibration learned a monotonic mapping between predicted probabilities and actual event frequencies, ensuring that the final outputs better reflected true mowing event probabilities under different cloud contamination conditions. To achieve this, the mowing event detection training-validation set was divided into *G* groups based on anomaly score quantiles, capturing different levels of cloud interference. Within each group, an isotonic regression model was fitted to the original model predictions, learning a piecewise-constant function to correct probability estimates. The calibrated probability for a given time step *i* was computed as:

$$p'_i = \hat{f}_{g(i)}(p_i)$$

Where:

$\hat{f}_{g(i)}$: Isotonic regression function fitted for anomaly score group $g(i)$.
$g(i)$: Group assignment function mapping time step *i* to an anomaly score quantile group.

This approach provided a data-driven correction that adjusted predictions dynamically based on observed error patterns, ensuring that confidence estimates aligned more closely with actual event frequencies. Isotonic regression-based calibration has been widely used to improve model reliability in fields where prediction probabilities must be well-calibrated, such as medical diagnosis and weather forecasting (Guo et al., 2017, Jiang et al., 2012). Compared to the linear scaling approach, this method offers greater flexibility, adapting correction strengths to different levels of cloud contamination while maintaining computational efficiency.

Apart from the previously mentioned experimental changes, all other data preprocessing steps and model hyperparameters remained identical to the original training setup. This ensured that the observed changes in performance could be attributed directly to the incorporation of anomaly scores and the mentioned improvement strategies.

3.4.3 Model Evaluation on Cloudy Test Cases

To evaluate the performance and improvements of the newly trained mowing event detection models compared to the original approach, a specific test case dataset was manually compiled. This dataset included time series from the 2023 and 2024 Danish agricultural monitoring season, where KappaZeta's geodata specialists had flagged cloud-related issues during quality assurance checks conducted prior to model prediction deliveries. The flagged fields were identified through comments containing the keyword 'cloud', resulting in a total of 125 unique fields. While not all flagged cases were directly linked to false positives induced by cloud contamination, the cloud-related anomalies in these time series were significant enough to be highlighted during quality control.

The compiled test dataset included quality-checked time series across three agricultural markers - mowing, ploughing and crop detection - indicating that not all cases were specific to grasslands but rather spanned broader agricultural applications.

Both the original 1D CNN mowing event detection model and several updated versions, each incorporating cloud anomaly score information through different integration strategies, were evaluated on the described test case dataset. There were no restrictions on the number of events predicted per time series, however, to ensure comparability between all model versions, only the first predicted event per field was considered in cases with multiple detections. This adjustment reflects the operational reality of agricultural monitoring workflows, where a field may reappear in subsequent prediction deliveries until a valid event is detected. If a correct prediction occurred before an incorrect one, the field-level evaluation was still treated as accurate. All predictions across models were manually reviewed and categorized as true positive (TP), false positive (FP), true negative (TN), or false negative (FN). In cases where a false positive aligned with or was near a cloud-contaminated Sentinel-2 observation, a comment stating "cloud" was added to the prediction. In instances where the results were particularly challenging to visually evaluate, such as cases involving very small areas or complex geometries of agricultural parcels, a label of ‘Not sure’ was assigned. The resulting confusion matrices were used to calculate F1 scores for each model variant, providing a basis for comparing model performance and understanding the added value of integrating anomaly scores in handling cloud-related challenges.

4 Results

This chapter presents the results of the experiments conducted to detect semi-transparent cloud anomalies in Sentinel-2 time series and to integrate this information into a mowing event detection framework. Section 4.1 evaluates the performance of five different cloud anomaly detection models using multiple classification and calibration metrics. Section 4.2 examines the impact of incorporating anomaly scores into the mowing event detection model through several enhancement strategies, including loss function modifications, architectural changes, and post-processing methods. Finally, section 4.3 provides a comparative evaluation of model performance on a cloud-affected test case dataset to assess the real-world effectiveness of the proposed modifications.

4.1 Cloud Anomaly Detection

The performance of the five different semi-transparent cloud anomaly detection models was evaluated on all optical data points for both the train-validation ($N = 10539$) and test ($N = 2474$) datasets. The models' performance was assessed based on metrics like F1-score, area under the curve (AUC) and mean squared error (MSE). User's accuracy (UA), which measures the proportion of correctly identified points out of all points predicted as a particular class, and producer's accuracy (PA), which evaluates the proportion of correctly identified points out of all actual points belonging to a particular class, were also calculated to better evaluate the model's performance on both the positive and negative class. These metrics are defined as:

$$\text{UA}_i = \frac{TP_i}{TP_i + FP_i} \quad \text{PA}_i = \frac{TP_i}{TP_i + FN_i}$$

where $i \in \{0, 1\}$ corresponds to the class being evaluated.

Model	Dataset	UA_0	UA_1	PA_0	PA_1	MSE	AUC	F1
IF	Train-Val	0.940	0.669	0.874	0.819	0.140	0.846	0.736
	Test	0.942	0.716	0.889	**0.835**	0.124	0.862	0.771
RF	Train-Val	0.949	0.840	0.965	0.779	0.070	0.872	0.809
	Test	0.941	**0.846**	**0.963**	0.769	**0.077**	0.866	0.794
LR	Train-Val	0.942	0.767	0.946	0.755	0.091	0.850	0.761
	Test	0.944	0.816	0.955	0.781	0.081	**0.868**	**0.798**
SVM	Train-Val	0.942	0.771	0.947	0.754	0.090	0.850	0.763
	Test	**0.945**	0.798	0.949	0.786	0.084	0.867	0.792
KDE	Train-Val	0.930	0.688	0.863	0.822	0.148	0.843	0.749
	Test	0.899	0.730	0.909	0.707	0.143	0.808	0.719

Table 2. Evaluation results for the semi-transparent cloud anomaly detection models - IsolationForest (IF), RandomForest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Kernel Density Estimation (KDE). UA represents user's accuracy and PA represents producer's accuracy. Class 0 corresponds to normal points, while class 1 corresponds to cloud anomalies.

The semi-transparent cloud anomaly detection models were evaluated across multiple supervised and unsupervised approaches, with results summarized in Table 2. As expected, supervised methods outperformed their unsupervised counterparts, with Logistic Regression (LR) and Random Forest (RF) achieving the highest accuracy across both training-validation and test datasets. Interestingly IsolationForest (IF), despite being unsupervised, performed surprisingly well, achieving a (PA_1) of 0.835 and an F1-score of 0.771 on the test set, only slightly lower than the best-performing supervised models. In the end the primary comparison was between RF and LR models, where RF achieved the highest anomaly class user's accuracy (UA_1) of 0.846 and lowest test set mean squared error (MSE) of 0.077. However, given that RF models often have a tendency to overfit and that LR model provided a balanced performance with an F1-score of 0.798 and the highest AUC of 0.868, LR was selected as the primary anomaly detection model for subsequent model calibration step. This decision was backed up with facts LR models overall exhibiting good efficiency, robustness, and interpretability. An example of the

model's anomaly scores applied to a grassland time series is shown in Figure 5, where predicted anomalies mostly align with cloud-affected time steps.

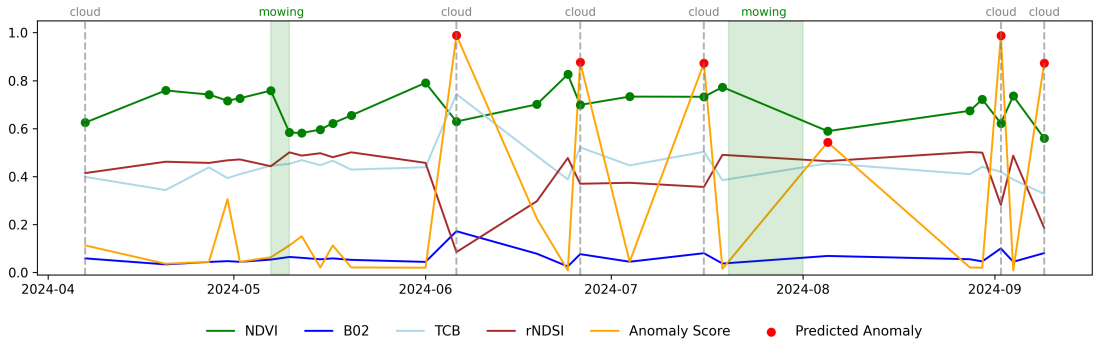


Figure 5. Example of a grassland time series together with the logistic regression output probabilities e.g. anomaly scores. labeled cloud events are shown with grey lines and labels, while mowing events are marked with green areas.

The calibration of the logistic regression model using isotonic regression improved the reliability of its predicted anomaly scores, as demonstrated in Figure 6. Prior to calibration, the model exhibited a slight tendency toward overconfident probability estimates, with an Expected Calibration Error (ECE) of 0.0651. After calibration, the ECE decreased to 0.0386, indicating a better alignment between predicted probabilities and actual event frequencies. This improvement suggests that the calibrated model provides more reliable probability estimates, mitigating overconfidence in high-anomaly cases while preserving the ability to distinguish between normal and anomalous observations. Given the potential importance of well-calibrated anomaly scores for downstream mowing event detection, this adjustment was expected to contribute to improved model robustness, particularly when integrating and implementing the anomaly score-based modifications discussed in previous sections.

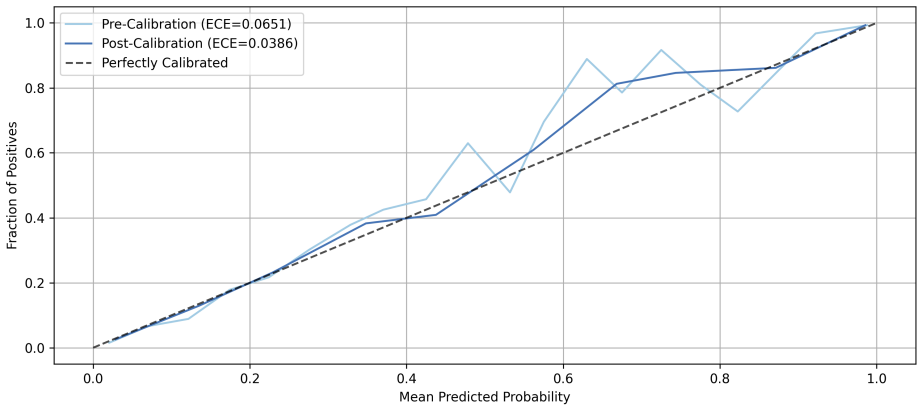


Figure 6. Test set calibration curves for the logistic regression semi-transparent cloud anomaly detection model before and after isotonic calibration on the validation set.

4.2 Mowing Event Detection

4.2.1 Model Training

To assess the impact of integrating anomaly scores into the mowing event detection framework, the previously trained and calibrated logistic regression cloud anomaly detection model was applied to the mowing event detection dataset. This allowed the extraction of anomaly scores, which were incorporated into the detection model through various enhancement strategies. Table 3 summarizes the training-validation results for the original mowing event detection model alongside all anomaly-enhanced modeling strategies, excluding post-processing approaches.

	$\mathbf{PR_{train}}$	$\mathbf{RC_{train}}$	$\mathbf{F1_{train}}$	$\mathbf{PR_{val}}$	$\mathbf{RC_{val}}$	$\mathbf{F1_{val}}$
Original	0.942	0.903	0.921	0.887	0.824	0.854
Original + Anomaly Score	0.959	0.931	0.945	0.891	0.840	0.864
Loss Function Modifications						
Weighted Cross-Entropy	0.944	0.930	0.937	**0.905**	**0.851**	**0.877**
Anomaly-Based Regularization	0.936	0.918	0.927	0.892	0.844	0.867
Model Architecture Modifications						
S2 Scaling Layer	0.941	0.923	0.932	0.893	0.844	0.868
Bottleneck Layer	0.934	0.909	0.921	0.894	0.845	0.868

Table 3. Evaluation results for the train and validation sets of the mowing event detection models. PR represents model precision and RC represents model recall.

It appears that integrating anomaly scores into the mowing event detection framework led to varying degrees of improvement across different enhancement strategies. Directly incorporating anomaly scores as an additional input feature resulted in a modest performance gain over the original model, with a slight increase in validation F1-score from 0.854 to 0.864. Among the loss function modifications, the weighted cross-entropy approach yielded the highest validation F1-score (0.877), primarily due to improved recall ($RC_{\text{val}} = 0.851$), suggesting that weighting loss based on anomaly scores helped the model better differentiate between true mowing events and cloud-induced anomalies. Anomaly-based regularization, while beneficial, exhibited a smaller improvement ($F1_{\text{val}} = 0.867$), indicating that although regularization helped stabilize training, its effect was less pronounced than direct loss weighting. In terms of model architecture modifications, both the Sentinel-2 scaling layer and the bottleneck layer achieved similar performance ($F1_{\text{val}} = 0.868$), demonstrating that feature-level adjustments and dimensionality reduction strategies could provide moderate enhancements in robustness. Overall, the results suggest that while all tested modifications contributed to improvements, loss function modifications, particularly weighted cross-entropy, offered the most substantial gains in handling cloud contamination effects.

The test set results, summarized in Table 4, provide further insight into the generalization performance of the anomaly-aware enhancements applied to the mowing event detection model. Similar to the training-validation phase, loss function modifications yielded the most notable improvements, with weighted cross-entropy achieving the highest overall performance ($F1 = 0.863$) by balancing precision ($PR_1 = 0.895$) and recall ($RC_1 = 0.833$) for mowing events. Anomaly-based regularization also improved upon the baseline, albeit to a lesser extent ($F1 = 0.855$). The model architecture modifications, including the Sentinel-2 scaling layer and the bottleneck layer, resulted in minor but consistent improvements, with F1-scores of 0.855 and 0.857, respectively. These improvements are also reflected in Matthews Correlation Coefficient (MCC), which increased slightly across all configurations, indicating enhanced agreement between predicted and true labels. MCC is particularly valuable for evaluating imbalanced datasets as it considers all elements of the confusion matrix, providing a more balanced view of model performance.

Post-processing modifications, applied separately to the original model predicted probabilities, showed mixed results. While conditional isotonic calibration improved precision at the cost of recall, yielding a moderate performance increase ($F1 = 0.848$), the more straightforward linear scaling approach led to a notable reduction in recall ($RC_1 = 0.765$), resulting in a notably lower overall F1-score ($F1 = 0.821$). Compared to the other modifications, the effects of post-processing methods were relatively small or even negative, suggesting that integration of anomaly scores into the model architecture or loss function may be more effective for addressing cloud-induced distortions in mowing event detection.

	PR_0	RC_0	PR_1	RC_1	MCC	F1
Original	0.979	0.984	0.866	0.827	0.828	0.846
Original + Anomaly Score	0.978	0.986	0.881	0.823	0.834	0.851
Loss Function Modifications						
Weighted Cross-Entropy	0.980	0.988	0.895	0.833	**0.847**	**0.863**
Anomaly-Based Regularization	0.979	0.987	0.888	0.828	0.841	0.855
Model Architecture Modifications						
S2 Scaling Layer	0.979	0.987	0.887	0.825	0.838	0.855
Bottleneck Layer	0.979	0.989	0.884	0.833	0.841	0.857
Post-Processing Modifications						
Linear Scaling	0.961	0.983	0.886	0.765	0.796	0.821
Conditional Isotonic Calibration	0.964	0.989	0.925	0.783	0.829	0.848

Table 4. Evaluation results for the mowing event detection models on the test set. Class 0 corresponds to time series points with no event labels, while class 1 corresponds to points with mowing events. MCC corresponds to Mathew's Correlation Coefficient.

4.2.2 Model Performance on Cloudy Test Cases

Out of the 125 fields included in the test case dataset, gathered from the quality assurance reports of the 2023 and 2024 monitoring season, 6 were deemed to status 'Not sure', due to challenges in visual evaluation, such as very small areas or complex geometries of the field parcels. These cases were excluded from further results evaluation. The remaining 119 fields were used to assess the performance of both the original and anomaly-aware mowing event detection models, as shown in Table 5.

	FNR	FPR	F1
Original	0.288	0.467	0.712
Original + Anomaly Score	0.333	0.283	0.722
Loss Function Modifications			
Weighted Cross-Entropy	0.225	**0.132**	**0.844**
Anomaly-Based Regularization	**0.224**	0.238	0.814
Model Architecture Modifications			
S2 Scaling Layer	0.263	0.286	0.778
Bottleneck Layer	0.293	0.222	0.784
Post-Processing Modifications			
Linear Scaling	0.316	0.381	0.722
Conditional Isotonic Calibration	0.284	0.295	0.757

Table 5. Evaluation results for the mowing event detection models on the cloud-affected test case dataset. FNR (False Negative Rate) represents the proportion of missed mowing events, while FPR (False Positive Rate) indicates the proportion of incorrectly predicted events.

The performance evaluation on the cloud-affected test case dataset provides insights into the impact of different anomaly-aware modifications. The original mowing event detection model exhibited a false negative rate (FNR) of 0.288 and a false positive rate (FPR) of 0.467, indicating a relatively high occurrence of both missed and incorrectly detected mowing events in cloud-contaminated time series. Incorporating anomaly scores as an additional input feature reduced the FPR to 0.283 but resulted in a slightly higher FNR of 0.333, suggesting that while this method helped suppress some false detections, it also led to an increased number of missed events. Among the loss function modifications, weighted cross-entropy stood out as the most effective, achieving the lowest FPR of 0.132 and the highest F1-score of 0.844, demonstrating notably improved differentiation between actual mowing events and cloud-induced anomalies compared to the original model. In contrast, the anomaly-based regularization approach, while also reducing the FNR to 0.224, had a higher FPR of 0.238, indicating a less pronounced

improvement in false positive suppression. Architectural modifications such as the S2 scaling layer and the bottleneck layer produced mixed results, with F1-scores of 0.778 and 0.784, respectively, and less significant reductions in false detections. Post-processing techniques had a more limited effect, with the linear scaling approach increasing the FNR to 0.316 while achieving an F1-score similar to the original model. Conditional isotonic calibration performed better, with an FPR of 0.295 and an F1-score of 0.757, but still showing only a modest improvement compared to the original approach.

The updated 1D-CNN mowing event detection model, enhanced with anomaly score feature and the weighted binary cross-entropy loss function, demonstrated a more cautious approach in its probability curves while still displaying a better overall performance compared to the original model. This behavior is particularly evident in cases where cloud contamination is likely to cause prediction errors. Out of all anomalies where the original model's event probability exceeded the decision threshold of 0.6, approximately 80.9% of the points showed a reduction in probability with the updated anomaly-aware model. On average, the anomaly-aware model assigned probabilities approximately 24.2 percentage points lower than those of the original model for these points, reflecting its heightened caution. This adjustment not only reduces the risk of false positives in cloud-affected regions but also contributes to the overall improvement in F1-score and prediction accuracy, as seen in the test case results. Figures 7 and 8 illustrate the updated model's ability to downweight high probabilities in ambiguous regions with high anomaly scores, while retaining its sensitivity to true event occurrences, highlighting its improved robustness against semi-transparent cloud contamination.

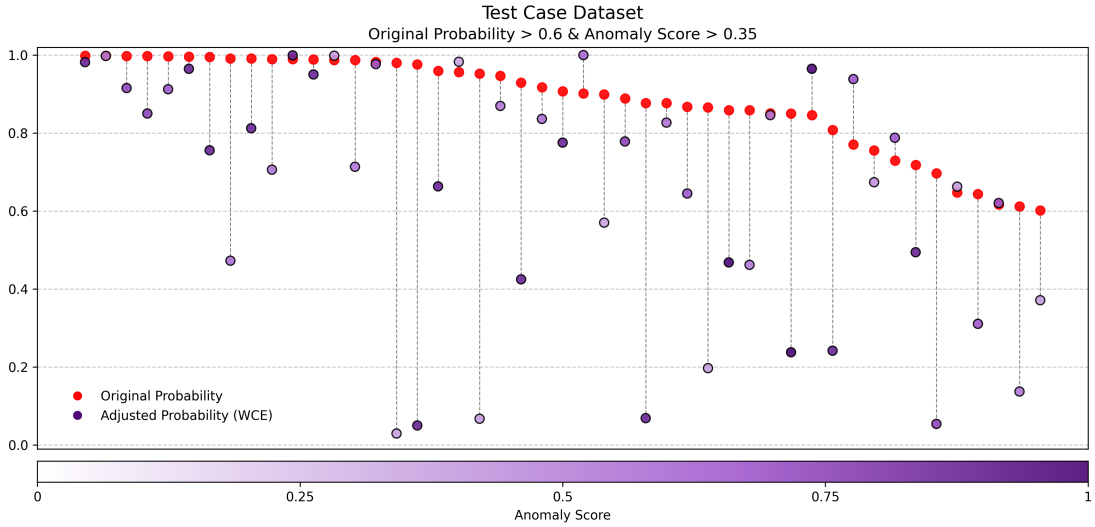


Figure 7. Original probabilities and adjusted probabilities for the anomalous optical points of the cloudy test case dataset, where the original event probability was above the decision threshold of 0.6.

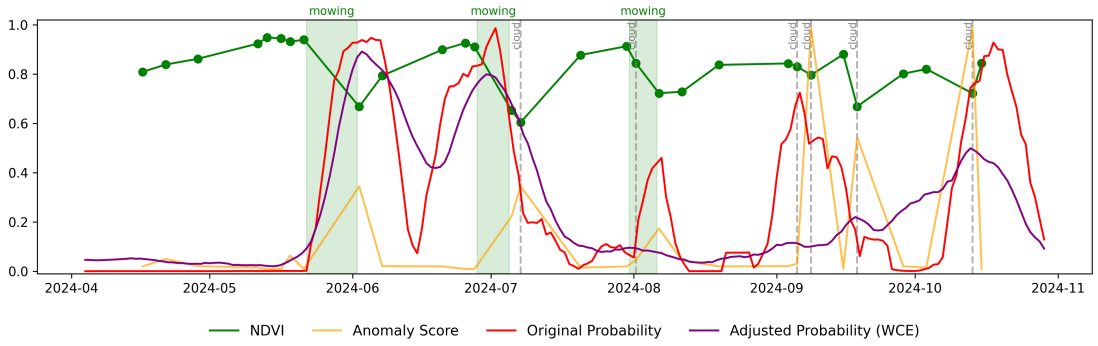


Figure 8. Example of 1D-CNN mowing event detection model output probabilities of with and without the integration of cloud anomaly scores for one field.

5 Discussion

This thesis presents a novel approach to tackling cloud-related challenges in remote sensing by integrating multiple techniques into a unified pipeline for mowing event detection on grasslands. Unlike traditional methods that often treat cloud contamination as an isolated issue, this work incorporates it into a time series analysis framework, integrating the semi-transparent cloud anomaly detection with a 1D-CNN mowing event detection model. By combining machine learning-based anomaly detection with domain-specific insights, this research aims to bridge the gap between theoretical advancements and practical applications. By specifically addressing semi-transparent cloud contamination and ensuring the pipeline is tailored to the operational needs of mowing event detection, this work offers a possible pathway for tackling interconnected challenges in modern remote sensing workflows. The following discussion briefly examines the strengths, limitations and broader implications of the three main components of this work - data gathering and labeling, cloud anomaly detection and anomaly-aware mowing event detection.

One of the most significant challenges in addressing cloud contamination in remote sensing is the lack of robust training data that is free from geographical, categorical or labeling biases, as highlighted by several researchers before (Domnich et al., 2021, Shtym et al., 2025, Stubenrauch et al., 2013). This issue is particularly pronounced for time series-based methods, which have remained largely underexplored so far, resulting in an absence of suitable datasets. The progress of this work highlighted that labeling large volumes of point-based data is not only labor-intensive, but also fraught with challenges. For example, minor semi-transparent cloud anomalies are often extremely difficult for human labelers to identify visually in satellite images, due to varying ground and lighting conditions. These factors reduce label quality and in turn also hinder the development and evaluation of cloud anomaly detection models tailored for specific use cases.

A key challenge in developing effective time series-based cloud anomaly detection models lies in the quality of labeled training data, as inconsistencies in cloud contamination characteristics can introduce disruptive variability into the learning process. One possible approach to mitigating this issue is the partial automation of the labeling process, as demonstrated in this thesis through a modified KappaMask workflow. While KappaMask

outperforms other cloud segmentation tools in detecting semi-transparent clouds, it still remains insufficient as a standalone time series cloud labeling solution, occasionally missing cloud-contaminated points or misclassifying thick cloud cover and shadows. To address this, an iterative labeling process using the IsolationForest model was introduced, providing a reasonably good initial label quality that enabled the development of supervised anomaly detection approaches. Although IsolationForest was ultimately outperformed by other supervised models, its role in the initial labeling process was crucial, as those models relied on its outputs for training. The success of such an approach highlights the importance of scalable and high-quality cloud labeling solutions, which remain a bottleneck in advancing time series-based cloud anomaly detection. Future improvements in pixel-based cloud segmentation tools are needed to enhance labeling accuracy across diverse environments, reducing geographical and categorical biases. Such advancements would facilitate the creation of extensive, high-quality labeled datasets, ultimately paving the way for globally adaptable time series-based cloud anomaly detection methods.

The logistic regression model emerged as the most suitable choice for detecting semi-transparent cloud anomalies in Sentinel-2 time series data within the scope of this study. While the model's performance was not ideal on its own, with a user's accuracy (UA) of 0.816 for anomalous points in the test set, it proved sufficiently accurate to provide meaningful input for the downstream mowing event detection model. Notably, many false positives were found to occur immediately before or after actual cloud anomalies in the time series (Appendix 2). This however is not necessarily problematic, as the purpose of the cloud anomaly detection model in this study was to provide information about ambiguous and potentially problematic time series regions, rather than pinpointing individual anomalies with absolute precision. Furthermore, the model effectively identified major anomalies, with the false positives typically corresponding to points with lower anomaly scores (Appendix 3) and also smaller deviations in the actual time series, minimizing their potential impact on the mowing event detection.

The logistic regression model's simplicity, interpretability, and low computational cost made it well-suited for integrating anomaly detection into large-scale time series mowing event detection pipeline. However, the model's inability to account for temporal patterns

represents a notable limitation. Since logistic regression does not inherently incorporate time-based dependencies, this information had to be introduced through extensive feature engineering, including the addition of lagged features, future features and rolling averages for example. While these engineered features generally produced satisfactory results, the approach was less effective at the boundaries of the time series. The lack of sufficient past or future information at the start and end of the series led to many actual anomalies being undetected (Appendix 4). This is exemplified with the fact that approximately 10% of false negatives in the test set corresponded to either the first or the last point of the time series. Moreover, the reliance on temporal features occasionally also resulted in false positives (Appendix 4), particularly at the beginning of time series, where indices like NDVI often exhibit sharp increases during early spring. These limitations underscore the challenges of adapting robust, yet simple model types to applications requiring high temporal sensitivity in the field of remote sensing.

One possible alternative approach that could address some of these limitations involves the use of encoder-decoder networks, which are gaining prominence in time series anomaly detection (Audibert et al., 2020, Malhotra et al., 2016). These models are trained on anomaly-free time series to learn the inherent temporal patterns of the data. When presented with anomalous data points, encoder-decoder networks should generate reconstructions that deviate from the input at anomalous points, resulting in elevated reconstruction errors that can then be classified as anomalies. This approach holds promise for semi-transparent cloud detection, as it naturally incorporates temporal information without requiring extensive feature engineering. Furthermore, encoder-decoder networks could be well-suited to capturing the complex temporal dependencies present in remote sensing time series.

However, despite their potential, encoder-decoder networks also face several major challenges when applied to cloud anomaly detection. One significant issue is the lack of cloud-free time series data, which is essential for training the model to recognize normal patterns. Generating such datasets would likely require some calibration or smoothing techniques, introducing their own potential biases, that could compromise the integrity of the analysis (Cai et al., 2017). Additionally, grassland time series exhibit considerable variability due to differences in management practices, environmental

conditions and historical usage (De Vroey et al., 2022, Reinermann et al., 2020). This diversity complicates the task of training encoder-decoder networks, as the models must generalize across highly heterogeneous data. Constructing a representative training dataset that accounts for this variability, while ensuring sufficient coverage of different grassland types, is already one of the major challenges in developing a reliable mowing event detection model, and would thus be a complex and resource-intensive task in training an encoder-decoder network as well.

The integration of anomaly scores into the mowing event detection framework offered a way to reflect spectral distortions caused by semi-transparent clouds. However, one key observation from this study is that simply including these scores as an additional input feature had a limited impact on model behavior. While small reductions in false positives were observed, this strategy did not meaningfully guide the network in handling cloud-affected time steps. This outcome highlights a broader challenge in uncertainty-aware learning, which is that deep learning models might not inherently interpret auxiliary inputs like anomaly scores as signals of reduced reliability. A likely explanation lies in the nature of semi-transparent cloud effects, meaning they often produce spectral fluctuations that resemble real mowing events. As a result, the model may treat high anomaly scores not as cautionary signals, but as informative features that align with expected event-like patterns. Rather than suppressing predictions in uncertain regions, the model might even reinforce them, inadvertently amplifying cloud-induced noise. This underscores the need for explicit mechanisms, such as uncertainty-aware objectives or loss-based constraints, when attempting to embed uncertainty into a model's reasoning process.

The most effective way to integrate anomaly scores into the model's decision-making turned out to be designing custom loss functions. Unlike simply adding anomaly scores as input features, which did not meaningfully alter the model's behavior, adjusting the loss function ensured that the model actively accounted for uncertainty during training. The weighted cross-entropy (WCE) approach, where anomaly scores influenced the loss assigned to each time step, proved to be particularly effective. By penalizing overconfident predictions in ambiguous regions, this approach encouraged the model to be more conservative when encountering data points with high anomaly scores. The

impact of this modification was not only evident in the improved performance on the cloud-affected test cases but also in the overall gains observed in the cleaner mowing event detection test set. This suggests that integrating uncertainty-aware objectives does more than just mitigate cloud-induced errors - it also enhances the model's ability to generalize more effectively across different observational conditions. Furthermore, the probability distributions of mowing event predictions shifted in a way that aligned with expected uncertainty levels, meaning that in cloud-affected regions the model exhibited lower confidence, reducing the occurrence of overconfident misclassifications. This behavior is consistent with broader trends in uncertainty-aware deep learning, where weighting mechanisms have been shown to improve reliability in conditions with inherent observational noise (Jiang et al., 2021, Kendall and Gal, 2017).

Architectural modifications such as anomaly-based feature scaling and the introduction of a bottleneck layer showed promise in mitigating cloud-induced contamination effects on the optical time series, but yielded less pronounced improvements compared to loss function modifications. While these approaches were designed to help the model suppress cloud-contaminated information at the input level, their full potential may not have been fully realized within the scope of this thesis. A possible explanation is that convolutional layers inherently smooth input variations, meaning that subtle feature adjustments might not have had as strong an influence on learning dynamics as direct loss penalties (Kiranyaz et al., 2021). However, these methods remain an interesting direction for future research, as more refined architectural adaptations - such as dynamic feature gating mechanisms or attention-based uncertainty modeling - could further enhance the model's ability to differentiate between genuine and cloud-induced spectral changes.

Another key observation made during the mowing event detection model enhancement experiments was that event probability post-processing techniques had a much smaller impact than expected. While they refined probability estimates in some cases, they could not fundamentally change model behavior, reinforcing the idea that uncertainty needs to be accounted for earlier in the modeling pipeline rather than after predictions are already made. Even with relatively complex post-processing strategies like conditional isotonic calibration, it proved difficult to meaningfully adjust probability estimates. Particularly when modifications were limited to only probability downscaling options rather than a

more adaptive recalibration. This suggests that if uncertainty-related biases are embedded in the model's outputs, they may not be easily corrected through post-hoc probability transformations alone. These findings have broader implications for uncertainty modeling in remote sensing applications, where approaches relying solely on post-hoc corrections may not be as effective than embedding uncertainty-aware mechanisms directly into model training process.

While the integration of anomaly scores into the model's input and loss function led to improved sensitivity to cloud-related ambiguities, its most notable effect was in reducing false positives and making probability outputs more conservative in uncertain time series regions. This resulted in a mowing event detection model that was less prone to overconfident misclassifications in cloud-affected time series, leading to improvements in both accuracy and interpretability. It is worth noting that the different anomaly integration strategies explored in this thesis are not mutually exclusive and could potentially be combined to further improve model performance. In this study, each approach was tested in isolation to better understand its individual contribution, leaving the exploration of their combined effects as a direction for future work.

Another natural extension of the current work would be to apply the cloud anomaly detection approach to other optical satellite data sources beyond Sentinel-2. This study focused on Sentinel-2 to establish the potential methodology, but the promising results indicate that expansion to other platforms could be a viable next step. In the current mowing event detection setup, Landsat data is also used alongside Sentinel-2, providing complementary observations. While its revisit frequency and spatial resolution are lower compared to Sentinel-2 (Wulder et al., 2022), each observation still remains valuable in sparsely sampled time series. Like Sentinel-2, Landsat data is also affected by semi-transparent cloud contamination, which can distort spectral signals and reduce the reliability of time series analysis. Although Landsat bands differ slightly in their composition and spectral coverage, the relatively simple feature requirements of the cloud anomaly detection model suggest that integration would be feasible with minimal adaptation.

Despite certain limitations and opportunities for further enhancement, this thesis has successfully shown that custom loss functions, tailored to incorporate cloud anomaly

information, can be effectively applied to the complex domain of time series-based agricultural monitoring using remote sensing data. This marks a significant step forward in addressing the challenges of cloud contamination in optical satellite input while maintaining the reliability and robustness required for practical applications.

6 Conclusion

This thesis addressed the challenge of semi-transparent cloud contamination in Sentinel-2 optical time series, which is a key source of uncertainty in remote sensing based monitoring applications such as agricultural event detection. Existing pixel-based cloud detection methods often fail to reliably identify subtle cloud anomalies, resulting in degraded data quality for workflows that rely solely on aggregated time series inputs. To mitigate this, a time series–based anomaly detection framework was proposed to identify cloud-contaminated points and integrate uncertainty into a downstream mowing event detection model.

A labeled training dataset was constructed by selecting Danish grassland parcels across multiple seasons and manually annotating cloud-contaminated points in their Sentinel-2 time series. This process was supported by an iterative labeling strategy using an unsupervised IsolationForest model to identify potential anomalies for verification. The labeled dataset enabled the training and evaluation of multiple different model types for cloud anomaly detection. Logistic regression was ultimately selected for its balance of accuracy, simplicity, and interpretability, as well as its ability to produce well-calibrated anomaly scores suitable for downstream integration.

Multiple strategies were explored to incorporate cloud anomaly information into a 1D-CNN mowing event detection model. These included direct feature augmentation, loss function modifications, model architecture adjustments, and post-processing techniques. The most effective approach involved weighting the training loss by anomaly scores, which substantially improved model reliability by reducing the number of false positives and promoting more cautious predictions in ambiguous time series segments.

The findings demonstrate that uncertainty-aware learning, when explicitly integrated into model training, can significantly enhance the robustness of remote sensing applications affected by cloud contamination. The time series-based anomaly detection framework developed in this thesis offers a scalable and practical solution for modern Earth observation pipelines, where the quality of pixel-level cloud masking alone is insufficient, and maintaining temporal data integrity is critical for operational performance.

References

F. Alonso-Sarria, C. Valdivieso-Ros, and F. Gomariz-Castillo. Imagery Time Series Cloud Removal and Classification Using Long Short Term Memory Neural Networks. *Remote Sensing*, 16(12):2150, Jan. 2024. doi: 10.3390/rs16122150. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga. USAD: UnSupervised Anomaly Detection on Multivariate Time Series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 3395–3404, New York, NY, USA, Aug. 2020. Association for Computing Machinery. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403392.

A. Bakumenko and A. Elragal. Detecting Anomalies in Financial Data Using Machine Learning Algorithms. *Systems*, 10(5):130, Oct. 2022. doi: 10.3390/systems10050130. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

M. Belgiu and L. Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114: 24–31, Apr. 2016. doi: 10.1016/j.isprsjprs.2016.01.011.

E. Benami, Z. Jin, M. R. Carter, A. Ghosh, R. J. Hijmans, A. Hobbs, B. Kenduiywo, and D. B. Lobell. Uniting remote sensing, crop modelling and economics for agricultural risk management. *Nature Reviews Earth & Environment*, 2(2):140–159, Feb. 2021. doi: 10.1038/s43017-020-00122-y. Publisher: Nature Publishing Group.

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, Aug. 2013. doi: 10.1109/TPAMI.2013.50. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

M. Berger, J. Moreno, J. A. Johannessen, P. F. Levelt, and R. F. Hanssen. ESA's sentinel missions in support of Earth system science. *Remote Sensing of Environment*, 120: 84–90, May 2012. doi: 10.1016/j.rse.2011.07.023.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. doi: 10.1023/A:1010933404324.

Z. Cai, P. Jönsson, H. Jin, and L. Eklundh. Performance of Smoothing Methods for Reconstructing NDVI Time-Series and Estimating Vegetation Phenology from MODIS Data. *Remote Sensing*, 9(12):1271, Dec. 2017. doi: 10.3390/rs9121271. Number: 12

Publisher: Multidisciplinary Digital Publishing Institute.

J. B. Campbell and R. H. Wynne. *Introduction to Remote Sensing, Fifth Edition*. Guilford Press, June 2011. ISBN 978-1-60918-177-2. Google-Books-ID: NkLmDjSS8TsC.

M. Campos-Taberner, F. J. García-Haro, B. Martínez, E. Izquierdo-Verdiguier, C. Atzberger, G. Camps-Valls, and M. A. Gilabert. Understanding deep learning in land use classification based on Sentinel-2 time series. *Scientific Reports*, 10(1): 17188, Oct. 2020. doi: 10.1038/s41598-020-74215-5. Publisher: Nature Publishing Group.

J. Choi, D. Seo, J. Jung, Y. Han, J. Oh, and C. Lee. Cloud Detection Using a UNet3+ Model with a Hybrid Swin Transformer and EfficientNet (UNet3+STE) for Very-High-Resolution Satellite Imagery. *Remote Sensing*, 16(20):3880, Jan. 2024. doi: 10.3390/rs16203880. Number: 20 Publisher: Multidisciplinary Digital Publishing Institute.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sept. 1995. doi: 10.1007/BF00994018.

X. Cui, X. Guo, Y. Wang, X. Wang, W. Zhu, J. Shi, C. Lin, and X. Gao. Application of remote sensing to water environmental processes under a changing climate. *Journal of Hydrology*, 574:892–902, July 2019. doi: 10.1016/j.jhydrol.2019.04.078.

K. W. Cunningham, J. Little, and M. N. Montgomery. Satellite Remote Sensing & Carbon Market Transparency. *Advances in Forestry Letters*, 2(3), 2013.

M. De Vroey, L. de Vendictis, M. Zavagli, S. Bontemps, D. Heymans, J. Radoux, B. Koetz, and P. Defourny. Mowing detection using Sentinel-1 and Sentinel-2 time series for large scale grassland monitoring. *Remote Sensing of Environment*, 280: 113145, Oct. 2022. doi: 10.1016/j.rse.2022.113145.

L. Di and E. Yu. Challenges and Opportunities in the Remote Sensing Big Data. In L. Di and E. Yu, editors, *Remote Sensing Big Data*, pages 281–291. Springer International Publishing, Cham, 2023. ISBN 978-3-031-33932-5. doi: 10.1007/978-3-031-33932-5_18.

M. Domnich, I. Sünter, H. Trofimov, O. Wold, F. Harun, A. Kostiukhin, M. Järveoja, M. Veske, T. Tamm, K. Voormansik, A. Olesk, V. Boccia, N. Longepe, and E. G. Cadau. KappaMask: AI-Based Cloudmask Processor for Sentinel-2. *Remote Sensing*, 13(20): 4100, Jan. 2021. doi: 10.3390/rs13204100. Number: 20 Publisher: Multidisciplinary

Digital Publishing Institute.

S. Dutta, S. Dey, and L. D. Girolamo. Chapter 9 - Remote Sensing of Cloudiness: Challenges and Way Forward. In A. Kumar Singh and S. Tiwari, editors, *Atmospheric Remote Sensing*, Earth Observation, pages 157–170. Elsevier, Jan. 2023. ISBN 978-0-323-99262-6. doi: 10.1016/B978-0-323-99262-6.00018-3.

A. Ehteshami Bejnordi and R. Krestel. Dynamic Channel and Layer Gating in Convolutional Neural Networks. In U. Schmid, F. Klügl, and D. Wolter, editors, *KI 2020: Advances in Artificial Intelligence*, pages 33–45, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58285-2. doi: 10.1007/978-3-030-58285-2_3.

A. F. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, ODD '13, pages 16–21, New York, NY, USA, Aug. 2013. Association for Computing Machinery. ISBN 978-1-4503-2335-2. doi: 10.1145/2500853.2500858.

P. Garg. *Remote Sensing: Theory and Applications*. Walter de Gruyter GmbH & Co KG, June 2024. ISBN 978-1-5015-2284-0. Google-Books-ID: um4NEQAAQBAJ.

J. Gawlikowski, P. Ebel, M. Schmitt, and X. X. Zhu. Explaining the Effects of Clouds on Remote Sensing Scene Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:9976–9986, 2022. doi: 10.1109/JSTARS.2022.3221788. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, July 2017.

R. Gupta and S. J. Nanda. Cloud detection in satellite images with classical and deep neural network approach: A review. *Multimedia Tools and Applications*, 81(22): 31847–31880, Sept. 2022. doi: 10.1007/s11042-022-12078-w.

D. Hand and K. Yu. Idiot's Bayes: Not So Stupid after All? *International Statistical Review*, 69:385–398, May 2007. doi: 10.1111/j.1751-5823.2001.tb00465.x.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006. doi: 10.1126/science.1127647. Publisher: American Association for the Advancement of Science.

Y. Ho and S. Wookey. The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. *IEEE Access*, 8:4806–4813, 2020. doi: 10.1109/ACCESS.2019.2962617. Conference Name: IEEE Access.

A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, Feb. 1970. doi: 10.1080/00401706.1970.10488634. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1970.10488634.

C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A Practical Guide to Support Vector Classification. 2010.

S.-H. Hwang, M. Kim, and S. E. Whang. RC-Mixup: A Data Augmentation Strategy against Noisy Data for Regression Tasks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1155–1165, Barcelona Spain, Aug. 2024. ACM. ISBN 9798400704901. doi: 10.1145/3637528.3671993.

X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, Mar. 2012. doi: 10.1136/amiajnl-2011-000291.

Z. Jiang, J. Araki, H. Ding, and G. Neubig. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, Sept. 2021. doi: 10.1162/tacl_a_00407.

P. P. Joshi, R. H. Wynne, and V. A. Thomas. Cloud detection algorithm using SVM with SWIR2 and tasseled cap applied to Landsat 8. *International Journal of Applied Earth Observation and Geoinformation*, 82:101898, Oct. 2019. doi: 10.1016/j.jag.2019.101898.

D. W. H. Jr, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Feb. 2013. ISBN 978-1-118-54835-6. Google-Books-ID: bRoxQBIZRd4C.

A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

B. K. Kenduiywo, M. R. Carter, A. Ghosh, and R. J. Hijmans. Evaluating the quality of remote sensing products for agricultural index insurance. *PLOS ONE*, 16(10): e0258215, Oct. 2021. doi: 10.1371/journal.pone.0258215. Publisher: Public Library

of Science.

S. Khanal, K. Kc, J. P. Fulton, S. Shearer, and E. Ozkan. Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities. *Remote Sensing*, 12(22): 3783, Jan. 2020. doi: 10.3390/rs12223783. Number: 22 Publisher: Multidisciplinary Digital Publishing Institute.

S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman. 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, Apr. 2021. doi: 10.1016/j.ymssp.2020.107398.

V. Komisarenko, K. Voormansik, R. Elshawi, and S. Sakr. Exploiting time series of Sentinel-1 and Sentinel-2 to detect grassland mowing events using deep learning with reject region. *Scientific Reports*, 12(1):983, Jan. 2022. doi: 10.1038/s41598-022-04932-6. Publisher: Nature Publishing Group.

V. Kristollari and V. Karathanassi. *Convolutional neural networks for detecting challenging cases in cloud masking using Sentinel-2 imagery*. Aug. 2020. doi: 10.1117/12.2571111.

J. Li, Y. Pei, S. Zhao, R. Xiao, X. Sang, and C. Zhang. A Review of Remote Sensing for Environmental Monitoring in China. *Remote Sensing*, 12(7):1130, Jan. 2020. doi: 10.3390/rs12071130. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

Z. Li, H. Shen, Q. Weng, Y. Zhang, P. Dou, and L. Zhang. Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:89–108, June 2022. doi: 10.1016/j.isprsjprs.2022.03.020.

K. Liang, G. Yang, Y. Zuo, J. Chen, W. Sun, X. Meng, and B. Chen. A Novel Method for Cloud and Cloud Shadow Detection Based on the Maximum and Minimum Values of Sentinel-2 Time Series Images. *Remote Sensing*, 16(8):1392, Jan. 2024. doi: 10.3390/rs16081392. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

T. Lillesand, R. W. Kiefer, and J. Chipman. *Remote Sensing and Image Interpretation*. John Wiley & Sons, Feb. 2015. ISBN 978-1-118-34328-9. Google-Books-ID: AFHDCAAAQBAJ.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal Loss for Dense Object

Detection. pages 2980–2988, 2017.

K. M. Lind and S. M. Pedersen. Perspectives of Precision Agriculture in a Broader Policy Context. In S. M. Pedersen and K. M. Lind, editors, *Precision Agriculture: Technology and Economic Perspectives*, pages 251–266. Springer International Publishing, Cham, 2017. ISBN 978-3-319-68715-5. doi: 10.1007/978-3-319-68715-5_12.

C.-a. Liu, Z.-x. Chen, Y. Shao, J.-s. Chen, T. Hasi, and H.-z. Pan. Research advances of SAR remote sensing for agriculture applications: A review. *Journal of Integrative Agriculture*, 18(3):506–525, Mar. 2019. doi: 10.1016/S2095-3119(18)62016-7.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data*, 6(1):3:1–3:39, Mar. 2012. doi: 10.1145/2133360.2133363.

Y. Liu, P. Rao, W. Zhou, B. Singh, A. K. Srivastava, S. P. Poonia, D. V. Berkel, and M. Jain. Using Sentinel-1, Sentinel-2, and Planet satellite data to map field-level tillage practices in smallholder systems. *PLOS ONE*, 17(11):e0277425, Nov. 2022. doi: 10.1371/journal.pone.0277425. Publisher: Public Library of Science.

Y. Ma. Advancing Disaster Management through Remote Sensing: Applications, Challenges, and Future Prospects. *Science and Technology of Engineering, Chemistry and Environmental Protection*, 1(9), Oct. 2024. doi: 10.61173/ye1me637. Number: 9.

S. Mahajan and B. Fataniya. Cloud detection methodologies: variants and development—a review. *Complex & Intelligent Systems*, 6(2):251–261, July 2020. doi: 10.1007/s40747-019-00128-0.

P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection, July 2016. arXiv:1607.00148 [cs].

A. E. Maxwell, T. A. Warner, and F. Fang. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39 (9):2784–2817, May 2018. doi: 10.1080/01431161.2018.1433343. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431161.2018.1433343.

T. Miyato, A. M. Dai, and I. Goodfellow. Adversarial Training Methods for Semi-Supervised Text Classification, Nov. 2021. arXiv:1605.07725 [stat].

S. Mohajerani and P. Saeedi. Cloud-Net: An end-to-end Cloud Detection Algorithm for Landsat 8 Imagery, Jan. 2019. arXiv:1901.10077.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learn-

ing. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 625–632, New York, NY, USA, Aug. 2005. Association for Computing Machinery. ISBN 978-1-59593-180-1. doi: 10.1145/1102351.1102430.

S. Oehmcke, T.-H. K. Chen, A. V. Prishchepov, and F. Gieseke. Creating cloud-free satellite imagery from image time series with deep learning. In *Proceedings of the 9th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, BigSpatial '20, pages 1–10, New York, NY, USA, Nov. 2020. Association for Computing Machinery. ISBN 978-1-4503-8162-8. doi: 10.1145/3423336.3429345.

M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, Jan. 2005. doi: 10.1080/01431160412331269698. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431160412331269698.

E. Perez, F. Strub, H. d. Vries, V. Dumoulin, and A. Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11671. Number: 1.

L. Pipia, S. Belda, B. Franch, and J. Verrelst. Trends in Satellite Sensors and Image Time Series Processing Methods for Crop Phenology Monitoring. In D. D. Bochtis, M. Lampridi, G. P. Petropoulos, Y. Ampatzidis, and P. Pardalos, editors, *Information and Communication Technologies for Agriculture—Theme I: Sensors*, pages 199–231. Springer International Publishing, Cham, 2022. ISBN 978-3-030-84144-7. doi: 10.1007/978-3-030-84144-7_8.

V. H. R. Prudente, V. S. Martins, D. C. Vieira, N. R. d. F. e. Silva, M. Adami, and I. D. Sanches. Limitations of cloud cover for optical remote sensing of agricultural areas across South America. *Remote Sensing Applications: Society and Environment*, 20: 100414, Nov. 2020. doi: 10.1016/j.rsase.2020.100414.

S. Qiu, Z. Zhu, and B. He. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sensing of Environment*, 231:111205, Sept. 2019. doi: 10.1016/j.rse.2019.05.024.

M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, Feb. 2019. doi: 10.1038/s41586-019-0912-1. Publisher: Nature Publishing Group.

S. Reinermann, S. Asam, and C. Kuenzer. Remote Sensing of Grassland Production and

Management—A Review. *Remote Sensing*, 12(12):1949, Jan. 2020. doi: 10.3390/rs12121949. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

D. Samariya and A. Thakkar. A Comprehensive Survey of Anomaly Detection Algorithms. *Annals of Data Science*, 10(3):829–850, June 2023. doi: 10.1007/s40745-021-00362-9.

A. H. Sanchez, M. C. A. Picoli, G. Camara, P. R. Andrade, M. E. D. Chaves, S. Lechler, A. R. Soares, R. F. B. Marujo, R. E. O. Simões, K. R. Ferreira, and G. R. Queiroz. Comparison of Cloud Cover Detection Algorithms on Sentinel–2 Images of the Amazon Tropical Forest. *Remote Sensing*, 12(8):1284, Jan. 2020. doi: 10.3390/rs12081284. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

S. Schmidl, P. Wenig, and T. Papenbrock. Anomaly detection in time series: a comprehensive evaluation. *Proc. VLDB Endow.*, 15(9):1779–1797, May 2022. doi: 10.14778/3538598.3538602.

D. W. Scott. Feasibility of multivariate density estimates. *Biometrika*, 78(1):197–205, Mar. 1991. doi: 10.1093/biomet/78.1.197.

K. Shaukat, T. M. Alam, S. Luo, S. Shabbir, I. A. Hameed, J. Li, S. K. Abbas, and U. Javed. A Review of Time-Series Anomaly Detection Techniques: A Step to Future Perspectives. In K. Arai, editor, *Advances in Information and Communication*, pages 865–877, Cham, 2021. Springer International Publishing. ISBN 978-3-030-73100-7. doi: 10.1007/978-3-030-73100-7_60.

T. Shi and H. Xu. Derivation of Tasseled Cap Transformation Coefficients for Sentinel-2 MSI At-Sensor Reflectance Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(10):4038–4048, Oct. 2019. doi: 10.1109/JSTARS.2019.2938388. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

T. Shtym, O. Wold, H. Trofimov, I. Sünter, A. Kostiukhin, M. Veske, K. Voormansik, N. Longepe, and A. Francis. KappaMaskv2: Going Global. Submitted for publication, 2025.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, New York, Feb. 2018. ISBN 978-1-315-14091-9. doi: 10.1201/9781315140919.

C. J. Stubenrauch, W. B. Rossow, S. Kinne, S. Ackerman, G. Cesana, H. Chepfer, L. D. Girolamo, B. Getzewich, A. Guignard, A. Heidinger, B. C. Maddux, W. P. Menzel,

P. Minnis, C. Pearl, S. Platnick, C. Poulsen, J. Riedi, S. Sun-Mack, A. Walther, D. Winker, S. Zeng, and G. Zhao. Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel. July 2013. doi: 10.1175/BAMS-D-12-00117.1. Section: Bulletin of the American Meteorological Society.

M. Tabassum, S. Mahmood, A. Bukhari, B. Alshemaimri, A. Daud, and F. Khalique. Anomaly-based threat detection in smart health using machine learning. *BMC Medical Informatics and Decision Making*, 24(1):347, Nov. 2024. doi: 10.1186/s12911-024-02760-4.

W. Tang, G. Long, L. Liu, T. Zhou, J. Jiang, and M. Blumenstein. *Rethinking 1D-CNN for Time Series Classification: A Stronger Baseline*. Feb. 2020. doi: 10.48550/arXiv.2002.10061.

R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, Jan. 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Z. Tsiropoulos, G. Carli, E. Pignatti, and S. Fountas. Future Perspectives of Farm Management Information Systems. In S. M. Pedersen and K. M. Lind, editors, *Precision Agriculture: Technology and Economic Perspectives*, pages 181–200. Springer International Publishing, Cham, 2017. ISBN 978-3-319-68715-5. doi: 10.1007/978-3-319-68715-5_9.

M. A. Wulder, D. P. Roy, V. C. Radeloff, T. R. Loveland, M. C. Anderson, D. M. Johnson, S. Healey, Z. Zhu, T. A. Scambos, N. Pahlevan, M. Hansen, N. Gorelick, C. J. Crawford, J. G. Masek, T. Hermosilla, J. C. White, A. S. Belward, C. Schaaf, C. E. Woodcock, J. L. Huntington, L. Lymburner, P. Hostert, F. Gao, A. Lyapustin, J.-F. Pekel, P. Strobl, and B. D. Cook. Fifty years of Landsat science and impacts. *Remote Sensing of Environment*, 280:113195, Oct. 2022. doi: 10.1016/j.rse.2022.113195.

C. Xu, X. Du, X. Fan, Z. Yan, X. Kang, J. Zhu, and Z. Hu. A Modular Remote Sensing Big Data Framework. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. doi: 10.1109/TGRS.2021.3100601. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

J. Yang, P. Gong, R. Fu, M. Zhang, J. Chen, S. Liang, B. Xu, J. Shi, and R. Dickinson. The role of satellite remote sensing in climate change studies. *Nature Climate Change*,

3(10):875–883, Oct. 2013. doi: 10.1038/nclimate1908. Publisher: Nature Publishing Group.

Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, May 2020. doi: 10.1016/j.rse.2020.111716.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 694–699, New York, NY, USA, July 2002. Association for Computing Machinery. ISBN 978-1-58113-567-1. doi: 10.1145/775047.775151.

V. Zekoll. *Cloud shadow detection and removal for high spatial resolution optical satellite data*. thesis, Universität Osnabrück, 2023.

V. Zekoll, M. Main-Knorn, K. Alonso, J. Louis, D. Frantz, R. Richter, and B. Pflug. Comparison of Masking Algorithms for Sentinel-2 Imagery. *Remote Sensing*, 13(1): 137, Jan. 2021. doi: 10.3390/rs13010137. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

H. Zhai, H. Zhang, L. Zhang, and P. Li. Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 144:235–253, Oct. 2018. doi: 10.1016/j.isprsjprs.2018.07.006.

C. Zhang, H. Li, H. Shen, H. He, Y. Liu, P. Yi, and L. Xu. A General Thin Cloud Correction Method Combining Statistical Information and a Scattering Model for Visible and Near-Infrared Satellite Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023. doi: 10.1109/TGRS.2023.3296151. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

H. Zhang, Q. Huang, H. Zhai, and L. Zhang. Multi-temporal cloud detection based on robust PCA for optical remote sensing imagery. *Computers and Electronics in Agriculture*, 188:106342, Sept. 2021. doi: 10.1016/j.compag.2021.106342.

K. Zhang, P. Li, and J. Wang. A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Models, Comparisons and Future Directions. *Remote Sensing*, 16(21):4113, Jan. 2024. doi: 10.3390/rs16214113. Number: 21 Publisher:

Multidisciplinary Digital Publishing Institute.

L. Zhang and L. Zhang. Artificial Intelligence for Remote Sensing Data Analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):270–294, June 2022. doi: 10.1109/MGRS.2022.3145854. Conference Name: IEEE Geoscience and Remote Sensing Magazine.

Z. Zhang and M. Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Z. Zhu, S. Wang, and C. E. Woodcock. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sensing of Environment*, 159:269–277, Mar. 2015. doi: 10.1016/j.rse.2014.12.014.

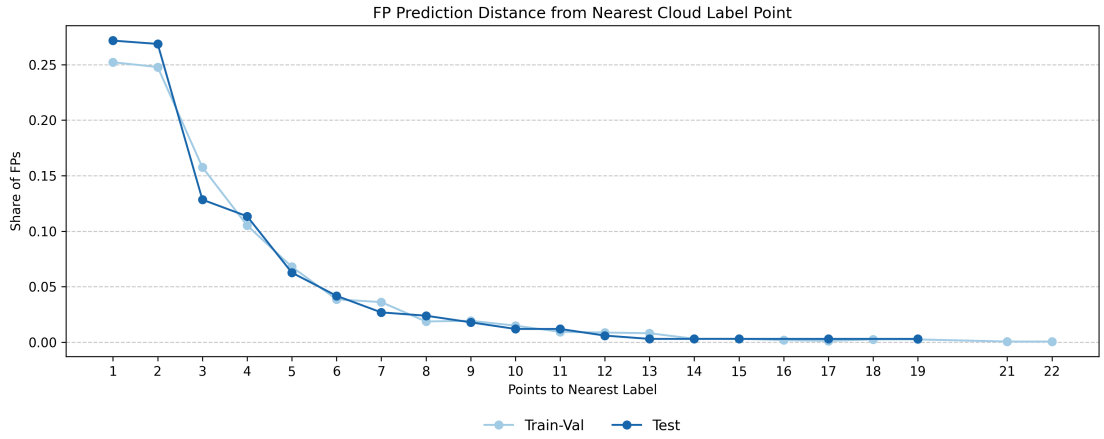
Appendix

I. Tables

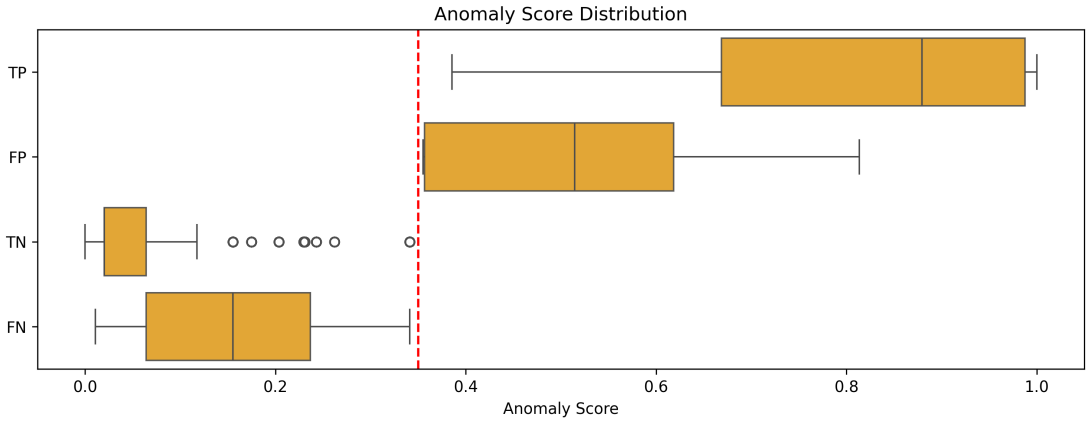
Model Type	Parameters Tested
IsolationForest	`n_estimators`: 75, **100**, 125 `max_samples`: 0.5, 0.7, **0.9** `max_features`: 0.8, **1.0** `contamination_rate`: 0.1, **0.1125**, 0.125 `bootstrap`: TRUE, **FALSE** `optical_only`: **0**, 1
RandomForest	`n_estimators`: 250, **300**, 350 `max_depth`: 5, **7**, 9 `min_samples_split`: **15**, 30, 45 `decision_threshold`: $\{0.25 + 0.05 \cdot n \mid n = 0, 1, 2, \ldots, 10\}$, $\mathbf{0.3}$ `optical_only`: 0, **1**
Logistic Regression	`reg_type`: **l1**, l2 `reg_strength`: 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, $\mathbf{10^{2}}$, 10^{3} `decision_threshold`: $\{0.25 + 0.05 \cdot n \mid n = 0, 1, 2, \ldots, 10\}$, $\mathbf{0.3}$ `optical_only`: 0, **1**
Support Vector Machine	`kernel`: linear, **rbf** `gamma`: **scale**, auto `reg_strength`: $\mathbf{10^{-1}}$, 10^{0}, 10^{1} `decision_threshold`: $\{0.2 + 0.05 \cdot n \mid n = 0, 1, 2, \ldots, 10\}$, $\mathbf{0.4}$ `optical_only`: 0, **1**
Kernel Density Estimation	`kernel`: **gaussian**, tophat, epanechnikov `bandwidth`: 0.1, 0.5, **1.0**, 2.0, 5.0 `density_percentile`: $\{0 + 0.025 \cdot n \mid n = 0, 1, 2, \ldots, 15\}$, $\mathbf{0.25}$ `optical_only`: 0, **1**

Appendix 1. Grid search parameters tested for each model type during the semi-transparent cloud anomaly detection model training. Best parameter set highlighted in bold.

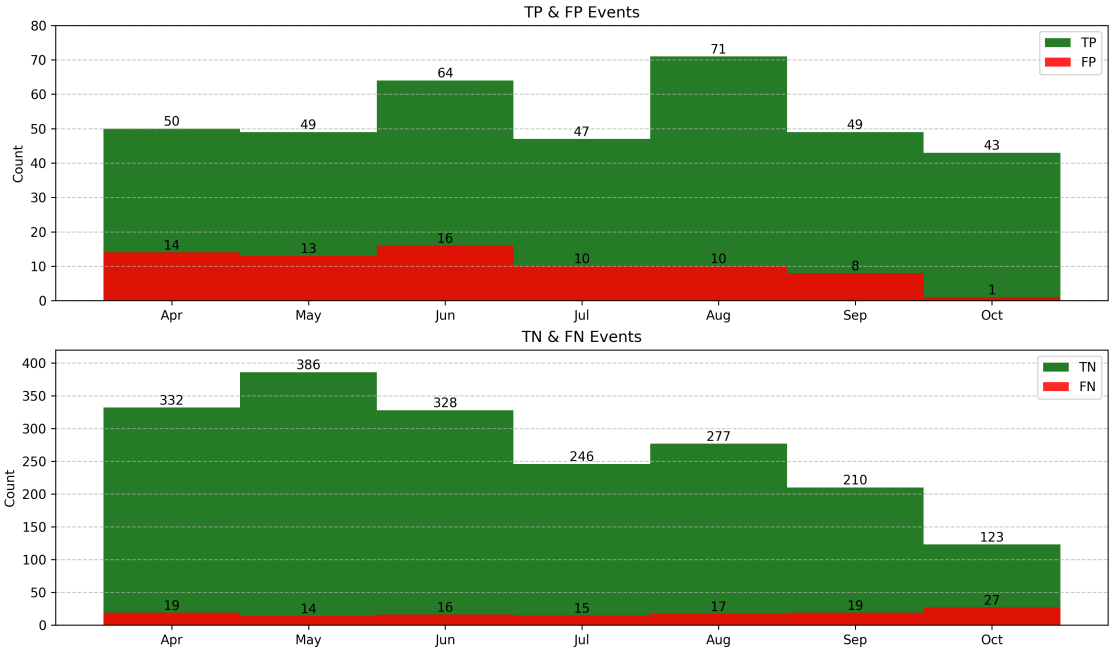
II. Plots



Appendix 2. Distribution of false positive (FP) predictions of the logistic regression model based on their distance, in terms of the number of original data points before linear imputation, from the nearest optical point with a cloud label.



Appendix 3. Boxplot of test set anomaly score distributions for true positive (TP), false positive (FP), true negative (TN), and false negative (FN) classifications. The red dashed line at 0.35 represents the calibrated logistic regression model's anomaly score threshold. Kruskal-Wallis test revealed significant differences among classification groups ($p \ll 0.001$), with post-hoc Dunn's tests confirming a significant difference between TN and FN distributions ($p \ll 0.001$), where false negatives had scores closer to the threshold. Although the FP and TP distributions did not differ significantly, the median for TP was 32.3 percentage points higher than for FP (0.879 and 0.556, respectively).



Appendix 4. Cloud anomaly detection test set distribution of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) with the logistic regression model. The top panel illustrates TP and FP counts, while the bottom panel highlights TN and FN counts across months. The results demonstrate the model's tendency towards reduced accuracy at the boundaries of the time series, particularly in early spring and late fall, where the lack of sufficient temporal context contributes to increased rate of misclassifications.

III. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Karl Hendrik Tamkivi**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Enhancing Mowing Event Detection by Mitigating Semi-Transparent Cloud Anomalies in Optical Satellite Image Time Series,

   supervised by Viacheslav Komisarenko and Tetiana Shtym.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Karl Hendrik Tamkivi
15.05.2025