Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering

Seyedramin Mortazavi

Ecosystem classification using machine learning

Master's thesis in Simulation and Visualization
Supervisor: Agus Hasan
Co-supervisor: Arron Wilde Tippett
January 2022

NTNU
Norwegian University of
Science and Technology

Seyedramin Mortazavi

Ecosystem classification using machine learning

Master's thesis in Simulation and Visualization
Supervisor: Agus Hasan
Co-supervisor: Arron Wilde Tippett
January 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering

NTNU
Norwegian University of
Science and Technology

ABSTRACT

Ecosystems are a crucial part of Earth and they provide essential resources and also affect human health in various ways. The planetary boundaries framework proposed by a team of scientists at at Stockholm Resilience Centre; it rooted in Earth system science, identifies nine pivotal processes crucial for maintaining the stability and resilience of the planet. According to the boundaries 6 boundaries have crossed by 2023. Europe has introduced a nature restoration law which sets targets for restoring ecosystems. In order to restore an ecosystem first we have to assess its health. Remote sensing can be used for the detection of ecosystems and their changes during time in order to assess the change in its health. Satellites can be used to get the data by measuring the amount of light reflected from the surface of the earth in various frequencies. Machine learning methods have been a popular tool in remote sensing helping with the detection of different ecosystems and studying them. In this project machine learning methods of Random Forests and SVM were used to classify land types in Norway. This algorithms were implemented on the data from sentinel-2 satellite, and Google earth engine platform was used for processing the data, training the models and visualization. The random forests model yielded the best result, with 90.8% accuracy for validation data and 89.6% for training.

SAMMENDRAG

Økosystemer er en avgjørende del av jorden, og de gir essensielle ressurser samtidig som de påvirker menneskers helse på ulike måter. rammeverket for planetariske grenser, foreslått av et team av forskere ved Stockholm Resilience Centre, forankret i jordens systemvitenskap, identifiserer ni avgjørende prosesser som er avgjørende for å opprettholde planetens stabilitet og motstandskraft. Ifølge rammeverket har 6 grenser blitt krysset innen 2023. Europa har innført en lov om naturgjenoppretting som fastsetter mål for å gjenopprette økosystemer. For å gjenopprette et økosystem må vi først vurdere helsen. Fjernmåling kan brukes til å oppdage økosystemer og deres endringer over tid for å vurdere endringen i helsen deres. Satellitter kan brukes til å skaffe data ved å måle mengden lys som reflekteres fra jordoverflaten i ulike frekvenser. Maskinlæringsmetoder har vært et populært verktøy innen fjernmåling for å hjelpe med å oppdage ulike økosystemer og studere dem. I dette prosjektet ble maskinlæringsmetodene Random Forests og SVM brukt til å klassifisere landtyper i Norge. Disse algoritmene ble implementert på data fra Sentinel-2-satellitten, og Google Earth Engine-plattformen ble brukt til å behandle data, trene modellene og visualisere resultatene. Random Forests-modellen ga det beste resultatet, med en nøyaktighet på 90,8% for valideringsdata og 89,6% for treningsdata.

PREFACE

I want to thank everyone who helped me finish my master's thesis. This project involved a lot of research, thinking, and looking back on things, and I couldn't have completed it without the support and guidance of many people. Firstly, a big thank you to my supervisors. They were always there for me, sharing their knowledge and encouraging me. Their advice helped shape my thesis into its final version. And, of course, a shoutout to the teachers and professors at NTNU for creating a great academic environment and providing resources that really helped with my research. I also appreciate my family and friends for supporting me throughout this tough journey. Their encouragement and belief in me kept me going. I hope this thesis adds something valuable to the field of study and inspires future research. I'm grateful and honored to have had the chance to do this research, and I want to thank everyone who supported me along the way.

Seyedramin Mortazavi

CONTENTS

LIST OF FIGURES

CHAPTER

ONE

INTRODUCTION

1.1 Ecosystems

Odum [1] defines ecosystems as complex, interrelated systems that comprise both living and nonliving components. They may self-sustain and are impacted by a range of circumstances, including human activity [2]. The term ”ecosystem” has replaced numerous other terminologies to characterize biotic and inorganic interactions in nature over time [3]. Today, human activities affect ecosystems, causing some areas to degrade. It’s vital to protect these ecosystems to keep them healthy, benefiting all living things, including humans.

1.1.1 Planetary boundaries

A team of scientists at Stockholm Resilience Centre introduced the planetary boundaries. The planetary boundaries framework, rooted in Earth system science, identifies nine pivotal processes crucial for maintaining the stability and resilience of the planet. Human actions have greatly disturbed important Earth processes. To address this, limits have been set to control human impact and maintain Earth in a stable state similar to the last 10,000 years. This helps protect global environmental functions and life-support systems as seen throughout human history.
The framework addresses the challenge of interconnected environmental issues, advocating for a comprehensive approach to human caused impacts on the Earth system. While current environmental concerns are often treated as separate problems (e.g., climate change, biodiversity loss, pollution), the planetary boundaries framework underscores the nonlinear interactions and aggregate effects of these issues on the overall state of the Earth system. Human activities on a planetary scale act as forces that, when processed through interactions and feedback within the Earth system, contribute to its ongoing evolution. The framework establishes limits on the impact of the anthroposphere by defining a scientifically based safe operating space for humanity, safeguarding Earth’s interglacial state and resilience.[4]
The notion of planetary boundaries, first introduced by carries substantial implications for global sustainability policy, as highlighted by Steffen[5]. However, the identification and application of this concept pose considerable challenges, particularly in the realms of governance and democratic legitimacy, as discussed by Biermann [6] and Pickering [7]. In response to these challenges, Häyhä [8] suggests a framework that translates planetary boundaries into targets at the national level, taking into account biophysical, socio-economic, and ethical dimensions. This proposed approach seeks to bridge the gap between global limits and local decision-making, fostering a more resilient and equitable implementation of the planetary boundaries concept.

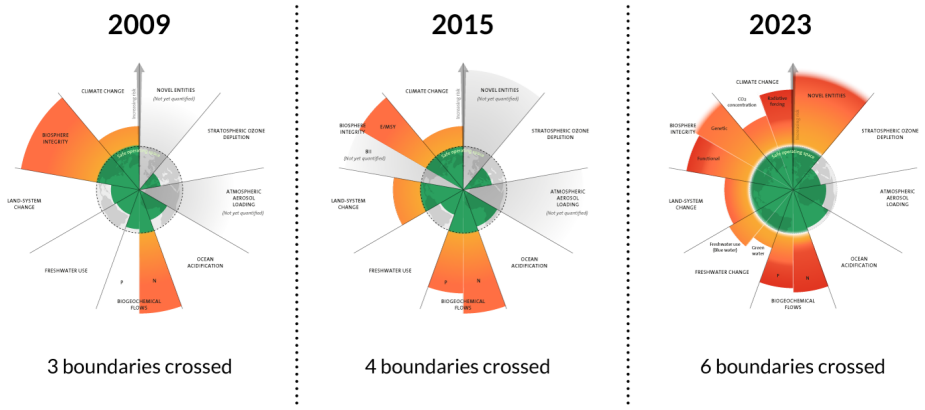


Figure 1.1.1: Planetary boundaries and its evolution from 2009 to 2023 [4]

1.1.2 Ecosystem accounting

Ecosystem accounting, a practical tool for evaluating ecosystem capital, has been developed to incorporate measurements of ecosystem services and assets into an accounting structure [9]. This method involves distinguishing between the capacity and flow of ecosystem services and using spatially explicit models to analyze various hydrological ecosystem services [10]. The System of Environmental Economic Accounting-Ecosystem Accounting (SEEA-EA) framework has been applied at the catchment scale to create ecosystem extent and condition accounts, utilizing nationally available datasets [11]. Additionally, the valuation and mapping of ecosystem services, including provisioning, regulating, and cultural services, have been explored to support the development of ecosystem accounts [12]).

1.1.3 Nature restoration law

The European Commission has introduced a Nature Restoration Law, a part of the EU Biodiversity Strategy. This law aims to set targets for restoring ecosystems, particularly those with high carbon capture potential and disaster prevention. Given that over 80% of European habitats are in poor condition, restoring various ecosystems is seen as essential for enhancing biodiversity, securing ecosystem services, mitigating global warming, and improving Europe's resilience to natural disasters and food security risks [13].
The implementation of the Nature Restoration Law will be monitored through two main indicators:

1. Restoration Measures: Member States are expected to implement restoration and re-establishment activities to facilitate ecosystem recovery.
2. Ecosystem Condition: Monitoring the condition and conservation status of ecosystems at the national or (biogeographic) regional level, assessing their trends against relevant baselines.

The definition of good ecosystem condition and restoration measures varies for different ecosystems. For other ecosystems lacking fully developed data and monitoring methods, the proposal suggests establishing an EU-wide methodology to assess their conditions. This allows for the later setting of specific targets and baselines. Information on indicators related to ecosystem condition already exists for some ecosystems, like urban, agricultural, and forest land, through schemes like Forest Europe or data collected by the European Environment Agency and the Commission via Copernicus [13].

1.1.4 Ecosystem health

Nature provides essential resources and recreation for humans, but global ecosystems are at risk due to human activities and climate change. To ensure stable and sustainable benefits for societies, it's crucial to maintain healthy ecosystems. Assessing and monitoring ecosystem health is key for early detection of environmental problems and their causes, serving as a vital step in ecological conservation.
Ecosystem health assessment (EHA), a component of environmental management since the late 1980s, integrates the concepts of ecosystem and health science. Initially focused on animal and plant health, EHA now considers ecosystems as complex systems, emphasizing the interplay between community processes and the physical environment. The definition evolved to measure system resilience, organization, and vigor.
Traditional EHA relies on field data and models, limiting widespread application and spatial-temporal specificity. Recognizing the need for spatial heterogeneity understanding, remote sensing emerges as a powerful tool. Remote sensing data can assess ecosystem health across large areas, detailing indicators like productivity, species richness, and resilience. Existing studies focus on single attributes, but a comprehensive approach integrating vigor, organization, and resilience is urgently required. Building such a system demands collaboration between remote sensing specialists and ecologists. While reviews exist on remote sensing applications in land cover, biodiversity, and ecosystem services, none have specifically addressed the opportunities and challenges of developing a comprehensive remote sensing-based, spatially explicit EHA and monitoring system [14].
The initial stage of this assessment involves the essential ability to easily identify and categorize various biomes and ecosystems. The study on land types and ecosystems is done using remote sensing data.

1.2 Remote sensing

Remote sensing is the acquisition of information about an object or phenomenon without making physical contact with the object. This is typically achieved through the use of sensors, such as those on satellites or unmanned aerial vehicles (UAVs), to collect data from a distance[15] .
There are two broad type of satellite sensors, passive and active. Passive satellite sensors detect natural emissions or reflections from Earth's surface without emitting their own signals, relying on ambient energy sources like sunlight. Examples include optical and thermal infrared sensors. In contrast, active satellite sensors emit their own signals, such as microwaves or lasers, and measure the return signal.
Remote sensing using satellite images involves the capture and analysis of data collected by orbiting satellites equipped with various sensors. These sensors detect electromagnetic radiation across different wavelengths, enabling the creation of detailed and comprehensive images of the Earth's surface. Satellite remote sensing offers a global perspective, allowing for the monitoring of large-scale environmental changes, land use patterns, and natural phenomena. The accessibility and regular revisit times of satellites contribute to their effectiveness in providing up-to-date information for applications such as environmental monitoring, agriculture, disaster management, and urban planning. Researchers and decision-makers leverage satellite imagery to gain insights into dynamic Earth processes and make informed decisions based on the analyzed data [16].
Remote sensing has a wide range of applications, including agriculture, environmental monitoring, disaster response, and military and defense[17]. It has revolutionized agriculture by increasing the spatial-temporal resolution of data collection, allowing for more precise monitoring and management of crops [15]. Additionally, remote sensing technology has been widely used in disaster emergency response and ecological environment monitoring [17].

The field of remote sensing draws from various disciplines such as computer science, statistics, and machine learning [18]. For instance, the automatic registration of multimodal remote sensing data is a challenging task due to significant nonlinear radiometric differences between these data, requiring advanced techniques from computer science and pattern recognition [19]. Furthermore, the development of knowledge-driven approaches is considered crucial in remote sensing research, alongside data-driven approaches based on innovative algorithms and enhanced computing capacities [20].
Remote sensing also plays a significant role in image analysis and classification techniques. Pixel-based and object-based image analysis are the two main techniques used in remote sensing for classifying and interpreting remote sensing images [21]. Moreover, the combination of multisource remote sensing data holds promise for improving urban land use and land cover classification accuracy, although challenges such as the ”semantic gap” persist [22].
The rapid development of remote sensing technology has led to the acquisition of vast amounts of remote sensing data, leading to the emergence of ”Big Data” in remote sensing. This not only refers to the high volume and fast generation velocity of data but also the variety and complexity of remotely sensed data [23]. As a result, there is a growing need for advanced machine learning and computing techniques to process and analyze massive amounts of remotely sensed imagery [24].

1.2.1 The electromagnetic spectrum and satellite bands

The science of remote sensing relies on the interpretation of measurements of electromagnetic energy reflected from or emitted from a target, which allows for the quantification of various characteristics of the Earth’s surface and atmosphere [25]. The electromagnetic spectrum is essential for remote sensing satellites, as it allows them to capture data about the Earth’s surface and atmosphere.
For instance, in agricultural remote spectral sensing, the incident electromagnetic radiation, typically sunlight, interacts with the surface of crops or soil, leading to reflection, absorption, or transmission of light, which can be captured and analyzed through remote sensing techniques [26].
Moreover, the development of remote sensing from satellites has significantly advanced the field, enabling the collection of data using optical and electronic systems operating within specific wavelength ranges [27]. This has led to the utilization of remote sensing data for environmental monitoring, including the assessment of desertification and vegetation recovery in grasslands after natural disasters such as wildfires[28].
Furthermore, the radiometric calibration of remote sensing systems is essential for establishing the functional relationship between the output signal value of remote sensing cameras and the input radiation, known as absolute radiometric calibration, which is crucial for accurate data interpretation and analysis [29].
The spectral bands used in remote sensing satellites play a crucial role in capturing and analyzing data from the Earth’s surface and atmosphere. These bands offer differential spectral frequencies and multi-featured information, which are essential for decision-making in remote sensing applications . Remote sensing relies primarily on reflected or emitted electromagnetic radiation from the Earth, encompassing optical and microwave wavelengths, to infer changes on the Earth’s surface and in the overlying atmosphere [27]. The advanced onboard sensors of satellites provide several spectral bands, each operating within specific wavelength ranges, enabling the collection of data for various applications such as environmental monitoring and land use analysis [27].

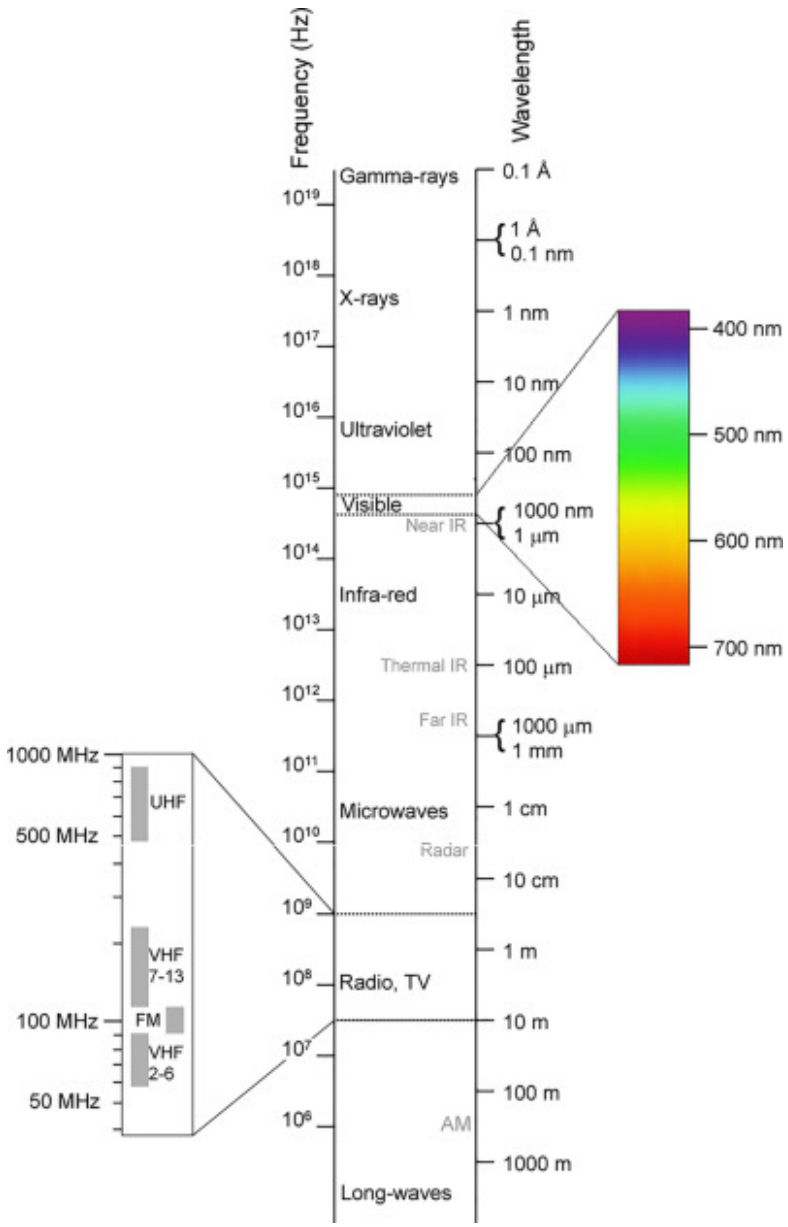


Figure 1.2.1: The electromagnetic spectrum [30]

The spectral bands of remote sensing satellites are used to capture raster data maps for further processing in different bands of the electromagnetic spectrum, allowing for the identification of various objects of interest and monitoring their changes [31]. These bands are also crucial for mapping impervious surfaces and assessing vegetation recovery after natural disasters, demonstrating the wide-ranging applications of spectral bands in remote sensing [32].
Furthermore, the spectral bands of satellite images cover hundreds of wavelengths, providing a wealth of data for analysis and interpretation. This extensive coverage enables the detection of alterations in minerals, the monitoring of photosynthetic phenology in vegetation, and the classification of land cover changes [33]. The spectral bands are also used for spectral reconstruction of surface types, contributing to the accurate interpretation of remote sensing imagery [34].
In addition, the spectral bands are essential for radiometric calibration, which establishes the functional relationship between the output signal value of remote sensing cameras and the input radiation, ensuring the accuracy of data interpretation and analysis [35][36]. Moreover, the spectral bands are utilized in sensor fusion and image classification, highlighting their significance in the interpretation and integra-

tion of remote sensing data [36][25].
In conclusion, the spectral bands of remote sensing satellites encompass a wide range of wavelengths, providing valuable data for environmental monitoring, land use analysis, mineral detection, and vegetation assessment. These bands are fundamental for radiometric calibration, sensor fusion, and image classification, making them indispensable for the comprehensive interpretation and analysis of remote sensing data. Different bands and their wavelength for landast 7, 8 and sentinet-2 sattelite can be seen in figure 1.2.2.

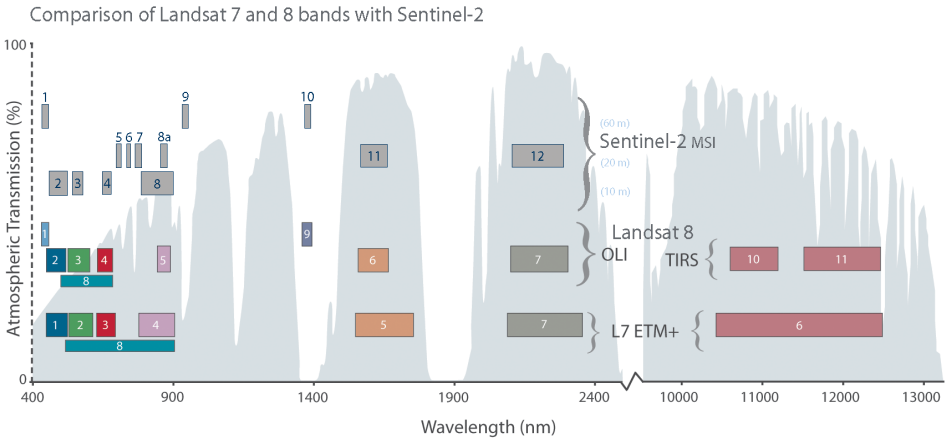


Figure 1.2.2: Landsat ans Sentinel 2 Bands [37]

1.2.2 Sentinel-2 satellite

The Sentinel-2 satellite, part of the European Union's Copernicus program, is equipped with the Multi-spectral Instrument (MSI) that provides radiometrically and geometrically superior multi-spectral high spatial resolution images over the global surface. It operates at a high revisit time (5 days at the Equator with two satellites in orbit) and covers a wide field of view, encompassing 13 bands in the optical NIR, SWIR parts of the electromagnetic spectrum [38].
The satellite's passive sensors are designed to capture data with more image channels, corresponding to more acquired spectral bands, allowing for the identification of various objects in an image based on the reflection spectrum of the objects [39].
The Sentinel-2 satellite's capabilities have been widely utilized in various applications, including land cover mapping, agriculture, environmental monitoring, and disaster response. Its high temporal interval and multispectral features have made it valuable for monitoring vegetation growth, analyzing land cover changes, and estimating aboveground biomass and carbon stock [40].
Additionally, the satellite's data has been effectively integrated with other remote sensing systems, such as Sentinel-1, for classification and land use/land cover mapping in urban areas and ecological environments [41]. Furthermore, the satellite's data has been used for bathymetry, inland excess water mapping, and shoreline position determination, demonstrating its versatility in environmental and meteorological applications [42].
In conclusion, the Sentinel-2 satellite, with its advanced MSI, has proven to be a valuable asset in the field of remote sensing, offering high spatial resolution, multispectral capabilities, and a high temporal interval. Its data has been instrumental in a wide range of applications, including land cover mapping, agriculture, environmental monitoring, and disaster response.

1.2.3 Sentinel-2 bands

The Sentinel-2 satellite is equipped with a Multi-spectral Instrument (MSI) that captures data in various spectral bands, providing valuable information for a wide range of applications. The satellite's spectral bands cover the visible, near-infrared, and shortwave infrared parts of the electromagnetic spectrum, enabling detailed analysis of the Earth's surface and vegetation.
The Sentinel-2 satellite's spectral bands have been extensively utilized in environmental monitoring, land cover mapping, agriculture, and disaster response. The satellite's 13 spectral bands, operating in the visible, near-infrared, and shortwave infrared spectrum, have been instrumental in monitoring vegetation growth, analyzing land cover changes, and estimating aboveground biomass and carbon stock [43]. Additionally, the satellite's data has been effectively integrated with other remote sensing systems for classification and land use/land cover mapping in urban areas and ecological environments [44].
The spectral bands of Sentinel-2 have also been used for various specific applications, such as estimating land surface temperature in agricultural lands [45], mapping invasive aquatic plants [46], and detecting oil spills on the sea surface[47]. Furthermore, the satellite's spectral bands have been found to be scalable with many optical sensors, such as MODIS, VIIRS, MERIS, Landsat, and RapidEye, demonstrating their compatibility and versatility in remote sensing applications[48].
The high spatial and temporal resolution, along with the wide spectral coverage of the Sentinel-2 spectral bands, has made the satellite's data commonly used in monitoring the evolution of terrestrial ecological environments, wetland landscape mapping, vegetation crop growth monitoring, and disaster warning mapping [49]. The spectral bands have also been used for snow cover mapping, ice avalanche monitoring, and assessing rangeland quality, showcasing their diverse applications in environmental and meteorological studies [50][51].

Band	Resolution	Central Wavelength	Description
B1	60 m	443 nm	Ultra Blue (Coastal and Aerosol)
B2	10 m	490 nm	Blue
B3	10 m	560 nm	Green
B4	10 m	665 nm	Red
B5	20 m	705 nm	Visible and Near Infrared (VNIR)
B6	20 m	740 nm	Visible and Near Infrared (VNIR)
B7	20 m	783 nm	Visible and Near Infrared (VNIR)
B8	10 m	842 nm	Visible and Near Infrared (VNIR)
B8a	20 m	865 nm	Visible and Near Infrared (VNIR)
B9	60 m	940 nm	Short Wave Infrared (SWIR)
B10	60 m	1375 nm	Short Wave Infrared (SWIR)
B11	20 m	1610 nm	Short Wave Infrared (SWIR)
B12	20 m	2190 nm	Short Wave Infrared (SWIR)

Figure 1.2.3: Sentinel 2 bands and their resolution

In conclusion, the Sentinel-2 satellite's spectral bands, with their wide coverage

and high resolution, have been pivotal in a multitude of remote sensing applications, ranging from vegetation monitoring to disaster response and environmental assessment. Depending on the application the data from different bands can be used for training machine learning models. Furthermore, there are various methods to combine the data from various bands into a new attribute; which can help increase the accuracy of the model.

1.2.4 Band combination

The combination of satellite bands plays a crucial role in extracting valuable information from remote sensing data for various applications. The combination of specific spectral bands has been utilized for diverse applications. For instance, the combination of different polarization bands in the Sentinel-1 satellite has been employed for oil spill detection, demonstrating the significance of band combinations in environmental monitoring [52].
Additionally, the combination of L-band, C-band, and X-band Synthetic Aperture Radar (SAR) images has been used for land cover classification and crop-type mapping, showcasing the potential of multi-sensor fusion for remote sensing applications [53].
Moreover, the combination of specific bands has been explored for applications such as Leaf Area Index (LAI) estimation and forest biomass inversion. For example, the combination of hyperspectral bands has been analyzed for LAI estimation, highlighting the importance of band combinations in vegetation parameter retrieval [54]. Similarly, the combination of bands for bidirectional reflectance distribution function (BRDF) analysis has been instrumental in fine vegetation classification and phenological studies [55].
One of the widely used band combinations is the Normalized Difference Vegetation Index (NDVI), which is derived from near-infrared (NIR) and red bands. The NDVI is a powerful indicator of vegetation health and density, providing insights into agricultural productivity, land cover changes, and environmental monitoring [56]. The combination of multispectral bands with composite indices such as NDVI, Enhanced Vegetation Index (EVI), and Soil Adjusted Vegetation Index (SAVI) has been effective in vegetation analysis and land cover mapping [57].

In summary, the combination of satellite bands, including the utilization of specific indices and multi-sensor fusion, is essential for a wide range of remote sensing applications, including vegetation analysis, land cover mapping, environmental monitoring, and disaster response.
For using the satellite data for machine learning one of the first steps is to remove the clouds from the images, this is done using a technique called cloud masking.

1.2.5 Cloud masking

Cloud masking is a crucial process in remote sensing, especially in the analysis of satellite imagery. Clouds can obstruct the view of the Earth's surface, affecting the accuracy of data analysis and interpretation. Various methods and algorithms have been developed to address cloud masking in satellite imagery, utilizing different techniques and spectral bands to effectively detect and mask clouds. These methods are essential for ensuring the quality and reliability of remote sensing data for a wide range of applications.
One approach to cloud masking involves the use of deep learning techniques, such as deep convolutional neural networks (CNNs). Hasan et al.[58] proposed a cloud detection method using deep CNNs, focusing on four spectral bands commonly available in multispectral satellite imagery. The study demonstrated that these bands show significant responses for cloud pixels, making them suitable for cloud detection. Similarly, Mateo-García et al.[59] explored the use of convolutional neural networks for multispectral image cloud masking, highlighting the potential of CNNs as a promising alternative for solving cloud masking problems.

Another widely used method for cloud masking is the utilization of specific algorithms designed for satellite imagery, such as the Function of Mask (Fmask) algorithm. Li et al [60] discussed the application of the Fmask algorithm for cloud detection in Landsat and Sentinel-2 satellite imagery, emphasizing its effectiveness in utilizing available band information for cloud masking. Additionally, Chen et al.[61] highlighted the operational use of the Fmask algorithm for Landsat data products, demonstrating its significance in cloud masking for satellite imagery.
Furthermore, the integration of spectral bands and advanced techniques has been explored for cloud masking. Meraner et al.[62] developed a deep residual neural network architecture to remove clouds from multispectral Sentinel-2 imagery, showcasing the potential of advanced neural network models for cloud removal. Additionally, Wu Shi [63] discussed the modification of the Automated Cloud Cover Assessment (ACCA) algorithm for cloud masking in Gaofen-1 wide field of view (GF-1 WFV) imagery, emphasizing the use of specific spectral bands for cloud mask generation.
In summary, cloud masking in remote sensing imagery involves a diverse range of methods, including deep learning techniques, specific algorithms, and the utilization of spectral bands. These approaches are crucial for ensuring the accuracy and reliability of satellite imagery data.

1.3 Machine learning

Machine learning is a technique in which a machine learns and improves on its own, based on past data [64]. It is an interdisciplinary approach in building mathematical models from known inputs to make data-driven predictions and decisions [65]. Machine learning has made enormous strides over the past two decades, and can now be easily used by everyone [66]. It is an algorithm that estimates an unknown dependency between system inputs and its outputs from the available data [67].
Machine learning is used in various fields such as ad placement, credit scoring, fraud detection, stock trading, drug design, natural language processing, image recognition, expert systems, simulation, manufacturing, and many other applications [67]. Machine learning methods are the most appropriate among the wide range of intrusion detection technologies [68].
Machine learning in remote sensing involves the application of advanced computational techniques to analyze and interpret remote sensing data for a wide range of applications. The integration of machine learning algorithms with remote sensing technologies has revolutionized the field, enabling the extraction of valuable insights from satellite imagery and other remote sensing data sources.

There are diverse applications of machine learning in remote sensing, ranging from urban mapping and infrastructure assessment to environmental monitoring and land cover classification. Zhang et al. [69] provided a technical tutorial on the state-of-the-art deep learning techniques for remote sensing data, covering topics such as preprocessor, classifier (uml), and discriminative model. Rukhovich et al. [70] discussed the use of deep machine learning for the automated selection of remote sensing data to determine areas of arable land degradation processes distribution. Li et al.[71] applied machine learning methods with VIIRS nighttime light and MODIS daytime NDVI data to map urban extent at large spatial scales.Xue et al. [72] utilized machine learning ensemble and satellite big data for large-scale high-resolution coastal mangrove forests mapping across West Africa. Kamusoko [73] explored geospatial machine learning in urban environments, focusing on challenges and prospects. Suharso et al. [74] conducted a systematic review on the role of machine learning in remote sensing for agriculture drought monitoring. Kelley et al.[75] used Google Earth Engine to map complex shade-grown coffee landscapes in Northern Nicaragua. Mondal et al. [76] evaluated combinations of Sentinel-2 data and machine-learning algorithms for mangrove mapping in West Africa. Chen et

al.[77] proposed a high spatial resolution PM2.5 retrieval using MODIS and ground observation station data based on ensemble random forest. Hird et al.[78] utilized Google Earth Engine, open-access satellite data, and machine learning in support of large-area probabilistic wetland mapping.

The integration of machine learning algorithms with remote sensing data has significantly enhanced the capabilities of analyzing and interpreting large-scale geospatial information, paving the way for innovative solutions in various domains. Classification is the common part of machine learning used in remote sensing, there are different classification methods and each can be used in a specific task.

1.3.1 Classifiaction methods

Classification in machine learning refers to the process of categorizing data into different classes or categories based on its features and attributes. It is a fundamental concept in machine learning and is widely used in various fields such as computer science, data mining, and artificial intelligence. Machine learning techniques for classification involve training a model on a labeled dataset to learn the patterns and relationships within the data, and then using this model to predict the class or category of new, unseen data.
Classification methods in remote sensing encompass a wide array of techniques that are essential for analyzing and interpreting remote sensing data. These methods play a crucial role in tasks such as land cover mapping, change detection, and environmental monitoring. There have been different research done in this field. Walter et al. [79] discussed the object-based classification of remote sensing data for change detection, highlighting the use of multispectral images for geology applications. Han et al. [80] combined 3D-CNN and Squeeze-and-Excitation Networks for remote sensing sea ice image classification, emphasizing the use of convolutional neural networks and support vector machine. Hamida et al. [81] proposed a 3-D deep learning approach for remote sensing image classification, demonstrating improved classification rates with lower computational costs. Liu et al. [82] conducted a comparative study on remote sensing image classification, comparing support vector machine and deep learning methods. Xia et al. [83] discussed the use of deep learning in ecological remote sensing images, focusing on artificial neural networks, support vector machines, and fuzzy theory for classification. Bruzzone et al. [84] proposed a novel transductive SVM for semisupervised classification of remote-sensing images, addressing ill-posed classification problems. Dópido et al. [85] developed a semisupervised self-learning approach for hyperspectral image classification, adapting active learning methods to a self-learning framework. Yao et al. [86] explored land use classification using deep convolutional neural network methods, focusing on the preservation of spatial features. Phillips et al. [87] introduced an SMP soft classification algorithm for remote sensing, emphasizing its high accuracy and automation. Alimjan et al. [88] performed remote sensing image classification using a combinatorial algorithm of SVM and KNN. Zhao and Pan [89] presented a remote sensing image feature selection method based on rough set theory and multi-agent system. Yin et al. [90] developed semi-supervised feature learning for remote sensing image classification, focusing on improving classification performance with unlabeled data.

These references demonstrate the diverse range of classification methods used in remote sensing, including random forests,support vector machines, deep learning, semi-supervised learning, and feature selection techniques. These methods are crucial for accurately analyzing and interpreting remote sensing data for a wide range of applications.

1.3.2 Random forests

Random Forests is an ensemble learning algorithm that builds multiple decision trees by training on random subsets of both data and features. It employs bag-

ging, using bootstrapped samples, and introduces diversity by randomly selecting features for each tree. Random Forest is a widely used machine learning algorithm in remote sensing for various applications.

Gislason et al. [91] discussed the application of Random Forest for land cover classification, highlighting its use in ensemble learning and boosting. Belgiu and Drăguţ [92] provided a comprehensive review of the applications and future directions of Random Forest in remote sensing, covering topics such as overfitting and classifier (uml). Pal [93] presented the use of the Random Forest classifier for remote sensing classification, emphasizing its application in artificial intelligence. Phan and Kappas [94] compared Random Forest with other classifiers for land cover classification using Sentinel-2 imagery, highlighting its effectiveness in producing high accuracies. Pierrat et al. [95] proposed the use of Random Forest models for predicting gross primary productivity in boreal forests based on remote sensing metrics. Saini and Ghosh [96] performed crop classification using Random Forest and Support Vector Machine, demonstrating the effectiveness of these machine learning algorithms. Walker and Hamilton [97] utilized Random Forest with remote sensing data to locate uncontacted indigenous villages in Amazonia, highlighting its accuracy and speed in handling high data dimensionality. Cheng et al. [98] achieved forest type classification in China using Random Forest based on spectral, spatial, and temporal features derived from remote sensing data.

These citations showcase the versatile applications of Random Forest in remote sensing, encompassing tasks such as land cover classification, crop identification, and forest type mapping. The algorithm's proficiency in managing high-dimensional data and generating precise outcomes establishes it as a valuable tool for the analysis and interpretation of remote sensing data.

1.3.3 Support vector machines (SVM)

Support Vector Machine (SVM) is a widely-used supervised machine learning algorithm used for classification and regression tasks. It works by finding an optimal hyperplane in a high-dimensional space to separate data into different classes. Its objective is to maximize the margin, the distance between the hyperplane and the nearest data points from different classes, promoting effective generalization to new, unseen data. SVM finds practical applications in tasks such as image and text classification.

Support Vector Machines (SVM) have been widely used in remote sensing for various applications, including image classification, object detection, and land cover mapping. Here are some references that discuss the use of Support Vector Machines in remote sensing:

Mountrakis et al. [99] provided a comprehensive review of the application of Support Vector Machines in remote sensing, covering topics such as feature extraction and classification accuracy. Melgani and Bruzzone [100] discussed the classification of hyperspectral remote sensing images using Support Vector Machines, emphasizing its effectiveness in handling high-dimensional data. Maxwell et al. [101] implemented machine-learning classification in remote sensing, highlighting the use of Support Vector Machines for accurate and stable classification. Nadzri [102] evaluated the effectiveness of Support Vector Machine and Random Forest classifiers in delineating green areas using remote sensing data. Chen et al. [103] discussed an improved multi-source data-driven landslide prediction method based on Support Vector Machines, highlighting its application in geotechnical engineering. Chen et al. [77] proposed a high spatial resolution PM2.5 retrieval method using MODIS and ground observation station data based on ensemble Support Vector Machines. [104] discussed the application of Support Vector Machines in object-based river extraction from multispectral remote sensing images. Foody and Mathur [105]

evaluated multiclass image classification using Support Vector Machines, highlighting its effectiveness in handling complex remote sensing data.

These references show how Support Vector Machines are used in remote sensing, demonstrating their effectiveness in handling high-dimensional data, accurate classification, and feature extraction.

1.4 Thesis goal

This thesis aims to use remote sensing data for detection and classification of different biomes in Norway. Machine learning methods, specifically SVM and Random Forests models, are employed and compared for analyzing satellite data and classify different biomes in Norway. The study also compares the effectiveness of combining bands and using different bands. The models developed can be utilized to track changes in ecosystems over time, evaluating their health status.

CHAPTER
TWO

THEORY

2.1 SVM

Support Vector Machines (SVM) is a powerful machine learning algorithm that is widely used for classification and regression tasks. It belongs to the family of supervised learning algorithms and is particularly effective in high-dimensional spaces. SVM is renowned for its ability to handle both linear and non-linear relationships in data.[106] This chapter delves into the fundamental theory and mathematical formulas that underpin SVM.

2.1.1 Basic Concepts of SVM

Linear Separation and Hyperplane

At the core of SVM lies the concept of linear separation. Given a set of data points, SVM endeavors to identify the hyperplane that optimally segregates the data into distinct classes. This hyperplane serves as the decision boundary, maximizing the margin between the classes, and is chosen based on the features present in the dataset, determining its dimensions in the n-dimensional space. The objective is to pinpoint the simplest decision boundary that effectively classifies the information points while ensuring a maximum margin between them.[106] Mathematically, a hyperplane can be represented as:

$$w \cdot x + b = 0$$

where w is the weight vector, x is the input vector, and b is the bias.

2.1.1.1 Margin and Support Vectors

The margin is the distance between the hyperplane and the nearest data point from either class. SVM aims to maximize this margin. Support vectors are the data points that lie closest to the hyperplane and are crucial in defining the decision boundary. [106]

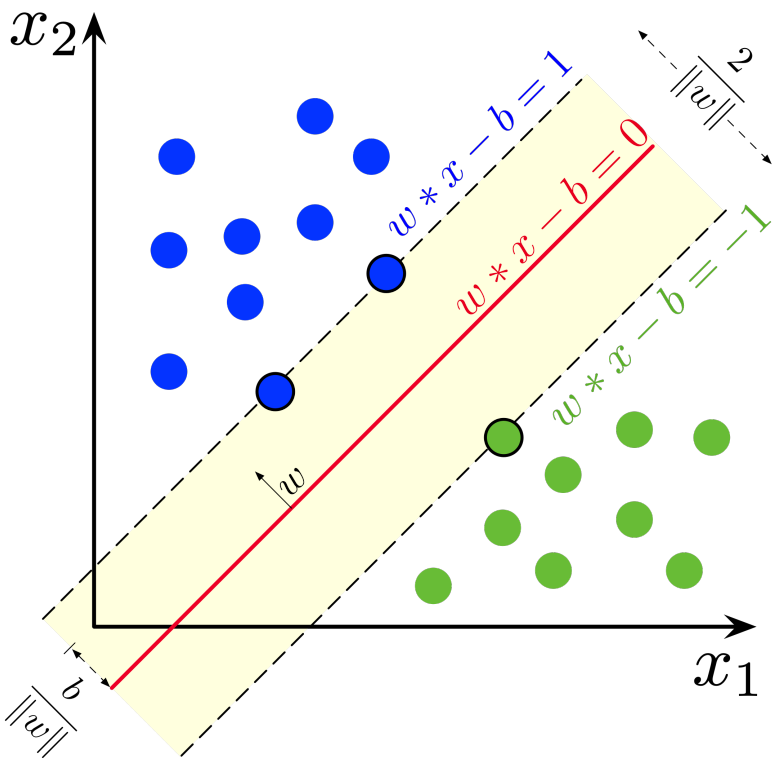


Figure 2.1.1: The linear SVM for two classes [107]

2.1.2 Kernel Trick and Non-Linear SVM

The ”Kernel Trick” in Support Vector Machines (SVMs) is a technique that transforms non-linearly separable data into a higher-dimensional feature space, facilitating linear separation. This allows SVM to identify a hyperplane with maximum margin, even for non-linearly separable data. Kernel functions compute inner products between points in the transformed space efficiently without explicitly calculating the transformation. Common kernels include linear, polynomial, and Gaussian (radial basis function), chosen based on the data’s nature. The linear kernel suits roughly linearly separable data, the polynomial kernel for complex curved borders, and the Gaussian kernel for data with unclear boundaries and intricate overlaps.[108]

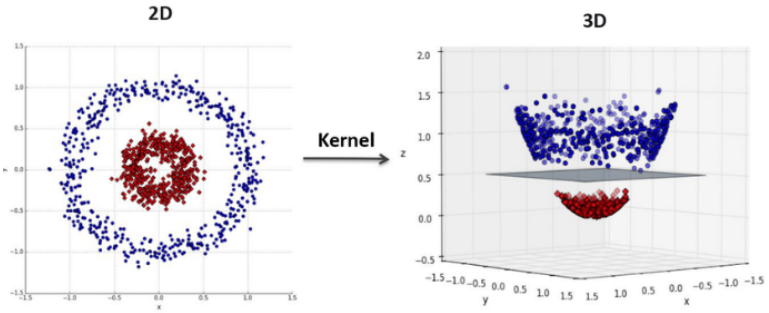


Figure 2.1.2: Kernel trick for separating the data [109]

2.1.3 Kernel Trick and Non-Linear SVM

2.1.3.1 Common Kernel Functions

- **Radial Basis Function (RBF) Kernel**: The Radial Basis Function (RBF) kernel, also known as the Gaussian kernel, is a popular kernel used in Support

Vector Machines (SVMs) and other machine learning algorithms. The RBF kernel is particularly effective for handling non-linear and complex relationships in the data.

The RBF kernel is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ are data points, $\|\mathbf{x}_i - \mathbf{x}_j\|$ represents the Euclidean distance between these points, and σ is a parameter that controls the width of the Gaussian distribution.

The RBF kernel has the property of mapping data into a higher-dimensional space, allowing SVMs to find non-linear decision boundaries in the original feature space. It is suitable for data with complex structures, where relationships between features are not easily captured by linear models. The choice of the σ parameter is crucial, as it influences the flexibility of the model; a smaller σ results in a more complex and flexible decision boundary.

- **Polynomial Kernel**: The Polynomial Kernel is another type of kernel frequently employed in Support Vector Machines (SVMs) and other machine learning algorithms. It is useful for capturing non-linear relationships in the data.

  The Polynomial Kernel is defined as:

  $$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + c)^d$$

  Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ represent data points, $\mathbf{x}_i^T$ denotes the transpose of $\mathbf{x}_i$, c is a constant term, and d is the degree of the polynomial.

  Similar to the RBF kernel, the Polynomial Kernel allows SVMs to map data into a higher-dimensional space, facilitating the discovery of non-linear decision boundaries in the original feature space. The parameters c and d play crucial roles in determining the kernel's behavior. The constant term c shifts the decision boundary, and the degree d controls the level of non-linearity.

- **Sigmoid Kernel**: The Sigmoid Kernel is a type of kernel commonly utilized in Support Vector Machines (SVMs) and other machine learning algorithms. It is particularly useful for capturing complex, non-linear relationships in the data.

  The Sigmoid Kernel is defined as:

  $$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\alpha \mathbf{x}_i^T \mathbf{x}_j + c)$$

  Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ denote data points, $\mathbf{x}_i^T$ represents the transpose of $\mathbf{x}_i$, α is a scaling parameter, c is a constant term, and tanh is the hyperbolic tangent function.

  Similar to other kernels, the Sigmoid Kernel allows SVMs to map data into a higher-dimensional space, facilitating the discovery of non-linear decision boundaries. The parameters α and c play crucial roles in determining the kernel's behavior. The scaling parameter α controls the degree of non-linearity, and the constant term c shifts the decision boundary.

2.1.4 Optimization Objective of SVM

2.1.4.1 Hard Margin SVM

In the case of linearly separable data, SVM aims to maximize the margin while ensuring that all data points are correctly classified. This leads to the following optimization objective:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

subject to $y_i(w \cdot x_i + b) \geq 1$ for all data points[110].

2.1.4.2 Soft Margin SVM

When data is not perfectly separable, a soft margin is introduced to allow for some misclassification. The optimization objective is then modified to penalize misclassifications:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \max(0, 1 - y_i(w \cdot x_i + b))$$

where C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing misclassifications[110].

2.1.5 Training SVM: Sequential Minimal Optimization (SMO) Algorithm

The training of SVM involves solving a quadratic optimization problem. The Sequential Minimal Optimization (SMO) algorithm is a popular approach to efficiently solve this problem. It iteratively selects two Lagrange multipliers and optimizes them while keeping the others fixed.

The key steps of the SMO algorithm include:

- Selecting two Lagrange multipliers based on a heuristic.
- Optimizing the selected multipliers subject to the constraints.
- Updating the threshold b and the error cache.

Support Vector Machines have proven to be versatile and effective in various machine learning applications. Understanding the theoretical foundations and mathematical formulations of SVM, including the optimization objectives and the kernel trick, is crucial for practitioners seeking to apply this powerful algorithm to real-world problems.

2.2 Random Forest

The Random Forest algorithm is a supervised learning technique that employs ensemble learning. Ensemble learning combines multiple algorithms or repeats a single algorithm to enhance the model's effectiveness. In the context of Random forest, numerous decision trees are constructed at random data points, and their predictions are averaged [111].

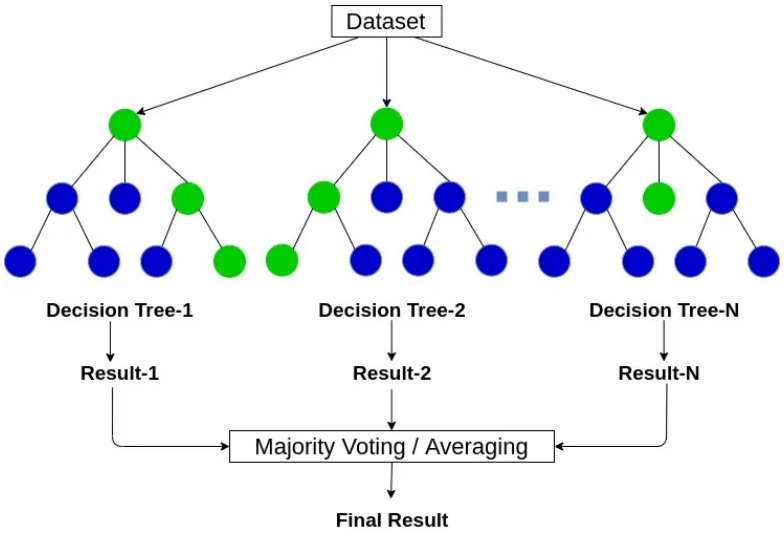


Figure 2.2.1: Random forest structure [111]

2.2.1 Theory of Random forest

The Random Forest algorithm is grounded in ensemble learning principles, aiming to enhance predictive accuracy and robustness. Through a technique called bagging, the algorithm creates multiple subsets of the training data using bootstrapping, and each subset is employed to train an individual decision tree. Crucially, random feature selection at each node introduces diversity, preventing trees from becoming overly similar. The final prediction is then derived by averaging (for regression) or voting (for classification) over the predictions of all trees in the forest. This ensemble approach significantly mitigates overfitting and enhances generalization performance. The reduction in variance is achieved by constructing multiple trees that collectively contribute to a more robust model. The out-of-bag (OOB) instances, excluded from the bootstrap sample, provide a convenient means of estimating model performance without requiring a separate validation set. Recognized for versatility and effectiveness across diverse datasets and tasks, Random Forest also offer insights into feature importance based on their contribution to reducing impurity across the ensemble. Overall, the Random Forest algorithm stands as a powerful and widely utilized method in machine learning, leveraging the strengths of individual trees to yield a robust and accurate predictive model [112].

2.2.2 Random Forest Algorithm

Beginning with the random selection of samples from a given dataset, the algorithm constructs a decision tree for each sample and obtains a prediction result from every decision tree. Following this, a voting process is conducted for each predicted result. Finally, the prediction result with the highest number of votes is selected as the ultimate prediction result [111].

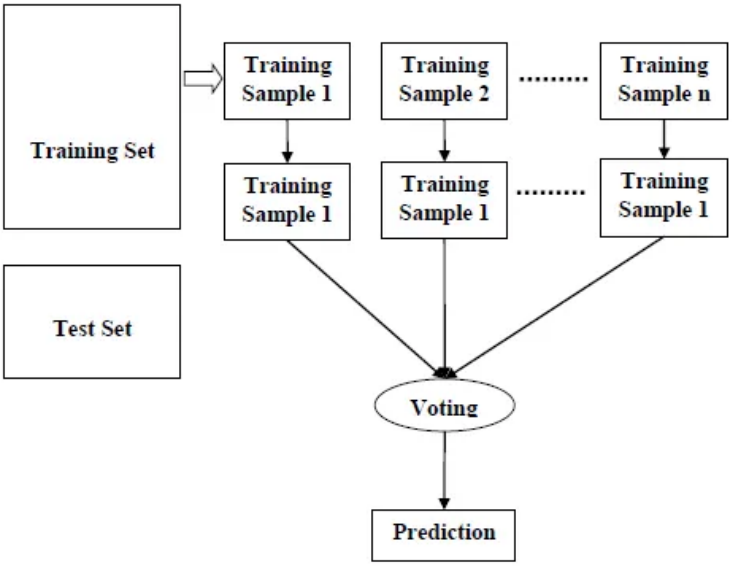


Figure 2.2.2: Random forest algorithm [111]

The Random Forest algorithm for regression or classification can be summarized as follows [112]:

1. For each index b from 1 to B:
   (a) Draw a bootstrap sample Z^* of size N from the training data.
   (b) Construct a random-forest tree T_b using the bootstrapped data. Recursively repeat the following steps for each terminal node of the tree until the minimum node size n_{min} is reached:
      i. Select m variables at random from the p variables.
      ii. Choose the best variable/split-point among the selected m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_{b=1}^{B}$.

To make a prediction at a new point x:

- For regression: $\hat{f}_{\text{rf}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.
- For classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}(x)$ is the majority vote of $\{\hat{C}_b(x)\}_{b=1}^{B}$.

2.2.3 Complexity Analysis of Random forest

Random forest have good computational performance and scalability. An original complexity analysis of random forest shows that they are computationally efficient and can scale well with increasing data size. The complexity of Random Forest depends on the number of trees in the forest, the number of features in the dataset, and the depth of the trees. The time complexity of training a single decision tree is $O(N \log(N) Pk)$, where N is the sample size, P is the feature size, and k is the depth of the tree [113]. The time complexity of training a Random Forest is $O(MN \log(N) Pk)$, where M is the number of trees in the forest.

2.2.4 Interpretability of Random forest

Random Forest's interpretability can be assessed by examining variable importance measures. The Mean Decrease of Impurity variable importance measure is commonly utilized, and its properties have been theoretically defined. Nevertheless,

variable importances computed from non-totally randomized trees (such as standard Random Forest) may experience various shortcomings arising from masking effects, misestimations of node impurity, or the inherent binary structure of decision trees [111].
Random Forests are a versatile tool in machine learning, offering a good balance between performance, interpretability, and computational efficiency. However, like all models, they have limitations. To use them effectively, it's important to understand their underlying theory and mechanisms an keep in mind potential challenges, such as the risk of overfitting, sensitivity to hyperparameters, and considerations with imbalanced datasets. Staying informed about practical aspects of Random Forest implementation helps optimize their use in various applications.

2.3 Remote sensing indices

Remote sensing indices are mathematical combinations or transformations of spectral bands or other remote sensing data to extract specific information about the Earth's surface or atmosphere. These indices are designed to enhance certain features or characteristics that might not be readily apparent in individual bands of satellite or airborne sensor data. The indices used in this research consist of [114]:

2.3.1 Normalized Difference Vegetation Index (NDVI):

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$$

NDVI is a widely-used vegetation index that assesses the health and density of vegetation. Higher NDVI values typically indicate healthier and more abundant vegetation.

2.3.2 Normalized Difference Built-Up Index (NDBI):

$$\text{NDBI} = \frac{\text{SWIR} - \text{NIR}}{\text{SWIR} + \text{NIR}}$$

NDBI is employed to identify built-up or urban areas in remote sensing imagery. Positive values of NDBI often indicate built-up areas, while negative values are associated with natural surfaces.

2.3.3 Modified Normalized Difference Water Index (MNDWI):

$$\text{MNDWI} = \frac{\text{Green} - \text{SWIR}}{\text{Green} + \text{SWIR}}$$

MNDWI is a water index used for the detection of water bodies. Higher MNDWI values typically represent water features due to the strong absorption of green light by water.

2.3.4 Normalized Difference Built-Up and Bareness Index (NDBaI):

$$\text{NDBaI} = \frac{\text{SWIR} - \text{Red}}{\text{SWIR} + \text{Red}}$$

NDBaI is designed to distinguish between built-up areas and bare land. Higher NDBaI values are associated with built-up areas, while lower values may indicate bare surfaces.

2.3.5 Modified Normalized Difference Built-Up Index (MNDbI):

$$\text{MNDbI} = \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}}$$

MNDbI is another index used for built-up area detection. Positive MNDbI values are indicative of built-up areas, while negative values are associated with natural surfaces.

2.3.6 Soil Adjusted Vegetation Index (SAVI):

$$\text{SAVI} = (1 + L) \cdot \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red} + L}$$

SAVI is a modified vegetation index that accounts for soil background reflectance. It is particularly useful in areas with sparse vegetation.

2.3.7 Burn Severity Index (BSI):

$$\text{BSI} = 2 \cdot \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}} + 1$$

BSI is utilized to assess the severity of burn areas following wildfires. BSI values provide information on the level of vegetation damage in burned areas.

2.3.8 Enhanced Vegetation Index (EVI):

$$\text{EVI} = 2.5 \cdot \frac{\text{NIR} - \text{Red}}{\text{NIR} + 6 \cdot \text{Red} - 7.5 \cdot \text{Blue} + 1}$$

EVI is an enhanced version of NDVI designed to minimize atmospheric influences and improve sensitivity in areas with dense vegetation. It is often preferred over NDVI for its improved performance in various environmental conditions.

CHAPTER

THREE

METHODS

3.1 Google earth engine

Google earth engine (GEE) is a cloud based tool for geospatial analysis developed by google. It gives access to various satellite imagery and remote sensing data. There is a built in coding environment which uses JavaScript programming language; as well as a python library that can be used in other editors. In this platform one can acquire satellite data process them, train machine learning models and visualize data on the map. The calculations run on the google cloud which makes it convenient and removes the need of a computer with high processing power.
In order to train machine vision algorithms labeled data is needed.Labeled data refer to images that are paired with corresponding annotations or labels that shows information about the objects in the image meaning the boundaries of an item and its label. In case of this project, the available data are the labels which show the location of a biome at name of the biome. Google earth engine matches the coordinates with the map, corresponds the label and takes and image from the map with data from the desired band of the selected satellite.

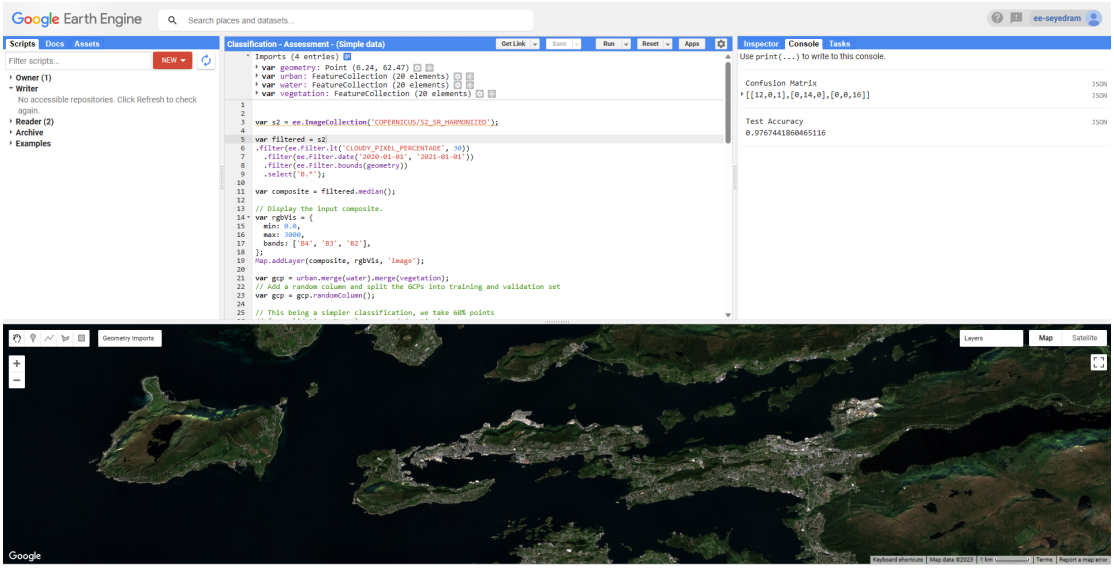


Figure 3.1.1: Main window of Google Earth Engine

The figure 3.1.1 shows the main windows of GEE and its panels. The panel on the left contains the list of scripts, assets and documentations. The middle panel is the code editor that can be used to write JavaScript code in the opened script and run the code. The right panel contains the console for the results and also inspector

and the ongoing tasks. And in the bottom part the map is available which can show the visualized results and data from the script.
In this thesis the built in coding environment was used to import training data, pre-processing, sampling, model training and visualizing the results.

3.2 The dataset

The training dataset contains a set of geometry coordinates which are labeled as different biomes classified based on the NIN (Nature in Norway) system. The dataset is prepared by the Norwegian Biodiversity Information Centre [115]. Established in January 2005 to bolster knowledge for nature diversity, The Norwegian Biodiversity Information Centre became an independent agency under the Ministry of Climate and Environment in 2018. Collaborating with scientific institutions, the Centre aggregates data spanning a century from entities like the Natural History Museum, aiming for easy accessibility through a technical infrastructure. Actively engaging with stakeholders, they emphasize using natural diversity knowledge as a basis for public discourse. Utilizing the internet as a primary channel, they provide expanding, user-friendly services. Their initiatives include mapping species diversity, documenting endangered species, and aiding in species identification, reflecting their commitment to promoting awareness of Norway's ecosystems [115]. The The system has 8 main biomes:

- **M:** Saltwater bottom systems
- **L:** Lakebed Systems
- **O:** Riverbed systems
- **T:** Mainland Systems
- **V:** Wetland Systems
- **H:** Marine waters
- **F:** Limnic bodies of water
- **In:** Snow and Ice Systems

The main focus of the data is mainland systems (T) which include all land-based ecosystems except those that are saturated, as those are categorized under Wetland Systems.This system has 45 subcategories, and the majority of terrestrial habitat types are part of the mainland systems.The dataset used for training in this thesis consists of 20 of the subcategories listed below:

1. T1 Bare rock
2. T2 Open shallow land
3. T3 Mountain heath, lee side and tundra
4. T4 Permanent forest land
5. T6 Strandberg
6. T12 Strand meadow
7. T13 Rasmark
8. T18 Open flood-resistant land
9. T21 Dune field
10. T29 Gravel and stone-dominated beach and shoreline

11. T30 Floodplain woodland
12. T31 Boreal hi
13. T32 Semi-natural meadow
14. T33 Semi-natural salt marsh
15. T34 Kystlynghei
16. T35 Strongly modified firm ground with loose soil cover
17. T38 Tree plantation
18. T39 Strongly changed and new firm ground in slow succession
19. T43 Strongly changed, permanent firm ground with an intensive character
20. T45 Cultivated permanent meadow

3.3 Data pre-processing

The pre-processing of the data consists of two parts, first the label dataset should be cleaned and then the satellite images should also be prepared. In case of the labels they should be in a format readable by google earth engine, the null values should be removed from them and also they should be split into training, test and validation sets.
For the satellite images, the area that is needed should be clipped, the satellite bands should be selected and clouds should be removed. Then the labels can be overlapped on the images to collect the input data for training machine learning models.

3.3.1 Preparing the labels

As discussed above the labels contains 43 different biomes in the mainland system (T) category. The table consists of different attributed (biome categories) and other columns which had the description of each biome. Since the categories are needed, the other columns were removed from the table. There were some properties with missing labels which were removed from the data.
Since the table was in Qgis format and converted to shape file which is readable in GEE. Then imported to the platform, now the data can be loaded and manipulated in GEE. The next step is separating the data into training and validation groups. In order to do this in GEE, a random column is added to the table which has a random number range in the range of [0, 1) based on a 'normal' distribution with the mean of 0 and standard deviation of 1. This column is used to randomly select the data for the split.
To prevent classed being removed from the input data, the separation ratio should be done for each biome class and not on the whole dataset. Therefore the table is separated to sub tables for each class, the the split is done and in the end the classes will join together resulting in two tables for training and validation.

3.3.2 Preparing satellite image collection

First step is loading the satellite image collection, which is in the GEE database and should be loaded. Then image collection can be filtered based on the requirements. Some of the filters consist of:

- **Metadata:** There are different metadata based on the satellite, here we filter by images that have less than 30% cloud coverage. So wen can mask these clouds later. Images with more cloud coverage can be be harder to mask.

- **Time:** Using this filter we can set a time range to gather the images taken by satellite during that period.
- **Location:** This filter is used to clip the images for the geometry that we are working on, which is Norway bounds.
- **Bands:** Depending on the use case certain bands of the satellite are needed, here we can select which bands we want to use. In this project all the bands where used.

When using different satellite bands, it is not the color data that we have, it is the amount of light that the satellite receives in that specific band. For example when we are showing the green band the lightness and darkness of different areas shows the amount of green color in the environment. We can then relate the values for red, green and blue to their colors and show a colored image.
As mentioned, we can also combine different bands to add to our image as a seperate band and have more input data. There are many indices introduced in various research that gives us a way to combine bands for a specific purose. Some examples of simple bands combinations are as follows.

- **Color Infrared (B8, B4, B3):** The color infrared band combination is meant to emphasize healthy and unhealthy vegetation.

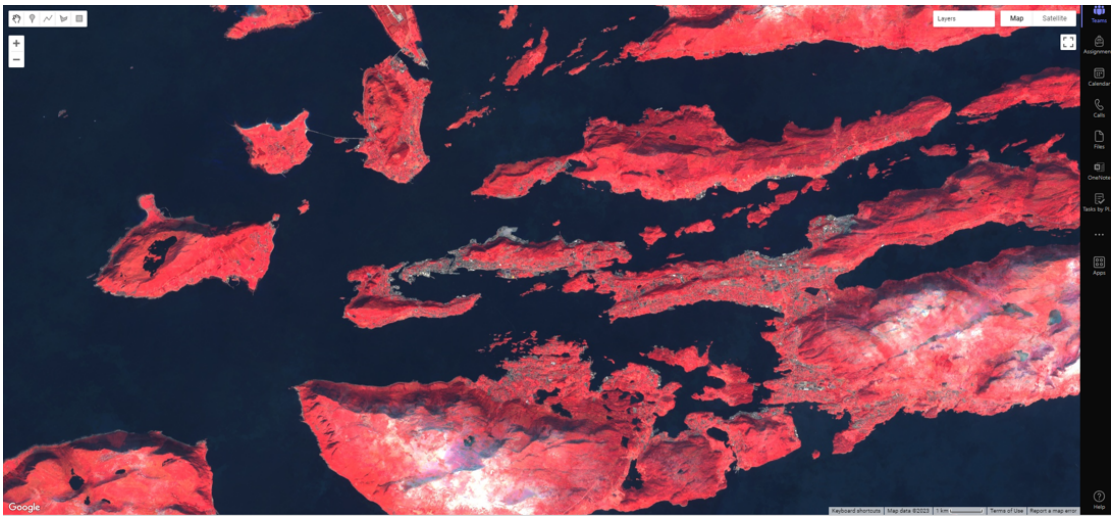


Figure 3.3.1: Color infrared bands visualized

- **Short-Wave Infrared (B12, B8A, B4):** This composite shows vegetation in various shades of green. In general, darker shades of green indicate denser vegetation. But brown is indicative of bare soil and built-up areas.

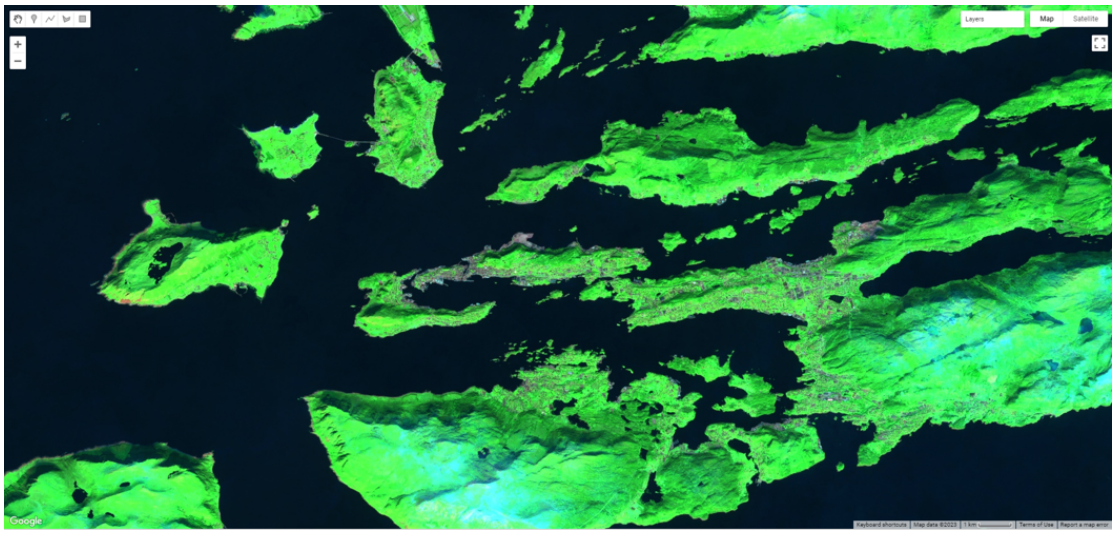


Figure 3.3.2: Short wave infrared bands visualized

- **Agriculture (B11, B8, B2):** It's mostly used to monitor the health of crops because of how it uses short-wave and near-infrared. Both these bands are particularly good at highlighting dense vegetation that appears as dark green.

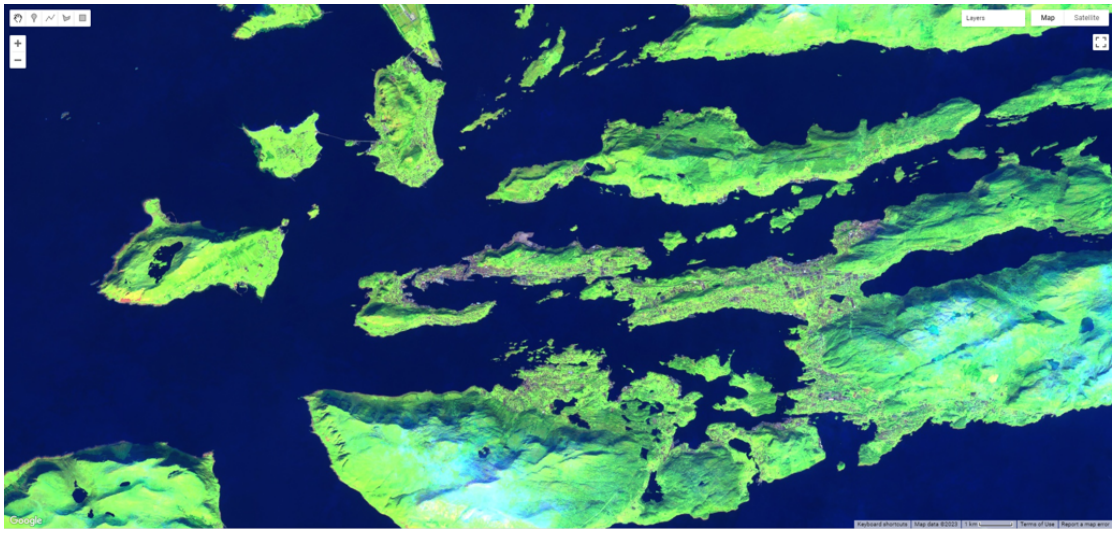


Figure 3.3.3: Agriculture wave infrared bands visualized

This can be also done for none visible bands. For example we can relate the ultra blue light values to color blue and have an image that the color blue represents the amount of ultra blue. This can be done for the sake of visualization.

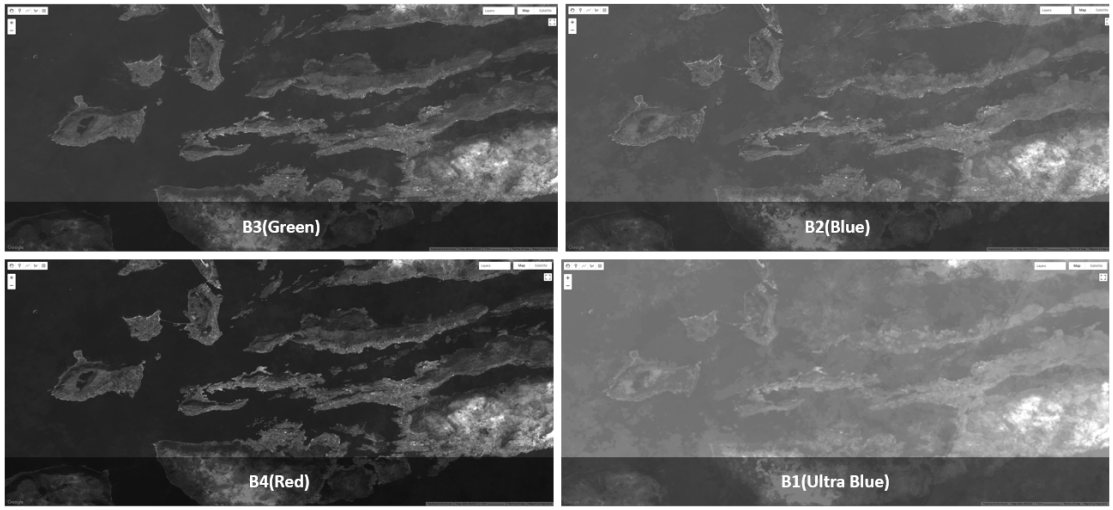


Figure 3.3.4: Comparing the data gathered from different bands

Since each pixel in the image has the data of different bands over a period of time, we can have the median value of the pixel. Since we get the median for a time period, which sometime has cloud over the image and sometimes not; this results in an image with less clouds.

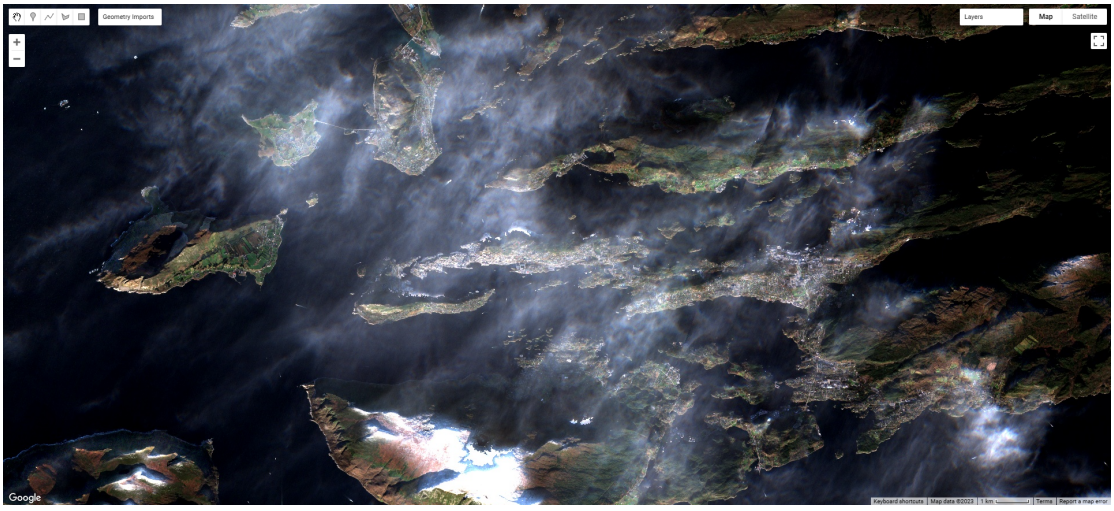


Figure 3.3.5: Before using median to remove clouds

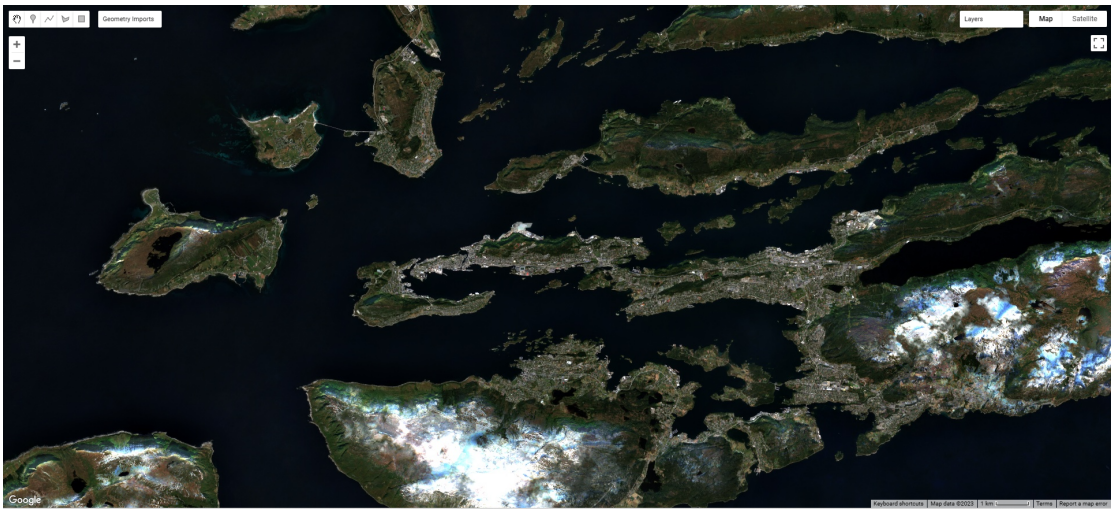


Figure 3.3.6: After using median to remove clouds

After preparing the image collection, the labels and be overlapped over the image collection to extract the training data. This is done using a sample function in GEE, this function gets the table including the labels, the name of the property that we want to classify (here land type) and scale which defines the amount of details needed. It is important that more details require more memory and computing power which may not be feasible or even unattainable. This function outputs a feature collection in which each feature has one property per band of input image. The feature collection can be used for training the desired model.

3.4 Model training

After preparing the data they can be fed to different classifiers for training. GEE uses Smile (Statistical Machine Intelligence and Learning Engine) library, which contains different classifier models as well as tools for assessing the results. Each model has different parameters that should be fine tunes to yield the best results.

- **Random forest:** This models gets the image collection as the input, the name of the column with classes (land type) the bands used in classification. For the model we can set the number of decision trees to create, the mean leaf population, number of variables per split, the fraction of input to bad per tree, the maximum number of nodes and a seed for the randomization. For hyperparamter tuning we can change the number of trees to check the results.
- **SVM:**The inputs for this function is the same as random forests (image collection, class and bands). Some of the parameters contain, the decision procedure (voting or margin), the SVM type, the kernel type.

CHAPTER
FOUR

RESULTS AND DISCUSSION

The input data involved 3000 labeled polygons for training and 500 for validation, which covered the 20 different biome categories. The training data was obtained by sampling satellite data from each pixel in these polygons. This resulted in a dataset of 33092 for training and 3751 for validation. The data was used to train different models and tune them. The properties used for training consist of all the bands in sentinel2 satellite, elevation and slope, and indices of, Ndvi, Ndbi, Mndwi, Nbal, Mnbal, Bsi and Evi. Random forests, SVM and KNN models were trained and their paramters were tuned to yield the best results. In this chapter the results are presented, compared and discussed.

4.1 Random Forests model

The random forest model was trained with with different number of trees (from 10 to 100) to find the best accuracy for the validation data, This can be observed in the figure 4.1.1.

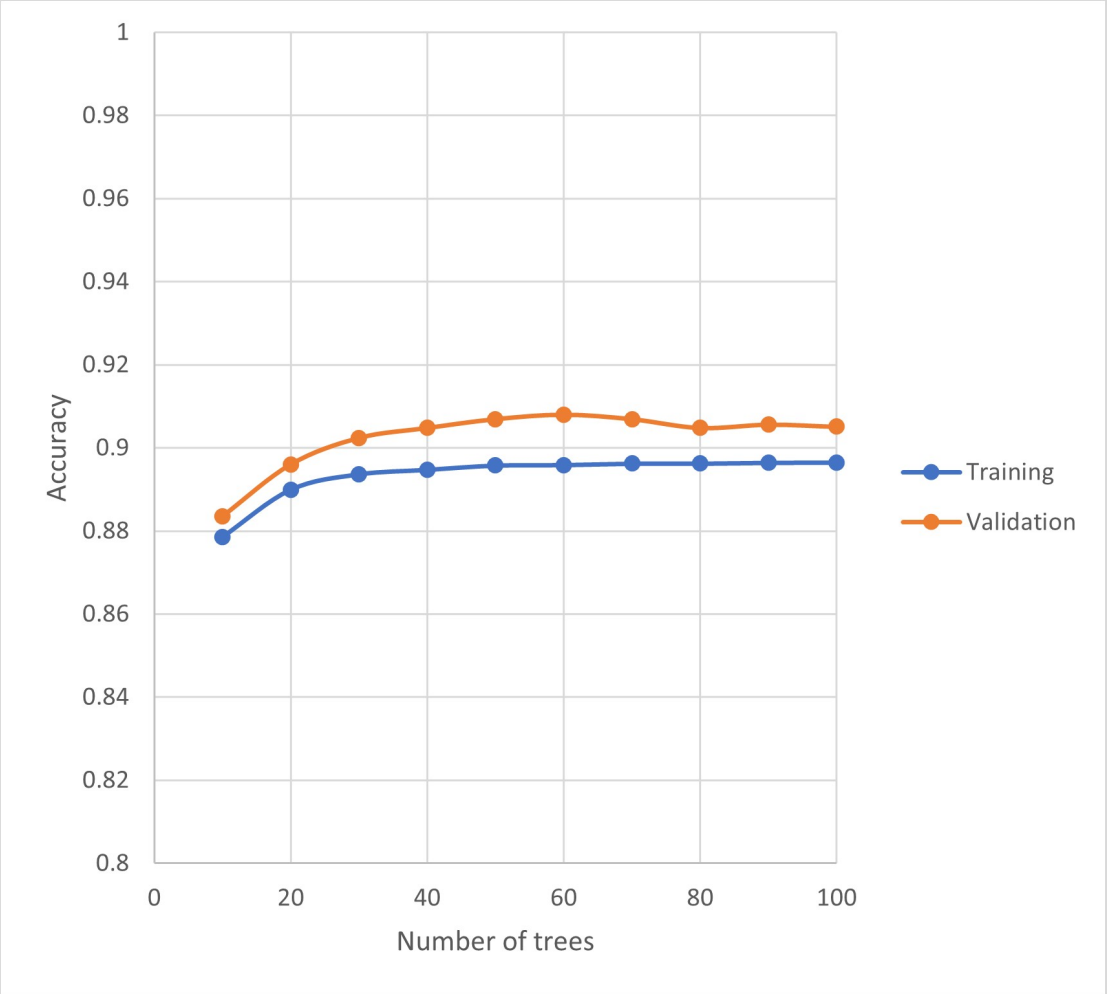


Figure 4.1.1: Random Forests

As it can be observed in the graph both the training and validation and have high accuracy and changing the number of trees doesn't increase the accuracy significantly.The highest value can be achieved using 60 trees for the model. This model has 89.6% accuracy for training and 90.8% for validation.
Increasing the number of trees doesn't always result in significant improvement.Since after a certain point adding more trees lead to the diminishing return, meaning the contribution of a new tree doesn't have a significant effect on the overall accuracy. This is due to the fact that the model uses ensemble learning and each tree votes for the best result so after a certain number of trees, the vote of a new tree can not change the result drastically. Also the number of trees also depends on the characteristics of the dataset, depending on how well the data can be classified we may need lower number of trees to get to an acceptable result.
WE must also take into account that increasing the number of trees will result in and increased training time, higher memory usage and also increased prediction time. Therefore it is desirable to select a number of trees that has a balance between model performance and computational efficiency.
Using the best parameters for the model we can calculate the confusion matrix for the validation data as seen in figure 4.1.2.

	T1	T2	T3	T4	T6	T12	T13	T18	T21	T29	T30	T31	T32	T33	T34	T35	T38	T39	T43	T45		
T1	327	0	0	1	0	0	15	0	1	0	0	0	0	0	0	0	0	0	0	0	95.1%	Producer's Accuracy
T2	2	87	0	3	0	0	6	0	0	0	0	4	1	0	1	0	0	0	0	0	83.7%	
T3	15	5	363	17	0	0	32	26	0	1	1	0	3	0	0	0	0	0	0	0	78.4%	
T4	0	3	1	378	0	0	0	0	0	0	2	3	7	0	0	0	0	0	0	0	95.9%	
T6	0	1	0	1	88	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	96.7%	
T12	0	1	1	0	0	39	0	1	5	0	0	0	0	1	2	0	0	4	0	0	72.2%	
T13	1	1	0	4	0	0	932	0	1	0	0	0	0	0	0	0	0	0	0	0	99.3%	
T18	0	0	0	0	0	0	0	55	0	0	0	0	0	0	0	0	0	0	0	0	100.0%	
T21	1	1	3	0	0	1	1	0	387	10	1	0	0	0	0	0	0	16	16	0	88.6%	
T29	1	2	0	0	6	7	0	0	2	42	0	0	0	0	0	0	0	0	0	0	70.0%	
T30	0	1	0	1	0	0	1	1	0	0	98	0	1	0	0	2	0	0	0	0	93.3%	
T31	1	0	0	0	0	0	0	0	0	0	0	111	0	0	3	0	1	1	0	0	94.9%	
T32	0	1	2	0	0	0	0	1	0	0	19	0	94	0	0	0	1	2	0	1	77.7%	
T33	0	0	0	0	0	1	0	3	2	0	0	0	0	21	0	0	0	0	1	2	70.0%	
T34	0	5	0	1	0	0	1	0	0	0	0	1	0	0	95	0	0	0	0	0	92.2%	
T35	0	1	0	0	0	0	0	0	0	0	0	1	2	0	0	27	0	0	0	1	84.4%	
T38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	113	0	0	0	100.0%	
T39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	0	0	100.0%	
T43	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	29	1	85.3%	
T45	0	2	0	2	0	0	1	0	0	0	5	0	3	0	2	3	0	3	14	62	63.9%	
	94.0%	78.4%	98.1%	92.2%	93.6%	81.3%	94.2%	63.2%	97.2%	79.2%	77.8%	92.5%	84.7%	95.5%	92.2%	84.4%	98.3%	66.7%	48.3%	92.5%	90.8%	
	Consumer's Accuracy																				Overall Accuracy	

Figure 4.1.2: Confusion matrix for validation data

Here we can see the correctly predicted values for each class in the main diagonal of the matrix. In the other cells we have the number of misclassified parameters in each class.

We have consumers accuracy which the measure that shows how well the model correctly identifies the data in a specific class in relation to the total number of data predicted at that class. For example in the first column; for class T1 out of the 348 data that the model predicted as this class, 94% (327) was correct and 4% belonged to the other classes (T2,T3,T13,T21,T29 and T31) that are misclassified. As it can be observed the model has high consumers accuracy across different classes. The T43 exhibits the lowest levels, primarily because it contains mostly regularly managed areas, such as lawns that undergo mowing, weeding, liming, spraying, etc. These areas are in close proximity to various other biomes from the remote sensing standpoint.

There is also the producers accuracy that is the measure of how well the model correctly classifies the data in a specific class in relation to the total number of data in that class. For example for the same class of T1, out of the 344 data, 327 were correctly classified which is 95.1% and 17 were classified as the classes of T4, T13 and T1, which is 4.9%.

Using the random forests model we can assess the features used and have the relative feature important (in percentage) chart as shown in the figure below.

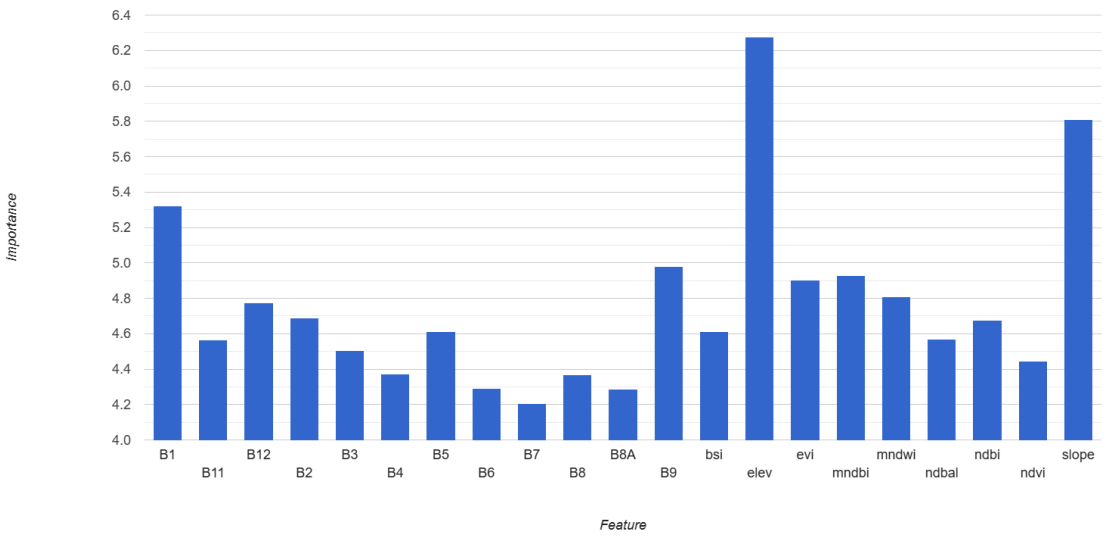


Figure 4.1.3: Feature importance chart

This gives a general idea of which feature have relatively more influence on the prediction. It can be observed that elevation and slope have the highest significance on the output of the model. This can be explained with topology of Norway, having a diverse topology in terms of elevation. From mountains, fjords to coastline. Elevation plays an important part in the biome found in different areas, this is due to the fact that with the change of elevation, the temperature, precipitation and thus vegetation will change.
The next most significant band it the first band (B1) which is the ultra blue band. This band is also known as the Coastal and Aerosol band and mostly used for the coastal studies. Taking into account that Norway has an extensive coastline along the Atlantic ocean, it could be expected that this band would have a high relative importance.

4.2 SVM model

The SVM model was trained with different kernels and gamma values to find the best parameter that fits the data.The choice of the kernel function has an important role in the performance of the classifier. Three different kernels were tested which consist of RBF, POLY and SIGMOID; the result from different kernels can be observer in figure 4.2.1.

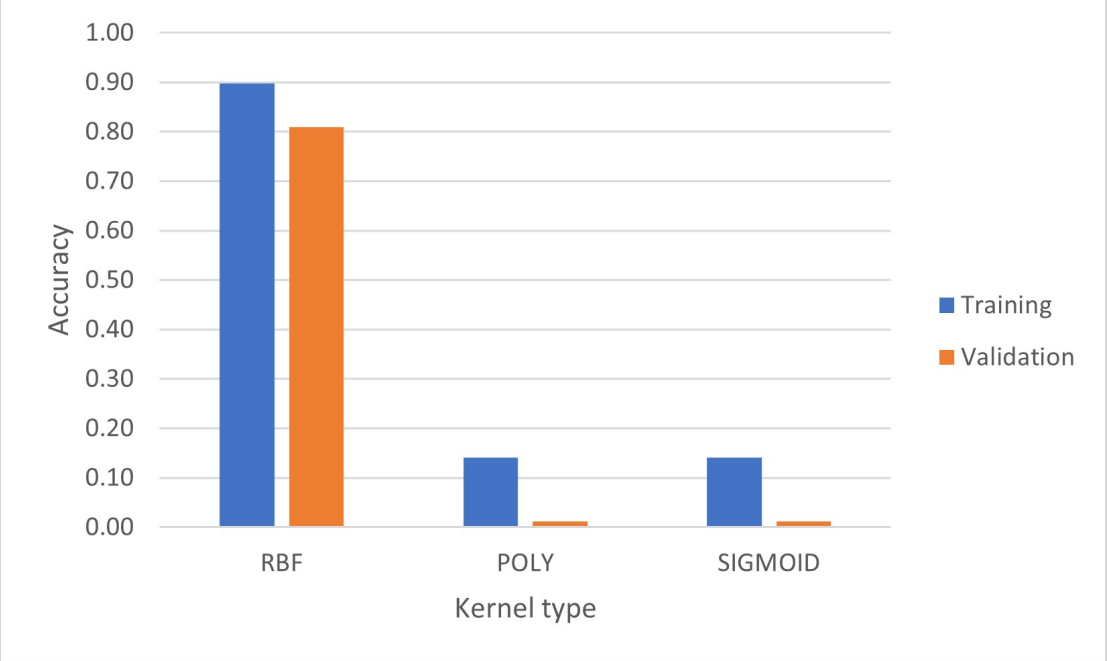


Figure 4.2.1: Comparing SVM kernels

The RBF kernel is often used in many problems due to its flexibility and the ability to make complex decision boundaries. Poly kernel is usually the choice in for complex non-linear relations and Sigmoid is useful when there is a clear separation in terms of logistic sigmoid function. Therefore since we have a complex classification problem RBF is the best choice as it can also be observed from the graph.
Therefore RBF kernel was selected and the gamma values were changed from 0.05 to 2 to detect its effect on the result (figure 4.2.2) .

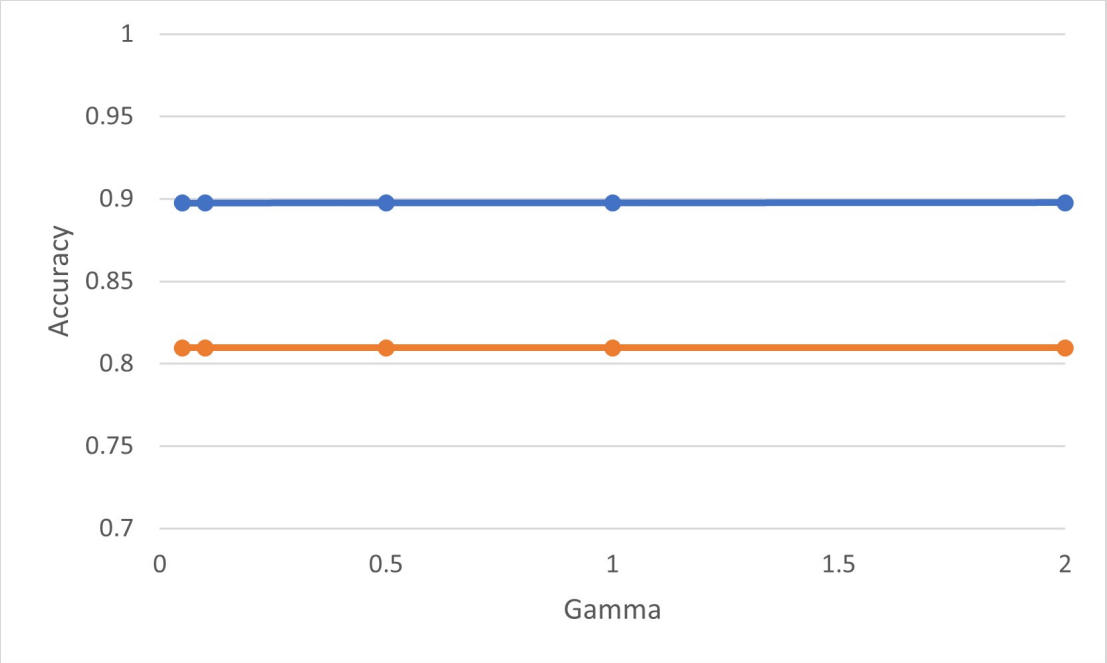


Figure 4.2.2: Comparing SVM gamma values

Changing the gamma value doesn't seem to have an effect on the accuracy of the model. It may be due to the fact that the inherent structure of the data may not require the fine tuning of this parameter and the model is already at the optimal result. The confusion matrix for validation data can be seen in figure 4.2.3.

	T1	T2	T3	T4	T6	T12	T13	T18	T21	T29	T30	T31	T32	T33	T34	T35	T38	T39	T43	T45		
T1	207	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9	0	95.4%	
T2	0	44	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	16	0	67.7%	
T3	0	0	155	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	120	0	56.4%	
T4	0	0	0	221	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	88.0%	
T6	0	0	0	0	53	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	94.6%	
T12	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	6	0	79.3%	
T13	0	0	0	0	0	0	558	0	0	0	0	0	0	0	0	0	0	0	31	0	94.7%	
T18	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	1	0	97.0%	
T21	0	0	0	0	0	0	0	0	167	0	0	0	0	0	0	0	0	0	119	0	58.4%	
T29	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	8	0	80.5%	
T30	0	0	0	0	0	0	0	0	0	0	54	0	0	0	0	0	0	0	19	0	74.0%	
T31	1	0	0	0	0	0	0	0	0	0	0	63	0	0	0	0	0	0	1	0	96.9%	
T32	0	0	0	0	0	0	0	0	0	0	0	0	56	0	0	0	0	0	16	0	77.8%	
T33	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	8	0	73.3%	
T34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	0	0	0	13	0	80.0%	
T35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	8	0	61.9%	
T38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	70	0	5	0	93.3%	
T39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	100.0%	
T43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	100.0%	
T45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	40	54.8%	
	99.5%	100.0%	100.0%	100.0%	100.0%	100.0%	98.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	5.7%	100.0%	81.0%	
	Consumer's Accuracy																				Overall Accuracy	

Producer's Accuracy

Figure 4.2.3: Confusion matrix for SVM

It can be observed that for many of the classes the accuracy is very high or near perfect. The only parameter that lowers the accuracy by large margin it the T43 class. The best achieved accuracy was 89.8% for training and 81% for validation data.

4.3 Model comparison

In the figure 4.4.1 the comparison of the models can be observed.

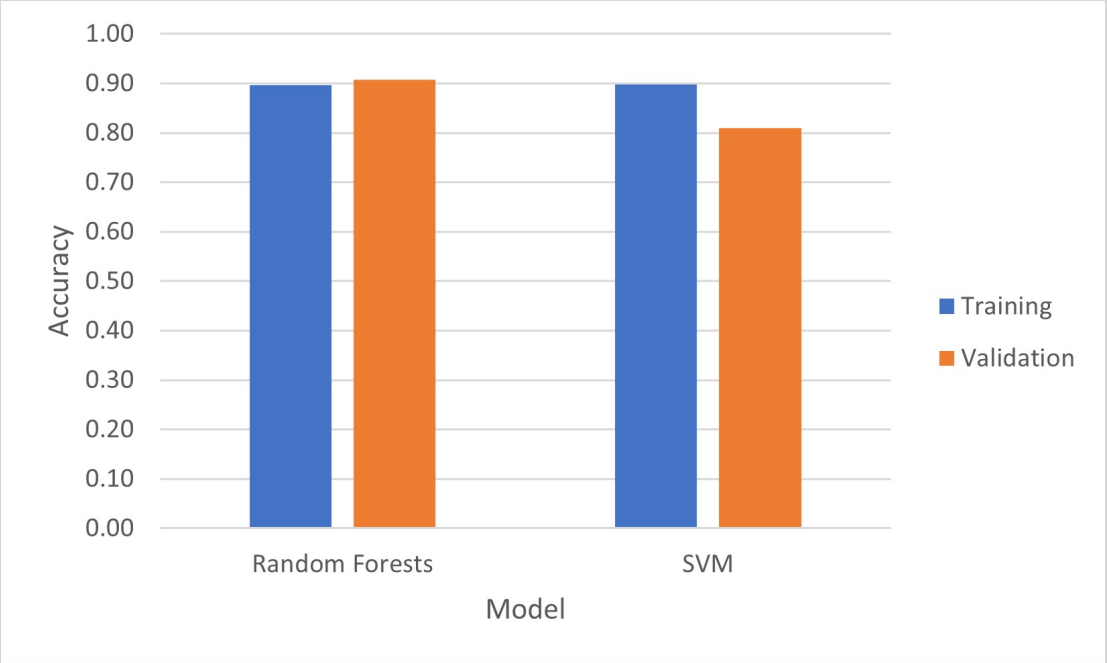


Figure 4.3.1: Comparing the trained models

As it can be observed from the chart, both models perform well with the tuned parameters. But random forests model performs slightly better. Also if we look back at the confusion matrices, for most individual classes SVM model has better performance, but since it is weaker in classifying the T43 class, it would be generally better to use the random forests model. It can also be considered that deploying the random forests in google earth engine requires less time and resources; thus we can use a better scale to have more detailed data and also classify the image with more accuracy.

4.4 Prediction results

Here the trained random forests model was utilized to classify the land for the whole country (figure 4.4.1).

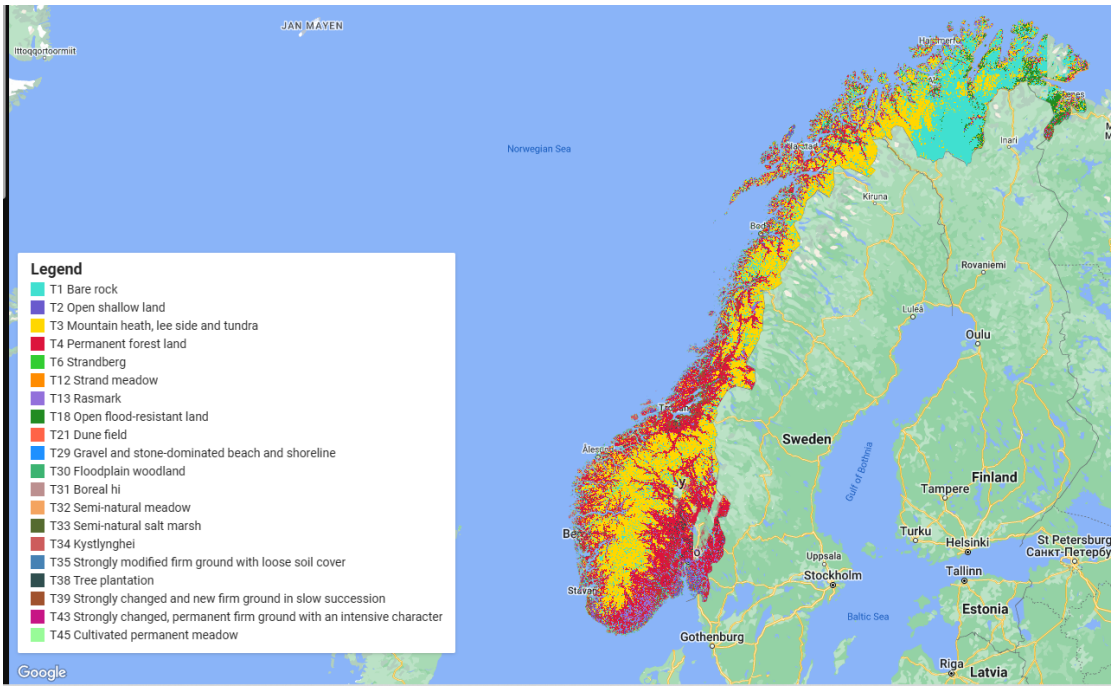


Figure 4.4.1: Classified map of Norway using Random forests model

The map reveals a dominant presence of yellow color, indicating extensive mountainous regions, particularly along the borders with Sweden and in the southern and southwestern parts of the country. Another noteworthy biome, T4, representing permanent forest lands, stands out prominently in the south and west. Through the examination of the map and considering the accuracy of the models, it can be inferred that the predictions for the 20 distinct land classes are reliable. These models offer a valuable tool for studying and assessing the condition and health of the various biomes. Utilizing this model allows for the extraction of diverse data, which proves useful for assessing each biome. The accompanying chart provides a practical overview of the area covered by different biomes in Norway, aligning well with the map observations. This combination of model-generated data and visual representation contributes to a more comprehensive understanding of each biome's characteristics in the Norwegian context.

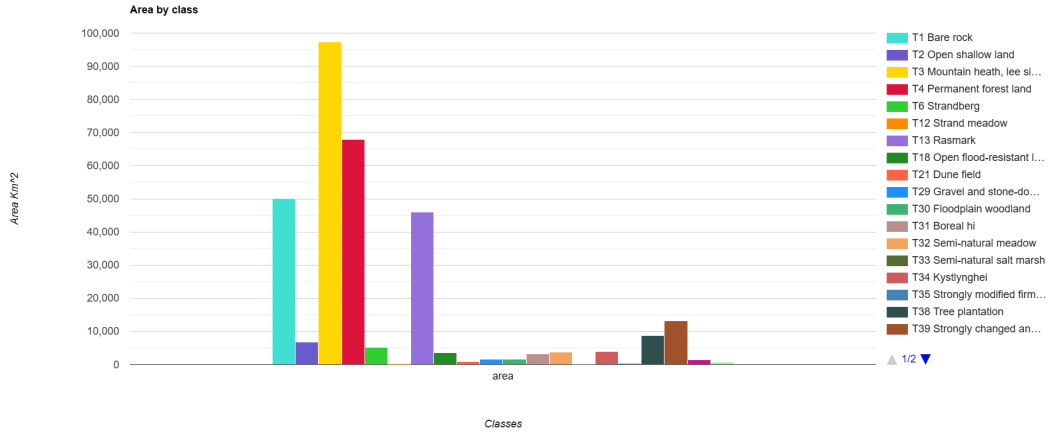


Figure 4.4.2: Distribution of Area Across Categories

Furthermore, a pie chart can be utilized to present the distribution of biomes across the country. This visual aid provides a straightforward and effective way to grasp the relative proportions of different biomes in Norway, offering a practical supplement to the data extracted from the model for a more accessible analysis of landscape characteristics.

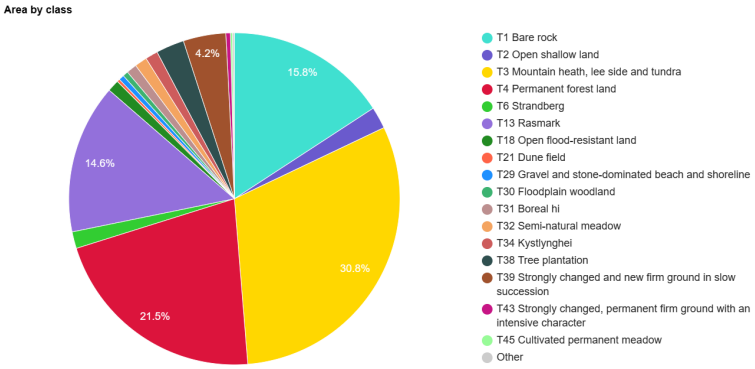


Figure 4.4.3: Proportional Area Distribution by Category

CHAPTER

FIVE

CONCLUSIONS AND FUTURE WORKS

The two reviewed models were both capable of the prediction with high accuracy. Random forests was selected because of its ability to classify the class that was hardest and also its lower training and modeling time. By using the model it becomes possible to better study each different ecosystem which gives the grounds for assessing the health of different biomes.

in future research data from other categories of the land can be used in the same method to train a model with wider range of applications. Different methods of working with the data such as running to calculations locally and using the python libraries can be studied. Since google earth engine is limited in terms of processing power and memory for each user. This may result in improved classification and also being able to use more input data for training.

REFERENCES

[1] Eugene P. Odum. "Ecosystem, Concept of". In: 2001. URL: https://api.semanticscholar.org/CorpusID:127046010.

[2] Michael Conrad. "The Ecosystem Process". In: 1983. URL: https://api.semanticscholar.org/CorpusID:80807648.

[3] Frank B. Golley. "The ecosystem concept: A search for order". In: *Ecological Research* 6 (1991), pp. 129–138. URL: https://api.semanticscholar.org/CorpusID:27458153.

[4] Katherine Richardson et al. "Earth beyond six of nine planetary boundaries". In: *Science Advances* 9.37 (2023), eadh2458. DOI: 10.1126/sciadv.adh2458. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.adh2458. URL: https://www.science.org/doi/abs/10.1126/sciadv.adh2458.

[5] Will Steffen et al. "Planetary boundaries: Guiding human development on a changing planet". In: *Science* 347.6223 (2015), p. 1259855. DOI: 10.1126/science.1259855. eprint: https://www.science.org/doi/pdf/10.1126/science.1259855. URL: https://www.science.org/doi/abs/10.1126/science.1259855.

[6] Frank Biermann. "Planetary boundaries and earth system governance: Exploring the links". In: *Ecological Economics* 81 (2012). Special Section: "Planetary Boundaries" and Global Environmental Governance, pp. 4–9. ISSN: 0921-8009. DOI: https://doi.org/10.1016/j.ecolecon.2012.02.016. URL: https://www.sciencedirect.com/science/article/pii/S0921800912000808.

[7] Jonathan Pickering and Åsa Persson. "Democratising planetary boundaries: experts, social values and deliberative risk evaluation in Earth system governance". In: *Journal of Environmental Policy & Planning* 22.1 (2020), pp. 59–71. DOI: 10.1080/1523908X.2019.1661233.

[8] Tiina Häyhä et al. "From Planetary Boundaries to national fair shares of the global safe operating space — How can the scales be bridged?" In: *Global Environmental Change* 40 (2016), pp. 60–72. ISSN: 0959-3780. DOI: https://doi.org/10.1016/j.gloenvcha.2016.06.008. URL: https://www.sciencedirect.com/science/article/pii/S0959378016300826.

[9] Lars Hein et al. "Progress and challenges in the development of ecosystem accounting as a tool to analyse ecosystem capital". In: *Current Opinion in Environmental Sustainability* 14 (2015). Open Issue, pp. 86–92. ISSN: 1877-3435. DOI: https://doi.org/10.1016/j.cosust.2015.04.002. URL: https://www.sciencedirect.com/science/article/pii/S1877343515000408.

[10] C. Duku et al. "Towards ecosystem accounting: a comprehensive approach to modelling multiple hydrological ecosystem services". In: *Hydrology and Earth System Sciences* 19.10 (2015), pp. 4377–4396. DOI: 10.5194/hess-19-4377-2015. URL: https://hess.copernicus.org/articles/19/4377/2015/.

[11] Catherine Anne Farrell et al. "Applying the System of Environmental Economic Accounting-Ecosystem Accounting (SEEA-EA) framework at catchment scale to develop ecosystem extent and condition accounts". In: *One Ecosystem* 6 (2021), e65582. DOI: 10.3897/oneeco.6.e65582. eprint: https://doi.org/10.3897/oneeco.6.e65582. URL: https://doi.org/10.3897/oneeco.6.e65582.

[12] Elham Sumarga et al. "Mapping monetary values of ecosystem services in support of developing ecosystem accounts". In: *Ecosystem Services* 12 (2015), pp. 71–83. ISSN: 2212-0416. DOI: https://doi.org/10.1016/j.ecoser.2015.02.009. URL: https://www.sciencedirect.com/science/article/pii/S2212041615000273.

[13] *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on nature restoration*.

[14] Zhaoqin Li, Dandan Xu, and Xulin Guo. "Remote Sensing of Ecosystem Health: Opportunities, Challenges, and Future Perspectives". In: *Sensors* 14.11 (2014), pp. 21117–21139. ISSN: 1424-8220. DOI: 10.3390/s141121117. URL: https://www.mdpi.com/1424-8220/14/11/21117.

[15] D. Anthony et al. "On Crop Height Estimation With UAVs". In: (2014). DOI: 10.1109/iros.2014.6943245.

[16] T.M. Lillesand and R.W. Kiefer. *Remote Sensing and Image Interpretation.* Wiley, New York, 2015.

[17] Zhiqin Zhu et al. "Remote Sensing Image Defogging Networks Based on Dual Self-Attention Boost Residual Octave Convolution". In: *Remote Sensing* (2021). DOI: 10.3390/rs13163104.

[18] John Ball, Derek T. Anderson, and Chee Seng Chan. "Comprehensive Survey of Deep Learning in Remote Sensing: Theories, Tools, and Challenges for the Community". In: *Journal of Applied Remote Sensing* (2017). DOI: 10.1117/1.jrs.11.042609.

[19] Yuanxin Ye et al. "Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity". In: *Ieee Transactions on Geoscience and Remote Sensing* (2017). DOI: 10.1109/tgrs.2017.2656380.

[20] Damien Arvor et al. "Ontologies to Interpret Remote Sensing Images: Why Do We Need Them?" In: *Giscience & Remote Sensing* (2019). DOI: 10.1080/15481603.2019.1587890.

[21] Dongliang Wang, Quanqin Shao, and Huanyin Yue. "Surveying Wild Animals From Satellites, Manned Aircraft and Unmanned Aerial Systems (UASs): A Review". In: *Remote Sensing* (2019). DOI: 10.3390/rs11111308.

[22] Yang Shi et al. "Urban Land Use and Land Cover Classification Using Multisource Remote Sensing Images and Social Media Data". In: *Remote Sensing* (2019). DOI: 10.3390/rs11222719.

[23] Jon Atli Benediktsson and Zebin Wu. "Distributed Computing for Remotely Sensed Data Processing [Scanning the Section]". In: *Proceedings of the Ieee* (2021). DOI: 10.1109/jproc.2021.3094335.

[24] Dalton Lunga et al. "Apache Spark Accelerated Deep Learning Inference for Large Scale Satellite Image Analytics". In: *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2020). DOI: 10.1109/jstars.2019.2959707.

[25] Amit Panwar. "Classifying Hyperspectral Imageries". In: *International Journal for Research in Applied Science and Engineering Technology* (2017). DOI: 10.22214/ijraset.2017.8037.

[26] Suiqiong Li, Aleksandr Simonian, and Bryan A. Chin. "Sensors for Agriculture and the Food Industry". In: *The Electrochemical Society Interface* (2010). DOI: 10.1149/2.f05104if.

[27] Wagh Sharad. "The Development of the Earth Remote Sensing From Satellite". In: *Mechanics of Gyroscopic Systems* (2021). DOI: 10.20535/0203-3771402020248768.

[28] Alibek Yussupov and Raya Z. Suleimenova. "Use of Remote Sensing Data for Environmental Monitoring of Desertification". In: *Evergreen* (2023). DOI: 10.5109/6781080.

[29] Huiwen Dong et al. "Development of on-Board Radiation Calibration Device for Water Color and Water Temperature Scanner". In: (2023). DOI: 10.1117/12.2661744.

[30] *Electromagnetic Spectrum*. commons.wikimedia.org/wiki/File:Electromagnetic-Spectrum.png. -.

[31] Angel Caroline Johnsy and Gilda Schirinzi. "A Lossless Coding Scheme for Maps Using Binary Wavelet Transform". In: *European Journal of Remote Sensing* (2017). DOI: 10.1080/22797254.2017.1274154.

[32] Assefa M. Melesse et al. "Remote Sensing Sensors and Applications in Environmental Resources Mapping and Modelling". In: *Sensors* (2007). DOI: 10.3390/s7123209.

[33] Seyd Teymoor Seydi and Mahdi Hasanlou. "Land Cover Change Detection Based on Genetically Feature Aelection and Image Algebra Using Hyperion Hyperspectral Imagery". In: *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2015). DOI: 10.5194/isprsarchives-xl-1-w5-669-2015.

[34] Weizhen Hou et al. "Study on the Spectral Reconstruction of Typical Surface Types Based on Spectral Library and Principal Component Analysis". In: (2019). DOI: 10.1117/12.2521743.

[35] Yonghua Jiang et al. "On-Orbit Radiance Calibration of Nighttime Sensor of LuoJia1-01 Satellite Based on Lunar Observations". In: *Remote Sensing* (2019). DOI: 10.3390/rs11182183.

[36] Dong Jiang et al. "Advances in Multi-Sensor Data Fusion: Algorithms and Applications". In: *Sensors* (2009). DOI: 10.3390/s91007771.

[37] *Comparison of Landsat 7 and 8 bands with Sentinel-2*. https://landsat.gsfc.nasa.gov/wp-content/uploads/2015/06/Landsat.v.Sentinel-2.png. -.

[38] Nicola Clerici, Cesar Augusto Valbuena Calderón, and Juan M. Posada. "Fusion of Sentinel-1a and Sentinel-2a Data for Land Cover Mapping: A Case Study in the Lower Magdalena Region, Colombia". In: *Journal of Maps* (2017). DOI: 10.1080/17445647.2017.1372316.

[39] Nguyen Thanh Doan. "Improving the Efficiency of Using Deep Learning Model to Determine Shoreline Position in High-Resolution Satellite Imagery". In: *E3s Web of Conferences* (2021). DOI: 10.1051/e3sconf/202131004002.

[40] J. L. E. Gesta et al. "Aboveground Biomass and Carbon Stock Estimation of Falcata Through the Synergistic Use of Sentinel-1 and Sentinel-2 Images". In: *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2023). DOI: 10.5194/isprs-archives-xlviii-4-w6-2022-117-2023.

[41] Paulo Amador Tavares et al. "Integration of Sentinel-1 and Sentinel-2 for Classification and LULC Mapping in the Urban Area of Belém, Eastern Brazilian Amazon". In: *Sensors* (2019). DOI: 10.3390/s19051140.

[42] Congshuang Xie et al. "Satellite-Derived Bathymetry Combined With Sentinel-2 and ICESat-2 Datasets Using Machine Learning". In: *Frontiers in Earth Science* (2023). DOI: 10.3389/feart.2023.1111817.

[43] Zhiwei Yi, Jia Li, and Qiting Chen. "Crop Classification Using Multi-Temporal Sentinel-2 Data in the Shiyang River Basin of China". In: *Remote Sensing* (2020). DOI: 10.3390/rs12244052.

[44] Laura Stendardi et al. "Exploiting Time Series of Sentinel-1 and Sentinel-2 Imagery to Detect Meadow Phenology in Mountain Regions". In: *Remote Sensing* (2019). DOI: 10.3390/rs11050542.

[45] Saeid Nasseri, Bahman Farhadi Bansouleh, and Arash Azari. "Estimation of Land Surface Temperature in Agricultural Lands Using Sentinel 2 Images: A Case Study for Sunflower Fields". In: *Irrigation and Drainage* (2023). DOI: 10.1002/ird.2802.

[46] Elena C. Rodriguez-Garlito and Abel Paz. "Mapping Invasive Aquatic Plants in Sentinel-2 Images Using Convolutional Neural Networks Trained With Spectral Indices". In: *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2023). DOI: 10.1109/jstars.2023.3257142.

[47] Pingkan Mayestika Afgatiani et al. "Determination of Sentinel-2 Spectral Reflectance to Detect Oil Spill on the Sea Surface". In: *Sustinere Journal of Environment and Sustainability* (2020). DOI: 10.22515/sustinere.jes.v4i3.115.

[48] Gourav Misra, Fiona Cawkwell, and Astrid Wingler. "Status of Phenological Research Using Sentinel-2 Data: A Review". In: *Remote Sensing* (2020). DOI: 10.3390/rs12172760.

[49] Jingfa Wang. "A Study of Forest Swamp Mapping in Hani Wetland Integrating Sentinel-1 and Sentinel-2 Satellite Images". In: (2021). DOI: 10.3233/faia210392.

[50] Shuangyi Wang et al. "Snow Cover Mapping and Ice Avalanche Monitoring From the Satellite Data of the Sentinels". In: *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2018). DOI: 10.5194/isprs-archives-xlii-3-1765-2018.

[51] R. Esmaeili Sarteshnizi, S. Sahebi Vayghan, and I. Jazirian. "Estimation of Soil Moisture Using Sentinel-1 and Sentinel-2 Images". In: *Isprs Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2023). DOI: 10.5194/isprs-annals-x-4-w1-2022-137-2023.

[52] Pingkan Mayestika Afgatiani, Argo Galih Suhadha, and A. Ibrahim. "The Capability of Sentinel-1 Polarization Combinations for Oil Spill Detection (Study Case: Karawang, Indonesia)". In: *Iop Conference Series Earth and Environmental Science* (2022). DOI: 10.1088/1755-1315/1109/1/012078.

[53] Mario Busquier et al. "Combination of Time Series of L-, C-, and X-Band SAR Images for Land Cover and Crop Classification". In: *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2022). DOI: 10.1109/jstars.2022.3207574.

[54] Yangyang Zhang, Jian Yang, and Lina Du. "Analyzing the Effects of Hyperspectral ZhuHai-1 Band Combinations on LAI Estimation Based on the PROSAIL Model". In: *Sensors* (2021). DOI: 10.3390/s21051869.

[55] Yanan Yan et al. "Application of UAV-Based Multi-Angle Hyperspectral Remote Sensing in Fine Vegetation Classification". In: *Remote Sensing* (2019). DOI: 10.3390/rs11232753.

[56] Fabrício B Silva et al. "Large-Scale Heterogeneity of Amazonian Phenology Revealed From 26-Year Long AVHRR/NDVI Time-Series". In: *Environmental Research Letters* (2013). DOI: 10.1088/1748-9326/8/2/024011.

[57] Yunita Nurmasari and Arie Wahyu Wijayanto. "Oil Palm Plantation Detection in Indonesia Using Sentinel-2 and Landsat-8 Optical Satellite Imagery (Case Study: Rokan Hulu Regency, Riau Province)". In: *International Journal of Remote Sensing and Earth Sciences (Ijreses)* (2021). DOI: 10.30536/j.ijreses.2021.v18.a3537.

[58] Amit Hasan et al. "Detection of Clouds in Medium-Resolution Satellite Imagery Using Deep Convolutional Neural Nets". In: *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2022). DOI: 10.5194/isprs-archives-xlvi-m-2-2022-103-2022.

[59] Gonzalo Mateo-García, L. Gomez-Chova, and Gustau Camps-Valls. "Convolutional Neural Networks for Multispectral Image Cloud Masking". In: (2017). DOI: 10.1109/igarss.2017.8127438.

[60] Xiaolong Li et al. "Cloud Detection of SuperView-1 Remote Sensing Images Based on Genetic Reinforcement Learning". In: *Remote Sensing* (2020). DOI: 10.3390/rs12193190.

[61] Xidong Chen et al. "A Novel Classification Extension-Based Cloud Detection Method for Medium-Resolution Optical Images". In: *Remote Sensing* (2020). DOI: 10.3390/rs12152365.

[62] Andrea Meraner et al. "Cloud Removal in Sentinel-2 Imagery Using a Deep Residual Neural Network and SAR-optical Data Fusion". In: *Isprs Journal of Photogrammetry and Remote Sensing* (2020). DOI: 10.1016/j.isprsjprs.2020.05.013.

[63] Xi Wu and Zhenwei Shi. "Utilizing Multilevel Features for Cloud Detection on Satellite Imagery". In: *Remote Sensing* (2018). DOI: 10.3390/rs10111853.

[64] M. S. Sadiq et al. "A Review on Machine Learning in Smart Antenna: Methods and Techniques". In: *Tem Journal* (2022). DOI: 10.18421/tem112-24.

[65] Erik Bründermann, Andrea Santamaria Garcia, and Anke-Susanne Müller. "Machine Learning Applications". In: (2020). DOI: 10.1515/9783110610987.

[66] None Anggi Rachmawati and None Yossaepurrohman. "Analysis of Machine Learning Systems for Cyber Physical Systems". In: *International Transactions on Education Technology (Itee)* (2022). DOI: 10.34306/itee.v1i1.170.

[67] "Machine Learning in Chemical Industry". In: *International Journal of Advances in Scientific Research and Engineering* (2017). DOI: 10.7324/ijasre.2017.32524.

[68] Shadman Latif et al. "Investigation of Machine Learning Algorithms for Network Intrusion Detection". In: *International Journal of Information Engineering and Electronic Business* (2022). DOI: 10.5815/ijieeb.2022.02.01.

[69] Liangpei Zhang, Lefei Zhang, and Bo Du. "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art". In: *Ieee Geoscience and Remote Sensing Magazine* (2016). DOI: 10.1109/mgrs.2016.2540798.

[70] D. I. Rukhovich et al. "The Use of Deep Machine Learning for the Automated Selection of Remote Sensing Data for the Determination of Areas of Arable Land Degradation Processes Distribution". In: *Remote Sensing* (2021). DOI: 10.3390/rs13010155.

[71] Xue Li, Alex de Sherbinin, and Yanni Zhan. "Mapping Urban Extent at Large Spatial Scales Using Machine Learning Methods With VIIRS Nighttime Light and MODIS Daytime NDVI Data". In: *Remote Sensing* (2019). DOI: 10.3390/rs11101247.

[72] Lei Xue et al. "Large-Scale High-Resolution Coastal Mangrove Forests Mapping Across West Africa With Machine Learning Ensemble and Satellite Big Data". In: *Frontiers in Earth Science* (2021). DOI: 10.3389/feart.2020.560933.

[73] Courage Kamusoko. "Geospatial Machine Learning in Urban Environments: Challenges and Prospects". In: (2021). DOI: 10.1007/978-981-16-5149-6_1.

[74] Aries Suharso et al. "The Role of Machine Learning in Remote Sensing for Agriculture Drought Monitoring: A Systematic Review". In: *International Journal of Advanced Computer Science and Applications* (2022). DOI: 10.14569/ijacsa.2022.0131290.

[75] Lisa C. Kelley, L. H. Pitcher, and Chris Bacon. "Using Google Earth Engine to Map Complex Shade-Grown Coffee Landscapes in Northern Nicaragua". In: *Remote Sensing* (2018). DOI: 10.3390/rs10060952.

[76] Pinki Mondal et al. "Evaluating Combinations of Sentinel-2 Data and Machine-Learning Algorithms for Mangrove Mapping in West Africa". In: *Remote Sensing* (2019). DOI: 10.3390/rs11242928.

[77] Xunlai Chen et al. "High Spatial Resolution PM_{2.5} Retrieval Using MODIS and Ground Observation Station Data Based on Ensemble Random Forest". In: *Ieee Access* (2019). DOI: 10.1109/access.2019.2908975.

[78] Jennifer N. Hird et al. "Google Earth Engine, Open-Access Satellite Data, and Machine Learning in Support of Large-Area Probabilistic Wetland Mapping". In: *Remote Sensing* (2017). DOI: 10.3390/rs9121315.

[79] Volker Walter. "Object-Based Classification of Remote Sensing Data for Change Detection". In: *Isprs Journal of Photogrammetry and Remote Sensing* (2004). DOI: 10.1016/j.isprsjprs.2003.09.007.

[80] Yanling Han et al. "Combining 3d-CNN and Squeeze-and-Excitation Networks for Remote Sensing Sea Ice Image Classification". In: *Mathematical Problems in Engineering* (2020). DOI: 10.1155/2020/8065396.

[81] Ahmed Ben Hamida et al. "3-D Deep Learning Approach for Remote Sensing Image Classification". In: *Ieee Transactions on Geoscience and Remote Sensing* (2018). DOI: 10.1109/tgrs.2018.2818945.

[82] Peng Liu et al. "SVM or Deep Learning? A Comparative Study on Remote Sensing Image Classification". In: *Soft Computing* (2016). DOI: 10.1007/s00500-016-2247-2.

[83] Bin Xia et al. "Land Resource Use Classification Using Deep Learning in Ecological Remote Sensing Images". In: *Computational Intelligence and Neuroscience* (2022). DOI: 10.1155/2022/7179477.

[84] Lorenzo Bruzzone, Mingmin Chi, and Mattia Marconcini. "A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images". In: *Ieee Transactions on Geoscience and Remote Sensing* (2006). DOI: 10.1109/tgrs.2006.877950.

[85] Yi Lin et al. "Performance Evaluation of Elm With a-Optimized Design Regularization for Remote Sensing Imagery Classification". In: *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2020). DOI: 10.5194/isprs-archives-xliii-b1-2020-45-2020.

[86] Xuedong Yao et al. "Land Use Classification of the Deep Convolutional Neural Network Method Reducing the Loss of Spatial Features". In: *Sensors* (2019). DOI: 10.3390/s19122792.

[87] Rhonda D. Phillips et al. "An SMP Soft Classification Algorithm for Remote Sensing". In: *Computers & Geosciences* (2014). DOI: 10.1016/j.cageo.2014.03.010.

[88] Gulnaz Alimjan et al. "A New Technique for Remote Sensing Image Classification Based on Combinatorial Algorithm of SVM and KNN". In: *International Journal of Pattern Recognition and Artificial Intelligence* (2018). DOI: 10.1142/s0218001418590127.

[89] Jian Zhao and Xin Pan. "Remote Sensing Image Feature Selection Based on Rough Set Theory and Multi-Agent System". In: (2015). DOI: 10.1109/fskd.2015.7382028.

[90] Xiaoshuang Yin et al. "Semi-Supervised Feature Learning for Remote Sensing Image Classification". In: (2014). DOI: 10.1109/igarss.2014.6946662.

[91] Pall Oskar Gislason, Jón Atli Benediktsson, and Jóhannes R. Sveinsson. "Random Forests for Land Cover Classification". In: *Pattern Recognition Letters* (2006). DOI: 10.1016/j.patrec.2005.08.011.

[92] Mariana Belgiu and Lucian Drăguţ. "Random Forest in Remote Sensing: A Review of Applications and Future Directions". In: *Isprs Journal of Photogrammetry and Remote Sensing* (2016). DOI: 10.1016/j.isprsjprs.2016.01.011.

[93] Mahesh Pal. "Random Forest Classifier for Remote Sensing Classification". In: *International Journal of Remote Sensing* (2005). DOI: 10.1080/01431160412331269698.

[94] Thanh Noi Phan and Martin Kappas. "Comparison of Random Forest, K-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery". In: *Sensors* (2017). DOI: 10.3390/s18010018.

[95] Zoe Pierrat et al. "Forests for Forests: Combining Vegetation Indices With Solar-Induced Chlorophyll Fluorescence in Random Forest Models Improves Gross Primary Productivity Prediction in the Boreal Forest". In: *Environmental Research Letters* (2022). DOI: 10.1088/1748-9326/aca5a0.

[96] Rajkumar Saini and S. K. Ghosh. "Crop Classification on Single Date Sentinel-2 Imagery Using Random Forest and Suppor Vector Machine". In: *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2018). DOI: 10.5194/isprs-archives-xlii-5-683-2018.

[97] Robert Walker and Marcus J. Hamilton. "Machine Learning With Remote Sensing Data to Locate Uncontacted Indigenous Villages in Amazonia". In: *Peerj Computer Science* (2019). DOI: 10.7717/peerj-cs.170.

[98] Kai Cheng, Juanle Wang, and Xinrong Yan. "Mapping Forest Types in China With 10 M Resolution Based on Spectral–Spatial–Temporal Features". In: *Remote Sensing* (2021). DOI: 10.3390/rs13050973.

[99] Giorgos Mountrakis, Jungho Im, and Caesar Ogole. "Support Vector Machines in Remote Sensing: A Review". In: *Isprs Journal of Photogrammetry and Remote Sensing* (2011). DOI: 10.1016/j.isprsjprs.2010.11.001.

[100] Farid Melgani and Lorenzo Bruzzone. "Classification of Hyperspectral Remote Sensing Images With Support Vector Machines". In: *Ieee Transactions on Geoscience and Remote Sensing* (2004). DOI: 10.1109/tgrs.2004.831865.

[101] Aaron E. Maxwell, Timothy A. Warner, and Fang Fang. "Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review". In: *International Journal of Remote Sensing* (2018). DOI: 10.1080/01431161.2018.1433343.

[102] Iklil Faqihah Mohamad Nadzri. "Analyzing the Effectiveness of Support Vector Machine and Random Forest Classifiers in Delineating the Green Area". In: *Iop Conference Series Earth and Environmental Science* (2023). DOI: 10.1088/1755-1315/1217/1/012032.

[103] Luanjie Chen et al. "An Improved Multi-Source Data-Driven Landslide Prediction Method Based on Spatio-Temporal Knowledge Graph". In: *Remote Sensing* (2023). DOI: 10.3390/rs15082126.

[104] G. Wang, J. Liu, and Guojin He. "Object-Based Land Cover Classification for ALOS Image Combining TM Spectral". In: *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* (2013). DOI: 10.5194/isprsarchives-xl-7-w2-263-2013.

[105] Giles M. Foody and Anupam Mathur. "A Relative Evaluation of Multiclass Image Classification by Support Vector Machines". In: *Ieee Transactions on Geoscience and Remote Sensing* (2004). DOI: 10.1109/tgrs.2004.827257.

[106] Dataaspirant. "SVM Kernel Functions: A Detailed Guide". In: *Dataaspirant* (May 2021). URL: https://dataaspirant.com/svm-kernel-functions/.

[107] Larhmam. *Maximum-margin hyperplane and margin for an SVM trained on two classes. Samples on margins are called support vectors.* Own work. https://upload.wikimedia.org/wikipedia/commons/7/72/SVM_margin.png. Oct. 2018.

[108] Suraj Yadav. "Support Vector Machines: The Kernel Trick". In: (2024). URL: https://medium.com/@Suraj_Yadav/support-vector-machines-the-kernel-trick-3f9c11875920.

[109] José Luis Rojo-Álvarez et al. "Support Vector Machine and Kernel Classification Algorithms". In: *Digital Signal Processing with Kernel Methods*. 2018, pp. 433–502. DOI: 10.1002/9781118705810.ch10.

[110] Unknown. "Difference between using a hard margin and a soft margin in SVM". In: (Unknown). URL: https://www.edge.com/%E2%80%A6.

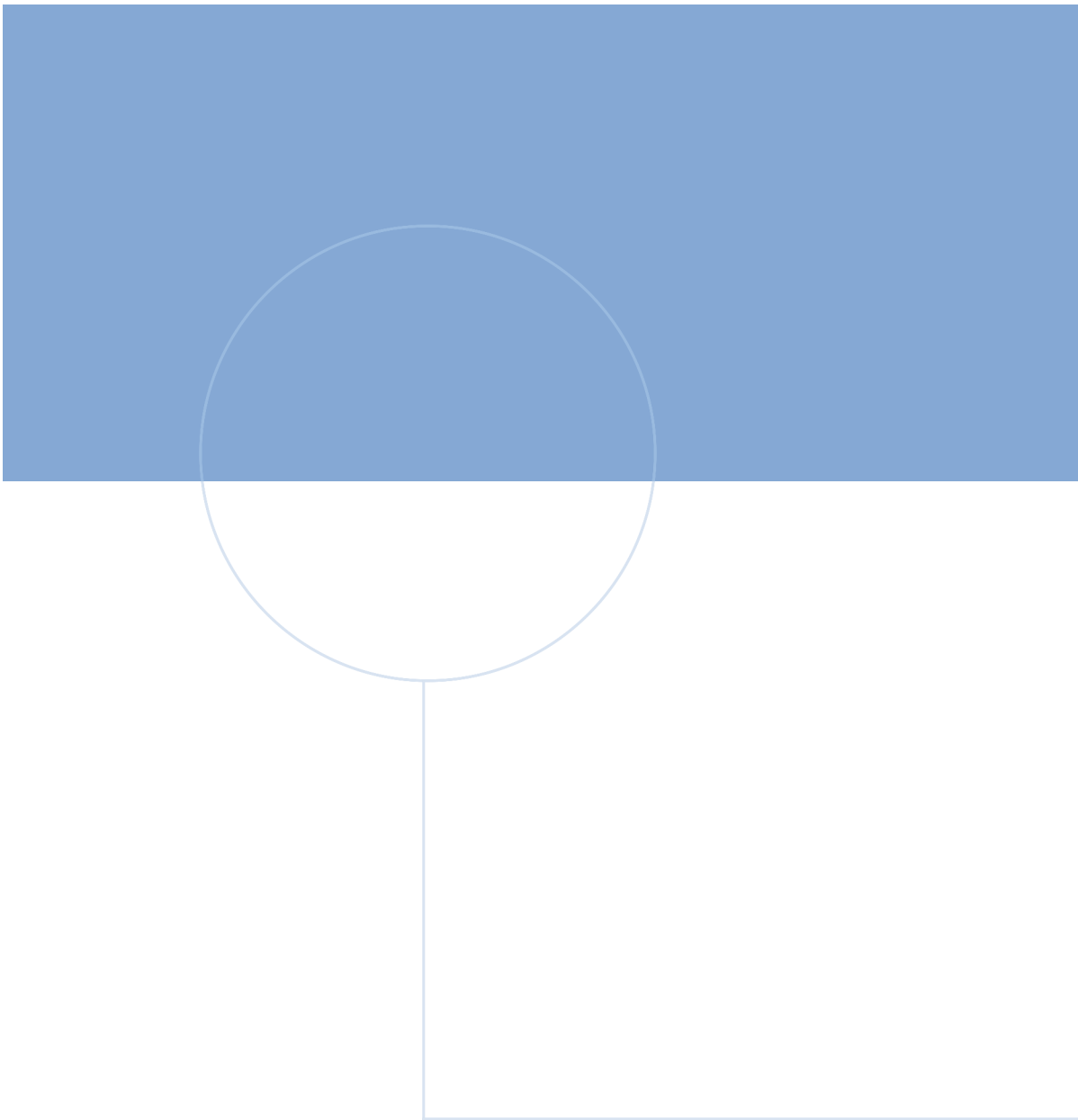
[111] Rahul Rastogi. *Random Forest classification and it's mathematical implementation*. July 2020. URL: https://medium.com/analytics-vidhya/random-forest-classification-and-its-mathematical-implementation-1895a7bb743e.

[112] URL: https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf.

[113] Feb. 2021. URL: https://www.aifinesse.com/random-forest/random-forest-complexity/.

[114] V. Henrich et al. "IDB - www.indexdatabase.de, Entwicklung einer Datenbank für Fernerkundungsindizes". In: *AK Fernerkundung, Bochum, 4.-5. 10. 2012*. (PDF). 2012.

[115] *The Norwegian Biodiversity Information Centre*. Org. number: 919 666 102. Postal address: Artsdatabanken, PO Box 1285 Torgarden, NO-7462 Trondheim.



NTNU

Norwegian University of
Science and Technology