Quality Assessment of Danish government geographical data using the gradient boosting algorithm Catboost

Mark Takacs

AALBORG UNIVERSITET
KØBENHAVN

A Thesis in the Field of Geoinformatics

For the Degree of Master of Geoinformatics

Alborg University Copenhagen

June 2018

Title:

Quality Assessment of Danish government geographical data using the gradient boosting algorithm Catboost

Keywords:

Remote Sensing, Sentinel, Catboost, Machine learning, Open government data

Author:

Mark Takacs

Principal Supervisor:

Jamal Jokar Arsanjani, PhD. AAU Geoinformatics

Secondary Supervisor: Mikkel Lydholm Rasmussen. DHI GRAS

Project period: 5th July 2018 – 13th September 2018

ECTS: 30 ECTS

Education: MSc in Geoinformatics

University: Aalborg University Copenhagen

AAU Master of Science Programme in Geoinformatics
A. C. Meyers Vænge 15
2450 København SV
Denmark

Preface

This thesis has been submitted as a final semester project for the Master of Science in Geoinformatics at Aalborg University Copenhagen.

The thesis was begun and completed in close collaboration with DHI GRAS, where I received a great deal of assistance from my colleagues throughout the data collection and processing stages of this project.

Many thanks to my two supervisors, Jamal and Mikkel. Jamal for his help and support throughout university, and Mikkel for his assistance in initiating and developing this project at DHI GRAS.

Mark Takacs

Copenhagen, September 2018.

Abstract

Geodata has become one the fundamental forces in the world today, valuable to all industries and government bodies. Just like any product or asset, its value or usefulness is reflective of its quality or validity. The field of remote sensing is continually growing, the data it provides is utilised in a multitude of applications across all sectors. This thesis focuses on Geographical spatial data, specifically water bodies. This data can be used in various land management operations such as disaster response, precision farming and even used to inform government policy decisions, to name a few.

This study explores the subject of data quality. Denmark is one of the leaders in open geographical data, its quality or accuracy is relatively unknown and subject to scrutiny. This thesis examines the 'lake' or water body dataset provided by GeoDanmark, the Danish framework for cooperation between municipalities and e Data Security and Enhancement Board (SDFE) on the establishment and maintenance of a nationwide public-sector geographic data.

The method used in assessing the quality of data is the combination of remotely sensed data provided by ESA's recently launched sentinel 2 satellite programme and Yandex's newly developed machine learning algorithm Catboost.

The results proved to be interesting, with a 28% of the sample data set to be wrong. This figure is subject to bias, due to the seasonal changes of water bodies. Catboost was shown to be an effective tool for assessing the quality of geo spatial datasets. In conclusion this is a significant result and opens the door for further examination of Denmark's geographic open datasets.

Table of Contents

List of Abbreviations

CERN - European Organization for Nuclear Research

CSV - Commas Separated Values

EMR - Electromagnetic Radiation

ESA - European Space Agency

ETM+ - Enhanced Thematic Mapper Plus

GBM - Gradient Boosting Machine

GST - Danish Geodata Agency

KMS - Kort & Matrikelstyrelsen

LEO - Low Earth Orbit

MidIR - Middle Infrared

MOIS - Moderate-Resolution Imaging Spectroradiometer

NASA - National Aeronautics and Space Administration

NDWI - Normalized difference water index

NIR - Near Infrared

OGD - Open Government Data

OLI - Operational Land Imager

PPGIS - Participatory Geographic Information Systems/Science

RGB - Red Green & Blue

SDFE - Styrelsen for Dataforsyning og Effektivisering

SDTS - Spatial Data Transfer Standard

TM - Thematic Mapper

TOA - Top of Atmosphere

VGI - Volunteered Geographic Information

List of Figures

List of Tables

1 Introduction

The Quality of Data in an information age is essential, being aware of the level of quality is a necessary parameter in determining the overall significance of a dataset. Errors in spatial data can result in a number of problems depending on the utilisation of the data, from inaccurate land management practices to flawed disaster prevention methods.

The field of machine learning provides boundless opportunity for the manipulation, editing and analysis of large data sets, a machine learning approach to assessing the quality of spatial data may provide a significant development within the area of spatial analysis.

The following chapter of this report will provide the background theory and research in the fields of remote sensing and machine learning relevant to this thesis study, concluding with the projects problem statement and research questions.

1.1 Background

Generally described, remote sensing is a method of "collecting and analysing data to acquire information about an object without the instrument used to collect the data being in direct contact with the object" (ESA, n.d.). Remote sensors acquire data by recording the energy that is reflected or emitted from Earth. These sensors can be installed on multiple platforms from satellites to aircraft (Figure 1).

Remote sensors can either be active or passive. Passive sensors measure external stimuli. They detect natural energy that is reflected or emitted from the Earth's surface. The most common source of radiation detected by passive sensors is reflected sunlight (NOAA, n.d.). In contrast, active sensors use internal stimuli for earth data collection, precipitation radars

are a form of active sensor systems. The precipitation radar measures the radar echo from rainfall to determine the precipitation rate on the Earth's surface (NASA, n.d.). This project will only use passively sensed data.

Remote sensing has a unique advantages over other forms of environmental measurement methods. These advantages include the estimation of parameters and surface/subsurface properties without direct contact with the area of measurement (i.e., non-invasiveness); the capability of making remote observations (figure 1), thereby preventing risks for the operator and reducing costs of in situ measurements; the possibility to revisit in time and carry out iterative workflows of data analysis for the purposes of monitoring and condition assessment (e.g., multi-temporal change detection) (Tapete 2018).

The number of fields applicable to remote sensing are many, from coastal, Oceanic, Hazard assessments and natural resource management. The technology is constantly advancing and provides the basis for copious amounts of advancement and analysis.

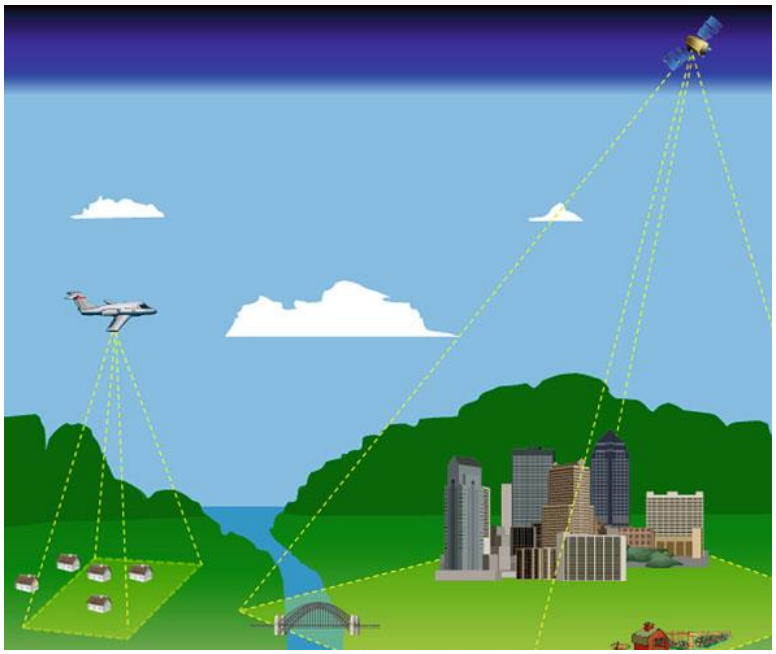


Figure 1 Remote sensing platforms. An instrument (i.e., sensor or scanner) is mounted on an aircraft or satellite that records data about the target scene or object, usually electromagnetic data (Khorram, 2012)

1.2 The Electromagnetic Spectrum

Electromagnetic radiation (EMR) is defined as all energy that moves with the velocity of light in a harmonic wave pattern (i.e., all waves are equally and repetitively spaced in time). Visible light is just one category of EMR; other types include radio waves, infrared, and gamma rays. Together, all of these types comprise the electromagnetic spectrum (Figure 2). As illustrated by Fig. 2, the different forms of EMR vary across the spectrum in terms of both wavelength and frequency (Khorram et al, 2012).

Wavelength is the distance between the positions in two wave cycles, while frequency is the number of wave cycles passing the same point in a given time period (1 cycle per s = 1 Hertz, or Hz). The mathematical relationship between wavelength and frequency is expressed by the following equation: C = km, where k is wavelength, m is frequency, and C is the speed of light (which is constant at 300,000 km per s in a vacuum). Visible light, represents only a small portion of the electromagnetic spectrum, it ranges in wavelength from about 3.9×10^{-7} m (violet) to 7.5×10^{-7} m (red), and has corresponding frequencies that range from 7.9×10^{14} to 4×10^{14} (figure 2).

In remote sensing, an instrument (i.e., sensor or scanner) is mounted on an aircraft or satellite that documents information about objects or areas on the ground. Usually, the data records the level of electromagnetic energy that the target has. The extent of the geographic area captured depends on the sensor's technical specifications and the altitude of the craft in which it is mounted.

When EMR comes into contact with matter (i.e., any object or material, such as trees, water, or atmospheric gases), there are a number of interactions that can occur: absorption,

reflection, scattering, or emission of EMR by the matter, or transmission of EMR through the matter. Remote sensing typically is concerned with the recording and detection of reflected and emitted EMR. Every object or material has particular emission and/or reflectance property, collectively known as its spectral signature, which distinguishes it from other objects and materials. Remote sensors are attuned to collect these ‘‘spectral signatures’’.

Spectral data can be recorded in two formats; analog (i.e., aerial photographs, popular before the digital era) or, more commonly, digital format (i.e., a two-dimensional matrix, or image, composed of pixels that store EMR values recorded by a satellite-mounted array) (Jensen 2005).

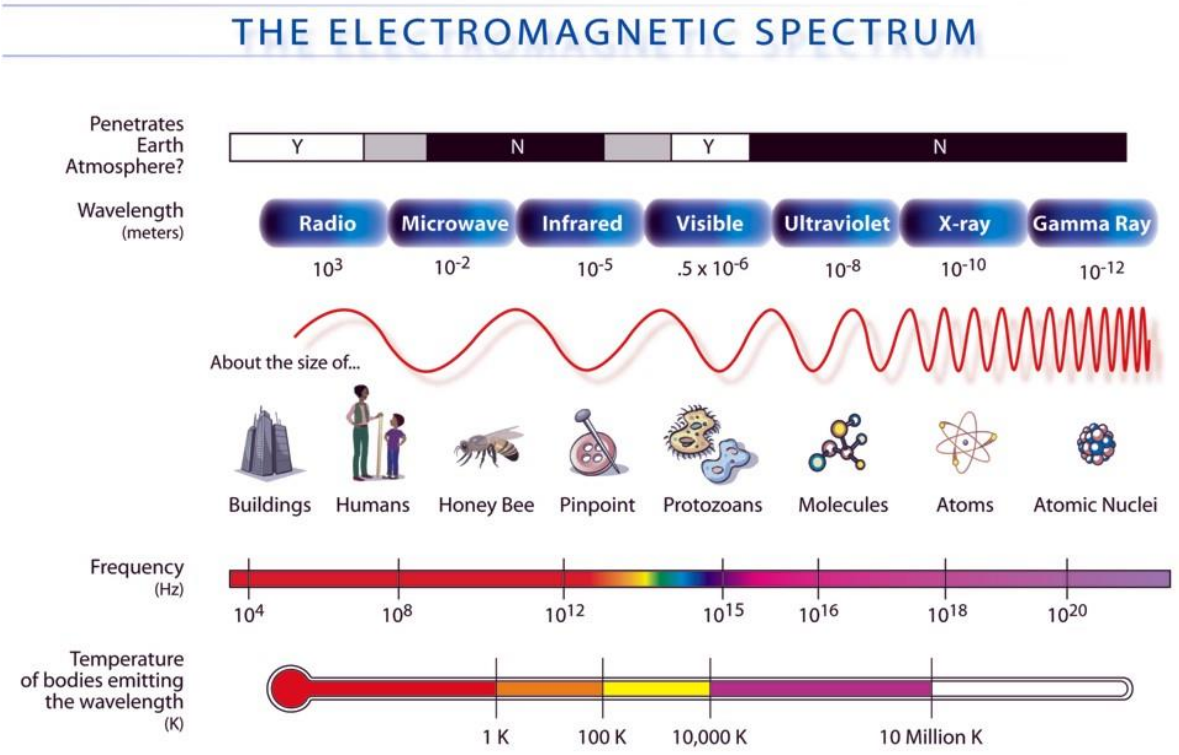


Figure 2 Electromagnetic spectrum (NASA,n.d).

Furthermore, sensors can be put into two categories; passive or active sensors. Passive sensors ; the more popular category of sensors presently in operation around the globe, record naturally occurring EMR that is either reflected or emitted from areas and objects of interest. In contrast, active sensors—such as microwave (i.e., Radio Detection and Ranging, or radar) systems—send artificial EMR toward the features of interest and then record how much of that EMR is reflected back to the system (Jensen 2005).

1.3 Data Resolutions

Remotely sensed data is primarily described by four types of resolutions:

1.3.1 Temporal Resolution

The temporal resolution stipulates the revisiting frequency of a satellite sensor for a target location. The following is an example of the temporal resolution categories.

- High temporal resolution: < 24 hours - 3 days
- Medium temporal resolution: 4 - 16 days
- Low temporal resolution: > 16 days

The revisit value refers to the period of time it takes a satellite to complete one complete orbit of the earth. The revisit period can range from 24 hours to over 16 days, consequently the complete temporal resolution of a remote sensing system is the equivalent to the period it takes the satellite to record the exact identical area at the same angle a second time. With an increase in overlap due to increasing latitudes and the amount of overlap in the imaging swaths of parallel satellite orbits, specific areas of the earth can be monitored more regularly. Satellites can correspondingly focus their sensors to the target area between different

satellites paths divided by periods from one to five days. The temporal resolution depends on multiple factors, including swath overlap, satellite capabilities and latitude.

The time of day or season has a large influence on satellite images. Specific target bodies can vary swiftly over time, for instance, the tides effect the sea, constantly expanding and withdrawing, or alternatively deciduous forests may lose their leaves during winter causing it to be harder to accurately distinguish green vegetation.

1.3.2 Spectral Resolution

The sensor's spectral resolution details the amount of spectral bands (red, green, blue, NIR, Mid-IR, thermal, etc.) in which the sensor can record EMR. However the number of bands is not the only fundamental characteristic of spectral resolution. The frequency of the bands in the electromagnetic spectrum is important, as mentioned in section 1.2. The following are examples of three spectral resolution levels: High spectral resolution with 220 bands, Medium spectral resolution containing 3 - 15 bands and Low spectral resolution with 3 bands.

The sensitivity of sensors to minor alterations in electromagnetic energy. The greater the radiometric resolution of a sensor, the more sensitive it is to detecting small variances in reflected or emitted energy.

1.3.3 Spatial resolution

The spatial resolution details the pixel dimensions of satellite images covering the earth surface (figure 3). In aerial photography, it is associated to the image detail and the level at which minor objects can be detected within the image. The spatial resolution of black and white (1 Band) aerial photographs ranges from 40 to 800 lines pairs per mm. The higher the

resolution of a sensing system, the more effectively the outline of objects on the ground can be observed. The spatial resolution of an image depends on:

- The image scale factor - spatial resolution decreases as the scale factor increases.
- The value of the optical system
- The grain assembly of the photographic film
- The contrast of the unique objects
- Atmospheric scattering effects – can cause reduced contrast and resolution
- Image motion – the relative motion between the ground and sensor can cause misrepresentation.

The most inconstant factor being the atmosphere, which is difficult to forecast and varies commonly (Jensen 2009).

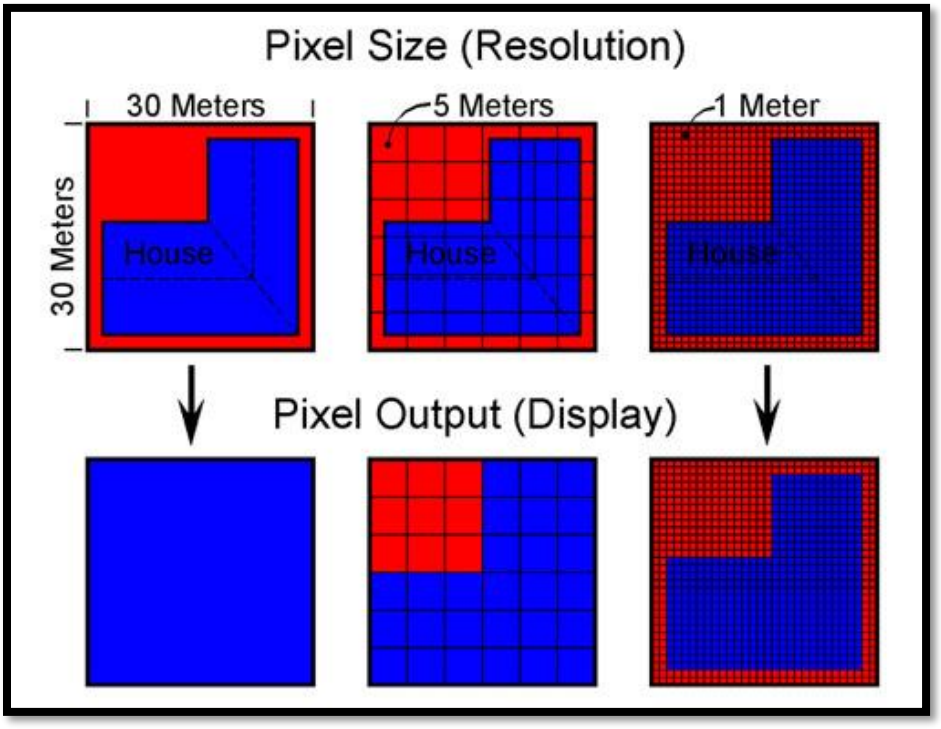


Figure 3 . Pixel size illustration (Satellite Remote Sensing Systems, n.d.)

1.3.4 Radiometric resolution

Radiometric resolution is the amount of data in a pixel and measured in units of bits. A single bit of information signifies a binary determination of yes or no, with a numerical value of 1 or 0 (Tempfli et al., 2009). Black and white images from digital cameras are usually in 8 bits, with a value range of 0-255 to denote the information. Colour images regularly have three channels in 8 bits, each channel has a value for red, green and blue. In unison they create the observed colour and the strength of each channel controls the shade, it is a technique of additive colour mixing.

A radiometric resolution of 11 means the pixel has 2048 possible intensities of blue, 12-bit resolution represents 4,096 shades of blue, and 14 bits represents 16,384 shades of blue. While increasing radiometric resolution equals a larger range for the pixel, it does not automatically mean that it is the best choice.

When designing a camera the equilibrium of the quality of the image against how many images you will be able to store, due to limited storage. The same balance is desired when determining the desired radiometric resolution for a satellite image, so that the image quality is balanced with its information capacity.

1.4 Earth Observation Satellite systems

Earth observation satellites vary according to their orbit path, the payload, and from the perspective of the imaging instrument, the data types, spectral characteristics and the swath width of the sensors. All of these parameters are set at the beginning of the mission, depending on the application of the satellite.

For example, in order to monitor the weather at large scales and high frequency, it is convenient for a satellite to be in a geostationary orbit. Enabling a continuous view of almost an entire hemisphere. However, as the orbit is a considerable distance above the earth (approximately 36,000 km) a high spatial resolution is difficult to obtain. But for applications such as the tracking of clouds over continents, a high spatial resolution is not required (ESA, n.d.).

For tasks requiring high resolution imagery of a specific area, such as the monitoring of a glacier lake, or the surveying of buildings destroyed by a natural disaster, a high spatial resolution instrument would be required. Such a sensor would typically have a narrow swath and be on a satellite at Low Earth Orbit or LEO (such as the QuickBird satellite which 600km above the earth). In such an orbit it is not possible to continuously monitor the same area, because of the constant motion of the satellite relative to the Earth. Images can only be acquired over the satellites path.

This analysis focuses on the study of land observation, specifically water body detection. For example, moderate-resolution Imaging Spectroradiometer (MODIS) images have been widely used to map water bodies on both global and regional scales. Carroll et al. (2009) developed a global raster water mask at 250-m resolution from a MODIS dataset. Feng et al

(2012) used MODIS images between 2000 and 2010 to estimate the flood changes of Poyang Lake. Huang et al. (2012) monitored water surface changes using long-term MODIS data time series. For regional studies, images provided by the Thematic Mapper (TM), the Enhanced Thematic Mapper Plus (ETM+) and the latest Operational Land Imager (OLI) from the Landsat series satellites are widely used.

Using multi-temporal Landsat TM and ETM+ images, Hui et al (2008) modelled the spatial and temporal change of Poyang Lake. Landsat OLI images were used by Du et al. (2014) to extract water body maps at subareas over the Yangtze River Basin and Huaihe River Basin in China. Additionally Rokni et al. (2014) extracted water features and monitored differences using Landsat TM, ETM+ and OLI images. When compared to MODIS, the Landsat TM, ETM+ and OLI images have much higher spatial resolutions (30 m) and can extract open water bodies with greater detail and accuracy. However, Landsat's spatial resolution images are still not high enough to adequately identify smaller-sized open water bodies, such as narrow drains and small pools. Commercial satellite systems such as SPOT6/7, IKONOS and Quick-bird, enable these small-sized water bodies to be mapped. But come at a substantial cost.

ESA launched a new optical high spatial resolution satellite, Sentinel 2 on 23 June 2015. Sentinel 2 can provide systematic global acquisitions of fine spatial resolution multispectral images with a high temporal resolution, meeting the requirements for the next generation of operational products, such as land cover maps, land cover change detection maps and geophysical variables (Drush et al., 2012, Pesaresi et al., 2016 & Immitzer 2016). The Sentinel 2 images has the potential to be of great significance for regional water bodies' mapping, due to its appealing properties (i.e., the 10-m spatial resolution for four bands and

the 10-day revisit frequency) and freely available data. As shown in Table 1, the Sentinel-2 multispectral image has 13 bands in total, in which four bands (blue, green, red and NIR) have a spatial resolution of 10 m and six bands have a spatial resolution of 20 m.

Table 1 Spectral bands for the SENTINEL-2 sensors (S2A & S2B) (ESA 2018)

	S2A		S2B		
Band Number	Central wavelength (nm)	Bandwidth (nm)	Central wavelength (nm)	Bandwidth (nm)	Spatial resolution (m)
1	443.9	27	442.3	45	60
2	496.6	98	492.1	98	10
3	560	45	559	46	10
4	664.5	38	665	39	10
5	703.9	19	703.8	20	20
6	740.2	18	739.1	18	20
7	782.5	28	779.7	28	20
8	835.1	145	833	133	10
8a	864.8	33	864	32	20
9	945	26	943.2	27	60
10	1373.5	75	1376.9	76	60
11	1613.7	143	1610.4	141	20
12	2202.4	242	2185.7	238	20

1.5 Water bodies and remote sensing

Water body extraction has become a very important part of remote sensing science as water monitoring plays an important role in water resource management. Due to their basic ability to retain, store, clean and evenly provide water, as well as their distinctive features such as still-water bodies, lakes, reservoirs and wetlands constitute essential components of the hydrological and biogeochemical water cycles influencing significant aspects of ecology, economy and human welfare. Knowledge of the distribution of lakes, reservoirs and wetlands is therefore of great interest to many scientific disciplines (Alderman et al., 2012, Bond et al., 2008, Sun et al., 2012).

Remote sensors have become a routine approach to land surface water monitoring because the acquired data can provide macroscopic, real-time, dynamic and cost-effective information, which is considerably different from conventional in situ measurements (Chen et al., 2004). Different methods, including single-band density cutting (Jain et al, 2005) unattended and monitored classification (Sivanpillai, 2010; Sheng et al., 2008) and spectral water indices (Ding, 2009; Feyisa et al., 2014; McFeeters, 1996 & Yuanzheng et al., 2016), were developed to extract water bodies from various remotely sensed images.

Among all existing water body mapping methods, the spectral water index-based method is considered a reliable method, because it is convenient, efficient and has low computational cost (Du et al, 2016). Different water indexes have already been proposed in the past few decades. McFeeters (1996) proposed the Normalized Difference Water Index (NDWI), using the green and Near Infrared (NIR) bands of remote sensing images based on the phenomenon that the water body has strong absorbability and low radiation in the range from visible to infrared wavelengths.

Indices minimize the problem of misleading information, caused by topographic shadows, cloud shadows, built-up areas, snow and ice. This misinformation comes from the difficulty in distinguishing water from other surfaces with a low albedo. Although the indexes have been improved over the years, there is still a need for more efforts in water body extraction (Li, Zhang & Xu, 2014).

In a recent study, a method that uses NDWI (McFeeters, 1996) and land surface temperature was developed, improving the results by more than 80% (Kaplan & Avdan, 2016). Highlighting the significance of temperature as a distinguishing characteristic of water bodies. Abdou et al., (2016) uses NDWI to visualise soil water content, pre and after flood.

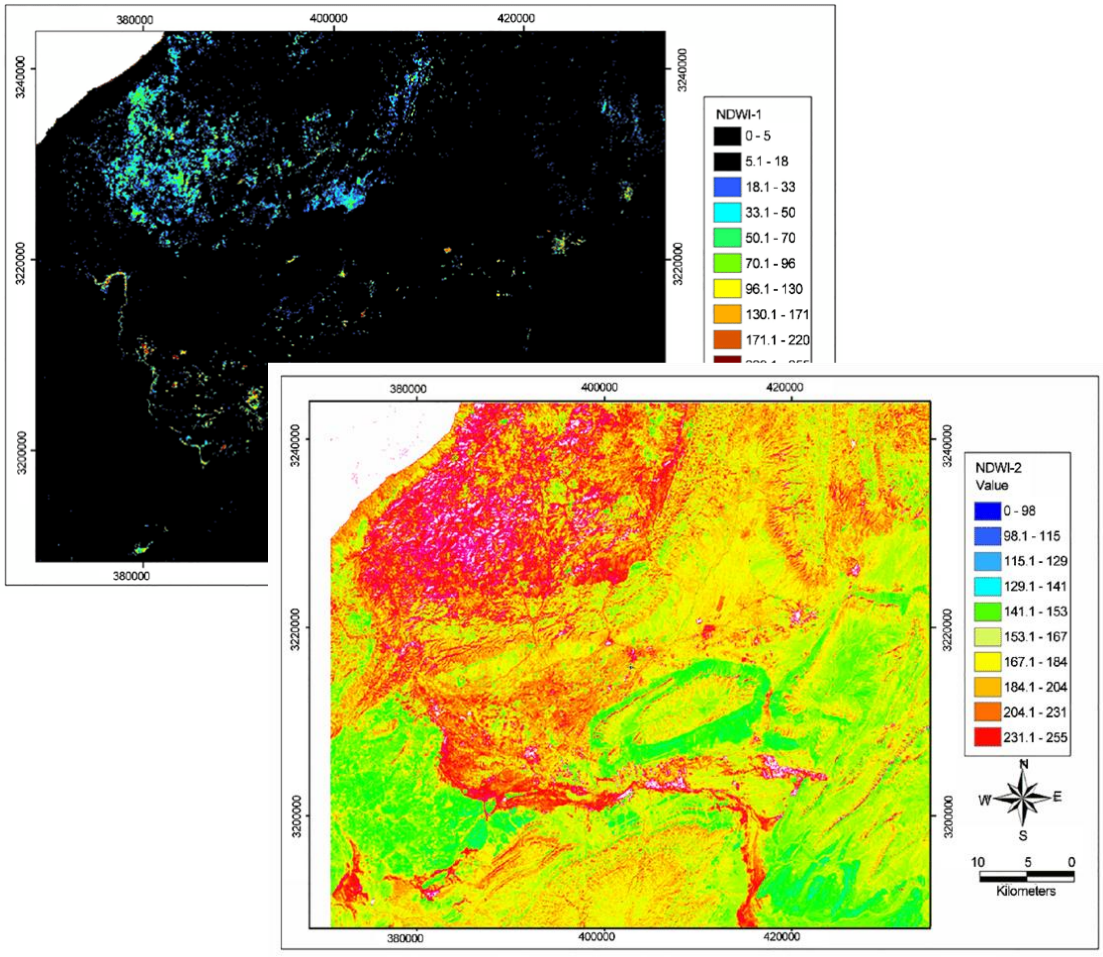


Figure 4 llustrate Soil Moisture and Ocean Salinity (SMOS) satellite data acquired before, during and after a flood-storm, t shows the soil moisture variability over the Guelmim region using NDWI (Abdou et al., 2016)

1.6 Machine learning

Machine learning is quickly becoming one of the most critical areas of general practice, research and development activity within computing science. This is amplified by the scale of academic research devoted to the subject and the active recruitment of machine learning specialists by major international companies such as Microsoft, Google and Amazon (Rogers & Girolami 2017).

Learning through knowledge and personal experience, which propagates from generation to generation, created the fundamentals of human intelligence. Also, at the centre of any scientific field lies the development of models (or theories) in order to explain the available experimental evidence at each period of time. In short, we always learn from data. Different data and different approaches to the data give rise to different scientific disciplines (Theodoridis & Sergios 2015).

Machine learning is an inter-disciplinary collation of widely distributed sub-branches of fundamental sciences: it incorporates countless paradigms of mathematical logic, multiple approaches to computational learning theory, artificial intelligence models and algorithm formalization methods; it has connections to statistics and mathematical optimization. It is often hard to determine which fundamental scientific discipline machine learning truly belong to.

Machine learning is employed in numerous industries in today's world: image analysis, computer network packet routing, system security aspects, digital search, spam filtering, autonomous car industry, big data analysis, optical character recognition, pattern matching, iris and human voice recognition to name a few. Remote sensed data analysis is no exception

to the lately growing adoptions of machine learning application in big data and image analysis.

1.6.1 Machine learning in remote sensing

Remote sensing first utilized machine learning methods in the 1990s. It was initially introduced to remote sensing as a way to automate knowledge-based building. The study authored by Huang and Jensen (1997) described how a knowledge – base was constructed with minimal input from human experts, and then decision trees were developed to infer the rules from the human input for the expert system. The generated rules were used at a study site on the Savannah River. Huang and Jensen (1997) concluded that the machine learning assisted approach, provided a higher accuracy when compared to conventional methods at the time. Subsequently similar developments in machine learning were made and was quickly adopted as an important tool by the remote sensing community. It is presently being used in a range of different projects, from an unsupervised satellite image scene classification (Li, et al. 2016) to the classification of Australian native forests (Shang & Chisholm, 2014).

1.6.2 Machine learning categories

Machine learning can be assigned to three categories, as seen below in figure 4.

- Supervised machine learning,
- unsupervised machine learning and,
- Reinforced learning.

The difference between supervised and unsupervised learning is that when using supervised models, the user has created a pre-defined label with a set of characteristics. Whereas the unsupervised algorithm, it interprets the data set by clustering the data into different classes

based upon the relation it has recognized between different records. Reinforcement learning is moderately different, the user provided the algorithm with an environment and the algorithm makes decisions within that environment. It is continually improving itself with each decision based on the result of the previous decision.

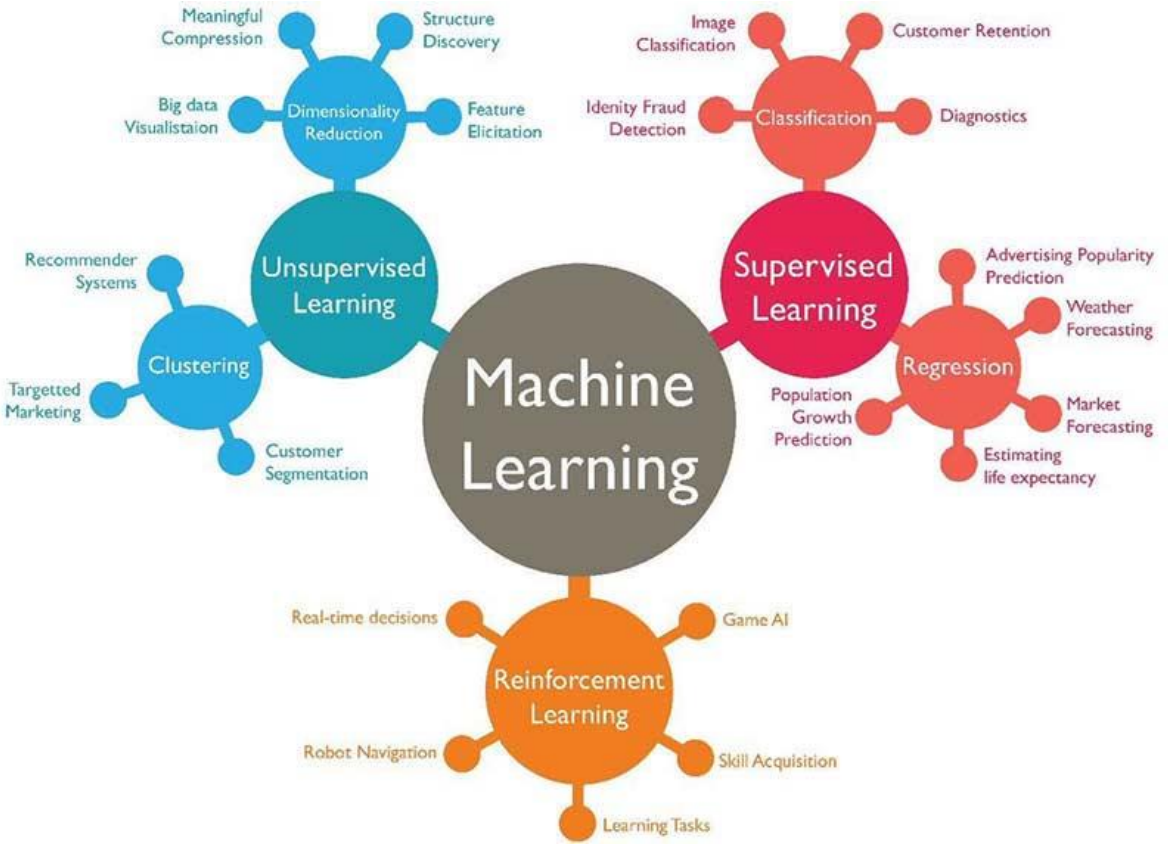


Figure 5 Machine learning and its three main categories Techleer (2017).

1.6.3 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification tasks. Commonly a task that appears in different machine learning applications is to construct a non-parametric model from the data .When designing the model, one strategy is to build a model from theory and adjust its parameters based on the observed data. Unfortunately, such models are not available in most real-life situations (Natekin & Knoll, 2013). The lack of a model can be circumvented if one applies non-parametric machine learning techniques like

neural networks or support vector machines to develop a model straight from the data. These are supervised learning algorithms (Natekin & Knoll, 2013).

One of the most common methods to data-driven modelling is to build only a single strong predictive model. An alternate approach would be to construct an ensemble of models for some specific learning task. Hypothetically, building a set of "strong" models like neural networks, which can be further combined to produce a better prediction (Figure 6).

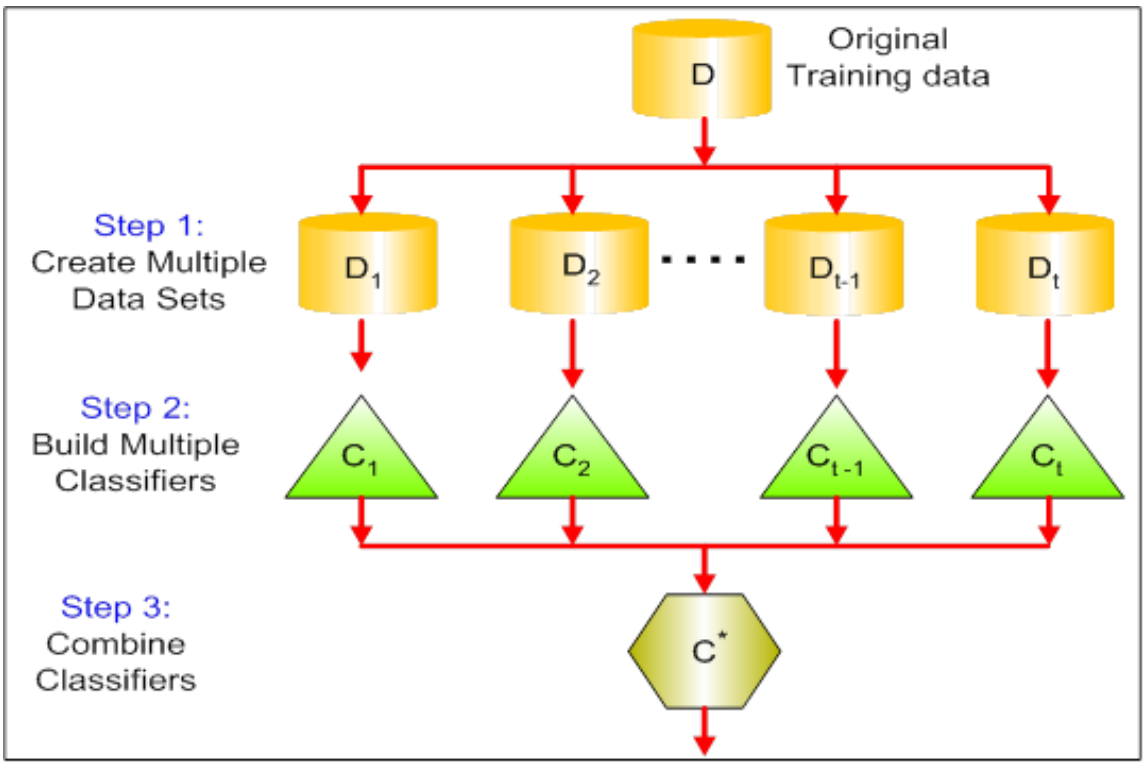


Figure 6 Ensemble learning basic concept (Srivastava, T., 2016)

In theory, the ensemble method depends on combining a sizable number of relatively weak models to obtain a stronger ensemble prediction. Some of the most prominent examples of machine learning ensemble techniques in remote sensing are random forests (Breiman, 2001) and neural networs (Hansen & Salamon, 1990).

The common ensemble techniques like random forests depend on the averaging of models within the ensemble. The family of boosting methods is based on a different, step by step strategy of ensemble structuring. The main concept of boosting is to add new models to the ensemble sequentially at each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far (Natekin & Knoll, 2013). The first popular boosting methods were solely algorithm-driven, which made the detailed analysis of their properties and performance rather difficult (Schapire, 2002). This led to a number of speculations as to why these algorithms either out performed every other method, or in reverse, were inapplicable due to severe overfitting (Sewell, 2011).

To create a connection with the statistical framework, a gradient-descent founded formulation of boosting systems was derived (Freund and Schapire, 1997; Friedman et al., 2000; Friedman, 2001). This design of boosting methods and the corresponding models were named gradient boosting machines. This structure also delivered the necessary justifications of the model hyper parameters and established the methodological foundation for subsequent gradient boosting model development.

In gradient boosting machines, or GBMs, the learning process sequentially places new models to deliver a more truthful estimate of the response variable. The primary concept behind this algorithm is to structure the newly created base-learners to be maximally correlated with the negative gradient of the loss function, connected with the entire ensemble. The loss functions applied can be random, but to give a better insight, if the error function is the classic squared-error loss, the learning process would end in successive error-fitting. In overall, the decision of the loss function is determined by the user, with together a high

diversity of loss functions resulting so far and with the possibility of applying the users own task-specific loss.

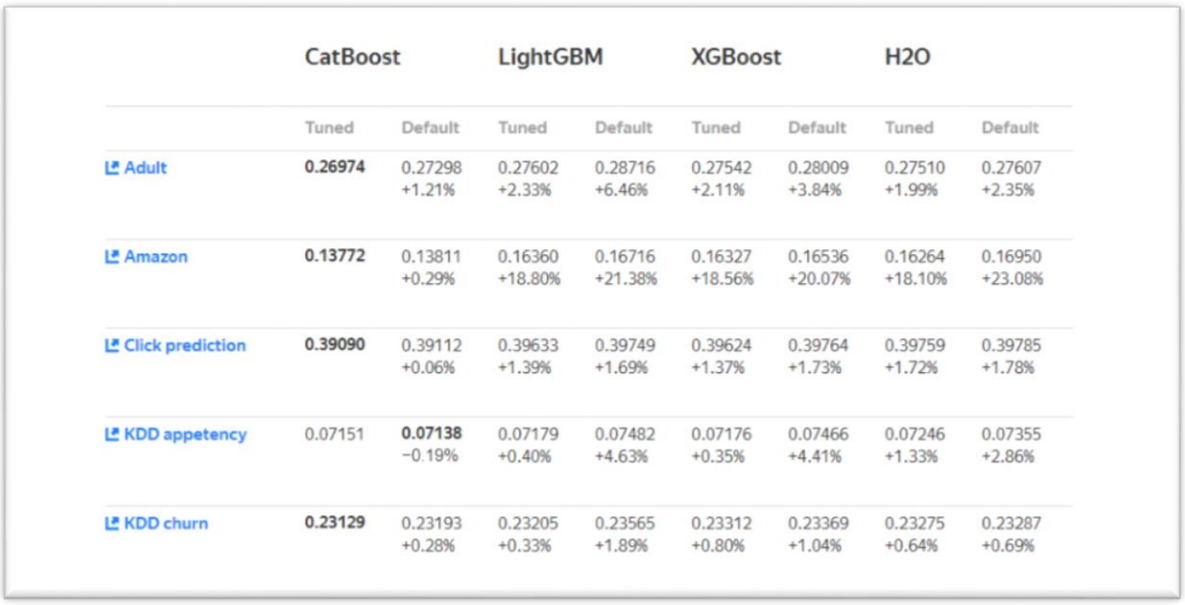
This high flexibility makes the GBMs highly customizable to many data-driven tasks. It introduces a level of autonomy into the model design thus making the selection of the more suitable loss function a matter of trial and error. However, boosting algorithms are relatively simple to implement (Natekin & Knoll, 2013), which permits the experimentation of different model designs. Furthermore the GBMs have presented extensive success in not only practical applications, but also in numerous machine-learning and data-mining trials (Bissacco et al., 2007; Hutchinson et al., 2011; Pittman and Brown, 2011; Johnson and Zhang, 2012).

1.6.4 Catboost

Catboost is a new open-sourced machine learning algorithm from Yandex. Catboost is intended for "open-source gradient boosting on decision trees," according to its GitHub repositories README. It delivers a means to perform classifications and rankings of data by using an assembly of decision-making models, or "learners," rather than a single one. Results produced by the learners are weighted and classified based on the strengths and weaknesses of each learner. By joining many learners, Catboost can produce greater results than decision-making systems that rely on single learners.

Catboost comes with support for Python and R, as well as a command-line interface to drive the machine learning library. Numerous machine learning libraries already implement some degree of gradient boosting algorithms. Python's Scikitlearn package has one version; XGBoost is available for several languages and data platforms; and Microsoft has the LightGBM library as part of its Distributed Machine Learning Toolkit project.

Catboost is intended to be astep up from other projects, according to Yandex, by being pre designed to operate in parallel with Yandex's own services. Yandex also states that it uses Catboost to generate predictions for its weather services, and that Catboost has been used at the European Organization for Nuclear Research (CERN) to cultivate values from the particle experiments conducted there (Catboost, n.d.). In figure 7 a table from catbbost's website compares performances on popular datasets, against other similar algorithms. Decimal values in this table represent Logloss values (lower is better) for Classification mode. The Percentages is the metric difference measured against tuned Catboost results.



	CatBoost		LightGBM		XGBoost		H2O	
	Tuned	Default	Tuned	Default	Tuned	Default	Tuned	Default
Adult	**0.26974**	0.27298 +1.21%	0.27602 +2.33%	0.28716 +6.46%	0.27542 +2.11%	0.28009 +3.84%	0.27510 +1.99%	0.27607 +2.35%
Amazon	**0.13772**	0.13811 +0.29%	0.16360 +18.80%	0.16716 +21.38%	0.16327 +18.56%	0.16536 +20.07%	0.16264 +18.10%	0.16950 +23.08%
Click prediction	**0.39090**	0.39112 +0.06%	0.39633 +1.39%	0.39749 +1.69%	0.39624 +1.37%	0.39764 +1.73%	0.39759 +1.72%	0.39785 +1.78%
KDD appetency	0.07151	**0.07138** −0.19%	0.07179 +0.40%	0.07482 +4.63%	0.07176 +0.35%	0.07466 +4.41%	0.07246 +1.33%	0.07355 +2.86%
KDD churn	**0.23129**	0.23193 +0.28%	0.23205 +0.33%	0.23565 +1.89%	0.23312 +0.80%	0.23369 +1.04%	0.23275 +0.64%	0.23287 +0.69%

Figure 7 Table comparing log loss values against alternative machine learning algorithms (CatBoost, n.d.)

The results in figure 7 suggest Catboost to be overall the more efficient algorithm when compared to its competitors.

1.7 Open Government Data

The majority of the data in this thesis is opened sourced. Open Government Data (OGD), Public participatory geographic information systems/science (PPGIS) and Volunteered geographic information (VGI) have emerged as important data contributors over the past decade (Hansen et al, 2013). These data sources provide the place component for built and natural environment data, however due to the methods by which this data is generated challenges arise regarding data reliability and therefore usefulness.

OGD is spatially referenced data made available for open use and can be freely used, reused and redistributed. Production is taxpayer funded and does not follow traditional pricing models where revenue is generated by selling data; therefore, benefits are realized through improved efficiency and cost savings to society (Hansen et al, 2013). It is the most reliable data source in most developed nations due to open data initiatives such as the INSPIRE directive and the EU Directive for the Re-use of Public Sector Information according to Hansen et al, (2013).

Open data consumers manipulate and utilise data in multiple ways, ranging from data integration to classification, also depending on the auxiliary assets they may obtain (Ferro & Osella, 2013). Considering this, legal and technical openness of datasets is not sufficient, on its own, to create an efficient reuse ecosystem (Helbig N. et al., 2012): failures in supplying sufficient quality information might impair not only the reuse of the data, but also the usage of the institutional portals (Detlor et al., 2014). Attempts to maximize quality and reusability of public sector data implies representing and exposing data so that they can be easily accessed, queried, processed and linked with other data with no restrictions (Sharon D.J., 2010).

1.8 Data Quality

The American Standard SDTS (Spatial Data Transfer Standard, 1997) has been the earliest to suggest a series of guidelines that details and documents GIS data quality, stating the basic scheme of the data quality report into five parameters: genealogy, positional accuracy, thematic accuracy, logical coherence and completeness (Bianchin, 2001). Data quality is subject to the scale, the accuracy, and the scope of the data set, as well as the quality of the other data sets that have to be used. The Open Geospatial Consortium's definition of data accuracy is as follows "Indications of the degree to which data satisfies stated or implied needs. This includes information about lineage, completeness, logical consistency and accuracy of the data" (OGC, n.d.). SDTS defines these five data quality elements, which are described in the rest of this section.

The data lineage refers to source materials, methods of origin and transformations applied to a database. It Includes temporal information (date that the information refers to on the ground) and is intended to be precise enough to identify the sources of individual objects (i.e. if a database was derived from different source, lineage information is to be assigned as an additional attribute of objects or as a spatial overlay) (Veregin,1999).

Positional Accuracy is the accuracy of a spatial component. Split into horizontal and vertical accuracy elements. Valuation methods are based on assessment with the source, comparison to a data set of greater accuracy, deductive approximations or internal data. Differences in accuracy can be described as quality layers or supplementary attributes (Veregin,1999). Attribute accuracy deals with the accuracy of a thematic component. Specific tests differ as a function of measurement scale. Attribute accuracy assessment methods are built on inferential estimates, sampling or map overlay.

The logical consistency of a dataset refers to the reliability of the relationships encoded in the database. This includes valid value tests for attributes, and detection of topological discrepancies based on graphical or precise topological assessments.

The completeness is the relationship concerning database objects and the abstract universe of all such objects. In the cluster Completeness the Data completeness is handling the completeness of an image, operating for example the effect of a shadowing object, sun flares on water surfaces or masking out by an object (Batini et al, 2017). This Study will only assess the positional accuracy of the dataset.

1.9 Problem statement

Quality assessment of current GeoDanmark lake data set, quality will be solely represented as by the positional accuracy of the data set. The method will use multi spectral satellite data and machine learning classification tools. This data is open government data and delineates the location and boundaries of the lakes in the country of Denmark.

Lakes have several benefits especially environmental, in an era of information, reliable geographical data is essential in the efficient management of our land masses. Geodanmark has created this lake dataset most likely by using several forms of validation data one of which could be the digitization of old data sources such as maps. This is not the most accurate approach as geographical objects such as lakes are constantly changing. An automated machine learning application of probabilities based on optical pixel values from sentinel imagery has the potential to provide a detailed quality assessment of GeoDanmarks lakes data set from the year of 2017.

This analysis could identify a number of errors in Geodanmarks lakes data set. This machine learning approach could be applied to any other geographic objects or features, for this study lakes are a more than adequate data type for this thesis, due to their unique spectral signature and high number within Denmark.

1.10 Research Questions

The project will focus on the implementation of a machine learning algorithm in assessing the quality of a current geographic dataset and the significance of the results. The following research questions are intended to adequately confront the problem stated in the previous section.

1. How accurate is Geodanmark's open government lake data set?
2. How effective is Catboost in classifying lake features in a geographic data set?
3. When implementing a machine learning approach to optical and geographical data, which features are the most influential?

1.11 Report Structure

This report has been structured to address a problem. Research questions have been identified based on the problem.

The report has been structured in the following way:

- Chapter 1: Introduction, theory, problem statement and research questions
- Chapter 2: Data and materials
- Chapter 3: Methods
- Chapter 4: Results and findings
- Chapter 5: Discussion and conclusion.

2 Data

In this Chapter the data being analysed, utilized and prepared will be discussed in Detail. This thesis aims to review open government geographical data provided by the board of geographical data In Denmark: GeoDenmark. Optical satellite 10m imagery values from the sentinel 2A and 2B missions, will provide the characteristics for each lake across the sample area for use within the machine learning model. This Chapter will review and describe this data; including the data's source, it's structure, and purpose within this quality assessment.

2.1 GeoDenmark & SDFE

Geodanmark is the framework for cooperation between all municipalities in the country of Denamark and the Board of dataforsyning and Efficiency Strategy (SDFE) on the establishment and maintenance of a country-wide open government geographical data. Geodanamarks open source geographical data consists of 59 object types classified in the following categories: buildings, construction, traffic, engineering, nature, hydro, administrative, topography and sundries. Data is freely available to download on the The Ministry of Environment (KMS): the Map Supply (kortforsyningen) website. The data object type that will be used in this thesis is a subset of hydro; lakes.

2.2 Sample area

The sample area will cover an area equal to that of one sentinel tile covering the mid-eastern region of Jutland Denmark as seen represented by the red box in figure 8. The area covers a total of 12,056.04km2 containing 30,777 lakes, sizes ranging from 20.30 m2 to 16,541,268.89 m2. The areas Topography is fairly flat and low lying like the rest of Denmark, minimising the chance of any topographic shadows or areas of snow or Ice.

Figure 8 Sample area represented in red

The average lake size is 5244.67m2 far below the median, equalling a very uneven size distribution (Figure 9). 30% of the lakes are under 500m2 and a maximum size of 16,541,268.89 m2. This uneven distribution should be taken into account especially when working with limited resolution imagery such as sentinel. There are 794 lakes above 10,000 m2, figure 9 below illustrates the distribution frequency of the lakes different sizes, with the 794 or 2.5% of the total lakes removed, allowing for a more descriptive interpretation.

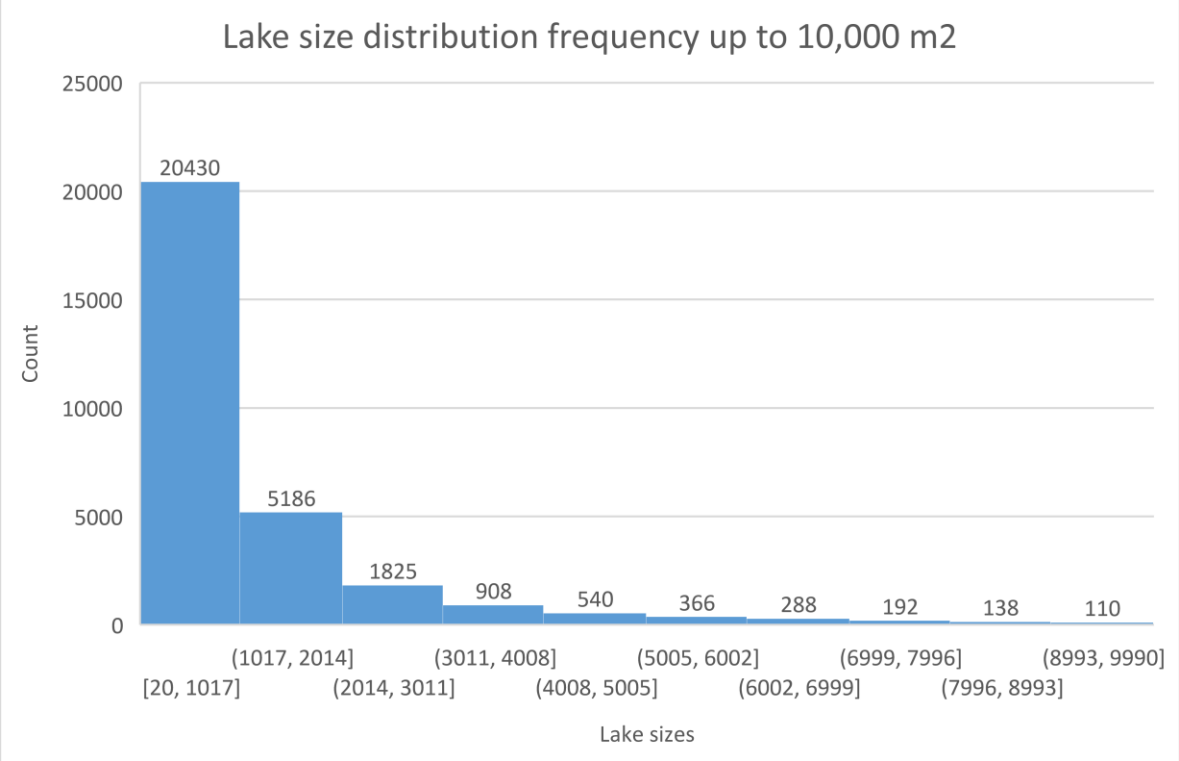


Figure 9 Lake Size distribution from 20 to 10,000 m2

Even with the extremely large lakes removed it is still hard to see any lakes above 7000 m2 compared to smaller sized lakes in figure 9. The majority of lakes being in the 20.00 – 1017.34 m2 size range. Lakes below 500m2 may be too small for sentinel to provide an accurate representation. The medium size lakes between 500m2 and 10,000m2 are still the majority at 55% (Figure 10). The smaller lakes under 500m2 size distribution (Figure 11).

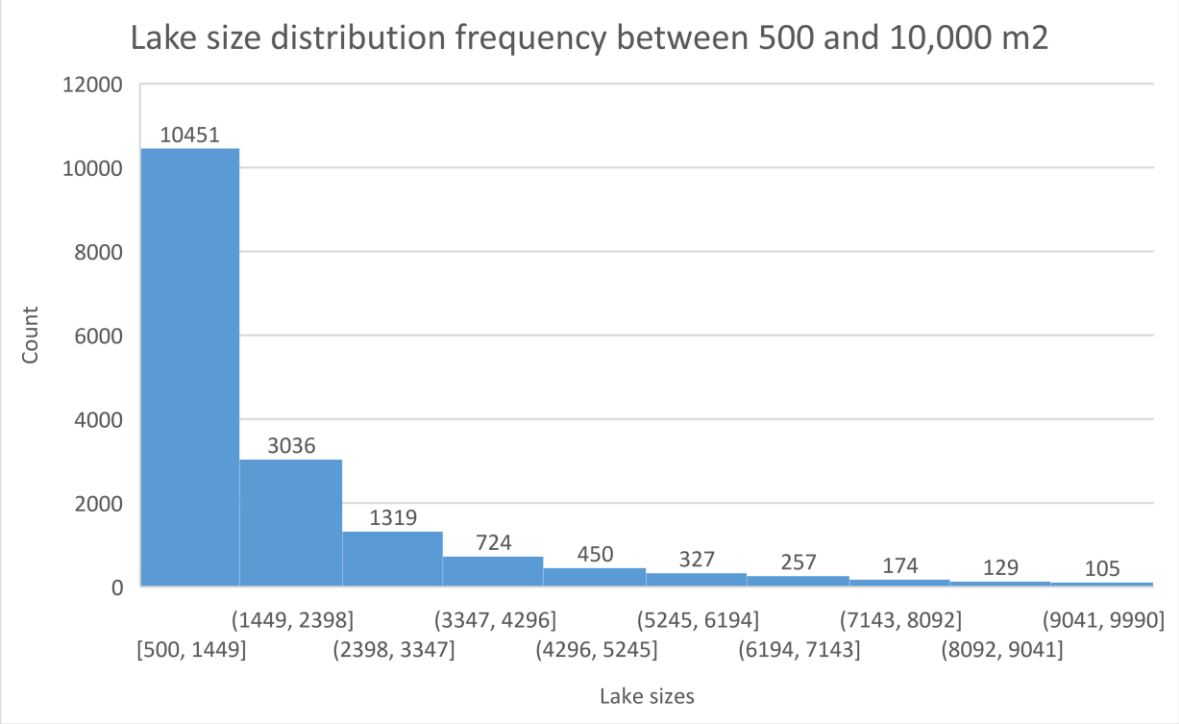


Figure 10 Lake size distribution between 500 and 10,000m2 (Medium size lakes)

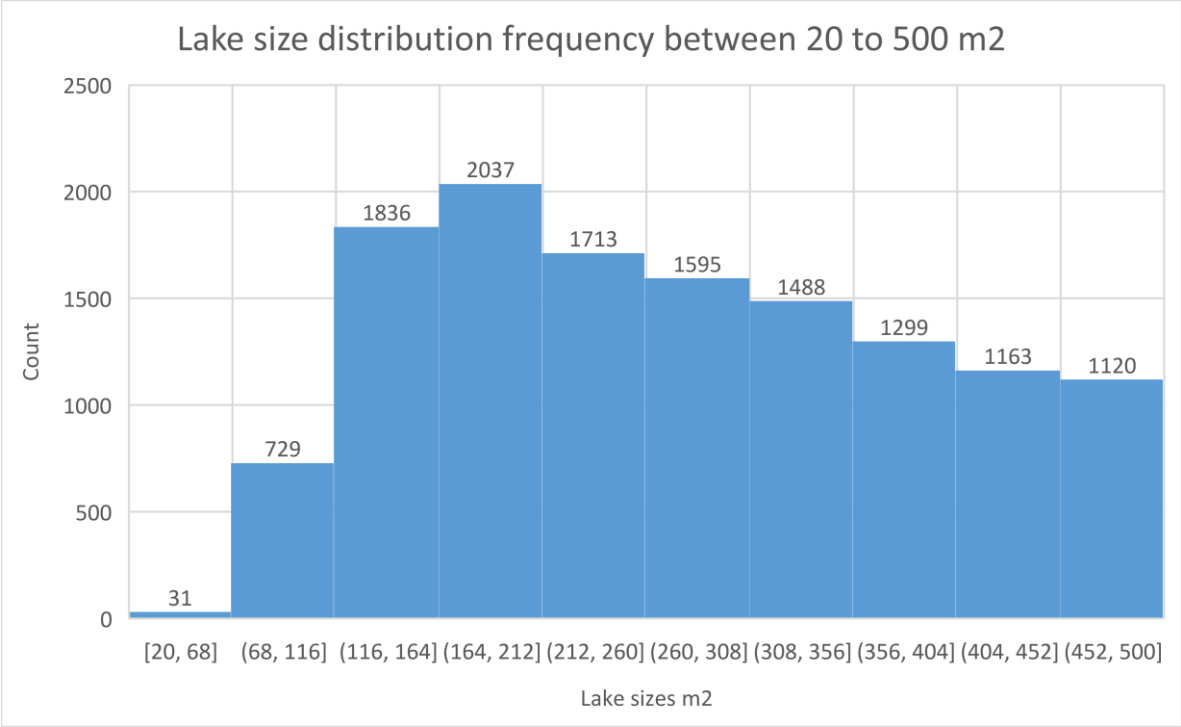


Figure 11 Lake size distribution between 20 and 500m2 (smaller lakes)

2.3 Sentinel

Sentinel 2 imagery has been chosen due to the images being freely available and easy to download from the European Space Agency website. The Sentinel 2 tiles covering Denmark can be seen in figure 9. Tile 32VNH was used as the boundary for our sample area seen in figure 12.

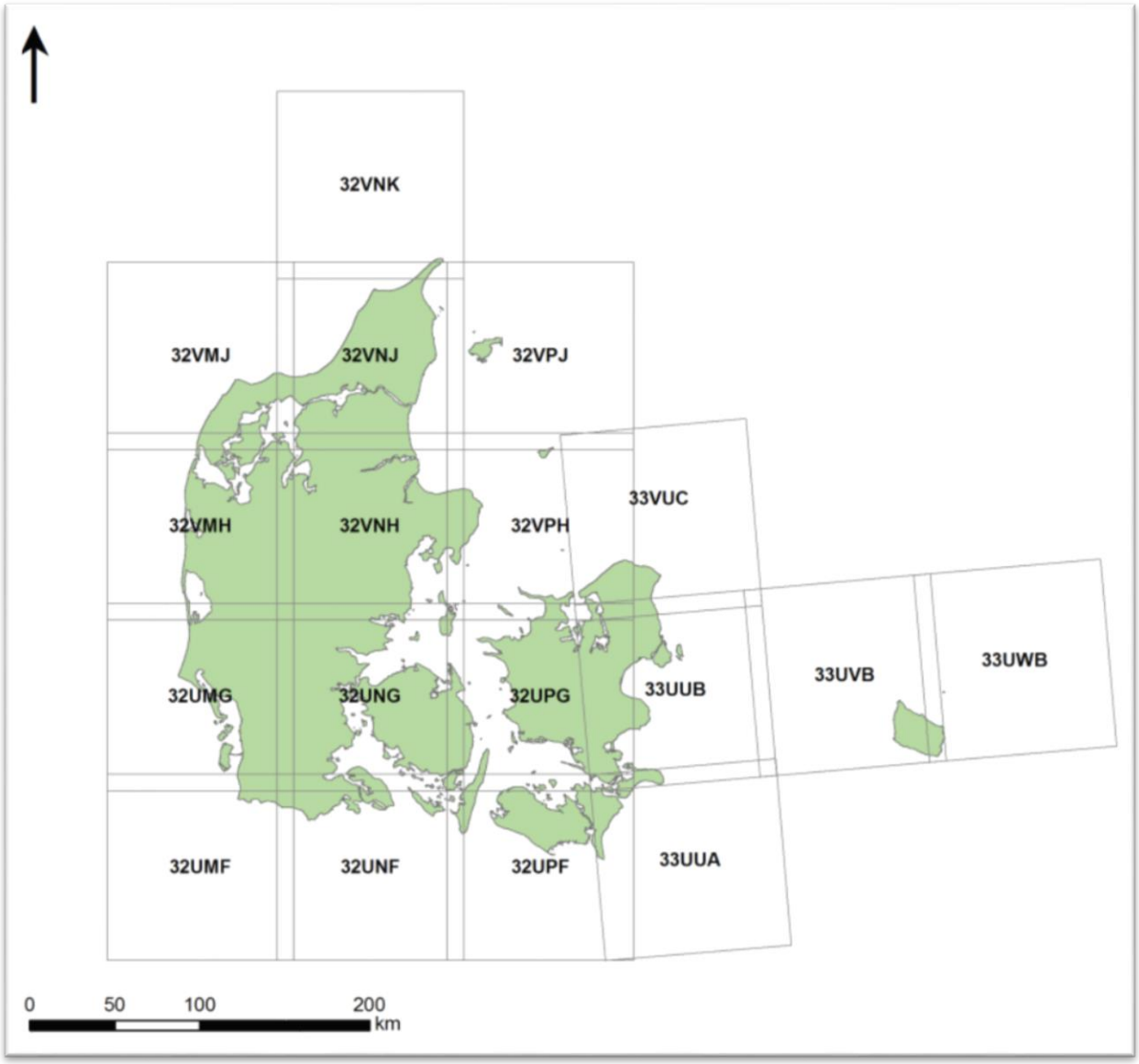


Figure 12 Sentinel 2 tile coverage of Denmark

The images provided have a high temporal resolution, and provide us with a range of dates throughout the year when selecting the image (Table 1). Sentinel RGB image T32VNH can be seen below in figure 13.

Figure 13 Sentinel RGB image tile T32VNH

A total of 40 sentinel bands where used, from 10 different dates from late 2016 to early 2018 (Table 2).

Table 2 Sentinel dates used in the project, with their tile number and sensor type.

Date	Tile	Sensor
2016/12/14	32VNH	S2A
2017/01/13	32VNH	S2A
2017/03/24	32VNH	S2A
2017/04/23	32VNH	S2A
2017/05/13	32VNH	S2A
2017/08/23	32VNH	S2A
2017/10/30	32VNH	S2A
2017/12/29	32VNH	S2A
2018/01/08	32VNH	S2A

Both raw L1C reflectance values (TOA, top-of-atmosphere) and L2A reflectance values (bottom of atmosphere) were evaluated, but it was noticed that the NDWI values were inconsistent and inaccurate when using the L2A scenes. As a result TOA reflectance values were only used.

Table 3 Sentinel 2A spectral bands used in the analysis.

Sentinel spectral number	Spatial resolution	Wavelength category
2	10m	Blue
3	10m	Green
4	10m	Red
8	10m	Near - Infrared

The Sentinel 2A bands used in this study all have a spectral resolution of 10m x 10m. Resulting in a pixel size of 100m. Below in figure 14 four different lake in shape and size can been seen with sentinel RGB. It can be quite difficult to distinguish a lake for the naked eye, especially the smaller lakes like in the bottom two images in figure 14.

Figure 14 Four different types of lakes with the Lakes.shp layer and sentinel RGB.

2.4 GoeDanmark Ortophoto

GeoDanmark 12.5 cm orthophoto taken in the summer of 2017 will be used to verify the training and testing set. The same lakes seen in figure 15 can be seen in figure 15 with 12.5 cm ortophoto's instead of the RGB sentinel image. A clear distinction can be seen between the two, the ortophoto's high resolution allows for a much clearer and visible image of the lakes.

Figure 15 Four examples of various lake types in the 12.5cm orthophoto's

Ortophoto's are some the highest resolution imagery available and is one of the best sources of validation data other that in situ data.

2.5 GeoDenmarks lake vector data

Quality of digitized lake data created by GeoDanmark is highly questionable this can be seen when comparing GeoDanmarks lake data with remotely sensed imagery taken in the past 2 years (figures 16 and 17 below). This can be seen when overlaying the Lakes.shp from GeoDanmark with 12.5 cm Ortophoto's taken in the spring of 2017. In figure 16 the lake.shp outlined in red, represents a lake boundary where there seems to be no indication of water in the Ortophoto.

Figure 16 Geodenmarks vector data indicates a lake (in red) where there clearly isn't one. Photo taken from 2017 summer 12.5cm ortophoto.

In second comparison it can be seen that the lake.shp files seem to be missing data as it does not illustrate boundaries around clearly visible lakes as seen in figure 17, although other neighbouring lakes are recognised.

Figure 17 GeoDanmarks's lave vector datset (in red), matching with 3 waterbodies in the ortohpoto, when there are actually 5.

These errors are widespread throughout GeoDanmark's data, all over Denmark. The GeoDenmark lake.shp file obtained from the website kortforsyningen.com can be seen in figure 18 below (In red), showing an abundant and evenly distributed amount of lakes.

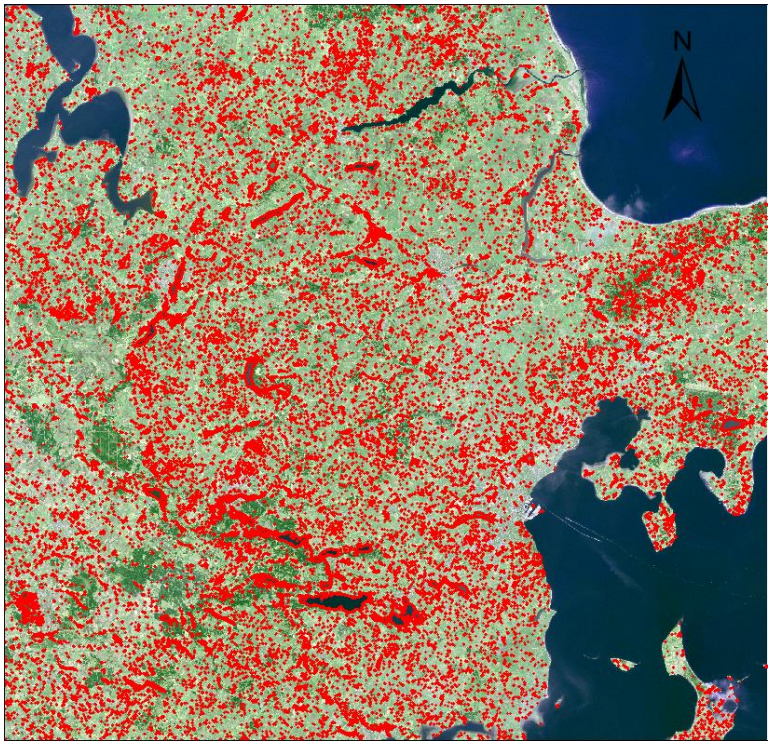


Figure 18 Lake Vector file overlapping with the sentinel 2 RGB

3 Methodology

This thesis aims to produce a probability data set for the lakes of Denmark throughout the given sample area. The original datasets which will be used are detailed in the previous chapter and the pre-processing and implementation of the methodology is described in the following chapter.

The chapter will be structured firstly detailing the pre-processing, application input, running the application and the output.

3.1 Data preprocessing

With the Vector and optical data gathered, further processing is required. A large portion of the assessment was spent extracting and formatting the data for use in the gradient boosting algorithm cat boost.

3.1.1 Lake characteristics

In addition to the RGB and NIR bands, a NDWI band for each date was created in python using the pre-mentioned formula (McFeeters, 1996):

$$NDWI = \frac{(Xgreen - Xnir)}{(Xgreen + Xnir\ )}$$

In addition, the QGIS field calculator provides the area m2 for each lake in the data set. Both The NDWI and area (m2) are added as features. Due to the pre mentioned limitations of sentinel, all lakes less than 500m2 were removed from the data set, leaving 21,520 lakes.

3.1.2 Data extraction in python

The first step was obtaining summary statistics for each band across all 9 dates, due to the large amount of data being extracted the programming language python was used, which has several packages that read and write raster data.

The python package rasterio was used to read write the raster data and numpy was used to format and create the data. Below in figure 19 is a section of the script which implements rasterio in extracting statistical values from the 45 different bands including the NDWI bands.

```
In [ ]: def get_band_stats(raster_file):
            stats = []
            with rasterio.open(raster_file) as src:
                for i, lake in enumerate(gdf_json['features']):
                    data_at_lake, transform = rasterio.mask.mask(src, [lake['geometry']], crop=True)
                    data_at_lake = data_at_lake[data_at_lake > 0]
                    if not data_at_lake.size:
                        continue
                    stats.append({
                        'id': lake['properties']['FOT_ID'],
                        'mean': np.mean(data_at_lake),
                        'min': np.min(data_at_lake),
                        'max': np.max(data_at_lake),
                        'std': np.std(data_at_lake),
                        'median': np.median(data_at_lake)
                    })
            return stats

In [ ]: bandfiles = ('Sentinel/T32VNH_20161214T104432_B02', 'Sentinel/T32VNH_20161214T104432_B03', 'Sentinel/T32VNH_20161214T104432_B04', 'Sentinel/T32VNH_20161214T104432_B08', 'Sentinel/T32\

In [ ]: outdir = r'/home/jovyan/exchange/lakes_tsv'
        skip_existing = True
```

Figure 19 Sentinel band statistics being read by the python package rasterio.

The values where then exported as csv files allowing them to be used in excel and QGIS/ARC. Any no data values are to be saved as a value -42, a value which does not have much weight in the decision tree due to its irregularity. Below in figure 20 is a screenshot of the csv file, in the first column is the lakes Id followed by their statistical values.

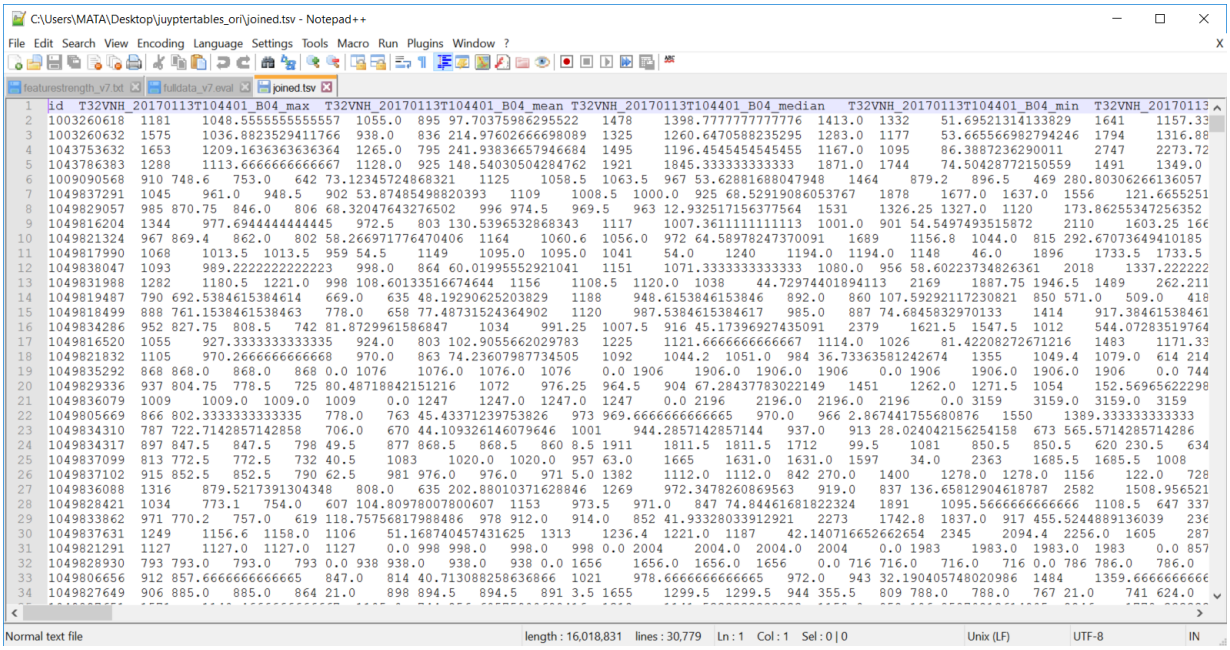


Figure 20 Rasterio statistical extraction script output as a csv file.

The summary statistics extracted are the mean, min, max, standard deviation and median value of each group of pixels within each of the lakes boundaries throughout the RGB, NIR and NDWI bands. The outputted CSV is then joined with GeoDanmarks lake.shp file in order to retrieve each of the lakes area (m2) alongside their summary statistics creating one Data set, comprising of 110 features for the algorithm to make its prediction.

3.1.3 Validation Set

The machine learning algorithm requires three data sets, training set, testing set and full set (total lakes). The training and testing sets are pre validated subsets of the full data set we created in the previous section. In order to validate these sets, 12.5cm orthophotos were used to confirm the presence and absence of 500 lakes within the sample area. This was done manually in QGIS, creating a new column within the lake data set and assigning the values '0' or '1' to each absent or present lake (figure 21 below).

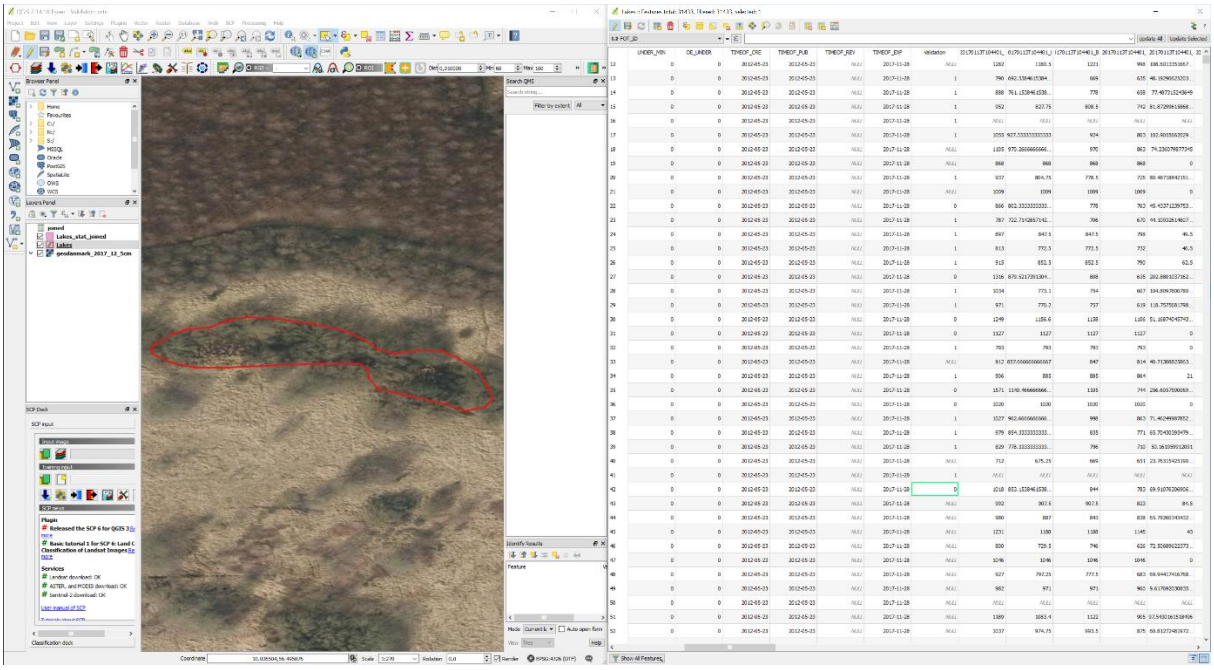


Figure 21 Validation process, Ortophoto and lake.shp on the left and the full data attribute table on the right.

The data which is used to train and test the model is completely subject to the user's interpretation of a lake. As a result a basic criteria was applied in determining whether a lake was a lake or not. In the context of this project lakes where determined to be a body of water within GeoDanmarks given boundary, the body of water may contain vegetation and or organisms which are often associated with lakes, such as algae, reeds and trees within reason.

Variation set distribution is also taken into account especially in terms of size as the full data set has a wide range of areas from 20m2 to 16km2.

A size filter is applied due to Sentinels mid-level resolution, a single pixel equalling to 100m2 would make all lake values under 100m2 distorted or unintelligible, so to avoid this, all lakes below 500m2 where removed from the dataset, reducing the amount of lakes by 30 % (9248). As a result the model is only effective to lakes over 500m2.

3.2 Catboost Input

With the Data set created including the 500 validated lakes, it requires some formatting and structure editing as is required by Catboost 9.1.1.0. Together with the three data set files. A column descriptor file (figure 23) is required to assign the target variable and numerical features. The target variable or label are the binary ‘0’ ‘1’ values and the rest of columns containing the pixel statistics and area (m2) will be the numerical features.

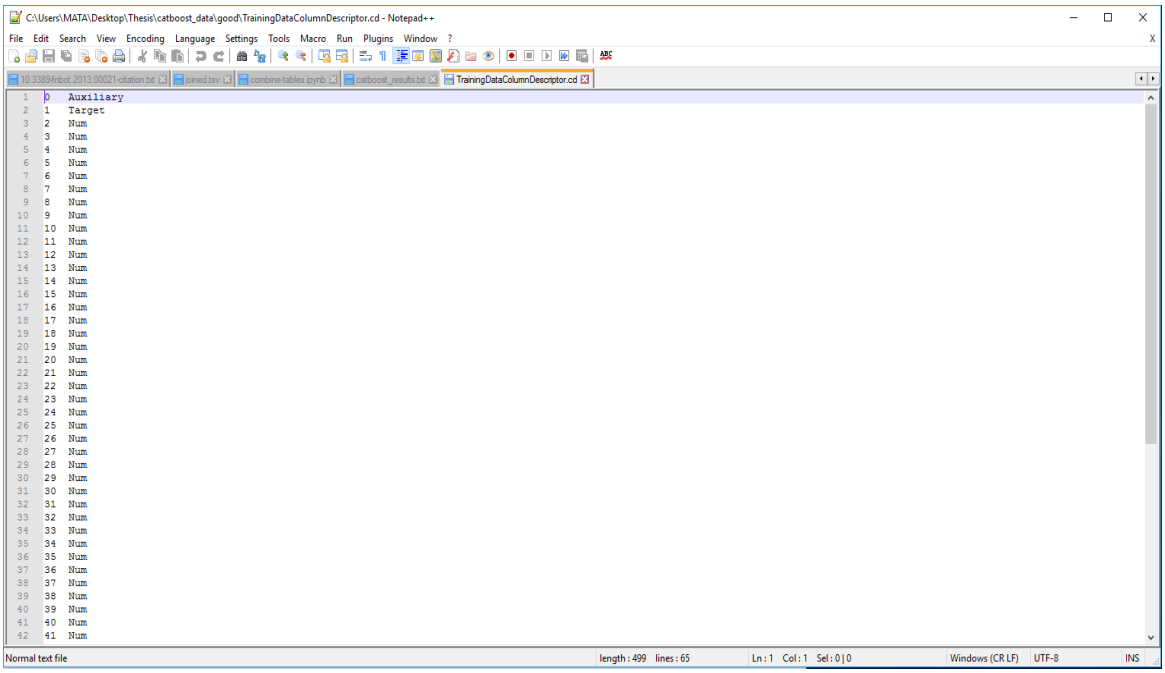


Figure 22 Screen shot of the Column descriptor file

3.3 Run Catboost

The command line version of Catboost was used. The first step was to train the model with the training and testing sets, with the console open in the work folder seen in figure 1, the following command line was used to train the model (figure 24):

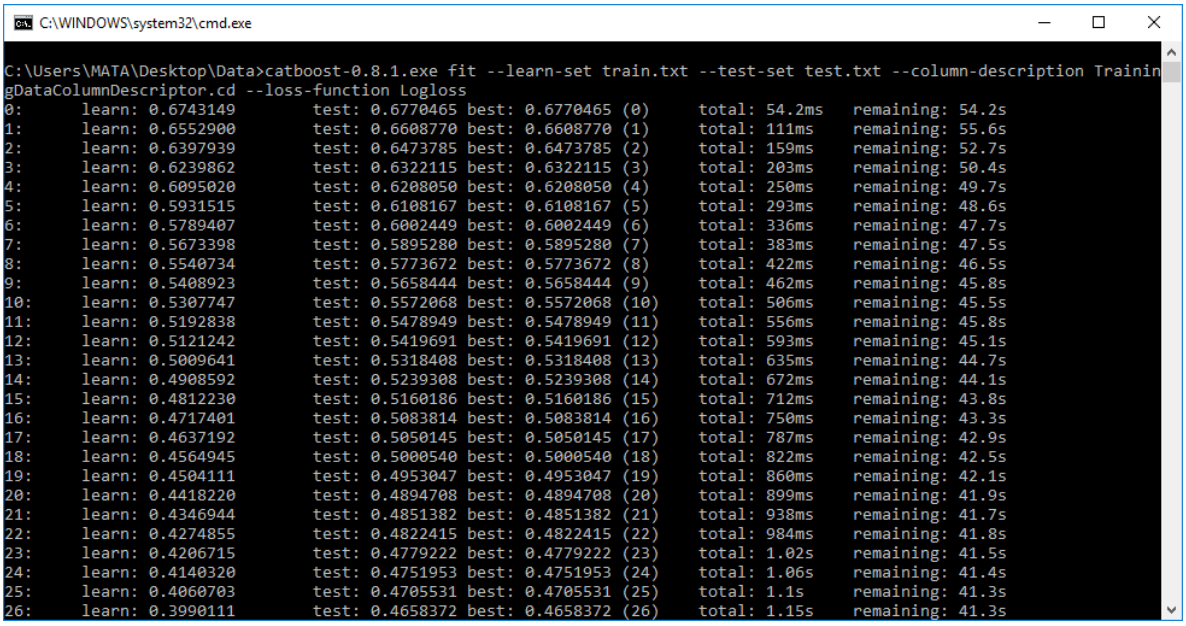


Figure 23 Screenshot of Catboost fitting the model in cmd.exe

Once the model had been created it is exported as a .bin file within the work folder. The model was then applied to the full data set in order to predict probability values for each lake in the sample area as seen in the screenshot (figure 25).

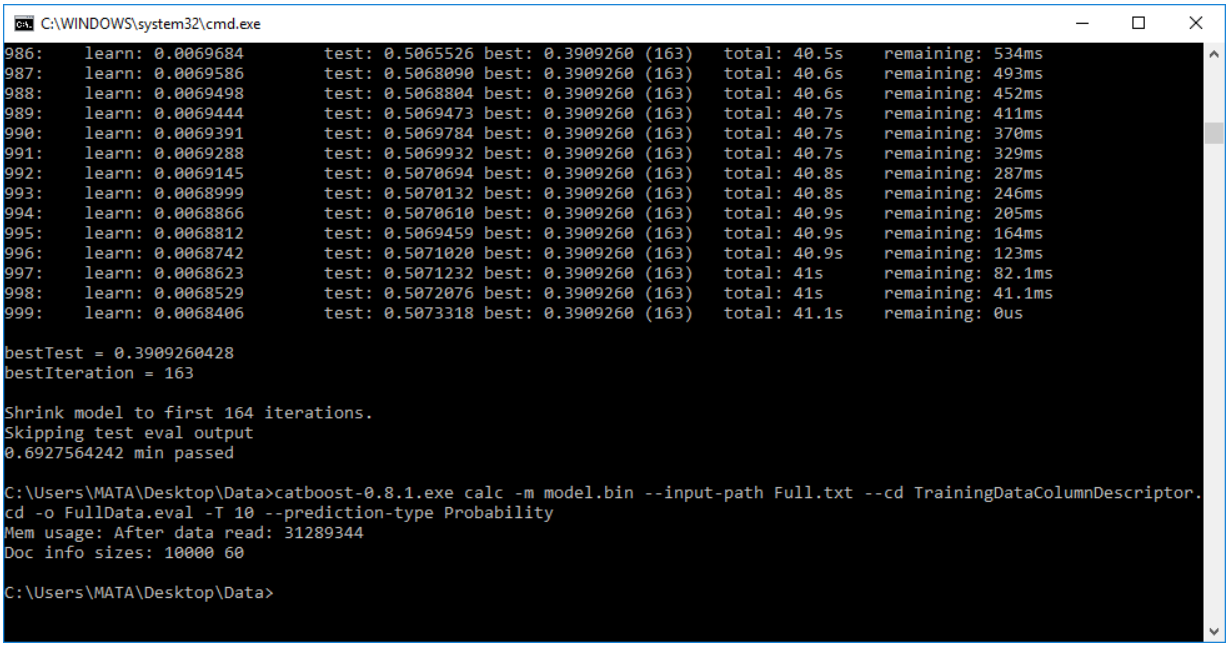


Figure 24 Applying the model to the total data set and creating a feature strenght file.

3.4 Catboost output

The resulting file ‘FullData.eval’ contains a probability value between 0 and 1 for each row, corresponding to the input file order as seen in figure 26 below.

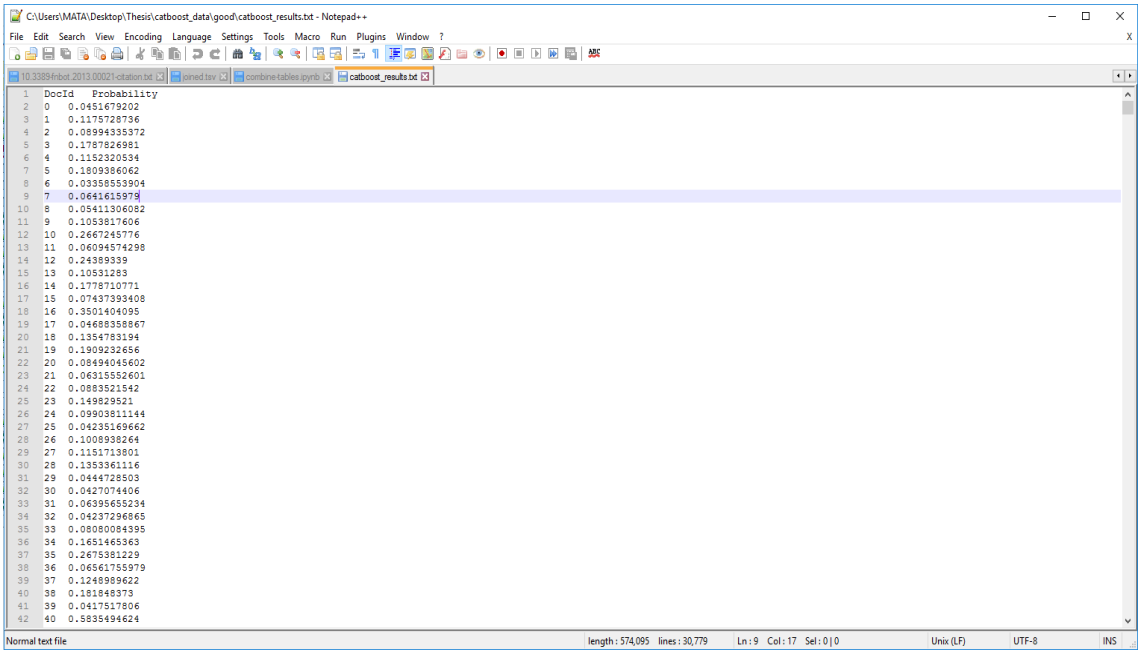


Figure 25 screenshot of output file containing probability values

The file is then joined to the full dataset, aligning the probability values with their respective ids. The Dataset is now ready for interpretation and analysis. Below is the visualization of the probabilities in QGIS; from a small sub sample containing 5 lakes. Probabilities have been split into 5 different classes and color graded (figure 27). Figure 28 is the same area but without the probability layer, illustrating the accuracy of the probability area.

Figure 26 Probability layer overlaying 12.5cm ortophoto, with legend.

Figure 27 Ortphoto without probability layer

4 Results and Findings

In this section the results of this project will be presented and analysed. Including any significant patterns within the data and the accuracy of the results.

4.1 Results

Table 4 splits the probability results into three categories; an area of low probability below 30 % indicating a very unlikely probability that the given lake is actually a lake, above 70% indicating a high likelihood of the given lake being a lake and the area in between these two values indicating an uncertainty.

Table 4 Probability category ranges.

Probability Factor	Count	Percentage of total %
< 30 %	7076	33
> 70%	10729	50
30 - 70%	3655	17

Fifty percent of the lakes are within the bracket of high probability, indicating that the other half of the dataset is incorrect or highly questionable. Thirty three percent of the lake data set has a probability value of thirty percent or under. Demonstrating a significant fault in the dataset. On visual analysis the accuracy of the model seem evident (figure 24).

Lakes with an uncertain probability value like in figure 29 with a probability value of 0.5 has visually different characteristics of that in figure 30. There appears to be a body of water, but it does not meet the boundary that GeoDanmark has provided and any water that is present is difficult to see, as it is predominantly consists of colours which are not associated with water. This could be due to surface vegetation such as algae hiding the underlying lake.

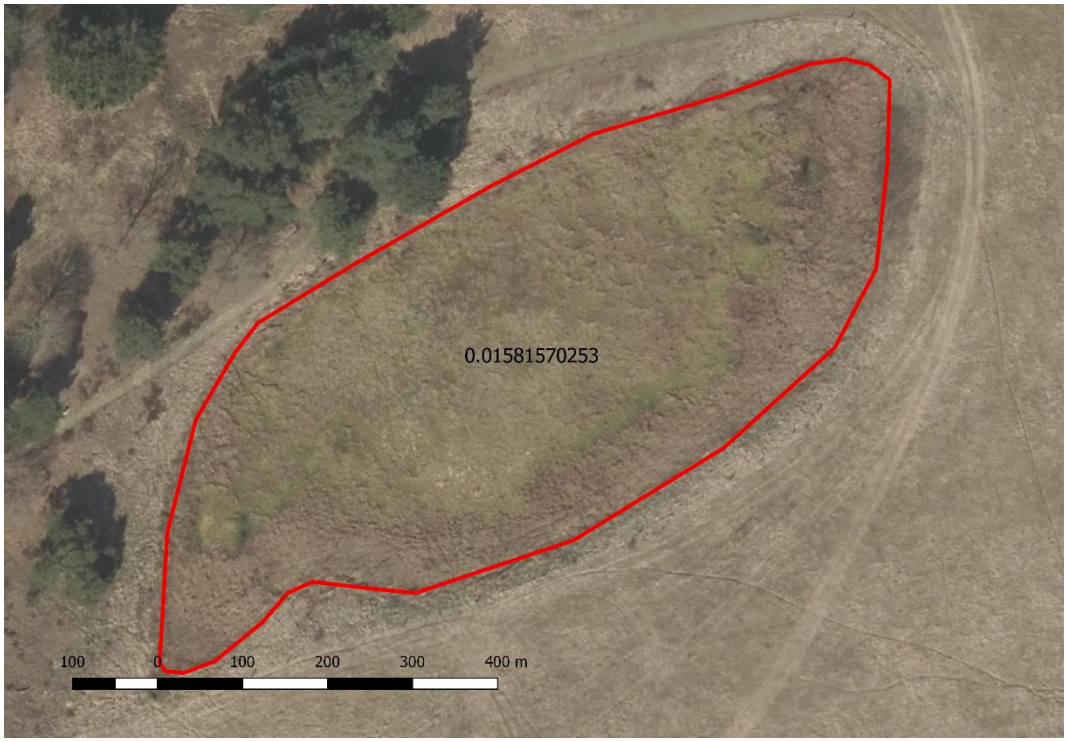


Figure 28 Lake Feature with a size of 4000m2 and a probability value of 0.016 or 1.6

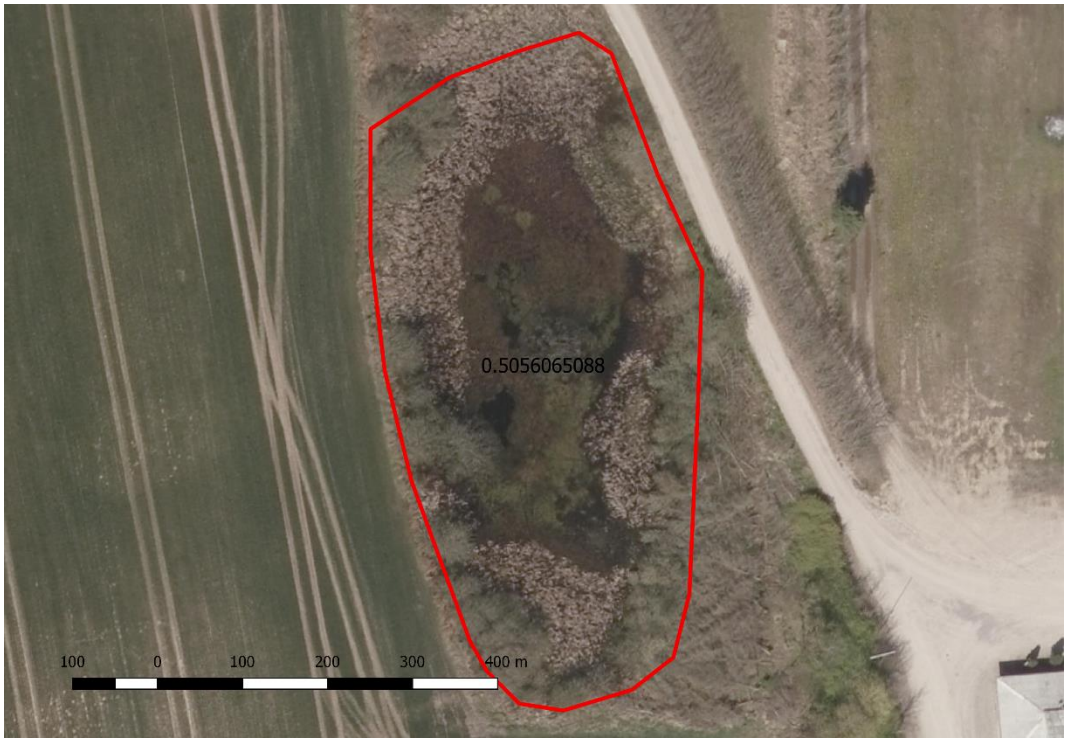


Figure 29 Lake Feature with a size of 3000m2 with a probability value of 0.51 or 51 %.

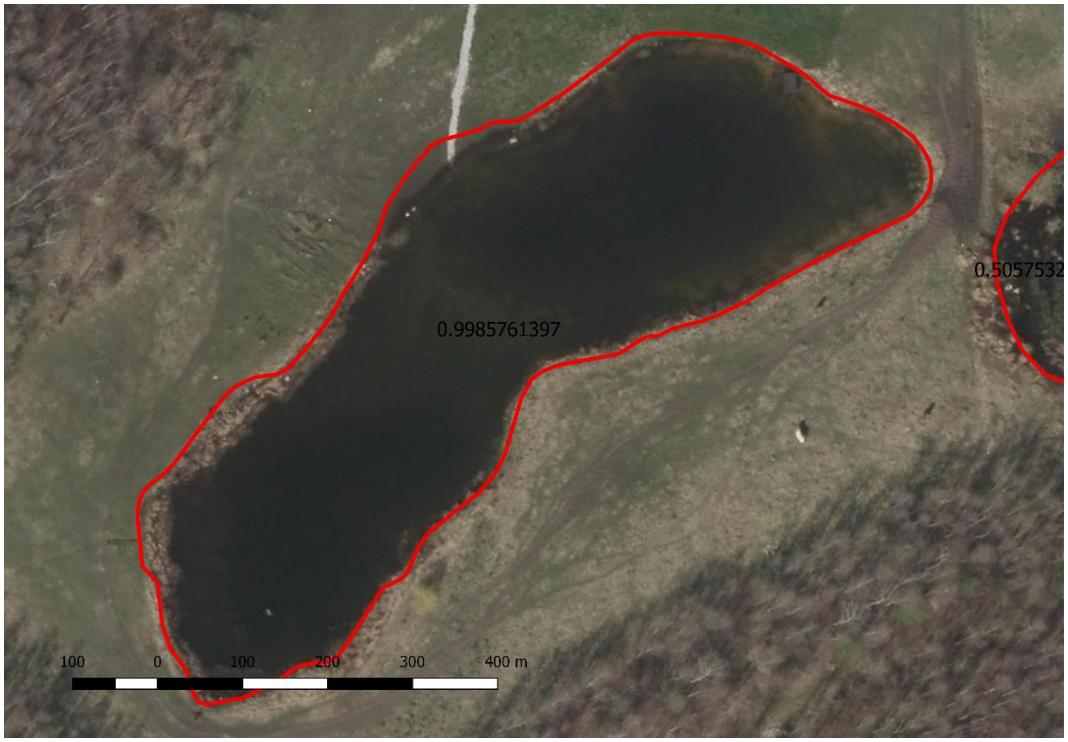


Figure 30 Lake Feature with a size of 3300m2 with a probability value of 0.99 or 99%.

Lakes which fell in the 70 – 100% probability range mostly consisted of bodies of water that fitted the given boundary and can be clearly identified as a lake in the ortphoto's, like seen in figure 31.

The uncertain value range is unavoidable as even with a strong training and data set the model will still struggle to clearly identify a lake due to the diversity of lake and not lake characteristics.

On visual analysis of the uncertain lakes, the majority are not in parallel with Geodanmarks lake boundary shapefile still suggesting an error within the dataset. In figure 32, the count of each lake for each of their probabilities can be seen. The histogram is largely 'U' shaped indicating a relatively decisive model.

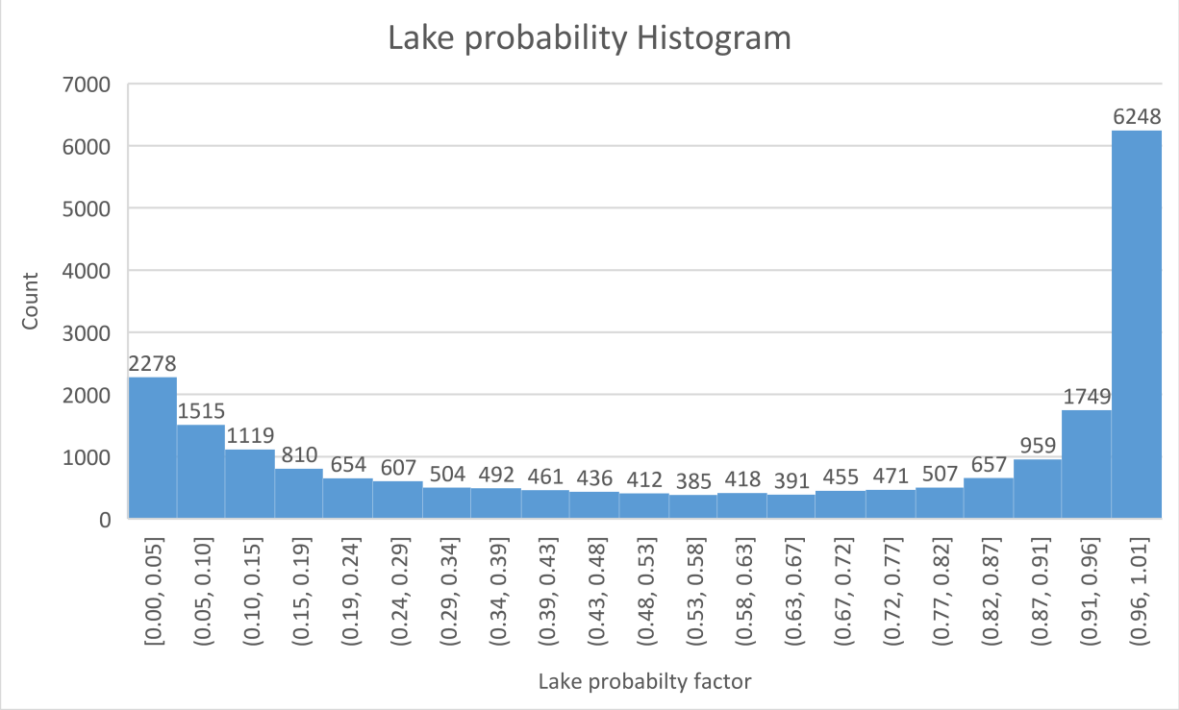


Figure 31 Histogram of the lake probability count

There appears to be a relationship between size and probability as seen in Figure 33, 34 and 35 which show an even distribution until the 95% probability region where the lakes area increases exponentially.

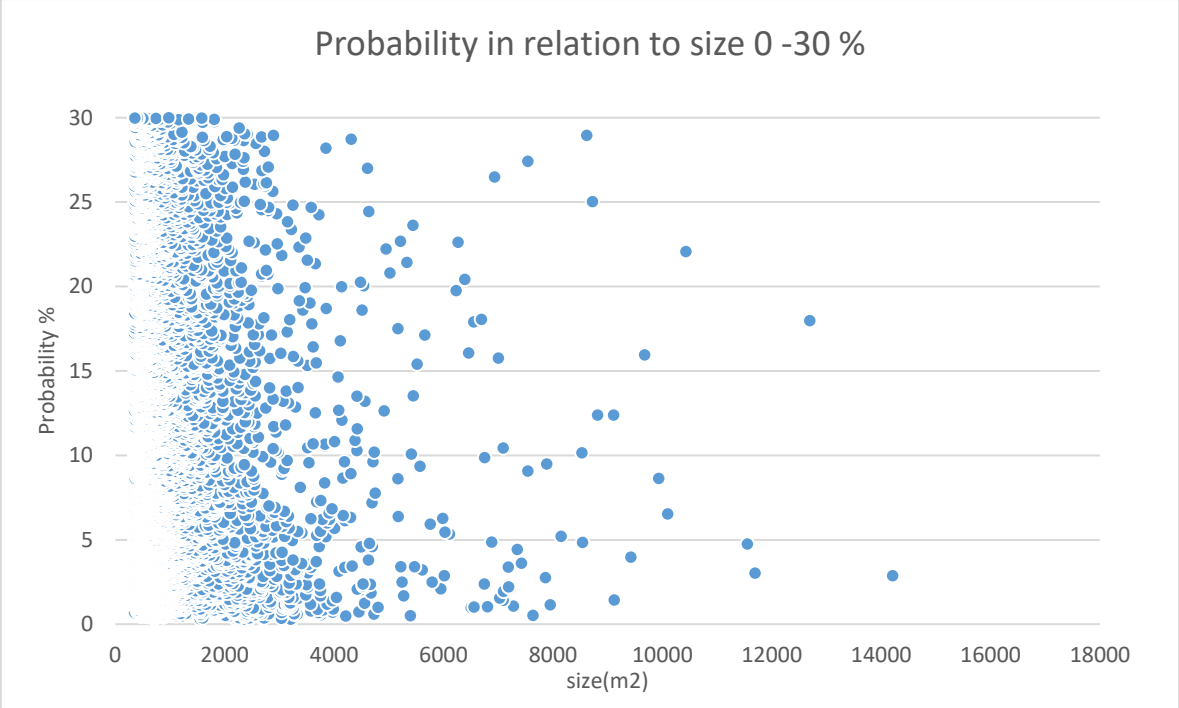


Figure 32 X and Y graph with the probability values between 0 and 30% and their sizes.

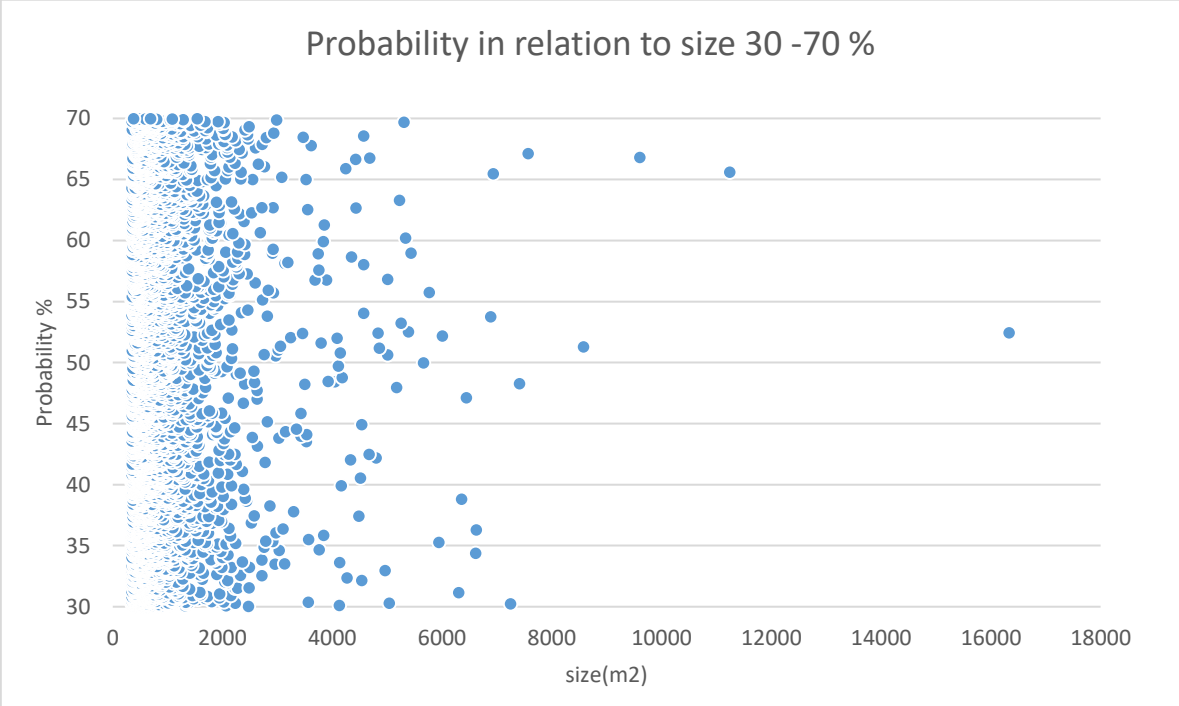


Figure 33 X and Y graph with the probability values between 30 and 70% and their sizes

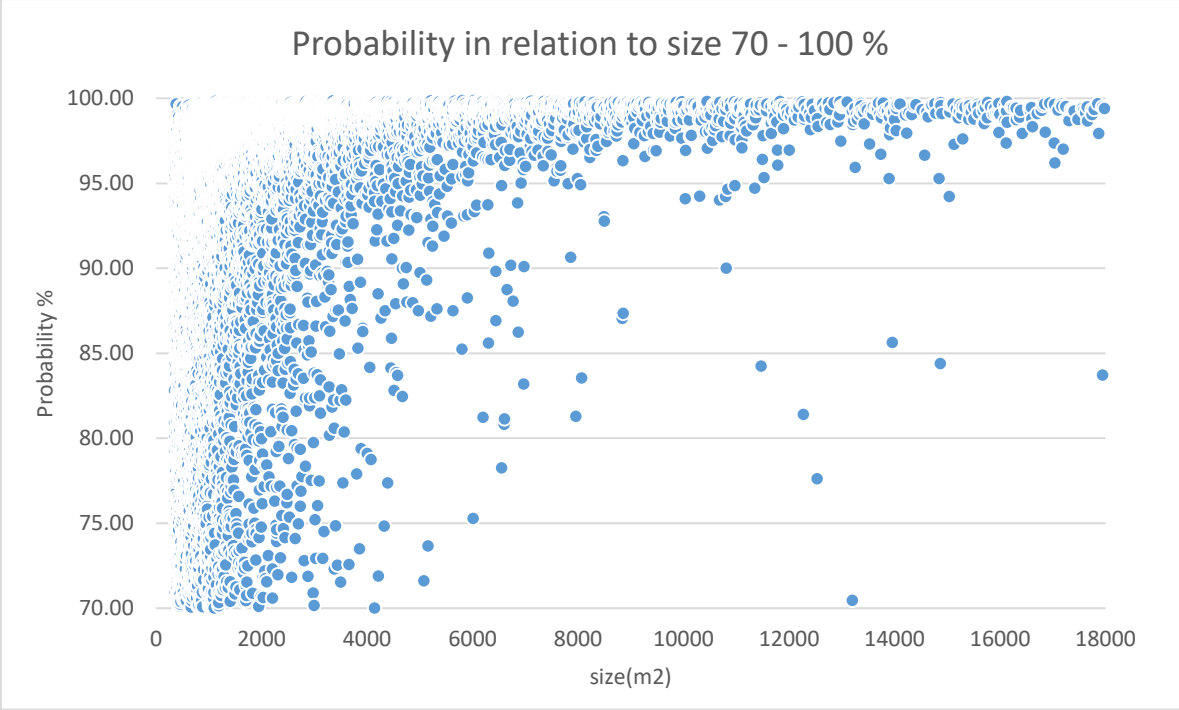


Figure 34 X and Y graph with the probability values between 70 and 100% and their sizes.

4.2 Accuracy assessment

The model's accuracy was measured using the following formula (Google Developers, n.d.):

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

In order to gather all correct and incorrect predictions a confusion matrix was created as seen below in table 5. The test set of 250 was visually verified and entered into the confusion matrix, as a result the accuracy figure is taken from a subsample of 250 from a total of 21,528.

Table 5 Confusion matrix of results, containing true positive, false positive, false negative and true negative.

Predicted class	Actual Lake		
	Positive	Negative	Total
Positive	97 (TPs)	11 (FPs)	108
Negative	28 (FNs)	114 (TNs)	142
Total	125	125	

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{97 + 114}{97 + 114 + 11 + 28} = 0.85$$

Accuracy comes to 0.85 or 85% of the predictions out the total 250 were correct. Thirty nine of the lakes were falsely classified. The balance between the false negative and false positive is relatively balanced, with more actual lakes being classified as not.

$$Precision = \frac{TP}{TP + FP}$$

$$Precision = \frac{TP}{TP + FP} = \frac{97}{97 + 11} = 0.89$$

The model has a precision of 0.89, according to the test set the when the model determines that a lake is a lake, it's correct 89% percent of the time. Vice versa the precision of the model successfully predicting false lakes is as follows.

$$Precision = \frac{TN}{TN + FN} = \frac{114}{114 + 28} = 0.8$$

5 Discussion

This chapter discusses the results of this thesis and details the strengths and weaknesses of the methodology implemented in this report. The possible applications and limitations of this research are also provided.

The results highlighted by this study are highly significant, revealing a large amount of errors within GeoDanmarks lake data set. Approximately 33% of all lakes within the sample area have a probability value of 0.3 or less. With an accuracy rate of 80% this still leaves 6015 lakes or 28% of the lake dataset is incorrect.

The reasons behind these errors could be due to a number of reasons such as seasonal changes in the environment, it should be noted that the ortophoto used to train and test the model, were taken in the summer of 2017. Shallow water bodies could be subject to drought resulting in water boundary change or even complete disappearance. As a result a portion of these errors discovered in the data set could be seasonal water bodies, present during other times of the year. Alternatively, human intervention may be another possible cause for incorrect features.

Such large errors within GeoDanmarks Lake Dataset could suggest further errors in other OGD datasets. This result questions the validity of GeoDanmark and other forms of open government data.

The gradient boosting library Catboost, proved to be an effective tool in identifying and predicting lake features using the multispectral satellite system sentinel 2 values. With an accuracy of 85% and relatively equally high distribution of precisions.

Size appears to have a correlation with lakes probability values, this is most likely due to the fact that larger lakes are more identifiable due to spatial resolution of Sentinel 2. The resulting feature strength file listed NDWI values to be one of the most influential features in determining the probability values. This result reinforces the value of NDWI (Mcfeeters,1996) in water body extraction and detection.

This study has shown the applicability of 'black box' machine learning algorithms and their use in remote sensing and GIS. The majority of time spent in this analysis was preparing the data for use within Catboost. Training and applying the model did not require an in depth knowledge of computer science and machine learning. This is just one of the many capabilities of Catboost and machine learning algorithms. However, this study is an encouraging beginning.

5.1 Limitations

The initial discovery in carrying out this project was the limitations of Sentinel 2's 10m resolution, thirty percent of the lakes within the sample area were under 500 m2 in size. It is not possible to acquire accurate pixel values for lakes under 100m2 with sentinel 2, so to ensure this problem from not occurring, lakes with an area of 500 m2 or less were excluded from the analysis.

5.2 Application

Catboost's use as a data analysis tool in this study, can be replicated on any other geographical data set and is not limited to the classification of just water bodies. The rest of Geodanmarks data set could be assessed, or any other, given sufficient training and testing sets can be obtained.

With the majority of countries incorporating some form of an open data framework, tools like these help to sustain a level of required quality and reusability, which a range of sectors and industries rely upon.

6 Conclusions

This thesis explores the area of machine learning and its application in remote sensing and geographic information systems. Using one of the latest and most competitive machine learning algorithms in the market today. Specifically, in the area of quality assessment regarding geographic datasets. In exploring this broad subject three research questions were created; firstly the question of GeoDanmarks data quality, secondly the effectiveness of the machine learning tool being used to assess this data quality, and lastly which feature value carried the most weight in the algorithms predictions. These questions and their answers give us some insight into the relatively new and highly ambitious field of machine learning and GIS. These three research questions have been answered and their significance discussed.

The quality of GeoDanmarks lake data set has been proven to contain a large percentage of errors, with a value of 28%. The effectiveness of the algorithm Catboost, in obtaining this error value was largely successful. With an accuracy value of 85% and precision values of around 80%, using a subsample of 250 validated lakes valued within a confusion matrix. The algorithm determined that NDWI was one of its strongest features in obtaining predictions.

In conclusion this study has adequately answered the research questions and explored the benefits of ‘black box’ machine learning tools and their application within the field of remote sensing. As mentioned in the limitations section this analysis focused on a niche subject, namely the validation and classification of an OGD dataset. While the results are promising,

future studies involving the use of Catboost within the field of remote sensing and OGD validation would be beneficial.

6.1 Future Directions

This analysis can be added to in a number of ways depending on the desired outcome. The quality of the training set is one of the primary factors in determining the models efficiency, the model will only know what a lake is, depending on the users input or the user's interoperation of what a lake is. In this case perhaps a more strict criteria should be developed and stated before carrying out a study of this kind. The data set contains the boundaries and locations of 'SOE' or 'lakes' when translated to English. This could refer to any body of water from small drains to massive inland bodies of water that are more than 10km2 in size. In future studies it may prove useful to categorize these into more homogenous groups, allowing for a better understanding and interpretation of the results.

This study was limited to lakes of 500m2 due to the limited mid – level resolution of sentinel. Higher resolution sensors Quickbird and worldview etc. would allow for a more in-depth analysis of water bodies, as large portion of the bodies of water were under 500m2 in size (30%). Depending on the study requirements, it may be necessary to upgrade to higher resolution imagery.

In addition to the abovementioned categorization of lake types, improved methods of lake validation would be beneficial. In this study, ortophoto's were solely used to validate the 500 training and testing lakes. In order to get a well distributed training set in terms of size, other sources of validation data (in situ) could be beneficial, as at times the smaller lakes were difficult to distinguish.

7 References

Abdou, Bannari & Mohammed, Ghadeer & El Battay, Ali & Mohamed, Nadir & Rouai, M. (2016). Assessment of Land Erosion and Sediment Accumulation Caused by Runoff after a Flash-Flooding Storm Using Topographic Profiles and Spectral Indices. Advances in Remote Sensing. 5. 315-354. 10.4236/ars.2016.54024.

Alderman, K., Turner, L. R., & Tong, S. (2012). Floods and human health: a systematic review. Environment international, 47, 37-47.

Batini, C., Blaschke, T., Lang, S., Albrecht, F., Abdulmutalib, H. M., Barsi, Á., ... & Kugler, Z. (2017). Data Quality in Remote Sensing. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 42.

Bianchin, A. (2001, October). Nuovi approcci alla validazione dei DB cartografici. In Atti 5a Conferenza Nazionale ASITA.

Bissacco, A., Yang, M. H., & Soatto, S. (2007, June). Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In 2007 IEEE conference on computer vision and pattern recognition (pp. 1-8). IEEE.

Bond, N. R., Lake, P. S., & Arthington, A. H. (2008). The impacts of drought on freshwater ecosystems: an Australian perspective. Hydrobiologia, 600(1), 3-16.

Caprioli, M., & Tarantino, E. (2001). Accuracy assessment of per-field classification integrating very fine spatial resolution satellite imagery with topographic data. Journal of Geospatial Engineering, 3(2), 127-134.

Carroll, M. L., Townshend, J. R., DiMiceli, C. M., Noojipady, P., & Sohlberg, R. A. (2009). A new global raster water mask at 250 m resolution. International Journal of Digital Earth, 2(4), 291-308.

CatBoost - state-of-the-art open-source gradient boosting library with categorical features support. (n.d.). Retrieved from https://catboost.ai/

Chen, Q., Zhang, Y., Ekroos, A., & Hallikainen, M. (2004). The role of remote sensing technology in the EU water framework directive (WFD). Environmental Science & Policy, 7(4), 267-276.

Dawes, S. S. (2010). Stewardship and usefulness: Policy principles for information-based transparency. Government Information Quarterly, 27(4), 377-383.

Detlor, B., Hupfer, M. E., Ruhi, U., & Zhao, L. (2013). Information quality and community municipal portal use. Government Information Quarterly, 30(1), 23-32.

Ding, F. (2009). Study on information extraction of water body with a new water index (NWI). Science of Surveying and Mapping, 34(4), 155-158.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., ... & Meygret, A. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25-36.

Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W., & Li, X. (2016). Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band. Remote Sensing, 8(4), 354.

European space agency (ESA). (2018). What is remote sensing? Retrieved from http://www.esa.int/SPE CIALS/Eduspace_EN/SEMF9R3Z2OF_0.html

Fanelli, G., Dantone, M., Gall, J., Fossati, A., & Van Gool, L. (2013). Random forests for real time 3d face analysis. International Journal of Computer Vision, 101(3), 437-458.

Feng, L., Hu, C., Chen, X., Cai, X., Tian, L., & Gan, W. (2012). Assessment of inundation changes of Poyang Lake using MODIS observations between 2000 and 2010. Remote Sensing of Environment, 121, 80-92.

Ferro, E., & Osella, M. (2013, April). Eight business model archetypes for PSI re-use. In Open data on the web workshop.

Feyisa, G. L., Meilby, H., Fensholt, R., & Proud, S. R. (2014). Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. Remote Sensing of Environment, 140, 23-35.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

Gao, B. C. (1995, June). Normalized difference water index for remote sensing of vegetation liquid water from space. In Imaging Spectrometry (Vol. 2480, pp. 225-237). International Society for Optics and Photonics.

Google Developers. (n.d.). Classification: Accuracy. Machine Learning Crash Course. Retrieved August 01, 2018, from https://developers.google.com/machine-learning/crash-course/classification/accuracy

Hansen, H. S., Hvingel, L., & Schrøder, L. (2013, August). Open Government Data–a key element in the digital society. In International Conference on Electronic Government and the Information Systems Perspective (pp. 167-180). Springer, Berlin, Heidelberg.

Helbig, N., Nakashima, M., & Dawes, S. S. (2012, June). Understanding the value and limits of government information in policy informatics: a preliminary exploration. In Proceedings of the 13th annual international conference on digital government research (pp. 291-293). ACM.

Huang, S., Li, J., & Xu, M. (2012). Water surface variations monitoring and flood hazard analysis in Dongting Lake area using long-term Terra/MODIS data time series. Natural hazards, 62(1), 93-100.

Huang, X., & Jensen, J. R. (1997). A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. Photogrammetric engineering and remote sensing, 63(10), 1185-1193.

Hui, F., Xu, B., Huang, H., Yu, Q., & Gong, P. (2008). Modelling spatial-temporal change of Poyang Lake using multitemporal Landsat imagery. International Journal of Remote Sensing, 29(20), 5767-5784.

Hutchinson, R. A., Liu, L. P., & Dietterich, T. G. (2011, August). Incorporating Boosted Regression Trees into Ecological Latent Variable Models. In AAAI (Vol. 11, pp. 1343-1348).

Immitzer, M., Vuolo, F., & Atzberger, C. (2016). First experience with Sentinel-2 data for crop and tree species classifications in central Europe. Remote Sensing, 8(3), 166.

Jain, S. K., Singh, R. D., Jain, M. K., & Lohani, A. K. (2005). Delineation of flood-prone areas using remote sensing techniques. Water Resources Management, 19(4), 333-347.

Jensen, J. (2005). Introductory digital image processing: A remote sensing perspective (3.nd ed., Prentice Hall series in Geographic Information Science). Upper Saddle River, N. J: Prentice Hall.

Johnson, R., & Zhang, T. (2014). Learning nonlinear functions using regularized greedy forest. IEEE transactions on pattern analysis and machine intelligence, 36(5), 942-954

Kaplan, G., & Avdan, U. (2017). Object-based water body extraction model using Sentinel-2 satellite imagery. European Journal of Remote Sensing, 50(1), 137-143.

Khorram, S., Koch, F. H., van der Wiele, C. F., & Nelson, S. A. (2012). Remote sensing. Springer Science & Business Media.

Lemma, R., Morando, F., & Osella, M. (2014). Breaking public administrations' data silos. The Case of Open-DAI, and a Comparison between Open Data Platforms. eJournal of eDemocracy and Open Government, 6(2).

Li, B., Zhang, H., & Xu, F. (2014). Water extraction in high resolution remote sensing image based on hierarchical spectrum and shape features. In IOP Conference Series: Earth and Environmental Science (Vol. 17, No. 1, p. 012123). IOP Publishing.

Li, Y., Gong, X., Guo, Z., Xu, K., Hu, D., & Zhou, H. (2016). An index and approach for water extraction using Landsat–OLI data. International Journal of Remote Sensing, 37(16), 3611-3635.

Li, Y., Tao, C., Tan, Y., Shang, K., & Tian, J. (2016). Unsupervised multilayer feature learning for satellite image scene classification. IEEE Geoscience and Remote Sensing Letters, 13(2), 157-161.

Lillesand, T., Kiefer, R. W., & Chipman, J. (2014). Remote sensing and image interpretation. John Wiley & Sons.

Liu, Y., Wang, Y., Li, Y., Zhang, B., & Wu, G. (2004, August). Earthquake prediction by RBF neural network ensemble. In International Symposium on Neural Networks (pp. 962-969). Springer, Berlin, Heidelberg.

McFeeters, Stuart K. "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features." International journal of remote sensing 17.7 (1996): 1425-1432.

N.M. Short, The Remote Sensing Tutorial [web site]. National Aeronautics and Space Administration (NASA), Goddard Space Flight Center (2010), http://rst.gsfc.nasa.gov/

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

National Ocean Service (NOAA). (2018). what is remote sensing? Retrieved from https://oceanservice.noaa.gov/facts/remotesensing.html

OGC (n.d.). Glossary of Terms - D. Retrieved July 13, 2018, from http://www.opengeospatial.org/ogc/glossary/d

Pesaresi, M., Corbane, C., Julea, A., Florczyk, A. J., Syrris, V., & Soille, P. (2016). Assessment of the added-value of Sentinel-2 for detecting built-up areas. Remote Sensing, 8(4), 299.

Pittman, S. J., & Brown, K. A. (2011). Multi-scale approach for predicting fish species distributions across coral reef seascapes. PloS one, 6(5), e20583.

Qi, Y. (2012). Random forest for bioinformatics. In Ensemble machine learning (pp. 307-323). Springer, Boston, MA.

Rogers, S., & Girolami, Mark. (2017). A first course in machine learning (2.nd ed., Chapman & Hall/CRC machine learning & pattern recognition series). Boca Raton, Fl: CRC Press.

Rokni, K., Ahmad, A., Selamat, A., & Hazini, S. (2014). Water feature extraction and change detection using multitemporal Landsat imagery. Remote Sensing, 6(5), 4173-4189.

Satellite Remote Sensing Systems. (n.d.). Retrieved September 01, 2018, from https://www.satimagingcorp.com/services/resources/characterization-of-satellite-remote-sensing-systems/

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In Nonlinear estimation and classification (pp. 149-171). Springer, New York, NY.

Sewell,M.(2011). Ensemble Learning. Technical Report, Department of Computer Science, University College London. Available online at: http://www.cs. ucl.ac.uk/fileadmin/UCL-CS/research/Research_Notes/RN_11_02.pdf

Shang, X., & Chisholm, L. A. (2014). Classification of Australian native forest species using hyperspectral remote sensing and machine-learning classification algorithms. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens, 7, 2481-2489.

Sheng, Y., Shah, C. A., & Smith, L. C. (2008). Automated image registration for hydrologic change detection in the lake-rich Arctic. IEEE geoscience and remote sensing letters, 5(3), 414-418.

Shu, C., & Burn, D. H. (2004). Artificial neural network ensembles and their application in pooled flood frequency analysis. Water Resources Research, 40(9).

Sivanpillai, R., & Miller, S. N. (2010). Improvements in mapping water bodies using ASTER data. Ecological Informatics, 5(1), 73-78.

Srivastava, T. (2016, August 03). Basics of Ensemble Learning Explained in Simple English. Retrieved September 01, 2018, from https://www.analyticsvidhya.com/blog/2015/08/introduction-ensemble-learning/

Sun, F., Sun, W., Chen, J., & Gong, P. (2012). Comparison and improvement of methods for identifying waterbodies in remotely sensed imagery. International journal of remote sensing, 33(21), 6854-6875.

Tapete, D. Remote Sensing and Geosciences for Archaeology. (2018). MDPI AG - Multidisciplinary Digital Publishing Institute.

Techleer.(2017). Machine Learning Algorithm - Backbone of emerging technologies. Retrieved from https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/

The National Aeronautics and Space Administration (NASA). (2017). What are passive and active sensors?.Retrieved from https://www.nasa.gov/directorates/heo/scan/communications/outreach/funfacts/txt_passive_active.html

Veregin, H. (1999). Data quality parameters. Geographical information systems, 1, 177-189.