UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

BUSINESS INTELLIGENCE THROUGH MACHINE LEARNING FROM SATELLITE REMOTE SENSING DATA

Diploma Thesis

Christos Kyriakos

Supervisor: Manolis Vavalis

July 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

BUSINESS INTELLIGENCE THROUGH MACHINE LEARNING FROM SATELLITE REMOTE SENSING DATA

Diploma Thesis

Christos Kyriakos

Supervisor: Manolis Vavalis

July 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΝΟΗΜΟΣΥΝΗ ΜΕΣΩ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΔΟΡΥΦΟΡΙΚΗΣ ΤΗΛΕΠΙΣΚΟΠΗΣΗΣ

Διπλωματική Εργασία

Χρήστος Κυριάκος

Επιβλέπων/πουσα: Μανόλης Βάβαλης

Ιούλιος 2022

Approved by the Examination Committee:

Supervisor **Manolis Vavalis**
Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Michael Vassilakopoulos**
Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Dimitrios Katsaros**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Mr. Manolis Vavalis for his valuable assistance in developing this thesis. I would, also, like to thank Mr. Michael Vassilakopoulos and Mr. Dimitrios Katsaros for their participation in the examination committee. I would like to thank the open-source community of Google Earth Engine, for providing open-access tools and guides that proved to be vital for the completion of the software proposed in this thesis. Last but not least, I would like to thank my family and friends for their support throughout my studies.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Christos Kyriakos

Diploma Thesis

BUSINESS INTELLIGENCE

THROUGH MACHINE LEARNING

FROM SATELLITE REMOTE SENSING DATA

Christos Kyriakos

Abstract

Remote sensing offers the opportunity to monitor natural and human driven phenomena directly and at a large scale. The rapid improvement of satellite imagery and their derived data products, provide valuable information for the public and private sectors. These ”big satellite data” have various properties (e.g enormous volume, noise etc) that make their direct use for geospatial and economic analysis inefficient. Artificial Intelligence has found successful application in numerous fields, since apart from being able to process large data volumes in short time frames, it also reduces bias and human error significantly. This thesis focuses on the use of satellite data and machine learning for providing insights for businesses and policy makers within Greece. Greece has been greatly affected by economic crisis, unregulated gentrification and more recently the pandemic resulting in increased vacancy rates, especially in major cities. Abandoned buildings have various negative implications on the area close to them, including increased chance for fire, crime and generally reduce its monetary value. First we introduce some important characteristics of satellites as well as some concepts related to machine learning. Numerous successful implementations of ML and RS within the research sphere that can improve the major Greek economic sectors (e.g agriculture) are also presented. Finally, we propose an open source software for the detection of abandoned or disused buildings based on Nighttime Lights and Built-Up Area Indices. This approach provided promising results, through the various, albeit not extensive, experimenters we performed and can be used by SMEs for Location Intelligence.

Keywords:

remote sensing, business intelligence,machine learning, abandoned buildings

Διπλωματική Εργασία

ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΝΟΗΜΟΣΥΝΗ ΜΕΣΩ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΔΟΡΥΦΟΡΙΚΗΣ ΤΗΛΕΠΙΣΚΟΠΗΣΗΣ

Χρήστος Κυριάκος

Περίληψη

Η τηλεπισκόπηση προσφέρει την ευκαιρία να ελέγχουμε τα φυσικά και ανθρωπογενή φαινόμενα άμεσα και σε μεγάλη κλίμακα. Η ραγδαία βελτίωση των δορυφορικών εικόνων και τα παραγόμενα από αυτές προϊόντα δεδομένων παρέχουν πολύτιμη πληροφορία για τον δημόσιο και ιδιωτικό τομέα. Αυτά τα ”μεγάλα δορυφορικά δεδομένα” έχουν διάφορες ιδιότητες (π.χ μεγάλο όγκο, θόρυβο κ.λ.π) οι οποίες καθιστούν την άμεση χρήση τους για γεωχωρικές και οικονομικές αναλύσεις μη αποτελεσματική. Η τεχνητή νοημοσύνη έχει εφαρμοστεί επιτυχώς σε πολλούς τομείς, καθώς εκτός από την δυνατότητα της να επεξεργάζεται μεγάλο όγκο δεδομένων σε περιορισμένα χρονικά διαστήματα, μειώνει ”προκαταλήψεις” και σφάλματα που προκύπτουν από τις ανθρώπινες ενέργειες. Αυτή η πτυχιακή εργασία επικεντρώνεται στη χρήση δορυφορικών δεδομένων και μηχανικής μάθησης για την παροχή πληροφοριών χρήσιμων σε επιχειρήσεις και υπεύθυνους χάραξης πολιτικής εντός του Ελλαδικού χώρου. Την Ελλάδα έχουν πλήξει σε μεγάλο βαθμό η οικονομική κρίση, η ανεξέλεγκτη αστικοποίηση και πιο πρόσφατα η πανδημία του COVID-19, με αποτέλεσμα την αύξηση του βαθμού κενής θέσης, ειδικότερα σε μεγάλες πόλεις. Τα εγκαταλελειμμένα κτίρια έχουν διάφορες αρνητικές επιπτώσεις στην περιοχή που τα περιβάλλει, όπως αυξημένη πιθανότητα πυρκαγιάς, αύξηση της εγκληματικότητας και γενικότερα υποβάθμιση της αντικειμενικής της αξίας. Αρχικά εισάγουμε τον αναγνώστη σε κάποια κύρια χαρακτηριστικά των δορυφόρων καθώς και σε κάποιες έννοιες σχετικές με τη μηχανική μάθηση. Επιπλέον, παρουσιάζουμε διάφορες επιτυχημένες εφαρμογές της μηχανικής μάθησης και της τηλεπισκόπησης στο πλαίσιο έρευνας, από τις οποίες θα μπορούσαν να ωφεληθούν τομείς της Ελληνικής οικονομίας (π.χ ο αγροτικός). Τέλος, προτείνουμε την δημιουργία ενός λογισμικού ανοιχτού κώδικα για τον εντοπισμό εγκαταλελειμμένων ή αχρησιμοποίητων κτιρίων βασισμένοι σε φώτα νυχτός και δείκτες χτισμένων περιοχών. Η μέθοδος αυτή που είχε υποσχόμενα αποτελέσματα, φανερά μέσα από διάφορα αλλά όχι εξαντλητικά πειράματα, μπορεί να χρησιμοποιηθεί από επιχει-

ρήσεις μικρού και μεσαίου μεγέθους για την απόκτηση χωρικών πληροφοριών.

Λέξεις-κλειδιά:

τηλεπισκόπηση, επιχειρηματική ευφυΐα, μηχανική μάθηση, εγκαταλελειμμένα κτίρια

Table of contents

List of figures

List of tables

Abbreviations

NDVI	Normalized Difference Vegetation Index
NDBI	Normalized Difference Built-up Index
EMBI	Enhanced Modified Bare Soil Index
VgNIRBI	Visible Green-Based Built-Up Index.
VrNIRBI	Visible Red-Based Built-Up Index
PISI	Perpendicular Impervious Surface Index
OSAVI	Optimized Soil Adjusted Vegetation Index
LAI	Leaf Area Index
Avg Rad	Average Radiance
ETM+	Enhanced Thematic Mapper Plus
USGS	United States Geological Survey
MCNN-seq	Multi CNN Sequence to Sequence
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network
RoI	Region of Interest
AoI	Area of Interest
PoI	Point of Interest

Chapter 1

Introduction

Remote sensing (RS), while having a lot of interpretations, mainly refers to the acquisition of earth surface features, objects or phenomena, and information about their geophysical and biophysical properties with the use of propagated signals such as electromagnetic radiation, without coming in contact with the object. RS systems have been growing rapidly due to developments in sensor system technology and digital processing and include satellites and aircraft-based sensor technologies. These systems allow data collection from all ranges of the electromagnetic spectrum, including energy emitted, reflected, and/or transmitted, which subsequently can be turned into information products. These products, have features that make them important for systematic and/or managerial decision-making, regarding local area studies or worldwide analyses, using either manual or machine-assisted interpretation. Taking this into account, RS has been applied successfully in a plethora of fields, ranging from commerce to public policy, such as land surveying, planning, economic, humanitarian and military applications. The complexity of these data, however, makes their use quite difficult since it requires background knowledge and the computational ability to process them. Machine learning, which has found successful application in different fields makes satellite data more accessible to even businesses that don't have experience working with them. Below, the main features of satellites as well as the key aspects of machine learning will be presented in order to better realize the importance and advantages of their combination.

1.1 Motivation

Apart from the traditional use of satellites in telecommunications, weather forecasting and military, satellite data seem to have a lot of interesting applications that can improve many fields that the majority of people takes no notice of. Satellite data are commonly used nowadays by governments, big enterprises or researchers in order to make informed decisions for large scale regions. However, albeit their importance, they were until recently mostly ignored by small and medium businesses (at least in Greece) due to lack of information, expertise and even costs. There's demand for a simple, easily accessible and user friendly environment in order to leverage satellite data. The platform should also be open source so that the costs of using it can be reduced to a big extent. There are numerous examples even among this field of successful open sourced geospatial analytic tools, such as Open Data Cube. This thesis aims to design and implement an open source platform that assists the decision making and management of small and medium scale businesses as well as policy making, through the use of satellite data and machine learning (proof-of-concept).

1.2 Structure

The rest of this thesis is organized as follows:

- **Chapter 2:** A brief introduction to remote sensing satellite data. Overview of important satellite characteristics and derived data products.
- **Chapter 3:** Introduction to Machine Learning concepts and model calibration. Issues regarding the implementation of ML with remote sensing are also discussed.
- **Chapter 4:** The various successful applications of Remote Sensing Machine Learning. Their importance for Greece is also discussed.
- **Chapter 5:** Extensive presentation of the proposed software, including model creation, experimentation and evaluation
- **Chapter 6:** Synopsis and Future Work

Chapter 2

Satellite Systems in Remote Sensing

An artificial satellite is an item that is placed into orbit, usually with the use of a rocket, in order to be used in a variety of fields. They are typically equipped with an antenna that enables communication with the space station and a source of power (i.e battery, solar panel etc.) and they can operate individually or within a larger system [1]. Their positioning varies as they are placed at different heights and follow different paths/orbits depending on their use case[2].

Geostationary orbit (GEO), Low Earth (LEO) and Medium Earth orbit (MEO), Polar orbit and Sun-synchronous orbit (SSO), geostationary transfer orbit (GTO) and Lagrange points (L-points) are some of the common satellite orbits. For reference, LEO is relatively close to earth with an altitude of less than 1000 km down to 160 km polar and orbit, while Geostationary orbit (GEO) entails where the satellite circles above equator.

Since the launch of the first satellite by the Soviet Union in 1957, they have been used for a lot of different purposes, such as earth monitoring or mapping planetary surfaces. They can also be categorized depending on the field they are being used into communications, navigation and weather satellites but also as military or civilian Earth observation satellites. Satellites, similarly to other remote sensing devices, have a variety of features the comprehension of which is important in order to select the best one for our use case. Below we will briefly present the main characteristics of satellites that can impact our area/type of study and the accuracy of our analyses.

[1] https://en.wikipedia.org/wiki/Satellite_formation_flying

[2] https://www.esa.int/Enabling_Support/Space_Transportation/Types_of_orbits

2.1 Satellite Sensors and Satellite Instruments

Satellite sensors are divided into active and passive depending on the way they transmit signals [3], [4].

2.1.1 Active Sensors

Active sensors are equipped with a radiation-transmitting equipment (e.g transponder) which allows the transmission of a signal directed towards a target (usually Earth) and the detection of the target-reflected radiation in order to be measured by the sensor, emulating a source of light. These devices usually employ microwaves due to their relative immunity to weather conditions. Active remote sensing, employs different techniques based on what is being broadcasted (e.g light or waves) and what is measured (e.g distance, height, atmospheric conditions, etc.). Some active sensors are:

- **SAR:** a transmitter that operates at microwave and radio frequencies. Its specific feature is a directional antenna emitting impulses in order to determine the distance to an object.

- **Lidar:** determines location and distance of an object by multiplying the time it took for the round trip by the speed of light. Lidars are able to determine atmospheric profiles or aerosols, clouds and other atmosphere constituents

- **Laser altimeter:** measures elevation using Light Detection and Ranging (LIDAR) in order to determine the topography of a surface.

- **Sounder:** provide vertical profiles of weather conditions by measuring infrared radiation.

- **Scatterometer:** is a microwave radar device performing at high frequencies, that measures backscattered radiation in the microwave spectrum. It is usually used in order to map wind direction and its speed at the planet surface level.

[3] https://earthobservatory.nasa.gov/features/RemoteSensing/remote_08.php

[4] https://www.nrcan.gc.ca/maps-tools-publications/satellite-imagery-air-photos/remote-sensing-tutorials/introduction/passive-vs-active-sensing/14639

Active sensors are fully functional anytime since they are relatively independent of atmospheric scatterings and sunlight. The absence of research conditions' restrictions makes them crucial for observations in inaccessible areas, such as those in marine sciences and rescue missions.

2.1.2 Passive Sensors

Passive Sensors measure naturally emitted energy such as reflected sunlight, since they do not streamline their own energy to the researched area. Passive sensors require proper weather conditions and sunlight as otherwise there will be nothing to reflect. Multispectral and hyperspectral sensors are used extensively in passive remote sensing so as to measure the desired quantity using various band combinations. Passive remote sensing devices include:

- **Spectrometer:** distinguishes and analyzes spectral bands. Typically they use grating or prisms for dispersion.

- **Radiometer:** measures the strength of electromagnetic radiation in some spectral band (visible, IR, microwave).

- **Spectroradiometer:** measures radiation intensity in multiple wavelength bands, usually of high spectral resolution designed for remote sensing applications(i.e., multispectral). They typically measure parameters regarding cloud features, sea color and temperature or atmosphere chemical traces and finally vegetation.

- **Imaging radiometer:** a radiometer that can scan, mechanically or electronically, and provides a 2D array of pixels for image generation.

- **Accelerometer:** detects changes in speed (e.g. linear or rotational). It is used in order to distinguish between the influence of gravity and those caused by atmosphere air drag on the satellite.

The need for improved, more accurate and versatile techniques has led to **Microwave Remote Sensing** which is the combination of active and passive remote sensing. Compared to the visible and infrared, the microwave part of the spectrum ranges from around 1cm to 1m in wavelength and as a result has special properties that can be used in remote sensing.

The main advantage of microwave radiation is that it can be used even on extreme environmental and weather conditions, since the longer wavelength microwaves are not susceptible to atmospheric scattering from cloud cover and haze rainfalls.

2.1.3 Resolutions of Satellite Instruments

Satellite imagery is a satellite product that is being used by businesses and governments. Their resolution varies based on the used instruments and the altitude of the satellite's orbit. Depending on their resolution they can be prohibitive for some applications. Satellite resolutions in the remote sensing context can be split into five types: spatial, spectral, temporal, radiometric and geometric and they are described as follows:

- **Spatial resolution** is a satellite image's pixel size that represents the size of the surface area (i.e. m2) being measured on the ground. This resolution refers to the smaller discernible feature of a satellite image and depends on the visibility of the sensor. For example, a passive sensor's spatial resolution depends on their Instantaneous Field of View (IFOV). A spatial resolution of 10m means that a pixel of the image represents a ground area of 10 x 10 meters.

- **Spectral resolution** is determined by the size of the wavelength interval and number of intervals measured by the sensor. In other words it refers to the sensor's capacity to detect specific wavelengths of the electromagnetic spectrum. Higher spectral resolution segments the electromagnetic spectrum into finer wavelength ranges allowing the identification of more specific classes(i.e rock type vs vegetation).

- **Temporal resolution** is defined by the time interval between imagery collection (e.g. days) for the same surface location. Satellites' ability to take photos of the same geographic region more regularly has drastically improved over the years, offering more accurate data.

- **Radiometric resolution** is a satellite sensor's capacity to capture a variety of brightness levels, such as contrast, and its actual bit depth. i,e the number of grayscale levels.

- **Geometric resolution** defines a sensor's capability to successfully display a patch of the Earth's surface within a single pixel. It is typically expressed in terms of Ground sample distance (GSD),

Sentinel-2 bands	Central wavelength (μm)	Resolution (m)
Band 1 – Coastal aerosol	0.443	60
Band 2 – Blue	0.490	10
Band 3 – Green	0.560	10
Band 4 – Red	0.665	10
Band 5 – Vegetation red edge	0.705	20
Band 6 – Vegetation red edge	0.740	20
Band 7 – Vegetation red edge	0.783	20
Band 8 – NIR	0.842	10
Band 8A – Vegetation red edge	0.865	20
Band 9 – Water vapour	0.945	60
Band 10 – SWIR – Cirrius	1.375	60
Band 11 – SWIR	1.610	20
Band 12 – SWIR	2.190	20

Figure 2.1: Sentinel 2 MSI Band Characteristics, source: [1]

Bands	Wavelength (micrometers)	Resolution (meters)
Band 1 - Coastal aerosol	0.43 - 0.45	30
Band 2 - Blue	0.45 - 0.51	30
Band 3 - Green	0.53 - 0.59	30
Band 4 - Red	0.64 - 0.67	30
Band 5 - Near Infrared (NIR)	0.85 - 0.88	30
Band 6 - SWIR 1	1.57 - 1.65	30
Band 7 - SWIR 2	2.11 - 2.29	30
Band 8 - Panchromatic	0.50 - 0.68	15
Band 9 - Cirrus	1.36 - 1.38	30
Band 10 - Thermal Infrared (TIRS) 1	10.60 - 11.19	100
Band 11 - Thermal Infrared (TIRS) 2	11.50 - 12.51	100

Figure 2.2: Landsat 8 OLI Band Characteristics, source: [2]

2.1.4 Spectral Bands and Spectral Indices

A **spectral band** is a limited range of values the sensor is set to detect along a electromagnetic spectrum. The electromagnetic spectrum is measured along a continuum of wavelengths (peak-to-peak distance). Depending on how this distance changes the wave has different properties regarding energy transmission (e.g heating).

We typically set spectral bands based on physical effects we understand and use them to differentiate between remotely-sensed objects, such as trees and bare soil or granite. In other words, a spectral band is a type of filter that allows only the desired wavelength acquired by the sensor to pass. This filter requires the use of an **atmospheric window**, which is a wavelength range of the electromagnetic spectrum that has the ability to pass through the atmosphere, come in contact with the target area and get back to the sensor without being absorbed or scattered. The optical, infrared and radio windows comprise the three main atmospheric windows.

Thus a spectral “band” is a defined portion of a spectral range usually for the purposes of attributing data collected from a sensor. Different satellite instruments measure the central

wavelength for each band differently. For reference, NASA and ESA use different algorithms to measure the CW of Landsat 8 and Sentinel-2. For Landsat 8, the center wavelength is calculated using the full width at half maximum (FWHM) method, which essentially uses the average from a large percent of the centered distribution. The ”lower and upper” values are FWHM boundaries. On the contrary, center wavelength values in the case of S-2, are calculated using the average derived from the metadata files for each satellite [5]. These metadata indicate the minimum, maximum, and central values for each band. Based on relative research [6], it seems that ESA uses a weighted average.

A **Spectral Index** is a mathematical equation that is applied on the various spectral bands of an image per pixel and is calculated using the band's values. The selected bands differ based on the index we want to calculate, however the majority of them are computed using the normalized difference formula 2.1:

$$(Bx - By)/(Bx + By) \tag{2.1}$$

In practical terms, it is the difference between two selected bands normalized by their sum. This methods minimizes the effects of illumination from shadows, clouds while also enhancing the spectral features that are not initially visible. There is a great variety of spectral indices used for different tasks. The Normalized Difference Vegetation Index (NDVI), and the Normalized Difference Water Index (NDWI) are some common spectral indices, which as their names suggest, are used to monitor vegetation health and the existence of water molecules respectively.

2.1.5 Brief Overview of major satellite missions and their derived RS products

In this section we will introduce some of the major Satellite missions across the United States and Europe. We will also introduce latency and processing levels related to data products and describe their use cases in remote sensing applications.

There are 77 government space agencies operating as of 2022 , 6 of which have launch capabilities, including: National Aeronautics and Space Administration (NASA) of the US, the European Space Agency (ESA), the China National Space Administration (CNSA), the In-

[5] https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/data-formats

[6] https://www.gisagmaps.com/landsat-8-sentinel-2-bands/

dian Space Research Organization (ISRO), the Japan Aerospace Exploration Agency (JAXA), and the Russian Federal Space Agency (RFSA or Roscosmos). [7]

The **United States** has three major federal agencies involved in the Earth Observation satellites: NASA, NOAA, and US Geological Survey (USGS).

The **USGS**, manages the Landsat missions which have provided long term land surface observations at fine spatial resolutions (i.e Landsat 8 offers a 30m spatial resolution). Since July 23,1972 that the Landsat 1 was launched till 2021 with the successful launch of Landsat 9 featuring the ETM+ sensor, the program has been observing the Earth continuously for half a century. ETM+, a multispectral scanning radiometer, provides images of the Earth's surface with a spatial Resolution of 30 m (60 m – thermal, 15-m pan). Currently, every two weeks or so, the Landsat satellites scan the whole surface of the Earth with a 30-meter resolution, including atmospherically corrected multispectral and thermal data and have been used widely in remote sensing for: shoreline mapping, forest monitoring, disaster management and precision agriculture to name a few.

NOAA operational satellite system for environmental monitoring consists of both geostationary and polar-orbiting satellites. The Geostationary Operational Environmental Satellite (GOES) server is mainly used for national, regional, short-range warning, and "nowcasting", while the polar-orbiting ones, such as Polar Operational Environmental Satellites (POES) and Suomi National Polar-orbiting Partnership (S-NPP), are used for long-term forecasting and environmental monitoring on a global scale. Both types of satellites contribute greatly to the global weather monitoring.

The Visible Infrared Imaging Radiometer Suite (VIIRS), one of the key sensors of the Suomi-Npp satellite, has been actively used for fire monitoring, urban expansion and economic development monitoring. It is a scanner radiometer that measures Earth radiation on the surface and atmosphere levels, in the visible and infrared spectra. VIIRS offers different spatial resolutions among the data that it collects from 22 different spectral bands, with 750m and 375m at Nadir. Its Day/Night band (DNB), is ultra sensitive to lowlight conditions and enables the generation of quality nighttime products with substantial improvements compared to older systems (DMSP/OSL).[8]

[7] https://en.wikipedia.org/w/index.php?title=Special:CiteThisPage&page=List_of_government_space_agencies&id=1079191100&wpFormIdentifier=titleform

[8] https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/

On the other hand, **Europe** features the EUMETSAT Polar System that consists of Metop, three polar orbiting meteorological satellites that go around the globe via the poles from an altitude of 817 km and continuously collect data. The satellites carry eight main instruments and their collected data are essential for climate monitoring and up to 10 days weather forecasting [9]. Moreover, The European Space Agency (ESA) has operated Cryostat and SMOs. CryoSat uses a SAR Interferometric Radar Altimeter, specially tuned to measure the thickness of polar sea ice and monitors ice sheets' changes.[10]. Soil Moisture and Ocean Salinity (SMOS) with a Microwave Imaging Radiometer with Aperture Synthesis (MIRAS) as its payload, is the first mission to provide global sea salinity and soil moisture observations. [11]

Furthermore, ESA in partnership with the European Commission are operating the Copernicus Program. This program includes the development of the sentinel satellites, a constellation responsible for various satellite missions of different purposes, such as Sentinel-1A and 1B, Sentinel-2A and 2B, Sentinel-3 and Sentinel-5p. For example, Sentinel-5p provides data for the quality of air while Senttinel-3 is responsible for climate and environmental monitoring. For revisit and coverage purposes, each Sentinel mission is built on a constellation of two satellites offering higher temporal frequency. In the case of Sentinel-2, for example, S2-A and S2-B offer a combined 5-day revisit time.

Apart from the the various successful government-launched satellite missions, many private organizations are also innovating the remote sensing space offering very high spatial and temporal resolution imagery.

Planet Labs is an American based company specializing in public Earth imaging that aims to provide daily monitoring of the entirety of Earth and pinpoint trends [5].

They design and manufacture Doves, which are Triple-CubeSat miniature satellites that get into orbit as payloads on other rocket launch missions and are equipped with high performance devices (i.e telescopes and cameras) that capture different swaths of Earth sending high quality data to a ground station. Dove collected images, provide information for climate monitoring, precision agriculture, urban planning, can be accessed online and sometimes fall under open data access policy. As of 2022 they have roughly 200 satellites in orbit offering services crucial for disaster management and decision making in general.

---

products/VNP46A1/

[9] https://www.eumetsat.int/our-satellites/metop-series

[10] https://www.esa.int/Applications/Observing_the_Earth/FutureEO/CryoSat

[11] https://earth.esa.int/eogateway/missions/smos

Airbus Defence and Space, a subsidiary of Airbus, is another company that is responsible for defense and aerospace products and services. They have launched more than fifty satellites such as TerraSAR-X NG, featuring the X-band radar sensor, an instrument that allows the acquisition of images with different resolutions, swath widths and polarisations, offering geometric accuracy unmatched by other spaceborne sensors. [12] Another satellite constellation, Pleiades, has been used successfully in remote sensing applications in fields like cartography, geological prospecting, agriculture and civil protection. Pleiades features the High Resolution Imager that delivers very high optical resolutions of 0.5m making it an ideal data source for civil and military projects. The combined effort from both the public and private sectors, has contributed greatly to the construction of satellites and remote sensing instruments with increased spatial, temporal and spectral resolution.

The data that are being collected from all these satellite missions can be distributed raw or after some basic preprocessing. Depending on the speed at which they become available to users they can be split into different categories. Since the terminology regarding **data latency** varies between the various earth science data providers (e.g NASA,NOAA etc), in the context of this paper we will use the one provided by NASA.2.1 One of the key differences between standard data products and Near Real Time and is that the latter make use of predictive orbit information for geolocation. Furthermore, the NRT processing algorithm can make use of ancillary data from other sources whose accuracy may vary. The Standard products, on the other hand, are processed utilizing precise geolocation and instrument calibration, and as a result, they offer a reliable, internally consistent record of Earth's geophysical characteristics that can aid scientific investigation. Even though NRT products may include information that makes their analysis' harder, they can be very important in applications such as Flood Disaster Response [6] and forecasting forest fire danger conditions [7].

Another aspect that differentiates the various data products is their **processing level**. NASA processes their data products at various levels, annotated in the range of 0 to 4. 2.2

[12] https://earth.esa.int/eogateway/missions/terrasar-x-and-tandem-x

Term	Latency
Real-time	Less than 1 hour
Near real-time (NRT)	1-3 hours
Low latency:	3-24 hours
Expedited :	1-4 days
Standard routine processing	Generally, 8 – 40 hours but up to 2 months for some higher-level products

Table 2.1: NASA Data Products Latency

Level 0	Reconstructed and unprocessed data from instruments and payload at full resolution. These data don't include artifacts such as communications headers or duplicates)
Level 1A	Reconstructed and unprocessed data from instruments at full resolution. These include time references and annotations with supplementary information (e.g georeferencing parameters)
Level 1B	L1A data processed to sensor units.
Level 1C	L1B data that include spectral descriptions to be evaluated by users.
Level 2	Derived geophysical variables at the same resolution and location as L1 source data.
Level 2A	Contains geolocation derived information describing the intercepted surface (e.g ground elevation).
Level 2B	L2A data that have been processed to sensor units.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
Level 3A	Usually periodic summaries of L2 products.
Level 4	Model output or results from analyses of lower-level data (e.g., variables derived from multiple measurements).

Table 2.2: NASA EOSDIS data products processing levels

Chapter 3

Introduction to Machine Learning

3.1 Machine Learning

Machine learning (ML), which is a subset of artificial intelligence, is the study of computer algorithms that can learn and develop on their own with medium to none human intervention. It seeks the automation of learning meaningful relationships and patterns from examples and observations. In other words, it describes the ability of a system to learn from problem-specific training data to automate the process of analytical model building in order to solve a task [8]. A machine learning pipeline consists of many stages: Data Collection, Data Preparation, Choosing a model, Training, Evaluation, Parameter Tuning and finally Predict.

3.1.1 Machine Learning Categories

Machine Learning can be divided into: Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning, depending on ”how they learn” from data [9].

- **Supervised Learning:** algorithms take labeled data as input and use them in their training process. Then based on these data a function is inferred, and is used in order to map new and unseen data. Supervised Learning is commonly used for Regression and Classification tasks and some algorithms include, Support Vector Machines (SVM), Linear and Logistic Regression and K Nearest Neighbors (kNN).

- **Unsupervised Learning:** is used when we don’t have labeled data. It discovers hidden patterns as probability densities or groups within the given dataset without human intervention. These models are utilized for tasks, such as Clustering, Association Mining,

Dimensionality Reduction and Anomaly Detection.

- **Semi Supervised:** is a mixture of the above methods, since it combines small labeled datasets (Supervised) and larger unlabeled ones (Unsupervised). The model evaluates other data based on its own interpretation of the data relationships, even though the algorithm takes mostly labeled training data as input. Fraud Detection, Data Labelling are some common semi supervised tasks.

- **Reinforcement Learning:** is an ML technique where the model learns by trial and error using feedback from its own actions. While the algorithm is programmed using positive or negative indications to fulfill a task, it may also determine on its own what steps to take during the process. Compared to supervised learning, RL ”rewards” and ”punishes” for positive and negative outcomes, instead of providing feedback to the agent in the way of correct actions for performing a task. Compared to unsupervised learning, it differs in terms of goals by finding a suitable action model that would maximize the total cumulative reward of the agent. It is commonly employed in resource management, personalized recommendations and robotics, with SARSA - Lambda and DQN being some the common algorithms.

3.1.2 Machine Learning Model Optimization

The performance of a machine learning model depends on the identification and selection of a suitable set of **hyperparameters**. Hyperparameters are parameters that control a model's learning process, acting like settings of the algorithm we want to implement and that when adjusted can optimize its performance. These settings differ from the usual model parameters that are learned during the training phase (e.g node weights), as they require manual tweaking. For example, Random forest hyperparameters are the variables and thresholds used to divide each tree node learned during training Scikit-learn provides default hyperparameters that may or may not be suitable for every problem. The process of finding hyperparameters for ML model optimization is called hyperparameter tuning/optimization, where we find a tuple of optimal parameters using a specified metric (e.g mean absolute percentage error). However, hyperparameter tuning can lead to overfitting, a statistical modeling error that happens when a function is too closely matched to a small number of data points. Because of this, while the model is able to work efficiently with the original dataset, it is unable to predict values

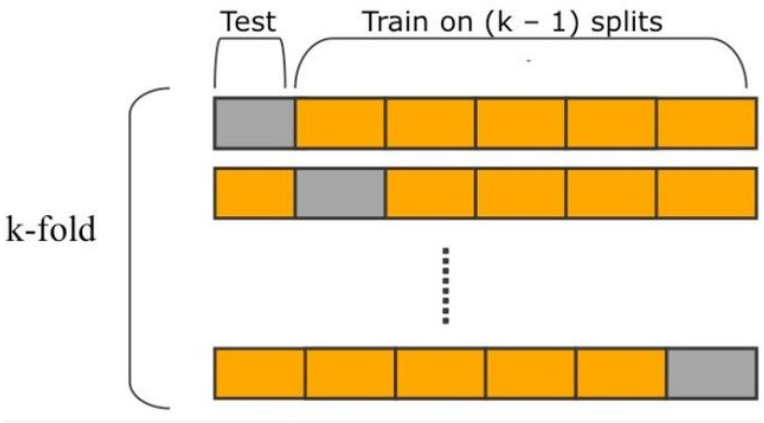

Figure 3.1: K-fold Cross Validation, source: [3]

for additional data sets. In order to avoid this the concept of cross validation should be also introduced.

Cross-Validation is a another statistical method that compares and evaluates algorithms by segmenting the given data into two parts: one used for model training and one for model evaluation. Typically, in cross-validation, these sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic iteration of cross-validation is k-fold cross-validation, where after the data are partitioned into k equally sized segments/folds, k iterations are performed for both the training and validation. These iterations are done in a way that allows a different part of the data to be held for validation while the rest k-1 folds are used for learning [10], 3.1.

Grid search, an exhaustive search over a manually chosen subset of the hyper - parameter space of a learning algorithm, has been the conventional method of executing hyperparameter optimization. In order to perform grid search, a performance metric, commonly determined by cross-validation on the training set or assessment on a hold-out validation set, is required.

3.1.3 General Issues Regarding Satellite Data Products

While satellite data have been widely accessible for a few years now and there are satellite derived products (i.e NASA Black Marble) that attempt to assist the monitoring of natural and human driven phenomena on earth such as urbanization, the analysis of said data is not an easy task. First of all, the huge amount of daily produced raw data except for being computationally demanding to process, are extremely hard for a human to evaluate and analyze in time and are also prone to human errors and bias. Thanks to the advancements of machine learning

and the advent of deep neural networks experts have been able to efficiently leverage satellite data for remote sensing. Machine learning algorithms reduce the need for human intervention considerably and valuable information about Earth phenomena can be extracted fairly easily. The combination of RS and ML can thus automate the analysis of satellite data, better use our data resources and also improve ”predictions” by discovering insights from these complex datasets. Some applications include environmental monitoring, disaster forecasting (e.g floods, wildfires), precision agriculture and even frameworks for economic assessment of a particular region.

However, wider implementation of this is often hindered by the availability of Analysis Ready Data (ARD), i.e processed data that reduce additional effort, otherwise required, from users that want to use them for analysis. In order to perform satellite data analytics on a large scale we rely heavily on the access to geometrically and radiometrically consistent observations, that do not include clouds or are of poor quality [11]. Fortunately, the USGS has compiled an archive of ARD for various Landsat instruments (e.g Landsat 7 EMT+, Landsat 9 OLI and TIRS etc), significantly reducing the data pre-processing burden on users. Also, it enables the user of Landsat products to create Land Use Land Cover Maps or perform change detection. The ARD are offered as tiled, georeferenced, top of atmosphere(TOA) and atmospherically corrected products. They are defined in common equal area projections, have quality assessment information and metadata that allow for further processing and retention of data lineage traceability. On the other hand, Sentinel ARD products are not available through the Copernicus Open Access Hub, limiting their usage and requiring different methodologies to take full benefit of them [12]. Additionally, the difficulties of merging various products, which are sometimes developed without taking into account, for example, mixed sensor use-cases, hinders applications that could use several observational streams. The lack of analysis ready data, introduces a level of uncertainty since the user has to pre-process the satellite data themselves, it requires geoprocessing expertise, which is not at all common, it is time consuming and requires additional computational power and storage to prepare. The lack of a validated dataset defeats the whole purpose of using machine learning in the first place, since if low quality data are fed into the model, the results will also lack the proper quality to assist decision making.

Chapter 4

Literature Review

4.1 State of the Art Research and Products for Satellite Remote Sensing Applications

In this section we will discuss some of the machine learning techniques that are being used in order to leverage remote sensing satellite data in a variety of industries. While satellite data have found application in a plethora of different fields we will focus on the ones of which the success is highly correlated with the Greek economy, such as the tourism and agriculture industries. Applications related to Nighttime lights and urbanization which are good proxies of economy will also be presented.

4.1.1 Scientific Literature

Satellite Data-driven Agriculture

Satellite images have been widely employed in **agriculture** since the emergence of space-borne remote sensing technologies. As discussed above, there has been huge progress in remote sensing enabling substantial spatial resolution, temporal frequency, and spectral availability of satellites. Currently, experts of the agricultural field are committed to the evolution of traditional agriculture into precision agriculture (PA), which is driven by data. Deep Learning (DL) has been employed extensively in tasks regarding identification, classification, detection, quantification, and prediction. At the field size, it can be used for predicting crop yield, and for land cover mapping and crop identification at the land scale.

One of the main goals of Precision Agriculture is maximizing crop yield while minimizing

the costs without harming the ecosystem in the process. Neural Networks architectures, such as Convolutional Neural Networks (CNNs) and more recently techniques including transfer learning make up for a powerful tool that makes the crop production prediction from RGB or spectral images in real time possible [13]. Transfer Learning is really important since it can be applied in order to transfer information gained from tasks where there's abundance of large labeled datasets, to others tasks where training data are scarce [14]. Particularly when the domains share a lot of similarities, TL offers a quick and affordable solution to the training data scarcity. The percentage of Greek land area used for agriculture was reported at 47.35% in 2018, according to the World Bank collection of development indicators, making it one of the most important sectors of the country's economy. According to the data collected by the Hellenic Statistic the main products of the Greek agriculture industry are: olives, cereals, fodder plants, industrial plants and vines. [1], [2]

The ability of CNN based frameworks and their capability to classify multispectral remote sensing data from SAT4 and SAT6 datasets was evaluated by Papadomanolaki M. et al.[15]. The proposed architectures achieved a classification accuracy of over 99% indicating the potential of deep layer architectures for designing operational remote sensing classification tools back in 2016. Since then, the rapid advances in network architectures and the computational power made the use of more deep layers feasible. Moreover, improvements in benchmark frameworks manage to eliminate the effects of overfitting (i.e the inabillity of a model to predict 'unseen' data) and to increase classification accuracy. One such example are Residual Neural Networks. Using raw imagery as input to the DNNs allows the model to learn from a wide range of nonlinear and complex features, possibly avoiding the time consuming feature extraction procedure, as well as reducing bias that can be generated during the feature selection process. It also alleviates the need for domain expertise related to image preprocessing.

Sagan V. et al. [16] WorldView-3 and PlanetScope derived raw multispectral and temporal satellite data in order to predict corn and soybean yield. Specifically 4 and 25 sets of WorldView-3 and PlanetScope cloud-free images respectively were acquired and fed into a 2D and a 3D model that explained roughly 90% variance in field-scale yield. The 3D CNN performed better, accuracy wise on PlanetScope data than the 2D CNN due to its ability to leverage the temporal features extracted from said data. The study suggests that food security

[1] https://www.statistics.gr/en/statistics/-/publication/SPG51/-

[2] https://www.statistics.gr/en/statistics/-/publication/SPG63/-

can be significantly increased by the utilization of Very High Resolution satellite data, as they offer real-time and efficient agricultural management.

However, the majority of freely available data originate from satellites with medium spatial and spectral resolution and may not be suitable for some scenarios of classification and monitoring [17],[18]. Sharma et.al. [19] developed a deep neural network model to predict the output of wheat crops using MODIS data. Modis has spectral bands with spatial resolution of 250m, 500m and 1000m, which are significantly coarser than PS or Worldview-3. The proposed CNN-LSTM model works without the need for dimensionality reduction or feature extraction due to its ability to ”digest” raw satellite images So how does higher resolution satellite images affect our estimation results?

Roznik et al. [20] investigated the accuracy increase of country level crop yield estimation using satellite imagery with high spatial resolution. The objective of their study was to quantify the accuracy boost that could be achieved by using NDVI derived from higher resolution images and masked specifically for cropland. They collected NDVI data of three different resolutions (250m, 500 m, 1km) different U.S states over an 11 year period, in order to monitor soybeans, corn, spring and winter wheat. They built regression models for each crop type which showed improved R-squared scores as the spatial resolution increased, confirming the more accurate yield estimates provided by higher resolution imagery.

Other studies [21], [22], [23] compared freely accessible satellite data(i.e Sentinel-2, Landsat-8) to paid ones (i.e Wordview) in different applications. Mudeneri et al. [21] compared and evaluated Sentinel-2 satellite vegetation and spectral data with that of dove nanosatellites such as Planetscope in detection and mapping of Striga hernonthica weed. While PS performed 5% better with a 92% accuracy in detecting the weed, the study demonstrates the ability of S2 instruments to provide near real time field level detection, a task that was quite difficult with previous multispectral sensors.

It is also worth mentioning that image quality acquired from multispectral, hyperspectral and RGB cameras depends on weather conditions making the Synthetic aperture radar (SAR) images an essential tool for RS in agriculture. However, backscatter noise for vegetation dynamics often results in difficulties regarding image interpretation. Deep Learning frameworks for object detection have been proposed in order to overcome this problem [24].

Zhao, W. Z. et al. [25] suggested an improved MCNN-Seq model to forecast optical time series using SAR data even when optical data are not available. They used Sentinel-1 SAR

and Sentinel-2 optical imagery collected over a period of two months each. However, the coarse spatial resolution and low temporal resolution of satellite imagery, makes it difficult to obtain multitemporal, large-volume, and high-quality datasets. This impedes the application of deep learning in low resolution satellite-borne remote sensing, making them insufficient for small scaled detailed observations.

Greece is among the top 10 European Countries, holding the 6th place, in wine production [3], [4]. Thus, grape yield prediction and vineyard monitoring are important in order to maintain product quality while not disrupting the supply chain. They provide farmers with the ability to better manage their field and obtain higher income. Yield prediction maps also allows them to view spatial variations across their field and determine the best harvesting time and marketing strategy, which are greatly affected by the grape's growth stages. Different methods are employed to estimate yield; nevertheless, because of constraints related to time and labor large-scale estimation is problematic. Machine learning and satellite remote sensing have the ability to provide quick and accurate assessments over wide areas for less money and in less time.

Tassopoulos et al. [26] conducted a study to monitor vine growth, in a Protected Designation of Origin (PDO) zone using freely available and high temporal resolution Sentinel-2 imagery. They selected 27 vineyards for their study and they calculated vegetation indices (i.e NDVI, EVI) for each one. The results indicated high negative correlation with the elevation topographic parameter during the vines' flowering stage. The performed ANOVA between the vegetation indices of each sub-region also showed that they have statistically significant differences, with most of them being able to detect the fruit at the flowering and harvest stage, and only NDVI and Red-Edge band Vis during veraison period (the onset of the ripening of the grapes). These data proved to be useful at monitoring at regional scale since S2 imagery captured all vineyards at the same time and in the same atmospheric conditions.

Arab et al. [27] developed machine learning models for yield prediction using vegetation index time-series. They leveraged Landsat 8 surface reflectance products within the 2017-2019 and built a regression analysis model as to map NDVI, LAI and NDWI. The exponential smoothing methods and moving averages used on satellite images detected different growth

[3] https://agridata.ec.europa.eu/extensions/DashboardWine/WineProduction.html

[4] https://ec.europa.eu/info/food-farming-fisheries/plants-and-plant-products/plant-products/wine

stages. The indices were highly correlated at the time that canopy expansion reached its maximum. The ANN approach indicated the superiority of NDVI, which had the highest accuracy across all years with R equal to 0.94, 0.95 and 0.92 respectively. The models were evaluated with ground truth yield datasets. The results of this research showcased that Landsat vegetation indicators can be used to calculate site-specific vineyard management and forecast yields. These studies suggest that machine learning integration with freely accessible and of high resolution satellite time-series data from Sentinel and Landsat can achieve reliable grape yield prediction models. The multispectral data of these satellites are robust, with a temporal frequency suitable for monitoring and assessing vine growth. These integrated models could be used for logistics and decision-making regarding table grape production. According to the Hellenic Statistical Agency, cereals are also on of the most produced crops in Greece.

Zhao Y. et al. [28] examined the capability of Sentinel-2 to infer field size dryland wheat yields, as well as how using a simulated crop water stress index could improve predictability (SI). The S2 derived VIs observed from 103 study fields over the period between 2016 and 2017 cropping seasons explained approximately 70% of variance, showing fairly high accuracy in predicting yield. The best model with RMSE = 0.54 t/ha featured a combination of OSAVI, CI and SI. This research shows how merging Sentinel-2 spectral indices with crop model ones may be used to create a more accurate yield forecasting model for fields with varying climate conditions.

Land related to Industrial non-food plant production has increased in recent years. Cotton specifically increased by 4.0% in 2019 compared with 2018, making it one of the most cultivated plants in Greece. Only three EU nations currently produce cotton on about 320,000 acres. With 80% of the cotton-growing land in Europe, Greece is currently the main grower. Spain's Andaluca region, with 20%, comes second and Bulgaria third with fewer than 1,000 acres. [5] Cotton is an economically important crop, being the main source for natural and sustainable fiber for textiles, but is highly susceptible to root rot. For the administration of cotton agriculture and international trade, accurate and prompt distribution monitoring are mandatory. RS can be an essential tool for the detection and portrayal of infestations of cotton root in cotton fields. Hence, it is important to be able to assess both the yield as well as monitor them of various diseases.

[5] `https://ec.europa.eu/info/food-farming-fisheries/plants-and-plant-products/plant-products/cotton_en`

Song et al. [29] used unsupervised classification in order to evaluate the potential of freely accessible satellite imagery for cotton root rot detection over methods involving aerial multispectral imagery. Sentinel-2A, although it missed some small rot patches, overall it outperformed the airborne images on both field and regional level. These findings show that images acquired through the Copernicus program may be utilized to identify cotton root rot and provide prescription maps for disease treatment at specific locations.

However, most prior studies on cotton identification using remotely sensed pictures have relied heavily on training samples, which are time-consuming and expensive to obtain. To get around this restriction, Xun et al. [30] attempted to develop a new index to identify cotton within an area of interest, termed as Cotton Mapping Index (CMI). Time series generated from S-1 SAR and S-2 MSI images were used for automatic cotton mapping and were assessed on both U.s and China locations achieving an accuracy of 81.20% for cotton classification. These findings suggest that CMI calculated from Sentinel imagery can be used for accurate cotton mapping. The advantage of the suggested index over traditional supervised classifiers such as Random Forests is that it requires no training samples and can obtain the map of cotton distribution before the harvest.

Reliable crop yield prediction at the field levels is crucial for managing difficulties and mitigating climate variability and change impacts during production. While the studies that were already presented achieved considerable results there is still a lack of accurate disaster vulnerability models that can be used to estimate yield losses and their pure insurance rate and which will ultimately assist the farmers and public sectors to plan their crops.

Urbanization: Effects on Greece and Solutions Provided by Satellite Imagery

Urbanization in developed countries is known to be accompanied by economic expansion and industrialization [31]. While urbanization is positively correlated with economic growth, the Greek urban system is characterized mainly by a growing dynamism of one or two metropolitan areas accommodating half of the country's population [6]. Moreover, lack of metropolitan governance, lack of land use regulation, lack of adequate infrastructure [32] and the unregulated urban development in general has led to multiple deprivation in the capital city of Athens and has made smaller cities unable to compete with the major ones in the majority of sectors [33]. It is thus mandatory to develop models that can assist policymak-

[6] https://urbact.eu/greece

ers in facilitating urbanization in a way that contributes to economic growth, employment growth, environmental sustainability, rather than the pursuit of speeding up the process of urbanization [34].

Akbar et al. [35] used Landsat 5 TM and Landsat 8 OLI images to investigate the spatial distribution and modelling of changes in urban landscapes as well as their economic impact. The Markov Model with NDVI masks they implemented achieved a score of 0.9 in differentiating LULC classes and correlating them with economic changes in their area of study, District Lahore.

Moreover, Chen C. et al [36] took advantage of Landsat time series characteristics and proposed a method for the extraction of economical features based on earth morphological changes due to regional economic growth. After collecting the Landsat data, they analyzed the correlation between economic indices and land use types. The proposed model showed the importance of construction land to inference Gross Domain Product of an area. Although, these studies prove the effectiveness of Landsat products to detect changes in land use and land cover, worldwide scale urbanization has brought about diverse types of urban LULC changes. These changes have have been mostly under studied with the focus of past research being urban growth [37].

Another area that has gained interest in the past years is nighttime lights and their capability of detecting changes in LULC as well as economy. Zheng Q. et al.[37] tried to fill gaps in the literature by proposing a framework using VIIRS monthly time series to characterize diverse urban land changes. They fit the VIIRS derived data to a Logistic-Harmonic model taking into account the uniqueness of urban land change and the temporal information of VIIRS time series. They produced BU area maps and disentangled the observed changes into five categories. The results show that classification based on temporal features can significantly improve the accuracy of mapping regions with heterogeneous BU and NBU landscapes and promotes temporal consistency and classification efficiency.

CH R et al. [38] tried to measure city growth using nighttime lights. After preprocessing the data with propriety techniques in order to correct blurring, saturation and compatibility issues with other satellite temporal and spatial resolutions, they developed a protocol to isolate stable nighttime lit pixels that constitute urban footprint. Their measured metropolitan area size can be used along with geo-referenced population datasets (i.e GHSL) and calculate the rate of urbanization and urban density.

Although the VIIRS instrument has outperformed the DMSP OLS nighttime lights in terms of image quality and has found extensive use in urban and economic studies, the fact that it's relatively new compared to the latter means there's still space for research and improvement.

Chen X et al. [39] proved the ability of VIIRS to estimate Cross Sectional and Gross Domain Product time series compared to NTLs derived from DMSP OLS across the US and Metropolitan Areas. VIIRS showcased better results at predicting GDP for MSAS suggesting the higher correlation with urban sectors than rural ones. This is common with the results of previous studies for DMSP OLS night lights. Additionally, VIIRS lights predict metropolitan statistical areas' GDP with higher accuracy compared with state GDP, suggesting that night lights may be related to a bigger extent to urban sectors than rural ones. The importance of taking possible biases that can impact hypothesis testing into consideration, when trying to understand socioeconomic phenomena based on nighttime lights.

Apart from the possible existence of bias in economic prediction through nighttime lights, there are also potential nonlinearities and measurement errors in the light production function. Bluhm et al. [40] studied DMSP nighttime lights for economic evaluation of small geographies across 6 counties with high statistical capacity, i.e their ability to gather, examine, and share high-quality information on their people and economy. Their results indicated the inability of nighttime lights to response to higher baseline's GDP changes, higher population densities and agricultural GDP. While changes in night luminosity do correlate with GDP changes even on small geographical spaces, the documented nonlinearity implies that some studies may be unable to identify policy-relevant effects or misinterpret this as the treatment effect of their model's variables in areas where lights don't react much to economic activity.

Urban areas can reflect, via nighttime lights, the spatial distribution of commercial activities but data collection difficulties make traditional methods unable to easily detect them. Duan X. et al. [41] proposed a method for urban commercial areas detection through the use of NTL satellite imagery. First, they preprocessed the images by setting the brightness value range between 0 and 255, a step necessary for improved cluster analysis efficiency. Then, they performed an exploratory data analysis where spatial patterns and optimal distribution characteristics were identified. Finally, after discerning hotspots through clustering analysis, they constructed standard deviation ellipses to detect the direction/ trend of the development of commercial areas. Comparing the results of their study with ground truth data, night lights

can indeed identify urban commercial areas, but the accuracy can be hindered by various factors, including weather conditions and vegetation coverage.

Gross Domain Product, enables policymakers and organizations to identify the state of the economy, i.e if it is contracting or expanding. Satellite imagery, offers the ability to estimate the gross domain product almost in real time and even in small geographical areas-compared with traditional statistical analyses- thus allowing businesses and economists to analyze the impact of changes (e.g taxes, economic shocks etc.) with relatively high precision. [7]

Tourism Intelligence through Satellite Data

Tourism has been a valuable source of revenue for the Greek economy, being one of its most important sectors. Greece has been a major tourist destination and attraction with the number of tourists increasing each year dramatically according to data from World Bank [8], hitting a record high of 33.1 mil international arrivals in 2018 [42].

Tourism generated Gross Domestic Product accounted for 6.8% of total Gross Value Added (GVA) in 2017 with 381.800 people being employed into the sector in 2018. amounting to 10% of total employment in Greece. In addition to that, Travel exports accounted for 43.3% of total service exports in 2018. This systematic growth was unfortunately hindered by the COVID-19 pandemic, however according to the provisional results of the Survey on Arrivals and Nights Spent in Hotels and Similar Establishments, conducted by the Hellenic Statistical Authority (ELSTAT), there was an 63.2% increase of arrivals and an increase of 87.0% in nights spent for the period of January – September 2021 compared with the corresponding period of 2020. [9] This increase indicates the recovery of Greek tourism after the lifting of the extreme safety measures such as the lockdown.

It is necessary to find ways to create an environment where tourism and tourism - related businesses can flourish while accounting for safety, enough accommodation and easy transportation. Tourism planning and administration can benefit from quick, affordable, and easy identification of popular tourist destinations. Using open-source, near real time data sources like social media, many studies have been conducted to examine and assess a lo-

[7] https://www.investopedia.com/articles/investing/121213/gdp-and-its-importance.asp

[8] https://data.worldbank.org/indicator/ST.INT.ARVL?end=2020&locations=GR&start=1995&view=chart

[9] https://www.statistics.gr/en/statistics

cation's tourism circumstances. Up until recently research heavily relied on annual stats of small sample size and/or on the integration of high resolution statistical datasets. These conventional big data sources ignore spatial heterogeneity and drivers of tourism demand while also being often unavailable [43], [44]. Remote sensing satellite data, offer near real time data over large scale geographies, are good indicators of economy, spatial distribution and have been used for tourism intelligence.

Tourism's spatial dispersion significantly affects both its operational effectiveness, its regional relevance and continuous satellite observation and in-depth study of night lights can pave the way to clarify human activities and socio-economic dynamics.

The ability of night light emissions to estimate the touristic activity within European countries was examined by Kirkigianni et al. [45]. The correlation of touristic activity with seasonal changes in nighttime lights satellite imagery collected between 2012 and 2013 from both DMSP-OLS and VIIRS was investigated. In order to evaluate their findings they used statistical tourism data on country level, which after proprocessing with GIS, were used in Linear and Geographically Weighted Regression. The results of their statistical tests show that there is a strong correlation between nighttime light emissions and tourist activity. The GWR has proven to be a useful tool for examining this relationship, but some additional factors should be taken into account before judging its ability and accuracy.

In a similar fashion, Wei J. et al. [46] used OLS and GWR to investigate nighttime light's seasonal changes and their relation to tourism in 112 regions of the Hunan Province, China. According to their results, the intensity of luminous radiation is highly correlated with tourism points of interest. Furthermore, spatial heterogeneity and seasonal differences of tourism activities were also observed across different regions. These findings related to social environment and resource allocation, can be helpful when studying tourism on the county level.

Devkota et al. [47] demonstrated the capability of nighttime lights and crowdsourced data (e.g OpenStreetMaps, Twitter) to detect tourism areas of interest. They generated active tweet clusters though DBSCAN clustering algorithm to identify the touristic places where essential facilites related to travelers needs were available. They then examined the adequacy of NTL remotely sensed data to recognize proper tourism areas in Nepal where the social media penetration is relatively low. They successfully detected important tourism areas in remote and urban regions with a F1 score of 0.72.

Greek tourist resorts and attractions are distributed mainly across coastal areas. The preser-

vation of these areas as well as measures to ensure safety conditions that won't hinder the tourists' accessibility are really important. Nearshore bathymetry estimation is crucial for understanding coastal processes and enabling many industries, including offshore construction, fishery and tourism among others. Common survey methods, based on monitoring via ships or airplanes, are costly and time consuming. Moreover, currently the estimation of bathymetry usually requires in-situ depth measurements in order to train inversion models, a difficult or even impossible process for many areas. Cover maps regarding seagrass and corals have been created using machine learning and Senintel-2 MSI data so as to support coastal management of small islands [48]. Apart from traditional Machine Learning techniques, deep learning techniques have been used in ocean remote sensing making the precise, efficient and intelligent mining of ocean data possible [49].

Biermann et al. [50] showcased the ability of Sentinel-2 in finding patches of plastic in Coastal Waters. Their study indicated the detectability of floating microplastics and their distinction from naturally occurring objects (i.e seaweed) in optical satellite data. They used a combination of novel spectral index, the Floating Debris Index (FDI) and Naive Bayes classifier to highlight the existence of plastic debris, with an accuracy of 86% in all their case study areas.

Basu B. et al. [51] implemented both unsupervised and supervised classification algorithms for the same task using multispectral sentinel-2 remote sensing imagery from Cyprus and Greece. Their models were built using a mix of six reflectance data bands and the NDVI and FDI indices, which were proven to be the most effective at spotting floating plastics. Their results varied depending on the algorithms they used, with Support Vector Regression having the highest accuracy. It should be noted though, that they used a small number of grids in order to train and evaluate their models, possibly hindering their performance.

Albright et al. [52] investigated the data fusion of S-2 and ICESat data for bathymetric inversion. Their results compared with data collected from an integrated lidar system called CZMIL, they had a RMSE of 0.35 m in waters with similarities in turbidity and bottom reflectivity. This demonstrates the ability of the fused imagery to estimate the depth of water of optically clear coastal waters.

Apart from the traditional, rural tourism has gained interest among native tourists. Rural tourism can support sustainable development but also combat the impeding economic growth that comes from the extreme urbanization that has plagued Greece [34]. Yang J. et

al. [53] studied the effects of rural tourism using multi source data, including satellite images. Specifically they investigated changes and their drivers in the morphology and social evolution of the countryside from a touristification perspective. Results from their study area Jinshitan, which showcased non-agricultural employment increase, support the notion that rural revitalization can be beneficial to rural communities enabling their economic growth.

Since tourists' preferences regularly shift in reaction to new risks, safety in tourist areas is of the utmost significance. Given how dangerous they can be, natural hazards must be taken into account in efforts to promote safe travel. Systems that monitor and forecast extreme natural phenomena in points of tourist and cultural interest allow for effective risk management and response.

Psaroudakis et al. [54] integrated satellite imagery and meteorological forecasts to develop an early warning and incident response system for the protection of tourists in outdoor Greece. The system includes wildfire, floods and extreme weather warning modules Their findings may be applied to the creation of further natural risk management strategies for cultural and natural heritage sites.

These studies and their outcomes for the implementation of satellite remote sensing in regions of touristic interest can provide valuable information for Greece's stakeholders. Tourism planners, policy makers as well as entities of the private sector can use satellite imagery for decision making regarding site selection, investment and generally improvement of the country's highest revenue generating economic sector.

Site Selection for Renewable Energy Platforms via Satellite Imagery

One of the goals of sustainable development is Affordable and Clean Energy (SGD7). Europe has consistently increased its renewable energy production and in 2020, it represented 22.1 % of energy consumed in the whole continent, which is almost 2 percentage points above the target of 20 %. Greece has seen an unprecedented increase in renewable energy consumption in the half past decade, with renewable energy sources shares being nearly double. According to World Bank and Eurostat data this increase went from 6.9% 2004, to 15.5 % in 2017 and 21.7% of gross final energy consumption exceeding the set goal.

The Greek energy sector, however, still largely depends on imported fossil fuels. As of 2017, almost half of its energy needs were covered by petroleum products, used in the transport sector but also converted into electricity. Greece's electrical independence relies heavily

to lignite making it one of the only 9 EU member states that still produce it endogenously and place it among the top 3 countries that still use it for electricity consumption and heat. [10] Although there has been concerns with the surge in solar plant installation including : possible instability in ecological fragile regions (i.e biodiversity loss, local climatic change and food sovereignty). In order to better monitor and prevent those problems, geo-referenced data are a necessity which is unfortunately lacking. High quality data to build accurate models for predicting the behavior of solar radiation are required for optimal solar energy systems management. The impact of the PV plants' installation can be monitored efficiently with the use of spatial distribution and dynamic data generated through Remote sensing.

The need for Greece to abide by the EU standards and the decreasing consumption of fossil fuels, the fact that the energy sector has a higher contribution to gross value added than most European countries and the generation potential of electricity due to its climate along with the govenrment support, make Greece a great place for investment and economic growth. A lot of public and private organizations have invested in solar and wind renewable energy. [11] In this section we will present some applications of data driven remote sensing for photovoltaic and wind turbine site selection as well as monitoring the environmental change detection due to the extensive use of lignite,

Narvaez et al. [55] proposed a ML methodology for site selection and solar radiation forecasting. Specifically they combined solar radiation ground truth data and satellite solar radiation data from geostationary meteorological satellites in order to obtain long-term solar information with improved spatio-temporal information (site adaptation). Then they construct a LSTM deep learning model that takes these improved data as input and makes accurate predictions regarding solar radiation over a particular region, offering almost 40% performance increase over traditional statistical methods.

Chen Z. et al. [56] used freely accessible Landsat 8 OLI images in order to identify and map the spatial distribution of Photovoltaic (PV) plants in a local and global scale. They combined spectral bands and indices for PV extraction, such as NDVI, BI and BUAI from the satellite imagery and fed them to high performant machine learning models. The XGBoost performed better in the raster-based extraction of PV plants with a 99.65% accuracy, but being

[10] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Production_of_lignite_in_the_EU_-_statistics

[11] https://www.enterprisegreece.gov.gr/en/invest-in-greece/sectors-for-growth/energy

unable to identify distributed power plants due to the images' limiting spatial resolution.

Apart from solar panel plantations, Greece has the appropriate topology for the installation of Offshore Wind Farms (OWF) and wind turbines in general for long term energy production [57], [58], Nezhad et.al [59]. The fact is that there's have been a growing interest towards the installation of OWFS since they offer available free space for large-scale construction, reduction and avoidance of environment disturbance due to noise, lights and change of topology. Greece has many uninhabited islands that could prove to be ideal for OFWs, however the research has been focused on the use of GIS software. Through observational activities and data from remote sensing, suitable locations across vast oceanic areas can be found, evaluated, and identified.

Majidi et al. [60] evaluated the potential of Sentinel 1 data to assess the wind source potential in Sardinia islands using a machine learning forecasting model. The model blends wind speed assessment, mapping, and forecasting to identify offshore and nearshore wind potential through the use of image processing methods, Adaptive Neuro-Fuzzy Inference and the Bat algorithm. Ten hotspots have been recognized as being particularly intriguing due to their high-energy potential, making them possible locations for the future installation of wind turbine generators (WTGs).

4.1.2 Existing Commercial Software

Apart from the various accomplished case studies within the research sphere, there are various companies that offer proprietary software in order to assist various businesses that try to moderzine and improve. Some of these companies are listed below:

- **Descartes Lab:** offer a plethora of business intelligence services though the use of satellite imagery and their fusion with their clients integral data. Their clients range from privately owned businesses to governments and they have solutions for mining, agriculture and others. An interesting case study of Descartes Labs is their use of Google Cloud Platform to provide accurate predictions for global food supplies and detect early warnings of famine. [12]

- **LiveEO:** Offers satellite monitoring solutions for various industries, including the monitoring of various infrastructure. They can detect changes in very high resolution

[12] https://cloud.google.com/customers/descartes-labs

images and create risk monitoring models.

- **Blackshark.ai:** is geospatial platform that combines satellite imagery and machine learning to provide insights at a global scale infrastructure. Their AI enriched methods have the ability to complement missing image attributes. They offer various enterprise solutions such as visualization, simulation and mapping, updated in real time. An interesting application off blackshark.ai is the display of their entire planet Earth in Microsoft's Flight Simulator game.

- **Iceye:** is a small and agile radar satellite constellation that provides effective change detection of any location on Earth multiple times during day and night, independently of weather conditions. IceEye has been used by the insurance industry but their SAR econsystem can benefit government agencies, environmental groups, emergency response units and companies in general.

- **AgroApps:** is a Greek company that provides crop and weather monitoring and forecasting services using ML and satellite imagery.

- **Agrotech:** similartly to AgroApps, is a Greek company, that uses satellite imagery to monitor crop growth cycle for optimal fertilizer use.

4.1.3 State of the Art Key Takeaways

Taking all of the above into consideration we can realize the importance of satellite remote sensing and machine learning. We presented various successful applications from the agriculture, tourism , energy production and urban studies domain that can greatly benefit Greece. These studies provide insights for the efficient combination of Data Science and Satellite Imagery by providing information about the many satellite data products, their benefits and limitations. Various spectral indices have been evaluated depending on the case study, including vegetation, water and imperviousness ones, with NDVI being the most commonly used across all sectors. This is probably due to its conceptual simplicity and ”tangibleness”. Moreover, nighttime light imagery even with their reduced spatial resolution can be a good estimator of economic growth and human activity in the majority of sectors. The various studies also provide methodologies for satellite data acquisition, data preprocessing techniques and machine learning model evaluation metrics. Through this, albeit not extensive, litera-

ture review, we can derive the growing interest in the use of Landsat and Sentinel products. These images are of high to medium spatial resolution and compared with private endeavors can be limiting in cases where extremely high precision is required (e.g military). On the other hand, they are free even for commercial usage and hence easily accessible by most businesses and organizations. Their accessibility and interoperability, that was showcased via their combination in several researches, coupled with their proved ability to estimate economic trends, detect and forecast changes on Earth's landscape as well as human activity make them ideal for our software. Additionally, the increasing use of Deep Learning Methods as well as re-thinking of older machine learning methods (e.g Random Forests) in the remote sensing sphere is is another main point of existing literature. Both approaches have their advantages and disadvantages (e.g computational constraints, simplicity, robustness, scalability etc) but both can automate several procedures while offering knowledge for decision making in the public and private sector. In the case of our approach, taking these characteristics into account, we chose to implement shallow machine learning methods that don't require huge computational capabilities. This is due to the fact, that our proposed software needs to be able to execute in completely free cloud platforms that often have restrictions. Finally, as far Greece is concerned, the solutions provided by 'local' companies are, to our knowledge, extremely limited to the agricultural and weather forecasting sectors. Furthermore, and that is prevalent to all mentioned commercial products, they are "locked" behind paywalls, making them prohibitive for the majority of SMEs.

Chapter 5

System

5.1 Case Study: Detection of Abandoned Buildings

5.1.1 Problem Statement

As we discussed above, the unregulated urbanization, the financial crisis that plagued Greece for almost a decade and even the economical implications of the COVID-19 health crisis has resulted in the emergence of disused and even abandoned buildings. Abandoned buildings and unmaintained structures are a regular occurrence in urban centers. These vacant areas are responsible for:

- increased crime rate (drug use, prostitution, etc.)
- increased danger for public health and safety, since they can be prone to collapsing and fires due to deterioration
- devaluation of nearby property
- generating low property taxes; increasing costs for local governments (to secure, inspect, provide additional police and fire services, etc.)

To put it another way, abandoned buildings contribute to a decline in the city's quality of life by providing an unappealing urban landscape for residents, visitors and potential investors.

While there are a lot of ways to find vacant properties such as driving around an area of interest, reaching out to local authorities and banks and even advertising, they cannot be

automated, require the cooperation of said parties and can often be a waste of time. This can have negative results especially in competitive markets where finding the property ahead of the competition may be crucial for a successful investment. Thus, a software that will be able to automatically and accurately find a vacant property in an area of interest could prove to be quite helpful for investors as well as local or government authorities that want to alleviate the affected neighborhoods.

Of course, even businesses could utilize the software as a means for better site selection. Business sites should, among other things, be placed in locations forecasting long-term economic growth and safety and also take advantage of financial incentives such as tax credits and tax breaks. Since the existence of abandoned buildings is correlated with the above, the ability to leverage satellite data to make educated guesses when census data, such as crime rates, regional GDP and population density are either outdated or unavailable can not only give a significant advantage over others but also reduce costs.

5.1.2 Implementation

We propose a software that can provide useful information that can be used for business intelligence and policy making through the use of spatial, satellite data and machine learning. The main aspect of this software is its ability to detect abandoned buildings over an area of interest, and display it to the user over map imagery. What differentiates it with similar software [61], [62] is the fact that it uses freely available satellite imagery and explores areas where ground truth data are scarce. Another key difference is its extensibility, since new applications can be added without the need for changes in its core architecture. Hnece, the software is suitable for SMEs that don't have the expertise to implement their own solutions, but also don't have the ability to invest in commercial ones. Leveraging our platform, they can make informed decisions about optimal site selection and thus reduce the costs of outsourcing it.

Data Collection

Our methodology leverages nighttime lights data from the VIIRS instrument aboard the Suomi NPP satellite and various spectral indices derived from Sentinel 2 Surface Reflectance imagery. It has been proved that both can be used for the successful detection of urbanization population density and economic activity but it was also confirmed through our own

experimentation.

The data were collected through Google Earth Engine, a cloud platform engineered by Google, that combines a huge catalog of satellite imagery from various sources, such as NOAA, NASA and USGS. It also features preprocessed geospatial datasets that make large scale analysis possible. Scientists, researchers, and developers use Earth Engine to detect changes, map trends, and quantify differences on the Earth's surface [63]. Earth Engine is available for commercial use while remaining free for academic and research use. We chose GEE over other platforms, such as Sentinel Hub, because of its ease of use, and documentation. GEE also provides an API for easier integration with Python, which was our programming language of choice for its data science libraries. Finally, the fact that GEE is capable of executing the majority of the needed computations on the cloud (albeit some limitations) makes it ideal for the platform we envisioned, i.e one that is accessible by anyone, without the need for a powerful computational system or abundance of storage. Finally, GEE is enhanced by contributions from the open source community, with libraries such as geemap and eemont being our main tools of use.

Using the eemont and Awesome Spectral Indices libraries, we calculated time series for: the average radiance from the VIIRS DNB collection, NDVI, NDBI, EMBI, PISI, VgNIRBI, VrNIRBI spectral indices. We chose these indices because they have been used extensively in applications where the evaluation of vegetation and Impervious surface over an area of interest was important. Eemont extends the original GEE capabilities by adding various automations such as histogram matching for data fusion, panchromatic sharpening and of course cloud masking.

Unfortunately there was no way to extract ground truth data regarding abandoned buildings for our study area, Volos, Greece. Ground truth data are essential in order to train and evaluate a machine learning model's performance. In order to combat that we reached two solutions. The first one involved the gathering and use of data from other regions. The second methodology focused on the acquisition of crowd-sourced data for not abandoned buildings and a really small hand labeled dataset for vacant ones. Both methodologies follow a similar structure based on the existing literature where, house and land vacancy in general have been found to be correlated with increased vegetation [61] as well as low nighttime lights' radiance [64].

For our first method, we selected datasets from the city of Chicago that had ground truth

labels, as well as the exact geometry features of each building. Specifically the Chicago dataset was pulled from the city's data portal and includes requests (such as water quality reports, illegal construction etc.) and is updated daily since 12/18/2018. This dataset contained requests for abandoned buildings and vacant land, however there was not a feature for not abanadoned buildings, so we assumed that requests related to water quality and illegal building were indications of their existence.

For the second method, we got data for non vacant buildings from OpenStreetMaps using the Turbo Overpass API. These data include spatial information for residential areas, public facilities (e.g schools) and amenities (e.g cafes). As for the abandoned buildings' data, we hand crafted a small dataset by searching through the city of Volos and generating their coordinates though geojson.io.

Preprocessing

The preprocessing pipeline consists of cloud masking, scaling the images retrieved via GEE, creating monthly image composites and finally combing our retrieved time series into one single dataset, ready to be fed into the classifiers.

Cloud Masking is a vital preprocessing step in any geospatial analysis application [1] since clouds captured in satellite imagery can interfere with the results of our analysis. In order to solve this problem we chose the Sentinel-2 Surface Reflectance product, which offers information about the cloud probability within an image. It's a collection of cloud probability images where for every image in the sentinel-2 archive a cloud probability per pixel at 10m scale was calculated, a joint effort between Sentinel Hub and Google. This provides a flexible method to mask cloudy pixels in order to create composites ready for classification tasks. While Sentinel-2 already offered the Quality Assurance band (QA60), a binary classifier for thick and cirrus clouds, the new algorithm called s2cloudless offers the ability to fine tune the cloud masking procedure by choosing a probability threshold between 0 and 100.

In order to perform cloud masking for our Sentinel 2 images, we used the a method provided by eemont[2]. We used the default option to filter all the images with more than 60% cloud probability, a moderate threshold that captures the majority of cloudy pixels while not removing clear pixels from the images

[1] https://medium.com/google-earth/more-accurate-and-flexible-cloud-masking-for-sentinel-2-images-766897a9ba5f

[2] https://eemont.readthedocs.io/en/0.1.7/guide/maskingClouds.html

The second preprocessing step involves the **Scale and Offset** operations on the GEE images. Most images in Google Earth Engine are scaled in order to fit into the integer datatype. In order to get the original values we multiplied the retrieved scalars. While the scaling procedure changes based on the bands and for the supported platforms (e.g Landsat, Sentinel etc.), an eemont method automates the scaling for all supported bands.[3]

Another important part of preprocessing is the **creation of Monthly Composites** Sentinel 2 SR has temporal resolution of 1 image every 5 days for the same region. On the other hand, VIIRS Monthly Composites are Monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB).In order to be able to match these products, we used a library called wxee, which also extends GEE's capabilities. Specifically, we generated time series for our selected spectral indices by performing a temporal aggregation from almost daily to monthly frequencies.

Machine Learning Models

We used two machine learning models for the detection of abandoned buildings, Random Forest and One Class Support Vector Machine depending on the aforementioned approaches. We constructed various classifiers provided by the sklearn python library in order to evaluate the effect of different band combinations as well as different parameter values on the final performance.

Of course, picking the indices manually is an exhausting and probably redundant process, so we created the rest of the combinations based on the results of their correlation matrix 5.1. High correlated features are more linearly dependant and affect the dependent variable virtually equally. Hence, when two features have high correlation, we can omit one of the two. While there are various techniques like dimensionality reduction, we prefer to select the features manually based on our intuition, regarding both their usage and correlation. In our case we disregarded the IBI index since it contained outliers as well as features that had an absolute value greater than 50.

Random forests are an ensemble based learning method that has found extensive application in different domains including Remote Sensing [65], [21], [66]. Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled.

[3] `https://eemont.readthedocs.io/en/0.2.0/guide/imageScaling.html`

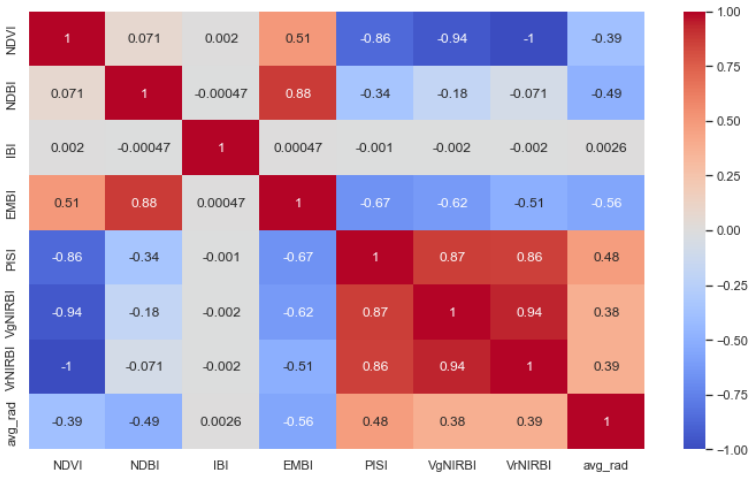

Figure 5.1: Correlation Matrix For Index Time Series generated for Chicago

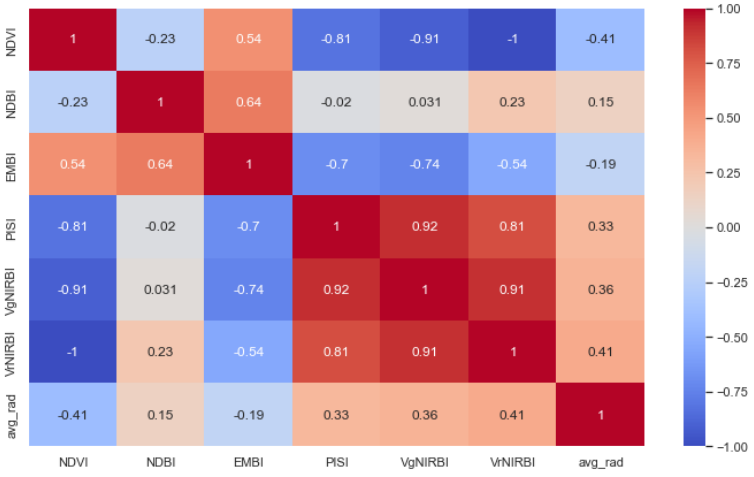

Figure 5.2: Correlation Matrix For Index Time Series generated for Volos

Features	Parameters (max depth, no of estimators)	RF Accuracy
NDVI,NDBI	20,400	0.663%
NDVI,NDBI,Average Radiance	20,350	0.6805%
VgNIRBI,Average Radiance	20,600	0.666%

Table 5.1: Selected Feature Combinations for Random Forest Experiments

The random forest algorithm is a combination of decision tree predictors, comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. One-third of training sample is used as test data (out-of-bag sample). Another instance of randomization is then injected into the dataset using feature bagging which increases diversity and decreases the correlation among decision trees. The forecast determination will differ depending on the type of task (i.e regression, classification). Individual decision trees will be averaged for a regression task, and a majority vote—i.e. the most common categorical variable—will produce the predicted class for a classification problem. Finally, the oob sample is used for cross-validation, which finalizes the prediction process. Because the averaging of uncorrelated trees reduces variation and prediction error, random forests lessen the danger of overfitting. Furthermore, it is adaptable because it can accommodate missing information and can be utilized in both classification and regression issues.

As we mentioned hyperparameter optimization is an essential task in machine learning, so we tried different combinations that would deliver the highest possible accuracy. Firstly, for each feature combination we tested the number of tree estimators from 50 to 600 at 50-tree intervals, and the depth parameter from 10 to 100 at increments of 10. The minimum number of data points placed in a node before the split, the minimum number of data points allowed in a leaf node as well as the bootstrap method were kept the same during the whole optimization. For example, in the case of using all available features to build the classifier, increasing the number of estimators resulted in the model's accuracy increasing till we reached the 450 threshold, and started decreasing slightly after the 500 mark. At the same time increasing the maximum number of levels in each decision tree increased the accuracy by a little till we reached a maximum of 0.697% at a max depth of 30 before the results stated fluctuating at smaller values.

A similar process was followed for the tuning of the other classifiers. The results are summarized in 5.1.

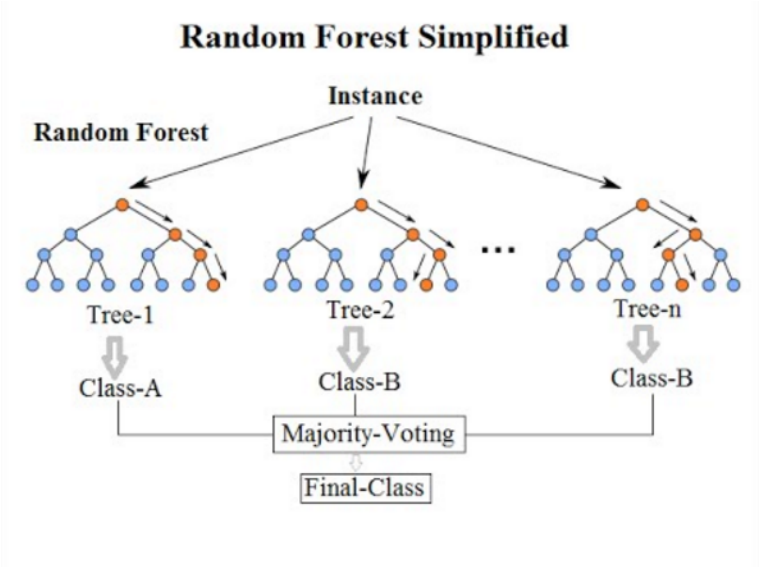

Figure 5.3: Random Forest Simplified Diagram, source: Wikipedia

Apart from the Random Forest method, we implemented an algorithm introduced by [67] and called **One class Support Vector Machine**. This technique is used in order to perform classification when the negative class is either absent, poorly sampled, or poorly characterized. The method solves this problem by defining a class boundary just with the knowledge of the provided positive class. This technique has found application in many fields, with concept learning and outlier detection being some of them [68].

In our case, the scarcity of abandoned buildings for our study area prevents us from using the aforementioned random forest classifier. In this context, and since data for non vacant buildings are available, we chose to use OCSVM. We set the not abandoned buildings as the positive class and assumed that buildings that didn't fulfill all the criteria set by the OSVM would be classified as negative, in other words abandoned. The model was trained on various combinations of spectral indices and NTLs with its accuracy varying.

In order to evaluate the performance of OCSVM we constructed a hand labeled geo-referenced dataset for abandoned and not abandoned buildings across the Municipality of Volos. The dataset consists of the buildings geometry which were extracted from Open-StreetMaps using the Overpass API or manually where exact coordinates were not available.

Similarly with the Random Forest approach, we performed some statistical tests on our selected parameters. Specifically we used the spearman and pearson coefficients to check if their correlation was significant at a 0.05 significance level (a=0.05). The Null Hypothesis, that the correlation coefficient was significantly different from 0 (no correlation), was rejected

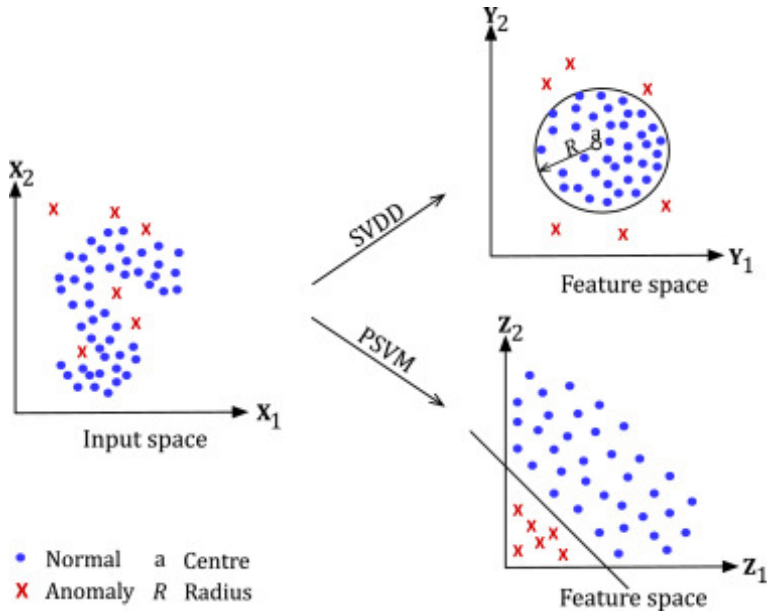

Figure 5.4: One Class Support Vector Machine source: [4]

in all index combinations, so their correlation matrix was also computed 5.2 in order to choose the ones that had a correlation less than 0.5.

5.1.3 Experiments

Before proceeding with our model experimentation, we wanted to validate the reaction of nighttime lights to human activity. Specifically we generated time series for the T. Oikonomaki street of Volos, which shelters mostly cafes and nightlife amenities. The data were collected for the period, between 2015 and and 2022 and evaluated their value changes in relation with various events. A.2 For example, the rise around the end of 2018 is connected with the increasing number of bars in the area, while the drop in 2020 is related to the COVID-19 pandemic and the lockdowns. This is also emphasized by the fact that we can see fluctuations after 2021, around the times that the lockdowns were more relaxed.

Apart from that we also created some map visualizations for the whole region of Magnisia, for 2016 and 2021. The average radiance values were normalized by subtracting the mean values and dividing the results by the variance in order to make the relatively low resolution viirs images more discernible. Although difficult to notice in the provided images A.3, A.4, there was an increase in nighttime lights between these four years possibly related to urban expansion and increasing human activity. Some lights became invisible in areas such as Trikeri, but that is probably an artifact of the normalization.

The ability of our methods to detect abandoned buildings was evaluated through the platform we have created. The experimentation followed a set number of steps. For each experiment, we selected areas with different characteristics, such as dense urban core, residential areas near city outskirts and coastal areas within the city of Volos. These areas were relatively big, containing roughly 2/3 of the whole city. Each region of interest was split into smaller equally sized polygons (grids), using a fishnet method. This method, takes as input the area of interest and splits it, according to user input, into n columns and k rows. As their number increases, the size of polygons decreases so that we can fit more inside the initial area. We experimented using different sizes, varying between 5x5, 15x15, 25x25. Afterwords, we evaluated the capability of different classifiers to detect a buildings vacancy property on the set grids. In some interesting cases, where the models seemed to perform worse due to their placing inside the polygon, we evaluated them in smaller respective subareas. The results varied greatly depending on the model, the grid size and the selected features.

Random Forest Experiments

Using the constructed classifiers for the various feature combinations we evaluated the ability of data from metropolitan areas of foreign nations, where data are available, to generalize in an average sized Greek city.

For our first experiment we used a random forest classifier trained on all the features of the Chicago dataset we've constructed. The first area we experimented on was the city center, which contains the majority of the city's cafes, pubs and shops and there's a bigger number of buildings in general. Selecting a 5 by 5 fishnet, the model detects lot of False Positives especially in not well-lit areas and while it detects many true negatives (e.g Koumoundourou street) most of these fell outside the selected Area of Interest. Increasing the grids increases the number of correctly classified areas, but so do the False Positives, with some of Koumoundourou's amenities being misclassified as abandoned being the most interesting example. The area near T.Oikonomaki was concerning as well, since the results were consistent with each grid size iteration, so we evaluated this area independently. Selecting a 5x5 fishnet, we noticed that the model mislassifies the majority of grids within this area. Grids with increased vegetation and no buildings were classified as not abandoned (False Negative) while the rest of the grids were completely falsely detected as abandoned (False Positives). Increasing the number of grids while it did increase the number of correctly de-

tected areas as not abandoned didn't have the ability to detect vacant places. Thus indicating inability of the model to assert a property in dense urban areas.

The next area that was assessed was the one expanding between the port, the Old Town and Epta Platania which is characterized for being mostly residential, with parks and the train station nearby. In this case, the model correctly detects the majority of areas with zero buildings such as roads or parks as abandoned, as well as schools and the train station as not abandoned. Many of the houses were also classified correctly as not abandoned indicating the ability to classify correctly non vacant buildings in not dense residential areas. While the various pubs and clubs of the Old Town were classified correctly there was a of part of the AoI that featured some restaurants that was detected as False Positive. This is probably due to the fact that this area is dimly lit and has increased vegetation.

Finally, we selected the coastal area of the city where we had the most true positives and negatives being detected in a fishnet of 25 by 25. The majority of areas without buildings being correctly categorized as abandoned and the majority of cafes by the coast as not abandoned. Although, there were instances that some hotels were misclassified as abandoned. Generally, the model can detect majority of non vacant properties but fails to detect abandoned buildings confidently B.1.

The second experiment involved the classifier with VgNIRBI spectral index and the average radiance The old town area when split into 15 by 15 fishnets, detected the previously misclassified restaurants and the Museum as not abandoned and the completely barren areas (e.g behind the Bus station) as abandoned. On the other hand areas that did include amenities but fell into the same grid with barren land, were misclassified, with the bus and train station being the most interesting. The majority of houses were also misclassified. Increasing the grid size alleviated some of the issues regarding amenities but the model couldn't identify the single storey houses as not abandoned. The coastal area consistently, across all different grid sets, had a majority of False Positives with most of amenities being classified as abandoned, areas without buildings with mostly vegetation as not abandoned. This model performed the worst in the city center case, with the majority of grids being classified as False Positives. Overall, the model didn't perform well and seems sensitive to changes in the landscape, making it inconsistent and unable to detect abandoned buildings.

The third experiment was performed using the NDVI, NDBI index combination. In all three areas and different fishnet sizes the model was unable to not only detect abandoned

houses but to also identify the non vacant ones correctly, a behaviour not shared by previous models. Parts of the seaside area were identified as not abandoned, while the nearby shops were. Moreover, the rest of the True Negative results, seem random, with the only common part being the inclusion of roads.B.2

In the fourth and final one, we evaluated the performance of NDVI, NDBI along with the average radiance. The model predicted the majority of split polygons in the various areas of interest as False Positives B.3.

All in all, while in most cases the Random Forest models using spectral index and nighttime light radiance from Chicago, seem to be capable of detecting the majority of not abandoned buildings, they fail dramatically to identify the abandoned ones, thus are not suitable for our needs. This can be due to various reasons, such as our wrong assumptions regarding the index selection and errors during the computation of said indices. Apart from bias generated by our decisions and assessment of the results, the data themselves may vary significantly compared to the ones we are trying to test our methods on. In order to test that we performed some statistical tests. The indices, derived from Chicago and Volos, were computed for the dates between January 1st of 2021 and May 1st of 2022, in order to counter possible outliers related to the pandemic and because of data availability.

Firstly, we performed a Levene test in order to check the variance in both cases. Since the returned p-value was equal to zero in all our cases, we rejected the Null Hypothesis and assumed that the two populations have significantly different variance. Afterwords, we performed a t-test in order to assess the means of the two independent samples. Since this method assumes by default that the samples have equal variance, something not relevant in our case, we used Welch's implementation. The calculated p-values were equal to zero, at a 0.05 level of significance (a=0.05), meaning that the samples have statistically significant mean difference. Spectral data derived from Chicago, were considerably different from the ones regarding our area of interest, and thus not suitable for accurate classification tasks.

One Class SVM Experiments

The second approach involves the use of One Class SVM models using indices derived from the previously mentioned correlation tests. The model's hyperparameters were kept to default for most of the classifiers, with only one case were we decided to reduce the nu

parameter by 0.1. [4]

The experiment using just the EMBI index was conducted first. Using this method with a 15 by 15 fishnet in the city center area, resulted in the detection of some abandoned buildings but not at a satisfying degree. Some of the bars and restaurants, especially the ones including some vegetation, were misclassified as abandoned. Increasing the number of polygons provided with similar results, with the only difference being that some buildings were classified correctly as True Negatives. Moving to the area with sparse buildings, the results were a bit more promising, with some unused storage buildings and pubs being classified correctly. Unfortunately, the model classified various houses as abandoned due to the increased presence of vegetation in their yards. contrary to our expectations, increasing the number of polygons in this case provided worse results, with errors in some previously correctly detected areas. In the case of the coastal areas, the model performed similarly where increased vegetation was present but classified some of the cafes as abandoned. Generally, the model can't provide adequate results for the detection of vacant buildings B.4.

Since the EMBI did make some correct predictions but failed in cases of various buildings that their activity is more pronounced during the night, we mixed it with the average radiance of nighttime lights. Checking the central area in various split instance, the majority of not abandoned buildings were classified correctly, including shops, pubs and restaurants independently of the vegetation presence. Some of the abandoned buildings or vacant lots were also categorized appropriately. When evaluating the area near the Old Town, the model's predictive capability was hit. Most of nightlife amenities as well as departments stores were False Positives. The only instance were the model was accurate was areas with no buildings. The results in the coastal area followed the same pattern. Areas that are less busy during the night such as restaurants or that had didn't have a lot of buildings were classified as abandoned regardless of other characteristics. However, that was also the case for some cafeterias. Splitting those larger areas in order to fit polygons more accurately didn't improve the results by a lot. Overall, the model seems sensitive to nightlights and since it performed better than the previous ones we decided to tune some of its parameters. To be precise we changed the nu parameter specified by Chang C. et al. [69], that fine tunes the trade-off between overfitting and generalization. Nu specifies a lower bound for the number of samples that are support

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html

vectors and an upper bound for the number of samples that are on the wrong side of the hyperplane. The default is 0.1. The nu parameter must be in the range [0,1]. For instance, for a nu value of 0.1, the decision boundary will allow a maximum of 10% of the training samples to be incorrectly classified or to be regarded as outliers. After experimenting with different values of nu with the constructed Volos dataset, we decided to use a nu of 0.4, in order to allow less of the training dataset to be misclassified and since it's less than the default of 0.5,

In the area that covers the Old Town and Epta Platania, we notice a big improvement in the classification of pubs and restaurants. All of the amenities in the old town and near the train station were True Negatives. Various points of interest were misclassified in bigger grids, but showed considerable improvements as they decreased in size. Issues regarding vegetation were still present, but at smaller degree than previous iterations. The identification of amenities across the coastal area was consistent in all different fishnet sizes. The park near Agios Konstantinos was classified as abandoned in its whole, in bigger sized grids, but as we increased the number of polygons, the buildings that were included were classified correctly as not abandoned. While there were cases of False Positives in areas with increased vegetation this is due to how they are situated in the grids. Since we were able to increase the grid number without exceeding the computational limits of GEE, we tested the area in a 30 by 30 fishnet. The results concerning the amenities were similar but there were also not abandoned buildings with increased vegetation that were classified correctly. Finally the central area, was classified correctly with low degree of identified False Positives or Negatives. The model was able to identify All cafes and restaurants that are prevalent in the area B.5 Hyperparameter tuning using more sophisticated methods seems mandatory in order to evaluate our method more efficiently.

5.1.4 Evaluation of our Methodology

Taking the above experiments into consideration we can reach a number of conclusions, vital for the improvement of the current methodology. First of all it can be derived that we can't use data originating from other cities or at least ones that don't share similar characteristics. In the case of the Random Forest approach. Chicago and Volos unfortunately had statistically significant differences in the majority of indices and night light emission. It is hence natural that the models created under the assumption that we could use these data failed to perform as intended across most experiments, and even when they did, their overall pre-

cision wasn't reassuring. On the other hand, in the case of One Class SVM, where data were collected from the case study area, the results were promising. In contrast with single urban extraction indices (e.g EMBI case), nighttime lights seem essential in detecting abandoned or disused buildings. Using the combined approach resulted in considerably less false positives or negatives in the classification scheme of dense and coastal areas, where the model was able to detect various vacant properties. The model while not adequately, performed better in suburban parts of the city than the rest. Both approaches were hindered, however, by the hyperparameter tuning process which was unfortunately manual, intuitive and possibly the best parameters were not selected. Moreover, all models seem to be affected by the size of fishnet that splits the area of interest. In most cases increasing the number of available polygons improved the model's precision but there were still various problems such as unwanted inclusion of roads, nearby barren land or vegetation. Equal sized polygons while easier to implement, oftentimes can't capture the characteristics of the roi efficiently creating "noise" which the classifiers were sensitive to. Thankfully we can combat that by either using other shapes to split the area with, or select the building's exact shape and create a mask. Goldblatt et al. [70] proved the ability of polygonal grids to cover the areas of interest in a more efficient way across regions with different topology and Zou S. et al. [61] used high resolution imagery to extract individual vacant house parcels. Finally, there is possible bias and error introduced with the way we evaluate the performance of these algorithms in both cases. Since there were no available data for abandoned buildings and the creation of a large dataset was a difficult process, we verified our results by either using google street view imagery, which can in many cases be outdated, or by visiting the sites ourselves.

User Interface

The user interface was based on the one provided in the geemap geospatial package [71], [72], and deployed via Streamlit. This approach not only offers additional capabilities apart from the ones we implemented but also highlights our software's openness and compatibility with existing ones. For the applications edited, but not created by us, you can refer to the original repository. [5] When the user first lands on the main page he is greeted by a short description of the app. On his left side he can choose between different applications including:

- Random Forest Classification of abandoned buildings

[5] https://github.com/giswqs/streamlit-geospatial

- One Class SVM classification of abandoned buildings
- Map Visualization
- Timelapse
- Marker Cluster for Greek Cities
- Population Heatmap for Greek Cities

In the case of our applications regarding the abandoned building detection, the user is asked to provide his desired geojson file. This file contains the spatial information for their area of interest. If they have no such file, they can generate one by simply drawing a polygon on the provided map, export it and then upload it A.1. After they select their area of interest they can specify in the number of equally sized grids that they want to divide it. However the number of polygons that the area can be split into is limited in the current implementation due to GEE's computational limits. The computation will take some time, but signal for each operation will be displayed to the main screen. When complete, the grids identified by our method, will be displayed on a map as green or red polygons depending on the absence or existence of abandoned buildings respectively. The source code can be accessed by following this link

Chapter 6

Conclusion

6.1 Synopsis

Satellite Remote Sensing has gained a lot of interest, with many successful applications in various different sectors and industries proving its importance. Satellite produced data are difficult for a human to analyze and use efficiently so machine learning has been employed in the majority of geospatial analyses. In this paper we presented how satellite imagery and machine learning can be used in order to improve the sectors that affect greatly the Greek economy, such as agriculture and tourism. Apart from these industries, we introduced means, though satellite imagery, for better policy making and management in Greece, that has been plagued with unregulated large scale urbanization. Motivated by the many remote sensing and machine learning achievements, we proposed a method in order to detect abandoned buildings over an area of interest. Abandoned buildings are a result of the daunt economic situation in Greece and their existence can impact the performance of close businesses. Finally, while our research wasn't extensive, mostly focusing on a single city, being unable to preprocess our data further and using inefficient validation methods, outlier detection seems promising in detecting vacant buildings, according to our results. Our proposed software can identify vacant and disused buildings in areas with different topology properties successfully, making it a great tool for Business/Location Intelligence. Small and Medium scale businesses can use it in order to validate a location for its economic sustainability (via nighttime lights) and possible dangers due to abandoned buildings, while policy makers can identify problematic areas and proceed to take necessary action.

6.2 Future Work

As far as future work is concerned, one of our software's fundamental issues, is the way we split our areas of interest using equal sized polygons. These shapes aren't able in a variety of cases to capture the characteristics of a building, resulting in misclassification. The use of alternative shapes such as hexagons should improve the model's predictability significantly. Moreover, the current hyperparameter tuning methods are manual and thus not tested extensively. Use of Gridsearch Cross Validation or similar but faster techniques (e.g TuneSearchCV Bayesian Optimization) would enhance our model's accuracy and counter overfitting. Furthermore, further experimentation regarding preprocessing should be conducted in order to increase the information we can gain from free of cost satellite imagery. Methods that increase image resolution with the use of GANs and techniques that reduce image noise could improve the software's performance. Alternatively, we could implement Deep Neural Networks, proposed in the existing literature, that take raw-unprocessed images as input and had promising results. Finally, the proposed software currently suffers from inefficient accuracy metrics, hence future work should use methods of automatic evaluation through existing very high resolution image datasets, such as SAT4, SAT5 and Google Street View [1].

[1] `https://github.com/Sardhendu/PropertyClassification`

Bibliography

[1] Gordana Kaplan and Ugur Avdan. Object-based water body extraction model using Sentinel-2 satellite imagery. *European Journal of Remote Sensing*, 50(1), 2017.

[2] Hashim Mazlan, Yahya Nurul Nadiah, Ahmad Samsudin, Teruhisa Komatsu, Misbari Syarifuddin, and Reba Md Nadzri. Determination of seagrass biomass at Merambong Shoal in Straits of Johor using satellite remote sensing technique. *Malayan Nature Journal*, 66(JUNE), 2014.

[3] Safae Sossi Alaoui, Yousef Farhaoui, and B Aksasse. Classification algorithms in Data Mining. *International Journal of Tomography and Simulation*, 31:34–44, 6 2018.

[4] Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58, 2016.

[5] Amy E. Frazier and Benjamin L. Hemingway. A technical review of planet smallsat data: Practical considerations for processing and using planetscope imagery. *Remote Sensing*, 13(19), 2021.

[6] Perry C. Oddo and John D. Bolten. The Value of Near Real-Time Earth Observations for Improved Flood Disaster Response. *Frontiers in Environmental Science*, 7, 2019.

[7] M. Razu Ahmed, Quazi K. Hassan, Masoud Abdollahi, and Anil Gupta. Processing of near real time land surface temperature and its application in forecasting forest fire danger conditions. *Sensors (Switzerland)*, 20(4), 2020.

[8] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3), 2021.

[9] Luminita Hurbean, Doina Danaiata, Florin Militaru, Andrei Mihail Dodea, and Ana Maria Negovan. Open data based machine learning applications in smart cities: A systematic literature review, 2021.

[10] Lei Refaeilzadeh Payam
}and Tang and Liu Huan. Cross-Validation. In M TAMER LIU LING
}and ÖZSU, editor, *Encyclopedia of Database Systems*, pages 532–538. Springer US, Boston, MA, 2009.

[11] John L. Dwyer, David P. Roy, Brian Sauer, Calli B. Jenkerson, Hankui K. Zhang, and Leo Lymburner. Analysis ready data: Enabling analysis of the landsat archive. *Remote Sensing*, 10(9), 2018.

[12] Gregory Giuliani, Bruno Chatenoux, Erica Honeck, and Jean-Philippe Richard. Towards Sentinel-2 Analysis Ready Data: a Swiss Data Cube Perspective. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 8659–8662, 2018.

[13] Thomas van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review, 2020.

[14] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning, 2021.

[15] M. Papadomanolaki, M. Vakalopoulou, S. Zagoruyko, and K. Karantzalos. BENCHMARKING DEEP LEARNING FRAMEWORKS FOR THE CLASSIFICATION OF VERY HIGH RESOLUTION SATELLITE MULTISPECTRAL DATA. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-7, 2016.

[16] Vasit Sagan, Maitiniyazi Maimaitijiang, Sourav Bhadra, Matthew Maimaitiyiming, Davis R. Brown, Paheding Sidike, and Felix B. Fritschi. Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174, 2021.

[17] Abubakarr S. Mansaray, Andrew R. Dzialowski, Meghan E. Martin, Kevin L. Wagner, Hamed Gholizadeh, and Scott H. Stoodley. Comparing planetscope to landsat-8 and

sentinel-2 for sensing water quality in reservoirs in agricultural watersheds. *Remote Sensing*, 13(9), 2021.

[18] Minkyu Moon, Andrew D. Richardson, and Mark A. Friedl. Multiscale assessment of land surface phenology from harmonized Landsat 8 and Sentinel-2, PlanetScope, and PhenoCam imagery. *Remote Sensing of Environment*, 266, 2021.

[19] Sagarika Sharma, Sujit Rai, and Narayanan C Krishnan. Wheat Crop Yield Prediction Using Deep LSTM Model, 2020.

[20] Mitchell Roznik, Milton Boyd, and Lysa Porth. Improving crop yield estimation by applying higher resolution satellite NDVI imagery and high-resolution cropland masks. *Remote Sensing Applications: Society and Environment*, 25, 2022.

[21] B. T. Mudereri, T. Dube, E. M. Adel-Rahman, S. Niassy, E. Kimathi, Z. Khan, and T. Landmann. A comparative analysis of planetscope and sentinel sentinel-2 space-borne sensors in mapping striga weed using guided regularised random forest classification ensemble. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 42, 2019.

[22] George Xian, Hua Shi, Jon Dewitz, and Zhuoting Wu. Performances of WorldView 3, Sentinel 2, and Landsat 8 data in mapping impervious surface. *Remote Sensing Applications: Society and Environment*, 15, 2019.

[23] Kristen L. Wilson, Melisa C. Wong, and Emmanuel Devred. Comparing Sentinel-2 and WorldView-3 Imagery for Coastal Bottom Habitat Mapping in Atlantic Canada. *Remote Sensing*, 14(5), 2022.

[24] Tianwen Zhang, Xiaoling Zhang, Jun Shi, and Shunjun Wei. HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 2020.

[25] Wenzhi Zhao, Yang Qu, Jiage Chen, and Zhanliang Yuan. Deeply synergistic optical and SAR time series for crop dynamic monitoring. *Remote Sensing of Environment*, 247, 2020.

[26] Dimitrios Tassopoulos, Dionissios Kalivas, Rigas Giovos, Nestor Lougkos, and Anastasia Priovolou. Sentinel-2 imagery monitoring vine growth related to topography in a protected designation of origin region. *Agriculture (Switzerland)*, 11(8), 2021.

[27] Sara Tokhi Arab, Ryozo Noguchi, Shusuke Matsushita, and Tofael Ahamed. Prediction of grape yields from time-series vegetation indices using satellite remote sensing and a machine-learning approach. *Remote Sensing Applications: Society and Environment*, 22, 2021.

[28] Yan Zhao, Andries B. Potgieter, Miao Zhang, Bingfang Wu, and Graeme L. Hammer. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. *Remote Sensing*, 12(6), 2020.

[29] Xiaoyu Song, Chenghai Yang, Mingquan Wu, Chunjiang Zhao, Guijun Yang, Wesley Clint Hoffmann, and Wenjiang Huang. Evaluation of Sentinel-2A satellite imagery for mapping cotton root rot. *Remote Sensing*, 9(9), 2017.

[30] Lan Xun, Jiahua Zhang, Dan Cao, Shanshan Yang, and Fengmei Yao. A novel cotton mapping index combining Sentinel-1 SAR and Sentinel-2 multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, 2021.

[31] Philippe Martin and Gianmarco I.P. Ottaviano. Growth and agglomeration. *International Economic Review*, 42(4), 2001.

[32] VITHLEEM HASTAOGLOU, Costis Hadjimichalis, NICOS KALOGIROU, and NICOS PAPAMICHOS. Urbanisation, Crisis and Urban Policy in Greece. *Antipode*, 19:154 – 177, 6 2006.

[33] Nikos Karadimitriou, Thomas Maloutas, and Vassilis P. Arapoglou. Multiple deprivation and urban development in athens, greece: spatial trends and the role of access to housing. *Land*, 10(3), 2021.

[34] Ha Minh Nguyen and Le Dang Nguyen. The relationship between urbanization and economic growth an empirical study on ASEAN countries. *International Journal of Social Economics*, 45(2), 2018.

[35] Tahir Ali Akbar, Quazi K. Hassan, Sana Ishaq, Maleeha Batool, Hira Jannat Butt, and Hira Jabbar. Investigative spatial distribution and modelling of existing and future urban land changes and its impact on urbanization and economy. *Remote Sensing*, 11(2), 2019.

[36] Chao Chen, Liyan Wang, Jianyu Chen, Zhisong Liu, Yang Liu, and Yanli Chu. A seamless economical feature extraction method using Landsat time series data. *Earth Science Informatics*, 14(1), 2021.

[37] Qiming Zheng, Qihao Weng, and Ke Wang. Characterizing urban land changes of 30 global megacities using nighttime light time series stacks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 2021.

[38] Rafael Ch, Diego A. Martin, and Juan F. Vargas. Measuring the size and growth of cities using nighttime light. *Journal of Urban Economics*, 125, 2021.

[39] Xi Chen and William D. Nordhaus. VIIRS nighttime lights in the estimation of cross-sectional and time-series GDP, 2019.

[40] Richard Bluhm and Gordon C McCord. What can we learn from nighttime lights for small geographies? measurement errors and heterogeneous elasticities. *Remote Sensing*, 14(5):1190, 2022.

[41] Xuzhe Duan, Qingwu Hu, Pengcheng Zhao, Shaohua Wang, and Mingyao Ai. An approach of identifying and extracting urban commercial areas using the nighttime lights satellite imagery. *Remote Sensing*, 12(6), 2020.

[42] Evangelia Dr. Kasimati. Economic Impact of Tourism on Greece's Economy: Cointegration and Causality Analysis. *International Research Journal of Finance and Economics*, 79, 6 2011.

[43] Filipe Batista e Silva, Mario Alberto Marín Herrera, Konštantín Rosina, Ricardo Ribeiro Barranco, Sérgio Freire, and Marcello Schiavina. Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. *Tourism Management*, 68, 2018.

[44] Xuankai Ma, Zhaoping Yang, and Jianghua Zheng. Analysis of spatial patterns and driving factors of provincial tourism demand in China. *Scientific Reports*, 12(1), 2022.

[45] Eleni Krikigianni, Chrysovalantis Tsiakos, and Christos Chalkias. Estimating the relationship between touristic activities and night light emissions, 2019.

[46] Juan Wei, Yongde Zhong, and Jingling Fan. Estimating the Spatial Heterogeneity and Seasonal Differences of the Contribution of Tourism Industry Activities to Night Light Index by POI. *Sustainability (Switzerland)*, 14(2), 2022.

[47] Bidur Devkota, Hiroyuki Miyazaki, Apichon Witayangkurn, and Sohee Minsun Kim. Using volunteered geographic information and nighttime light remote sensing data to identify tourism areas of interest. *Sustainability (Switzerland)*, 11(17), 2019.

[48] W. Lazuardi, P. Wicaksono, and M. A. Marfai. Remote sensing for coral reef and seagrass cover mapping to support coastal management of small islands. In *IOP Conference Series: Earth and Environmental Science*, volume 686, 2021.

[49] Xiaofeng Li, Bin Liu, Gang Zheng, Yibin Ren, Shuangshang Zhang, Yingjie Liu, Le Gao, Yuhai Liu, Bin Zhang, and Fan Wang. Deep-learning-based information mining from ocean remote-sensing imagery, 2021.

[50] Lauren Biermann, Daniel Clewley, Victor Martinez-Vicente, and Konstantinos Topouzelis. Finding Plastic Patches in Coastal Waters using Optical Satellite Data. *Scientific Reports*, 10(1), 2020.

[51] Bidroha Basu, Srikanta Sannigrahi, Arunima Sarkar Basu, and Francesco Pilla. Development of novel classification algorithms for detection of floating plastic debris in coastal waterbodies using multispectral sentinel-2 remote sensing imagery. *Remote Sensing*, 13(8), 2021.

[52] Andrea Albright and Craig Glennie. Nearshore Bathymetry from Fusion of Sentinel-2 and ICESat-2 Observations. *IEEE Geoscience and Remote Sensing Letters*, 18(5), 2021.

[53] Jun Yang, Ruxin Yang, Ming Hsiang Chen, Ching Hui (Joan) Su, Yin Zhi, and Jianchao Xi. Effects of rural revitalization on rural tourism. *Journal of Hospitality and Tourism Management*, 47, 2021.

[54] Chrysostomos Psaroudakis, Gavriil Xanthopoulos, Dimitris Stavrakoudis, Antonios Barnias, Vassiliki Varela, Ilias Gkotsis, Anna Karvouniari, Spyridon Agorgianitis, Ioannis Chasiotis, Diamando Vlachogiannis, Athanasios Sfetsos, Konstantinos Kaoukis, Aikaterini Christopoulou, Petros Antakis, and Ioannis Z. Gitas. Development of an early warning and incident response system for the protection of visitors from natural hazards in important outdoor sites in Greece. *Sustainability (Switzerland)*, 13(9), 2021.

[55] Gabriel Narvaez, Luis Felipe Giraldo, Michael Bressan, and Andres Pantoja. Machine learning for site-adaptation and solar radiation forecasting. *Renewable Energy*, 167, 2021.

[56] Zhenghang Chen, Yawen Kang, Zhongxiao Sun, Feng Wu, and Qian Zhang. Extraction of Photovoltaic Plants Using Machine Learning Methods: A Case Study of the Pilot Energy City of Golmud, China. *Remote Sensing*, 14(11):2697, 6 2022.

[57] Jacob Fantidis, D V Bandekas, N Vordos, and S Karachalios. Wind Energy Potential in Greece Using a Small Wind Turbine. 6 2013.

[58] Maria Margarita Bertsiou, Aimilia Panagiota Theochari, and Evangelos Baltas. Multicriteria analysis and Geographic Information Systems methods for wind turbine siting in a North Aegean island. *Energy Science and Engineering*, 9(1), 2021.

[59] M. Majidi Nezhad, M. Neshat, D. Groppi, P. Marzialetti, A. Heydari, G. Sylaios, and D. Astiaso Garcia. A primary offshore wind farm site assessment using reanalysis data: a case study for Samothraki island. *Renewable Energy*, 172, 2021.

[60] M. Majidi Nezhad, A. Heydari, D. Groppi, F. Cumo, and D. Astiaso Garcia. Wind source potential assessment using Sentinel 1 satellite and a new forecasting model based on machine learning: A case study Sardinia islands. *Renewable Energy*, 155, 2020.

[61] Shengyuan Zou and Le Wang. Individual Vacant House Detection in Very-High-Resolution Remote Sensing Images. *Annals of the American Association of Geographers*, 110(2), 2020.

[62] Shaojuan Xu and Manfred Ehlers. Automatic detection of urban vacant land: An open-source approach for sustainable cities. *Computers, Environment and Urban Systems*, 91, 2022.

[63] Theodomir Mugiraneza, Andrea Nascetti, and Yifang Ban. Continuous monitoring of urban land cover change trajectories with landsat time series and landtrendr-google earth engine cloud computing. *Remote Sensing*, 12(18), 2020.

[64] Luyao Wang, Hong Fan, and Yankun Wang. An estimation of housing vacancy rate using NPP-VIIRS night-time light data and OpenStreetMap data. *International Journal of Remote Sensing*, 40(22), 2019.

[65] Lan H. Nguyen, Deepak R. Joshi, David E. Clay, and Geoffrey M. Henebry. Characterizing land cover/land use from multiple years of Landsat and MODIS time series: A novel approach using land surface phenology modeling and random forest classifier. *Remote Sensing of Environment*, 238, 2020.

[66] José Antonio Valero Medina and Beatriz Elena Alzate Atehortúa. Comparison of maximum likelihood, support vector machines, and random forest techniques in satellite images classification. *Tecnura*, 23(59), 2019.

[67] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Piatt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, 2000.

[68] Shehroz S. Khan and Michael G. Madden. One-class classification: Taxonomy of study and review of techniques, 2014.

[69] Chih Chung Chang and Chih Jen Lin. Training ν-support vector classifiers: Theory and algorithms. *Neural Computation*, 13(9), 2001.

[70] Ran Goldblatt and Michelle Stuhlmacher and Beth Tellman and Nicholas Clinton and Gordon Hanson and Mate. Mapping Urban Land Cover : A Novel Machine Learning Approach Using Landsat and Nighttime Lights, 2017.

[71] Qiusheng Wu. geemap: A Python package for interactive mapping with Google Earth Engine. *Journal of Open Source Software*, 5(51), 2020.

[72] Qiusheng Wu, Charles R. Lane, Xuecao Li, Kaiguang Zhao, Yuyu Zhou, Nicholas Clinton, Ben DeVries, Heather E. Golden, and Megan W. Lang. Integrating LiDAR data and multi-temporal aerial imagery to map wetland inundation dynamics using Google Earth Engine. *Remote Sensing of Environment*, 228, 2019.

APPENDICES

Appendix A

Images regarding the Application

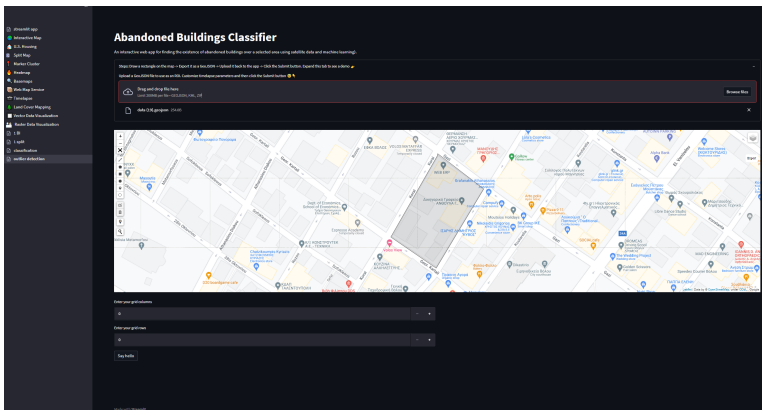

Figure A.1: User Interface Example

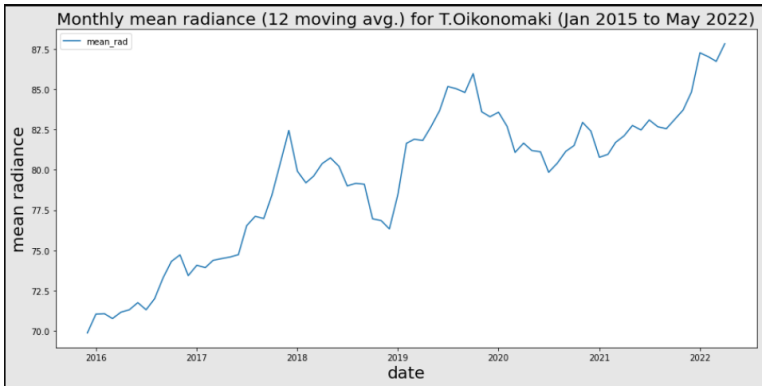

Figure A.2: Average Radiance Plot for T.Oikonomaki

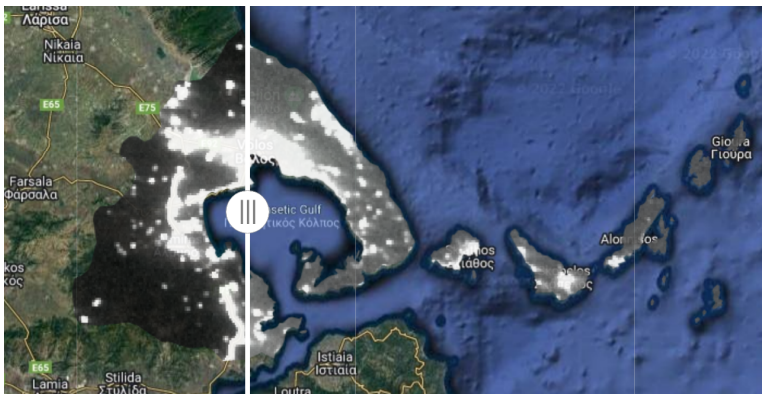

Figure A.3: Nighttime Lights over Magnisia, Before Preprocessing

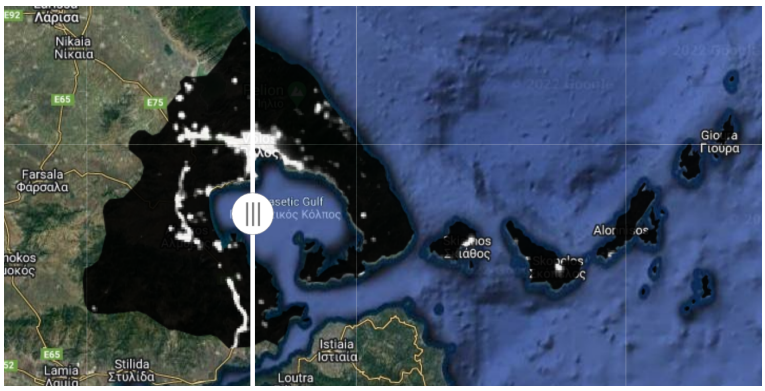

Figure A.4: Nighttime Lights over Magnisia, After Preprocessing

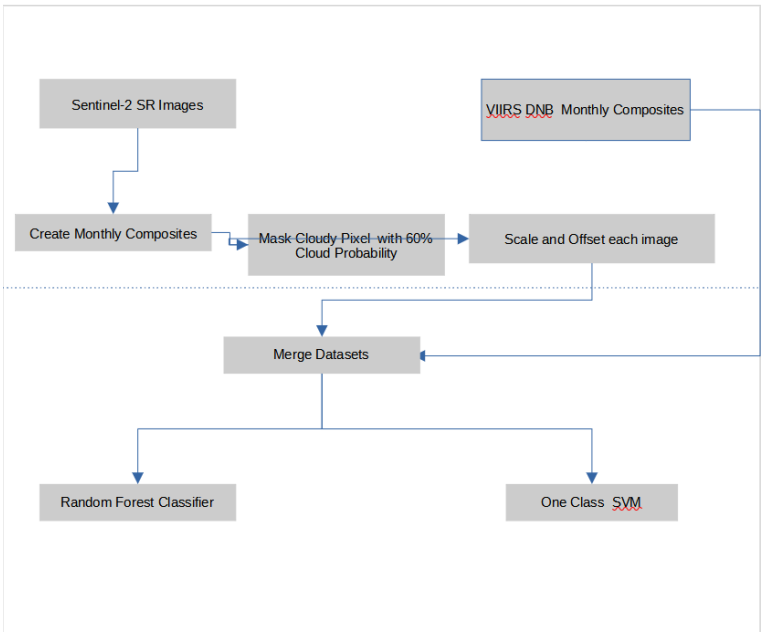

Figure A.5: Machine Learning Pipeline Diagram

Appendix B

Experimentation Results

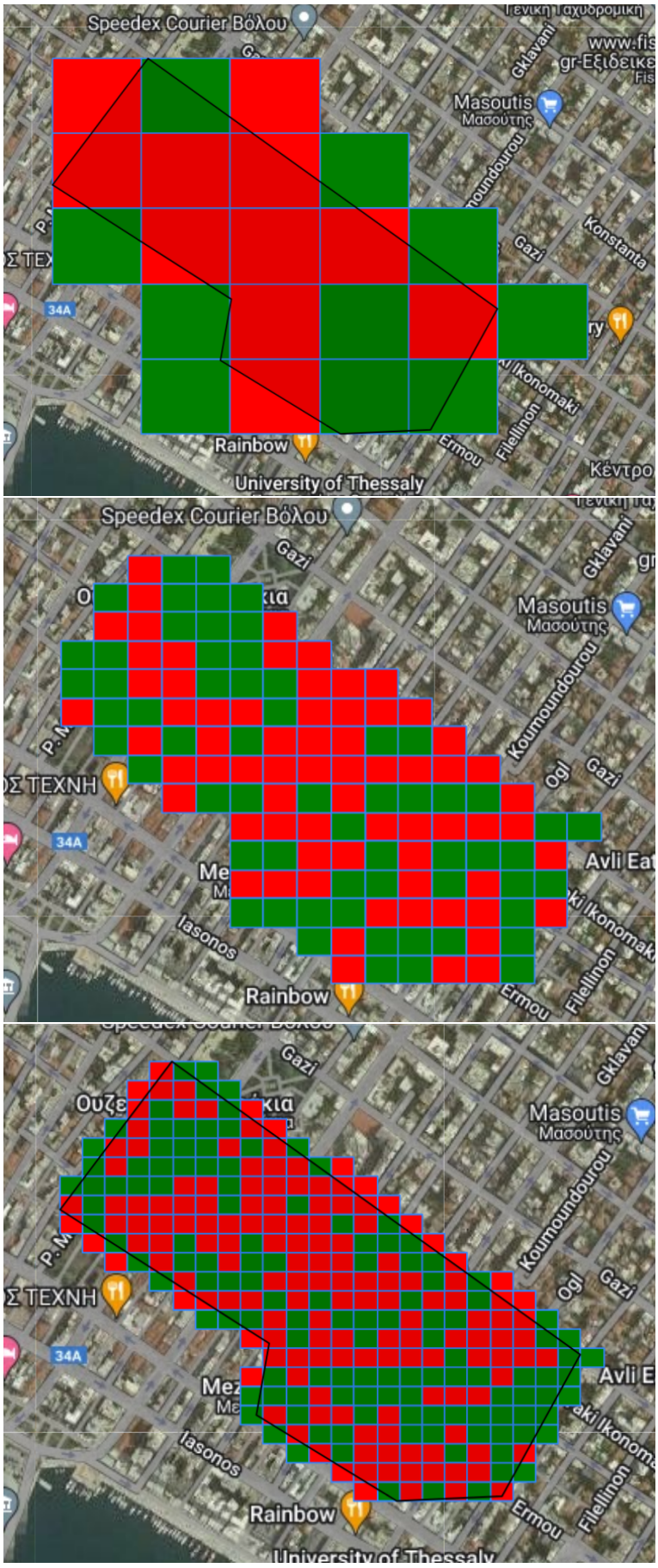

Figure B.1: Experimentaion Results on Central Area using All Spectral Indices in Random Forest

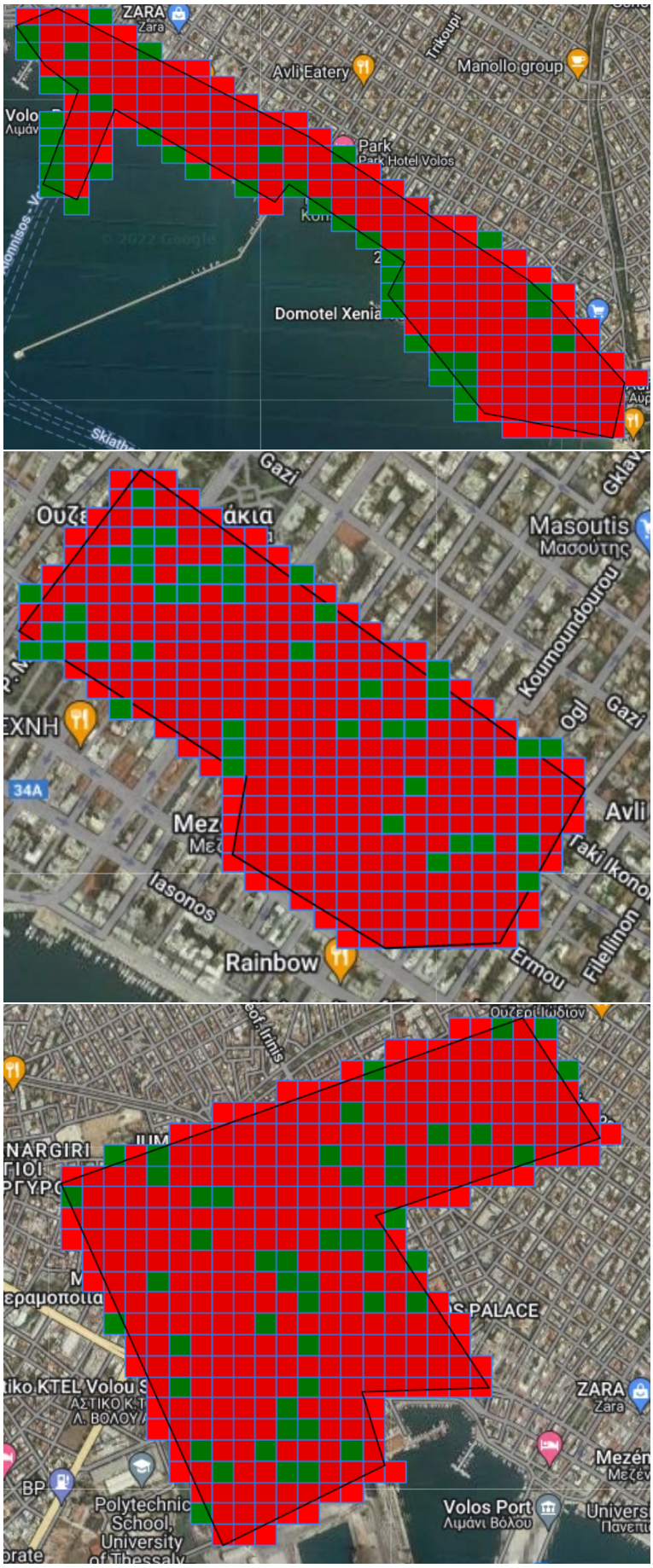

Figure B.2: Experimentaion Results on all areas(25x25) using NDVI/NDBI in Random Forests

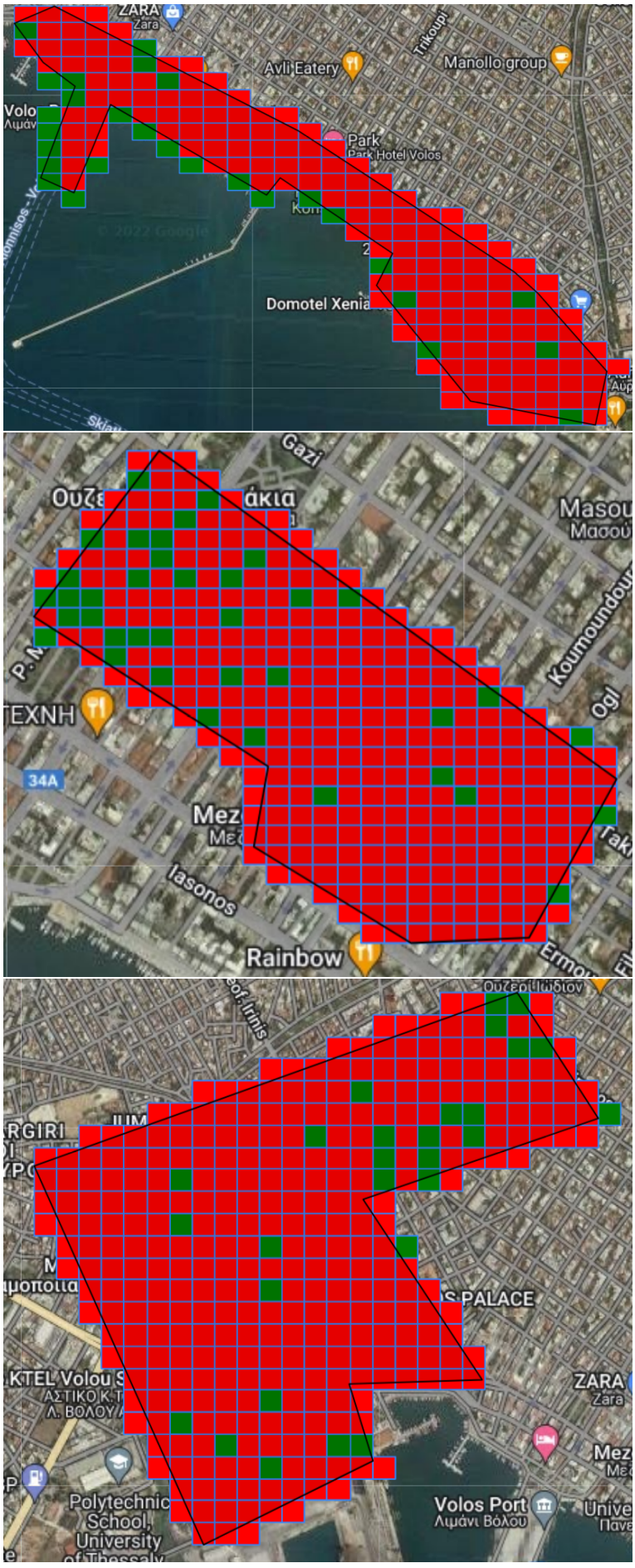

Figure B.3: Experimentaion Results on all areas(25x25) using NDVI/NDBI/Average Radiance in Random Forests

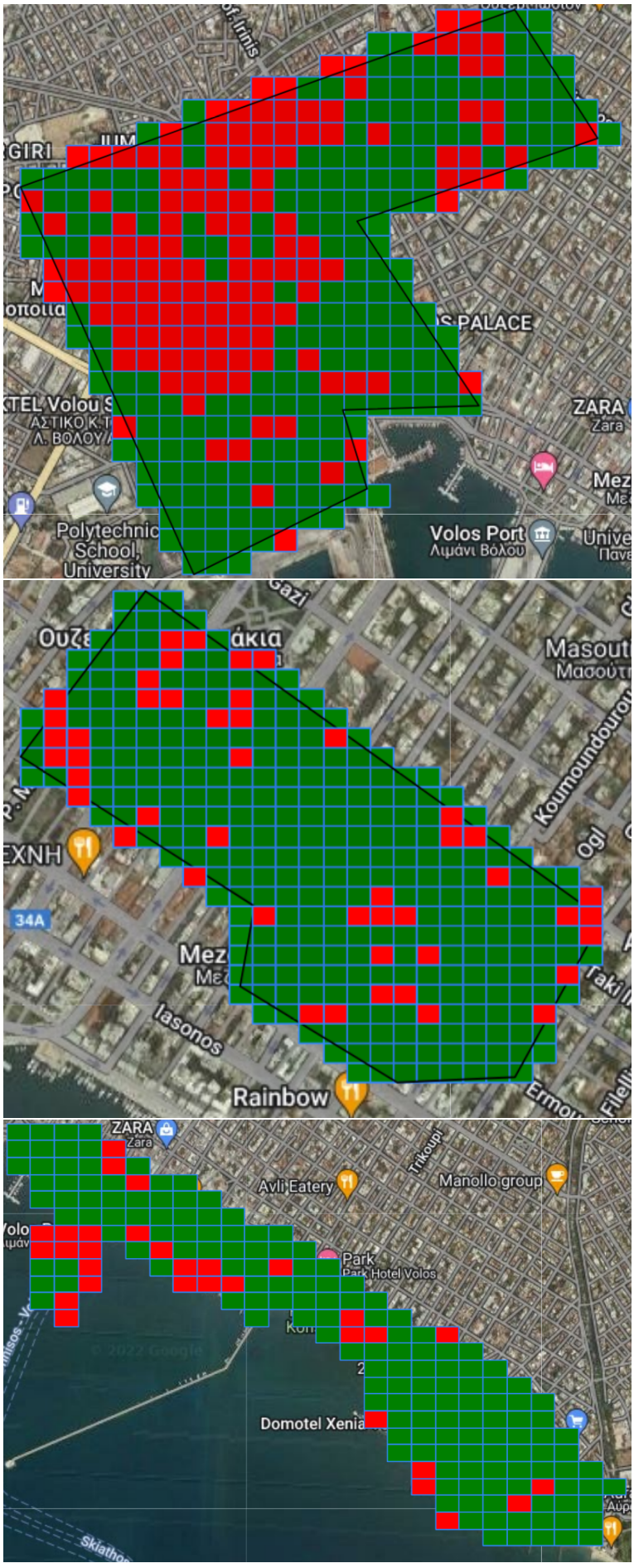

Figure B.4: Experimentaion Results on all areas(25x25) using EMBI index in One Class SVM

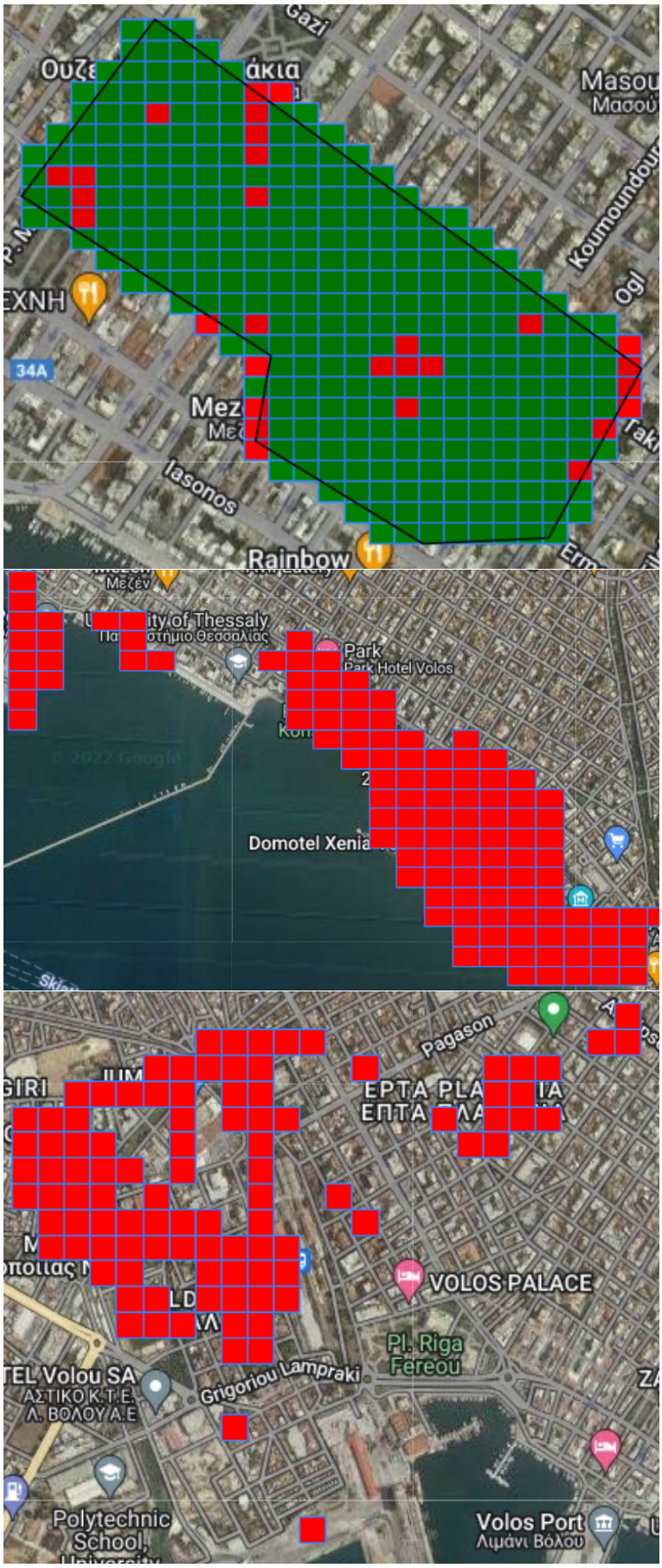

Figure B.5: Experimentaion Results on all areas(25x25) using EMBI and Average Radiance in One Class SVM