

Diesel GenSat: Using Satellite Data to Detect Diesel-Powered Irrigation for Guiding Electrification in Ethiopia

June Lukuyu*
jlukuyu@umass.edu
University of Massachusetts Amherst
Amherst, MA, USA

Gunther Bensch
RWI – Leibniz Institute for Economic
Research
Essen, Germany

Terence Conlon
Columbia University
New York, NY, USA

Anvita Patel
University of Massachusetts Amherst
Amherst, MA, USA

Vijay Modi
Columbia University
New York, NY, USA

Jay Taneja
University of Massachusetts Amherst
Amherst, MA, USA

ABSTRACT

In Sub-Saharan Africa, electricity access is progressing, but electricity use for economic growth remains stagnant. Powering economies sustainably is vital to enhancing livelihoods and is particularly challenging in agriculture-led rural economies. The financial viability of electrification hinges on identifying potential sources of demand to ensure sustainable revenues for utilities, which in turn provides economic benefits to consumers.

This paper presents a technique for identifying areas with diesel-powered irrigation activity in Ethiopia based on remotely sensed data. We develop and evaluate a supervised classification model based on data collected in the Western Ethiopia Highlands on irrigation practices. We find that a feature-based multivariate time series classification approach combined with a k -Nearest Neighbors model accurately predicts about 75% of areas with diesel-powered irrigation activity. Our results suggest that our technique could be valuable in identifying areas in Ethiopia with potential anchor loads for electricity grid extension by replacing existing diesel pumps for irrigation with electric pumps. Guidance on financially-viable areas to expand electricity networks, especially those with economically-vibrant activities like irrigation, is crucial for enabling electricity service providers to recover costs and expand access to more communities more quickly.

CCS CONCEPTS

• **General and reference** → **Empirical studies**; • **Computing methodologies** → **Supervised learning by classification**.

KEYWORDS

Diesel irrigation detection, Remote sensing, Machine learning, Electrification planning

ACM Reference Format:

June Lukuyu, Gunther Bensch, Terence Conlon, Anvita Patel, Vijay Modi, and Jay Taneja. 2022. Diesel GenSat: Using Satellite Data to Detect Diesel-Powered Irrigation for Guiding Electrification in Ethiopia. In *The Thirteenth ACM International Conference on Future Energy Systems (e-Energy '22)*, June 28–July 1, 2022, Virtual Event, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3538637.3538862>

1 INTRODUCTION

Over the past decade, the rate of electrification in sub-Saharan Africa has steadily increased. The proportion of people with access to electricity grew from 33% in 2010 to about 47% in 2019, outpacing population growth. Sub-Saharan Africa remains the region with the most significant electrification deficit despite this marked improvement. Ethiopia is one of the three largest deficit countries globally, accounting for 58 million people without access to electricity [32]. In response, the Government of Ethiopia, with support from the World Bank, launched the National Electrification Plan (NEP) in 2017. It includes a comprehensive plan to reach 100% electricity access by 2025, 65% of the population through the grid, and the rest through off-grid solutions [45].

To realize the full impacts of electrification, increasing access to electricity connections must be accompanied by affordable, reliable, and sustainable electricity consumption and financially sustainable power systems [19, 57]. The unfortunate reality is that in most countries in sub-Saharan Africa, increasing electricity access rates have not been accompanied by the same level of consumption growth [22]. Consequently, people often still rely on traditional fuels, especially for cooking, and utilities and mini-grid companies are struggling for financial viability. A 2016 study of utilities in the region found that only two sub-Saharan African countries have financially viable electricity utilities [37].

Electrification planning must address electricity demand constraints at every stage, not least the planning stage, to ensure that electrification stimulates economic growth. This concept of demand stimulation during electrification planning has historical precedence: Vietnam, an electrification success story, prioritized the electrification of areas with high potential for growth in productive uses of electricity, especially irrigation of the agricultural regions. In turn, government revenue and household incomes increased, leading to higher consumption and thus promoting the overall financial viability of rural electrification [24].

Demand stimulation has become an integral part of electrification planning in sub-Saharan African countries, including Ethiopia

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
e-Energy '22, June 28–July 1, 2022, Virtual Event, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9397-3/22/06...\$15.00
<https://doi.org/10.1145/3538637.3538862>

[45]. Ethiopia's NEP specifically prioritizes grid access to areas with the highest potential for irrigation and agricultural processing, considering the particular importance of agriculture for rural livelihoods [45]. From a utility's standpoint, the financial viability of rural electrification rests on its ability to generate sufficient revenue from the sale of electricity that outweighs the cost incurred by grid extension. Therefore, extending the grid to places with adequate demand from anchor customers, who offer utilities a consistent and substantial source of revenue to supplement the low demand from rural customers, is crucial. In turn, cost recovery for the utility means that they can set cost-reflective tariffs for rural electricity consumers at an affordable rate. Existing capital budgets can also stretch to provide access to more communities. Instead, if utilities spend capital extending the grid to communities where consumption remains low, their finances suffer, and as-yet-unconnected communities wait longer for electricity connections.

How can communities with significant potential electricity consumption be identified? Several previous studies that have addressed electrification planning have utilized Geospatial Electrification Models (GEMs) with a strong supply-side focus. They evaluate the most appropriate and cost-effective technologies for providing universal access while relying on assumptions regarding future electricity demand based on population growth and GDP [36, 38, 43]. Very few studies consider productive use activities to assess potential electricity demand for informing electrification planning and policy. A noteworthy study developed a tool for identifying potential electricity demand hotspots at a high spatial resolution in Kenya, including irrigation load, by combining techno-economic modeling and literature estimates [20]. Another previous effort to develop scalable tools for mapping areas with high potential for electric irrigation pumps in sub-Saharan Africa primarily focused on mapping potential for off-grid solar irrigation pumps through a multi-decision criteria model. Both studies consider groundwater levels, aquifer productivity, crop and land suitability, and population [53]. However, neither distinguishes the current non-electric energy consumption for irrigation.

Currently, about 2% of agricultural land in Ethiopia is irrigated, with smallholder farmers primarily relying on diesel-powered motorized pumps and manually operated pumps for irrigation. An estimated 200,000 diesel-powered motorized irrigation pumps were in use nationally in 2019 [28]. Reliance on these fuels is unsustainable in that it is expensive and detrimental to human health and the environment. Their combustion releases pollutants into the atmosphere, mainly nitrogen oxides (NO_x), carbon monoxide (CO), and particulate matter (PM) [21]. Therefore, moving from diesel-powered to electric pumps can increase the development impact of electrification.

This paper proposes a novel approach to identify areas with existing diesel-powered irrigation in Ethiopia by combining ground data from an agricultural survey with satellite-measured pollution, crop cover, elevation, and surface water data. We apply a supervised machine learning time-series classification technique that leverages the coincidence between irrigation seasons and seasonal variability of pollution and crop cover datasets. The goal of our study is thus to develop a model that can later detect fossil fuel-powered irrigation activity in Ethiopia based on publicly-available data alone. These areas can then be targeted for electrification and serve as a

first productive use load for the grid, improving environmental and financial sustainability for farmers and the utility, enabling acceleration of rural electrification. Publicly-available satellite-measured pollution data is becoming a reliable measure of surface-level pollution where real-time air pollution monitoring on the ground is scarce [31, 35, 72]. Using remote sensing data enables the development of transferable and scalable models like ours to estimate ground-level pollution.

Our specific research approach involves three steps: (1) We conduct on-the-ground surveys to collect the first-of-its-kind comprehensive dataset on the locations and measurements of cultivated plots in two regions of Ethiopia with high irrigation levels. In addition, we collect their crop cultivation and irrigation practices, including the source and method of obtaining water for irrigated plots. (2) We assess two approaches that classify cultivated areas into three classes: not irrigated, irrigated using diesel pumps, and irrigated with other non-diesel-based methods. The first approach directly classifies all observations into the three classes. The second approach is a two-step binary classification, first detecting irrigation activity and then classifying irrigated areas into two classes: areas irrigated using diesel pumps and those irrigated with other methods. (3) We evaluate the performance of four supervised classification algorithms and compare the efficacy of our classification approaches. In doing so, we discuss the limitations of this analysis.

We begin with a brief literature review in Section 2. Then, we present and evaluate our approach in Section 3 and Section 4. Finally, we highlight applicability issues, essential caveats, and future directions for the work in Section 5 before concluding.

2 RELATED WORK

There have been several previous efforts to detect irrigation signals based on time-series optical or radar satellite imagery. Some of these studies have used supervised decision tree, random forest, and support vector machine (SVM) based classification approaches [23, 46, 60]. In contrast, others have used unsupervised decision tree-based classification [4], achieving a good overall performance of over 80% accuracy. Other studies have achieved comparable accuracy with a k -means clustering algorithm [15, 16] and a deep learning approach of a convolution neural network (CNN) architecture [3]. These studies, however, apply these models to irrigation-related indices such as the Normalized Difference Vegetation Index (NDVI) to generate large-scale irrigation maps without distinguishing the type of irrigation method.

A notable study achieved high precision using a U-Net architecture deep learning approach to identify center-pivot irrigation systems [51]. However, these irrigation systems form large circular patterns that can be easily seen from satellite images. However, we cannot apply this method to irrigation systems that are not distinctly identifiable in satellite imagery.

Numerous studies have used satellite pollution measurements to identify spatial and temporal changes and patterns in surface emissions [27, 40, 42]. However, most of these studies have identified emissions patterns at a country-level scale or over large areas and have not distinguished individual or highly localized sources of emissions. Studies that have used satellite-measured pollution

data to detect individual sources of surface emissions have primarily focused on fossil fuel power plants, oil tanks, and ships. They expressly apply Gaussian plume air pollution models to CO_2 or NO_2 data to detect and quantify emissions from individual fossil fuel plants [7, 44, 52] and ships [26]. These models have only been applied to sources with visible smoke plumes and, to our knowledge, have not been applied to smaller sources of emissions, like diesel-powered pumps. In addition, this technique faces the challenge of distinguishing between multiple sources of emissions that are in proximity due to the effect of winds. To overcome this challenge, some studies have applied CNN deep learning techniques, combined with SVM-based classification to satellite imagery to detect fossil fuel power plants [68, 70], oil tanks [65, 71], and ships [69], achieving between 80% and 95% accuracy. These techniques, however, can only be applied to emissions sources that are distinguishable from their surrounding environment in satellite imagery, e.g., through the cooling towers of power plants. Some studies have leveraged the seasonal variability in time series tropospheric pollution measurements to identify the respective source of emissions. These studies, for example, match seasonal NO_2 patterns to harvesting seasons to identify biomass burning, especially in Africa and Southeast Asia [1, 33, 66]. However, no studies to our knowledge have explored matching seasonal pollution patterns to irrigation seasons to detect fossil fuel-powered irrigation pumps. A recent study applied an ensemble supervised machine learning classification technique, combining XGBoost learners, random forest, LASSO, and ridge regression models to remotely sensed rainfall and surface water data to identify functioning and non-functioning groundwater pumps [58]. The model leverages the relationship between surface water availability and groundwater pump use, correctly identifying pumps in use with about 80% accuracy. However, this study does not leverage the temporal patterns of the rainfall data and treats each time unit as a distinct feature. In addition, it does not distinguish the type of pumps or the activity for which the pump is used, so pumps commonly used for drinking water supply cannot be considered separately, for example.

This paper proposes a modified Multivariate Time Series Classification (MTSC) approach. We leverage the Spatio-temporal patterns from multiple time-series pollution and vegetation index variables and integrate some time-constant features. Some previous studies that have applied MTSC tasks have used Long Short Term Memory (LSTM) and recurrent neural network deep learning models [34, 64]. However, deep learning approaches require large training datasets, which is a limitation in our study. Combining distance-based methods such as dynamic time warping (DTW) with a k -nearest neighbors (k -NN) classification algorithm has been used successfully to classify multivariate time series [55]. However, we cannot apply these distance-based methods to a disparate dataset of time-series and time-constant features.

Another common alternative to distance-based methods is feature-based classification approaches, which extract features from the time series data. Then, standard classifiers such as SVM, k -NN, and Random Forest are applied to the extracted features [39, 49, 67]. We, therefore, employ a feature-based multivariate time series classification approach using standard classifiers that have proven successful in MTSC approaches and that are appropriate for our dataset based

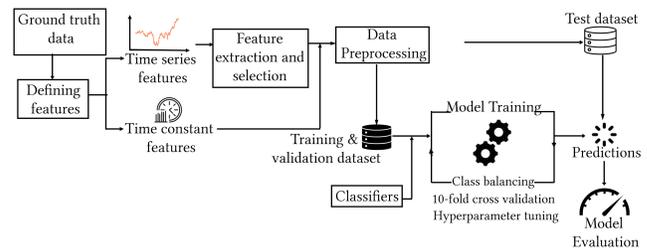


Figure 1: Overview of study methodology

on domain knowledge. This approach allows us to utilize a heterogeneous set of time-series and time-constant features.

3 METHODOLOGY

This section introduces our approach for detecting areas irrigated with diesel-powered water pumps, summarized in Figure 1. We begin by describing the study area, followed by a description of the ground reference primary data and features selected according to insights into the characteristics of diesel-powered pumps and irrigated land. Next, we choose supervised binary classification models based on the characteristics of our datasets and, finally, describe our classification approach.

3.1 Ground Truth Data

Our study area consists of two regions in Ethiopia, the area East of Lake Tana in the Amhara region and the central parts of the Oromia region, each covering around $3,000 \text{ km}^2$ (Figure 2). Ethiopia is located in the North-Eastern part of the African continent between 3°N and 15°N latitude and 33°E and 48°E longitude. Amhara and Oromia regions have the largest area of irrigated agriculture in Ethiopia, comprising over 70% of the existing irrigation schemes in the country [17]. In consultation with the Ministry of Agriculture, we selected six districts in Amhara and six districts in Oromia for data collection. We then divided satellite imagery over the study districts into square pixels of 5 km resolution, of which we selected 36 square pixels as study sites, 18 in each region. For the sample to capture the local variety in irrigation intensity, we based the study site selection on a stratified random sampling approach. The five strata referred to different levels of signal strength determined through an analysis of satellite imagery.

For our analysis, we use irrigation data collected in the study area during April and May 2021 using two primary data collection methods. First, at each of the 36 sites, we conducted an extensive farmer survey in the community lying closest to the center of the square pixel selected as the study site. We interviewed a random sample of about 1,000 farmers and elicited information on their socioeconomic and demographic status, crop cultivation, and irrigation practices (household survey component). If accessible, the farmers' locations and measurements of cultivated plots (four GPS coordinates on the boundary and one at the center of the plot) were also captured, together with the status of irrigation and crop cultivation on the plots at the time of data collection (plot measurement component). The second team of enumerators visited 140 randomly-sampled pixels of 250 m resolution at each study site.

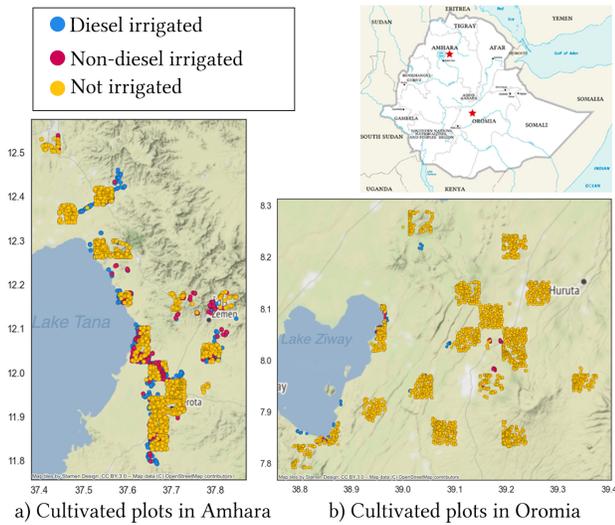


Figure 2: Location of cultivated plots in (a) East of Lake Tana in Amhara region and (b) East of Lake Ziway in central Oromia region, Ethiopia.

Table 1: Number and total area of irrigated plots by the method of irrigation

Irrigation status/method	Number	Area (km^2)
Irrigated, diesel water pump	561	2.69
Irrigated, mechanical water pump	31	0.09
Irrigated, gravity	326	2.07
Irrigated, hand held method	28	0.07
Not irrigated	4162	20.63

They measured the plot at the center of each pixel and up to two additional plots if multiple plots were present within a pixel. They also observed the plot features independently. For the sample to capture the local variety in irrigation intensity, we over-sampled in irrigated areas to provide more positive observations for algorithm training. Therefore, the distribution of our ground truth dataset does not mirror the underlying distribution of irrigation in the regions of study. About 6,100 plots were measured, among which 570 plots were measured as part of the farmer survey exercise. We collected information on the method used to obtain water for irrigation for about 1,100 plots irrigated during data collection. The remaining 5,000 plots were labeled as non-irrigated. We cleaned the dataset by removing the plot instances with invalid geometries or intersecting with other plot polygons. The number and area coverage of the resulting plots by the status and method of irrigation is shown in Table 1.

We also collected the coordinates of about 260 irrigation wells and the respective plots to which they supply water. Based on these data, we determined that the average distance of the irrigation wells from the plots is about 35 m. To account for the placement of the motorized pumps at the irrigation wells, we include a 35 m

Table 2: Number of $250 \times 250 m^2$ pixels per class

Class	Number of pixels
Diesel irrigated	558
Non-diesel irrigated	358
Non-irrigated	4763

buffer around the plots. The spatial distribution of the plots based on labels in the study regions is shown in Figure 2.

We define a $250 m$ resolution grid cell covering both of the study areas selected to align with the sampling pixel resolution. Then, we label and select pixels such that the pixels with a specific class label only contain plots with the same class label. We specify three classes: diesel irrigated, non-diesel irrigated, and non-irrigated (Table 2). The resolution of the pixels is selected such that the spatial resolution of the satellite-measured pollution data is improved without compromising the consistency of the satellite measurements with *in situ* surface measurements, as well as our ability to detect spatial heterogeneity in the pollution measurements [9, 29].

3.2 Defining Features

We identify and select features of the pixels based on knowledge of the characteristics of cultivated land, factors that would influence the choice of irrigation method, and the attributes of diesel-powered water pumps.

3.2.1 Time series pollution data. We use time-series satellite measured pollution data as one of the main input features of our model. We hypothesize that for primarily agricultural areas, there is statistically significantly higher pollution in areas with plots irrigated via diesel-powered pumps than in regions with plots irrigated via other methods during the irrigation seasons. In Ethiopia, these seasons are the so-called Belg and Meher seasons, which we determined to fall between September to March based on our survey data (Figure 3).

The Tropospheric Monitoring Instrument (TROPOMI), which was launched on October 13, 2017, aboard the European Space Agency’s Sentinel-5 Precursor, measures ultraviolet (UV), ultraviolet-visible (UV-VIS), ultraviolet near-infrared (NIR), and shortwave infrared (SWIR) spectral bands. From these bands, a wide range of pollutant gases, including nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), ozone, and formaldehyde, can be retrieved at a spatial resolution of 7 by $3.5 km^2$ (reduced to $3.5 \times 5.6 km^2$ on August 6, 2019) at nadir at a swath width of 2600 km, and daily overpasses at approximately 13:30 local solar time at the equator. We downloaded TROPOMI level-2 NO_2 and CO tropospheric vertical column density measurements from the ESA Copernicus Open Access Hub over the Amhara and Oromia regions of Ethiopia for the period of January 2018 to July 2021. A tropospheric vertical column density is the vertically integrated number of NO_2 (or CO) molecules per unit area between the surface and the tropopause in units of $molec./cm^2$. We filtered the data to remove cloud-covered scenes, errors, and problematic retrievals (i.e., measurements with a quality assurance value of less than 0.75). Since the angle of polar-orbiting satellites on a given area is slightly

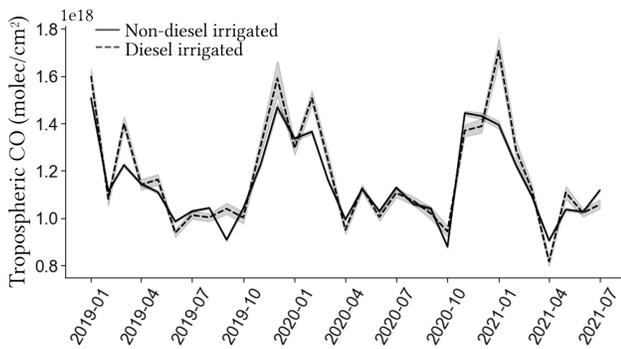


Figure 3: Patterns of average tropospheric CO measurements for diesel irrigated and non-diesel irrigated pixels

different for each overpass, we first accumulate daily observations from consecutive overpasses over each month in each study region to improve spatial sampling. We then arrive at average monthly NO_2 and CO values in each pixel using inverse distance weighted (IDW) interpolation, which predicts the value at an unmeasured location using a k number of measured values surrounding the prediction location. The measured values are weighted proportionally to the inverse of the distance between the measured data point and the prediction location [47, 63].

Previous studies [6, 31, 61, 72] have shown that TROPOMI satellite measurements are a good estimate of *in situ* surface concentrations of both NO_2 and CO in most parts of the world. However, none of these have performed ground-based validation in sub-Saharan Africa, likely due to the scarcity of ground-deployed air quality sensors. We obtained data for August 2020 from the NO_x sensor on a Real-time Affordable Multi-Pollutant (RAMP) monitor recently deployed in Accra, Ghana, [2] to validate the correlation between surface NO_2 concentrations and tropospheric NO_2 measurements. The RAMP was developed in a collaboration between Carnegie Mellon University and SenSevere. It incorporates Alphasense electrochemical sensors to measure CO , NO_2 , SO_2 and O_3 , and a non-dispersive infrared (NDIR) sensor to measure CO_2 [73]. We find that the TROPOMI NO_2 measurements of the 7 by 3.5 km^2 pixel that encompasses the NO_x sensor in Accra have both temporal agreement and a strong correlation (0.71) to measurements taken by the NO_x sensor (Figure 4), providing some confidence in the ability of this satellite to reflect air quality dynamics at ground level.

3.2.2 Time series vegetation indices. To capture the crop cover dynamics of these cultivated pixels, we obtained remotely sensed vegetation data: the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) measurements. These indices quantify vegetation greenness from measurements of light intensity coming off the Earth’s surface in visible and near-infrared wavelengths. EVI improves upon the quality of the NDVI by correcting for some atmospheric conditions and canopy background noise and is more sensitive in areas with dense vegetation. These data products are collected by the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA’s Terra spacecraft at a 16-day temporal resolution and a 250-meter spatial resolution. We

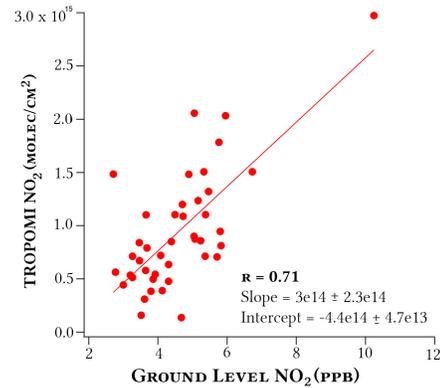


Figure 4: Correlation between tropospheric NO_2 column measurements and ground NO_2 concentrations in Accra, Ghana

extracted monthly averaged data from Google Earth Engine over the study regions for the same period as the pollution data (January 2018 to July 2021). We re-sampled the data to the 1 km^2 grid cells.

3.2.3 Time constant features.

Elevation: The choice of which method of irrigation to use is heavily influenced by topography. We hypothesize that areas irrigated with methods of irrigation that rely on gravity and slope will be at higher elevations than those irrigated with diesel-powered pumps (Figure 5a). We, therefore, use digital elevation data for our analysis, obtained from the NASA Shuttle Radar Topography Mission (SRTM) dataset at a 30 m spatial resolution. The SRTM was a primary component of the payload on the Space Shuttle Endeavour, which launched on February 11, 2000 [59]. We averaged the elevation in meters for each 1 km^2 pixel.

Water availability: Availability of water sources for irrigation impacts the type of irrigation method used by farmers. We, therefore, hypothesize that areas irrigated with diesel pumps will likely be closer to large surface water bodies such as lakes and rivers than those irrigated with other methods, especially non-mechanized methods (Figure 5b). Therefore, using geospatial data on water bodies in Ethiopia from the RCMRD GeoPortal [25], we define the distance of each pixel centroid to the nearest primary water source, that is, a lake or main river, as a feature to capture water availability.

Settlement patterns: Lastly, we consider settlement patterns as a possible feature characterizing the method of irrigation. We use data on population densities and proximity to road infrastructure to capture the variation in settlement patterns. We hypothesize that cultivated plots with non-motorized and non-mechanized irrigation methods that rely on human operation will likely have a higher population and be closer to roads than plots irrigated with diesel-powered pumps (Figures 5c and d). We use one arc-second (approximately 30 m) resolution population estimate of Ethiopia obtained from satellite imagery as part of the High-Resolution Settlement Layer (HRSL) population density dataset [18]. We sum the population values for each pixel and define the distance of each pixel centroid to the nearest major road.

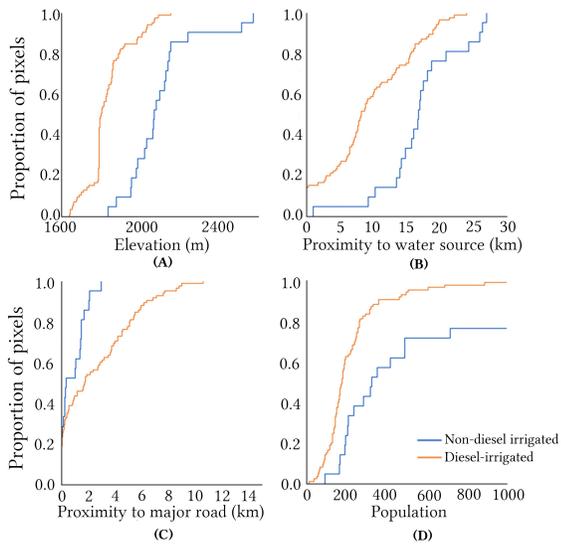


Figure 5: Empirical cumulative distribution functions of time constant features for pixels irrigated with and without diesel.

3.3 Time series feature extraction and selection

A common approach to time series classification is to treat each time point as a separate feature and directly apply a standard classifier. However, this approach is problematic because the classifier ignores information contained in the time order of the data. In our case, the classifier will ignore the seasonality of irrigation cycles in the pollution and crop cover features. To ensure that we capture all the information contained in the time series, we extract features that capture the overall properties of the time series and the correlation between the different measurements in the time series [12]. The algorithm calculates over 200 features for each time series, including the minimum, maximum, mean, median, 25th percentile, 75th percentile, standard deviation (stdev), the linear regression $y = ax + b$ coefficients a and b , and the area under the curve (AUC). A comprehensive list of the features is listed in [11]. The algorithm implements an additional step of feature significance testing and feature selection to avoid increased model and computation complexity and poor model accuracy by including irrelevant inputs. After feature extraction, each feature vector is individually and independently evaluated concerning its significance for the classification problem, and its importance is quantified as a p -value. Finally, a vector of all p -values is assessed based on the Benjamini-Yekutieli procedure [5] to decide which features to keep.

3.4 Data pre-processing

The features in our dataset vary in scale, range, and units. Therefore, to ensure that our models do not make assumptions about the distribution of our data, before training, we first employ a standardization technique to re-scale our features to have a standard normal distribution (mean of 0 and standard deviation of 1). Then we use a min-max normalization technique to re-scale the features into $[0,1]$.

3.5 Classifiers

We based the choice of classifiers for our analysis on the size and distribution of features of our dataset. We consider four classifiers that work well with small, complex datasets and support the nonlinear distribution of features: Random Forest, Support Vector Machine (SVM), k -Nearest Neighbor (k -NN), and Logistic Regression.

Random forest is an ensemble of tree-structured classifiers such that each tree is trained on the values of a random vector sampled independently and with the same distribution for all the trees in the forest. Using a random selection of features to split each node in a tree decreases the correlation between decision trees in the forest and thus reduces the possibility of overfitting [8]. It uses a bootstrapping technique which enables it to work well on relatively small datasets. In addition, it is simple to implement and robust to outliers. After training, the classification result is determined by averaging the most frequent prediction.

Support vector machine (SVM) is a widely used classifier due to its flexibility and robustness. It is based on maximizing the gaps between two classes by defining a hyperplane that splits the two classes [14]. In the case of multi-class classification, SVM uses a “one-versus-one” classification approach, whereby $n_{classes} * ((n_{classes} - 1)/2)$ classifiers are constructed, and each one trains data from two classes [54]. It is effective in high-dimensional spaces and can compute decision boundaries without assuming specific input data distributions. It also performs very well with limited training data. To adapt SVM for nonlinear classification to avoid overfitting, we use a Radial Basis Function (RBF) kernel to map the data into a higher-dimensional space.

k -Nearest Neighbor (k -NN) is one of the simplest classifiers, often used for its simplicity of interpretation and low computation time. It hinges on the assumption that similar observations exist in proximity in a multidimensional space. Therefore, it works by calculating the distance between observations and assigning each observation to the class most common among its k nearest neighbors [30]. Besides its simplicity, another advantage is that it is agnostic to the data distribution. However, its drawbacks are that it has low efficiency and its performance highly depends on the value of k .

Logistic regression is a widely used classification technique that uses a logistic function to model a binomial target variable. However, this technique can also be extended to model a multinomial target variable [41]. It is easy to implement, efficient, and is a high-bias model, which means it works well for small datasets. However, its major limitation is the assumption of linearity between the dependent and independent variables.

3.6 Diesel-powered irrigation detection

This section presents our approach to developing and training a model that detects areas with diesel-powered irrigation activity using the above list of input features.

3.6.1 Class balancing. Given the imbalanced nature of our dataset that reflects the ground reality of irrigation methods used in Ethiopia, we use a class balancing technique during model training. Datasets with imbalanced classes cause poor performance with traditional machine learning models and evaluation metrics that assume a balanced class distribution. Previous studies have used random

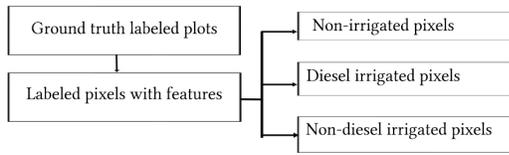


Figure 6: Single-stage ternary classification approach

under-sampling [62] to address the class imbalance by randomly removing some observations of the majority classes. However, this technique is only a good choice when working with a vast dataset, which is not the case for our study, where each ground data point is expensive and onerous to collect. Therefore, we use the Synthetic Minority Oversampling Technique (SMOTE) to address the class imbalance in our dataset [10]. This approach over-samples the minority classes by creating synthetic instances of the class, selecting a minority class at random, finding its five nearest minority class neighbors, randomly selecting k of the neighbors based on the amount of oversampling required, and generating a synthetic example in a randomly selected point between the two examples in feature space. This technique results in a balanced dataset that aims to reflect the characteristics of the underlying unbalanced classes.

3.6.2 Cross-validation. Since we have a limited dataset, removing part of our dataset for validation risks introducing a problem of under-fitting and bias. We, therefore, use a tenfold cross-validation approach, which works by dividing the data into ten subsets, and during each iteration, one of the subsets is used as the validation set, and the other subsets are combined into a training set, ensuring that each observation is a validation set once and is a training set nine times [48]

3.6.3 Hyperparameter tuning. We use a grid search strategy for different model hyperparameter values during training. We tune the *kernels* and *C* hyperparameters for the SVM classifier, *solver* and *C* hyperparameters for the logistic regression classifier, the *nearest neighbors*, *metric* and *weights* hyperparameters in the k -NN classifier, and the *number of estimators* and *maximum features* for the random forest classifier.

3.6.4 Classification approach. We propose to compare two classification approaches to detect diesel-powered irrigation. The first is a single-stage ternary classification approach, categorizing our specified classes directly. The second is a two-stage binary classification approach, which seeks to leverage and build on existing seasoned irrigation detection efforts. In both cases, we train and cross-validate the classifiers on 70% of the data and reserve the remaining 30% as a test dataset.

Single-stage ternary classification approach: In this approach (Figure 6), we use all the labeled observations to train multi-class classification models to categorize the feature vectors into three classes directly: diesel irrigated, non-diesel irrigated, and non-irrigated.

Two-stage binary classification approach: This approach breaks the classification technique into two stages: first, binary classification of irrigated and non-irrigated pixels. In an ideal case, we would then use the resulting true positive labels (correctly classified irrigated pixels) from the first step to train a binary classification

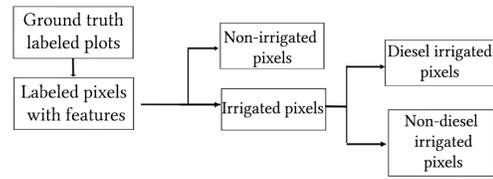


Figure 7: Two-stage binary classification approach

model to categorize the feature vectors into two classes: diesel irrigated and non-diesel irrigated (Figure 7). However, given the relatively small sample size in our study, for the first step of our approach, we validate the irrigation status of the pixels in our sample with an irrigation detection model [Blinded for review]. This model pairs Sentinel-2 imagery with labeled plot polygons to train an irrigation detector that achieves 95% accuracy in our study area. We then train a binary classification model to categorize the feature vectors of the irrigated pixels into two classes: diesel irrigated and non-diesel irrigated.

3.7 Performance Metrics

We evaluate the performance of the classifications based on the widely-used metrics for classification tasks: precision, recall, F1 score, the Matthews correlation coefficient (MCC), and the area under the Receiver Operating Characteristic (ROC) curve (AUC) [56].

4 ANALYSIS

We evaluate the performance achieved by the models in classifying pixels with diesel irrigation activity using a 10-fold cross-validation approach. We present results in terms of precision, recall, F_1 score, and the Matthews correlation coefficient (MCC). We also evaluate the sensitivity of the classification using the area under the ROC curve.

4.1 Single-stage ternary classification

This classification approach considers both the pollution and vegetation index time series. We extracted 3,156 features from the NO_2 , CO , NDVI, and EVI time series data. Feature significance testing and selection based on the Benjamini-Yekutieli procedure results in 1291 time-series features, 374 selected from EVI, 359 from NDVI, 307 from CO , and 251 from NO_2 , resulting in a total of 1295 features, including elevation, population, distance to a primary water source, and distance to a major road.

We find that Random Forest achieves the highest F_1 score of 0.77 for the diesel irrigated class, outperforming the other three classifiers in identifying diesel irrigated pixels, as shown in Table 3. The logistic regression model performs the worst. We note, however, that the SVM slightly outperforms the Random Forest model when we consider the MCC, with a value of 0.76 compared to Random Forests' MCC of 0.75. It is noteworthy that the F_1 scores do not vary significantly from the MCCs, suggesting that the models are not incorrectly favoring one class over the others.

The confusion matrices presented in Figure 8 show where incorrect classification is occurring. All the classifiers incorrectly

Table 3: Comparison of model performance and classification approaches

Classifier	Single-stage ternary				Two-stage binary			
	Precision	Recall	F_1 score	MCC	Precision	Recall	F_1 score	MCC
Random Forest	77%	80%	0.77	0.75	86%	92%	0.87	0.70
Logistic regression	64%	76%	0.66	0.67	85%	80%	0.84	0.57
SVM	75%	80%	0.76	0.76	81%	96%	0.84	0.67
k -NN	67%	81%	0.70	0.71	91%	86%	0.90	0.72

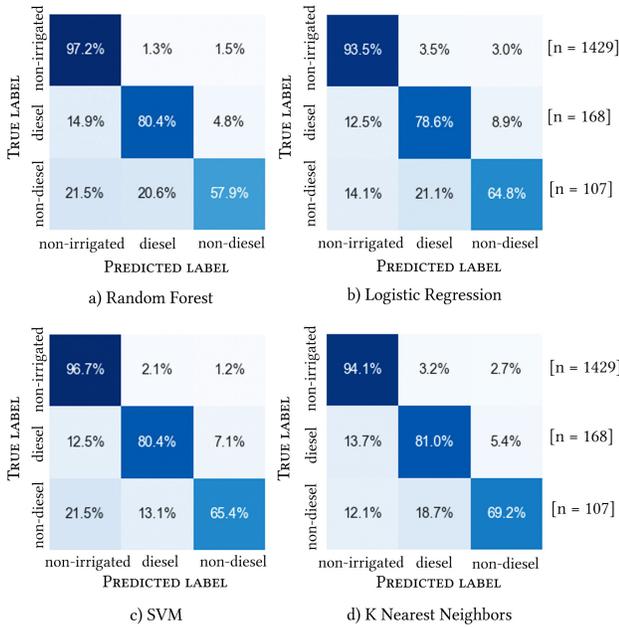


Figure 8: Confusion matrices of models in the single-stage ternary classification approach

classified most diesel irrigated observations as non-irrigated. This result suggests that the two-stage binary classification approach is likely to perform better as it separates the classification of irrigation activity and the classification of the type of irrigation method. It is also noteworthy that the models in this approach are better at classifying the non-irrigated observations than the other two classes—the results of the AUC analysis support this finding.

The ROC curve of the Random Forest model is shown in Figure 9. It is plotted using the "One vs. All" methodology, which means that the ROC for each class is classified against the other two. We find that the model has the highest ability to separate the non-irrigated class, with an AUC value of 0.72. The diesel irrigated and non-irrigated classes have AUC values under 0.7, suggesting that the model is struggling to distinguish these two classes.

4.2 Two-stage binary classification

We evaluate whether separating the classification task into two separate binary classification tasks outperforms the direct ternary classification approach presented in the previous subsection in identifying diesel irrigated activity. In the irrigation detection model

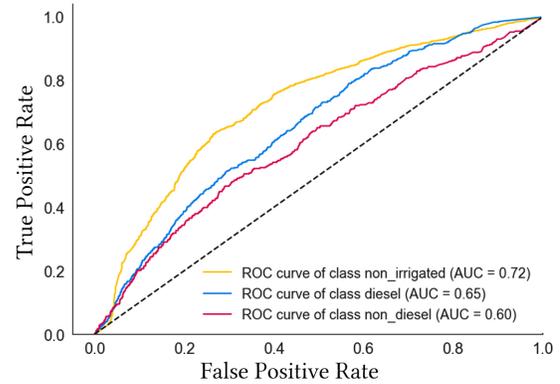


Figure 9: ROC curve for Random Forest model in a single-stage ternary classification approach

that we employ in the first step of our classification task, non-irrigated and irrigated labels are predicted with 98.3% and 95.5% accuracy [13]. Subsequently, we use the irrigated pixels in our sample that align with the predictions from the irrigation detection model as inputs to the second step of the classification task.

Out of 1,578 time-series features extracted from the NO_2 and CO time series, the algorithm selected 81 from NO_2 and 105 from CO as significant for the classification problem. Therefore, together with the time constant features, 190 features. Note that we do not use the vegetation indices time-series in this step, as it is the main feature used in the irrigation detection model in the first step of this approach.

Overall, we find that every model in the binary classification approach outperforms the models in the ternary approach. The k -NN model outperforms the other three models considering its F_1 score of 0.9, which captures how well the diesel-irrigated class is predicted, and its MCC of 0.72, which captures how well both classes are predicted. We can see the balanced performance of the k -NN in its confusion matrices presented in Figure 10, showing correct predictions in about 86% of the observations of both classes.

We find that the measure of separability of the models significantly improves with the two-stage binary classification approach. The k -NN model achieves a mean AUC of 0.93 (Figure 11) with 10-fold cross-validation, suggesting a 93% chance that the model will correctly distinguish a diesel irrigated observation from a non-diesel irrigated observation.

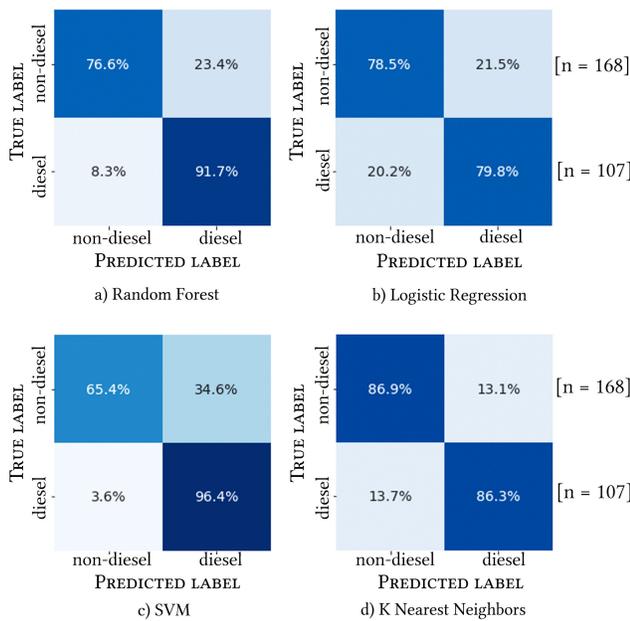


Figure 10: Confusion matrices of models in the two-stage binary classification approach

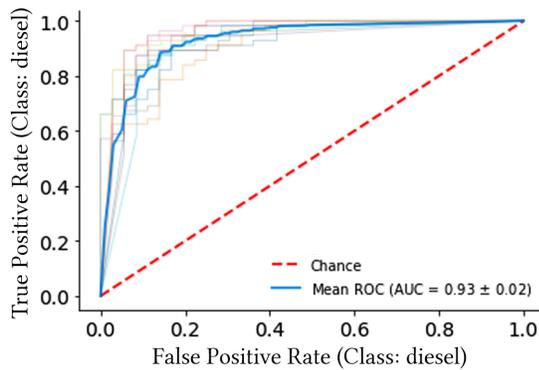


Figure 11: 10-fold cross-validation ROC curve of diesel irrigated class for k -Nearest Neighbor model in two-stage binary classification approach. Note: the value of k is 5.

4.3 Feature Contribution

We evaluate which features played a vital role in enabling our features to distinguish between diesel-irrigated and non-diesel irrigated pixels by comparing two techniques that estimate feature importance: Random Forest and Gradient Boosting. We find that both methods commonly identify several time-series features extracted from CO , elevation, and proximity to a primary water source as the key features that contribute to the success of our model, as shown in Table 4. It is noteworthy that some features from the CO pollution time series contribute more to the performance of our model than features from the NO_2 pollution time series.

Table 4: Top ten features contributing to model performance

Rank	Random Forest	Gradient Boosting
1	Elevation	Non linearity c3 statistic, CO
2	Fourier coefficient, CO	Elevation
3	Linear least-squares regression, CO	Linear least-squares regression, CO
4	Non linearity c3 statistic, CO	Fourier coefficient, CO
5	Partial autocorrelation function, CO	Proximity to major water source
6	Proximity to major water source	Autocorrelation, CO
7	Autoregressive coefficient, CO	Continuous wavelet transform, NO_2
8	Quantile, CO	Index mass quantile, NO_2 mean
9	Minimum, CO	Variance, CO
10	Fourier coefficient, NO_2	Standard error, CO

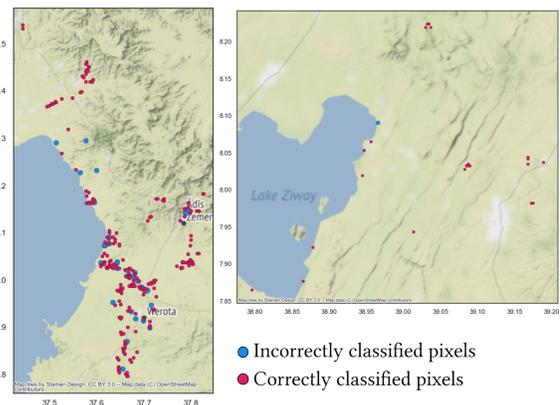


Figure 12: Locations of test set pixels by classification

4.4 Spatial Analysis

We also consider the areas where our model was unsuccessful in classification. We find that the most significant proportion of incorrectly classified pixels are in areas where the pixels of different classes are near each other, mainly to the East of Lake Tana in the Amhara region, as shown in Figure 12 (compared with Figure 2). One possible reason is that the pollution measurements, which have been re-gridded at a 250 m resolution, are hard to distinguish among pixels close together.

To that end, we explore the impact of the spatial placement of pixels of different classes on the performance of our k -NN model. We consider (1) only pixels of the two classes close to each other and (2) only pixels of the two classes far from each other. We determine this proximity threshold by considering the distribution of distances between diesel-irrigated pixels and non-diesel irrigated pixels and vice versa and categorizing those below the median distance value as being in proximity. The median distance of diesel-irrigated pixels from non-diesel irrigated pixels is 714 m, and that of non-diesel irrigated pixels from diesel-irrigated pixels is 357 m. As shown in Table 5, our model performs much worse when we only consider the pixels of two classes close to each other, achieving an MCC of only 0.23 compared to 0.72 when trained on the full dataset.

On the other hand, we also note that there is better performance when we only consider the pixels of two classes that are further than the median distance, achieving an improved MCC score of 0.79. While our model struggles to separate the diesel and non-diesel

Table 5: Performance of k -NN model considering the spatial proximity of class pixels

	Precision	Recall	F_1 score	MCC
Full dataset	91%	86%	0.90	0.72
Class pixels in close proximity	70%	68%	0.70	0.23
Class pixels far from each other	92%	90%	0.92	0.79

Table 6: Bootstrap Matthews correlation coefficient estimates and confidence intervals for classifiers in the two-stage classification approach

Classifier	Bootstrap estimate	Confidence interval	Standard error
Random Forest	0.75	[0.67,0.83]	0.030
Logistic regression	0.56	[0.48,0.64]	0.031
SVM	0.72	[0.64,0.80]	0.031
k -NN	0.74	[0.66,0.82]	0.032

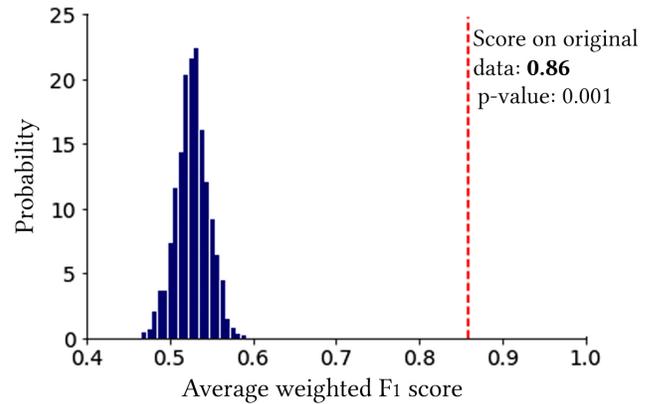
irrigated pixels in areas with multiple pixels of different classes, as evidenced by the MCC, electrification decisions are usually made atomically over areas that comprise heterogeneous activities. Therefore, a good precision and recall score for the diesel-irrigated class in these areas is sufficient.

4.5 Validation of Results

In our analysis, we use a sample of own-collected primary data given the absence of comprehensive data for Ethiopia that contains information on the location and type of water pumps for irrigation. We acknowledge that the small size of our dataset is challenging for classification but note the lack of any similar datasets in the literature. While our dataset is a good representation of the overall classification problem, there is a likelihood that it may not be enough to capture the complexity of the classification problem on a large scale. In the following, we present two efforts to give confidence to our results.

4.5.1 Bootstrap Confidence Intervals. We construct confidence intervals around the Matthews correlation coefficient of our models in the two-stage binary classification approach to evaluate its variability. We use a widely used statistical method, the bootstrap method, which randomly draws samples from the original dataset (with replacement) to obtain estimates of the MCC, thus creating a distribution. We generate a 99% confidence interval of coefficient values from 1000 bootstrap samples for the models in the two-stage binary classification approach. We find a 99% chance that the interval [0.66, 0.82] contains the MCC of our best performing model, the k -NN, and the MCC of the poorest performing model, the logistic regression, is included in the [0.48,0.64] interval with high certainty.

4.5.2 Permutation testing. Permutation testing helps to investigate whether or not the performance score obtained from the models is by chance. The algorithm generates a null distribution of the performance score of the classifier on 1000 different permutations of the dataset, where features remain the same but labels undergo

**Figure 13: Distribution of average weighted F_1 scores from permutation test of k -NN model in two-stage classification approach.**

various permutations [50]. We performed permutation testing on the F_1 score of the two-stage classification k -NN model. We find that the weighted average F_1 score obtained on the original data, 0.86, is statistically significantly higher than the scores obtained using the permuted data, as shown in Figure 13. This result indicates that our dataset contains actual dependency between features and labels, which our model was able to use to identify the observations of the diesel irrigated class.

When we apply our k -NN model to the predicted irrigated areas of the Amhara region [13], it classifies about 20% of the irrigation activity as using diesel-powered pumps. As shown in Figure 14, a significant proportion of the diesel-powered irrigation activity is in the region's western part.

5 DISCUSSION AND FUTURE WORK

Our model's high precision and recall suggest that our approach could be valuable in identifying areas with potential anchor loads for grid extension by replacing existing diesel pumps for irrigation with electric pumps in Ethiopia. Grid extension planning has myriad considerations, including the locations of populations without electricity, unelectrified productive use loads including mining and industry, proximity to the existing grid, and political considerations. However, information on where diesel-powered irrigation already is represents sites more likely to have stable revenue from electricity sales. This information can be of significant value to grid extension planning in settings like Ethiopia, where irrigation is central to the electrification strategy and little other economic activity exists in rural areas [45]. While capturing the relative importance of each consideration is difficult enough in planning grids, it becomes even more challenging in *integrated electrification planning* scenarios that also consider decentralized electrification via mini-grids and solar home systems.

Nonetheless, as a simple thought experiment to show the value of accurate predictions of diesel-powered irrigation areas, imagine that two grid planners each have enough budget to extend an existing grid to 100 additional sites. Each of them gains another site for every

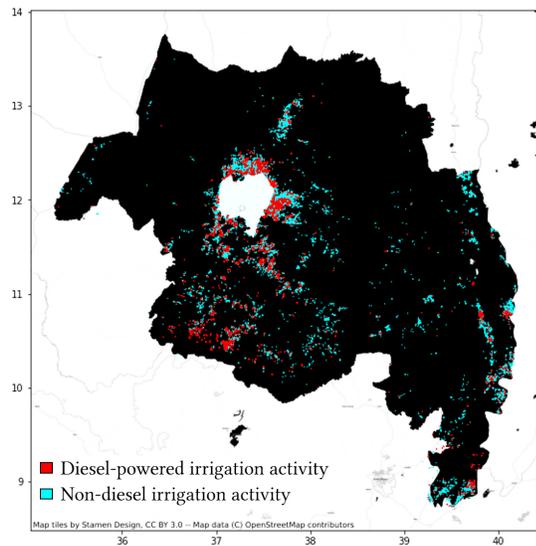


Figure 14: Predictions of diesel-powered irrigation activity in the Amhara region of Ethiopia.

two electrified diesel-powered irrigation sites (since these sites return more consistent revenues). Both planners have perfect maps of all irrigation sites in the country. Still, one has no information about diesel-powered irrigation sites, and the other can interrogate any location and receive a correct answer 75% of the time (similar to our algorithm's performance).

Given that only 20% of irrigation sites are diesel-powered, our second planner is nearly four times more likely to correctly select areas with diesel-powered irrigation. Consequently, the first grid planner uses their capital budget to electrify 112 sites, while our second planner can electrify 250 total sites (2.2x). While this exercise is highly simplified and any assumptions may be discounted, it aims to convey the enormous value to electricity planners in improving insight into which areas may yield higher stable revenues.

Future directions will consider the equity implications of this work by investigating the prioritization of electrification in more affluent areas where consumers can already afford diesel pumps. Second, we have yet to consider this technique's applicability in built environments that could have confounding pollution activity. Third, having a limited sample size makes it difficult to generalize our findings. However, collecting a larger ground truth sample is highly resource-intensive and logistically challenging. Therefore, there could be an opportunity to create hand-labeled ground truth data from satellite data to train novel and transferable deep learning models.

6 CONCLUSION

While universal electrification is beginning to come into view in Sub-Saharan Africa, the attendant livelihood gains from widespread adoption of electricity for economic benefits remain stubbornly far. Financially sustainable electricity service providers are crucial to an electrified economy, and identifying potential sources of sustainable

revenue for utilities that provide financial benefits to customers is a central challenge. This work presents a novel technique that leverages a raft of remote sensing datasets – including pollution, vegetation, and population as derived from satellites – to identify whether irrigation sites in Ethiopia are powered by diesel or non-diesel sources. We evaluate our technique via a unique household survey among farmers in two regions of Ethiopia that we collected. Our results show a more than 3.5x improvement over the random chance threshold for this problem (20% to $\approx 75\%$) that can enable enhanced business models for electricity service providers, environmentally-sustainable production for farmers, and accelerated electrification for the people of Ethiopia. The broad application of this technique can substantially aid in the successful expansion of electricity systems throughout agriculture-led developing regions.

ACKNOWLEDGMENTS

We acknowledge the generous UK Aid funding from the UK government under the Applied Research Programme on Energy and Economic Growth (EEG).

REFERENCES

- [1] R J van der A, H J Eskes, K F Boersma, T P C van Noije, M Van Roozendael, I De Smedt, D H M U Peters, and E W Meijer. 2008. Trends, seasonal variability and dominant NO_x source derived from a ten year record of NO₂ measured from space. *Journal of Geophysical Research* 113, D4 (2008), n/a–n/a. <https://doi.org/10.1029/2007jd009021>
- [2] AfriqAir. 2020. Africa qualité de l'air (AfriqAir). Retrieved January 28, 2022 from <https://www.cmu.edu/epp/afriqair/index.html>
- [3] Chitra Agastya, Sirak Ghebremusse, Ian Anderson, Colorado Reed, Hossein Vahabi, and Alberto Todeschini. 2021. Self-supervised Contrastive Learning for Irrigation Detection in Satellite Imagery. <https://doi.org/10.48550/ARXIV.2108.05484>
- [4] Hassan Bazzi, Nicolas Baghdadi, Ibrahim Fayad, Mehrez Zribi, Hatem Belhouchette, and Valérie Demarez. 2020. Near Real-Time Irrigation Detection at Plot Scale Using Sentinel-1 Data. *Remote Sensing* 12, 9 (2020), 1–12. <https://doi.org/10.3390/rs12091456>
- [5] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29, 4 (2001), n/a–n/a. <https://doi.org/10.1214/aos/1013699998>
- [6] T. Borsdorff, J. Aan de Brugh, H. Hu, I. Aben, O. Hasekamp, and J. Landgraf. 2018. Measuring Carbon Monoxide With TROPOMI: First Results and a Comparison With ECMWF-IFS Analysis Data. *Geophysical Research Letters* 45, 6 (2018), 2826–2832. <https://doi.org/10.1002/2018gl077045>
- [7] H Bovensmann, M Buchwitz, J P Burrows, M Reuter, T Krings, K Gerilowski, O Schneising, J Heymann, A Tretner, and J Erzingler. 2010. A remote sensing technique for global monitoring of power plant CO₂ emissions from space and related applications. *Atmospheric Measurement Techniques* 3, 4 (2010), 781–811. <https://doi.org/10.5194/amt-3-781-2010>
- [8] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. <https://doi.org/10.1023/a:1010933404324>
- [9] Angela Cersosimo, Carmine Serio, and Guido Masiello. 2020. TROPOMI NO₂ Tropospheric Column Data: Regridding to 1 km Grid-Resolution and Assessment of their Consistency with In Situ Surface Observations. *Remote Sensing* 12, 14 (2020), 2212. <https://doi.org/10.3390/rs12142212>
- [10] N V Chawla, K W Bowyer, L O Hall, and W P Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. <https://doi.org/10.1613/jair.953> arXiv:1106.1813
- [11] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. 2018. Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* 307 (2018), 72–77. <https://doi.org/10.1016/j.neucom.2018.03.067>
- [12] Maximilian Christ, Andreas W Kempa-Liehr, and Michael Feindt. 2016. Distributed and parallel time series feature extraction for industrial big data applications. arXiv:1610.07717
- [13] Terence Conlon, Christopher Small, and Vijay Modi. 2022. A Multiscale Spatiotemporal Approach for Smallholder Irrigation Detection. *Frontiers in Remote Sensing* 3 (2022), n/a. <https://doi.org/10.3389/frsen.2022.871942>
- [14] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297. <https://doi.org/10.1007/bf00994018>

- [15] Jacopo Dari, Luca Brocca, Pere Quintana-Seguí, Stefano Casadei, María José Escorihuela, Vivien Stefan, and Renato Morbidelli. 2022. Double-scale analysis on the detectability of irrigation signals from remote sensing soil moisture over an area with complex topography in central Italy. *Advances in Water Resources* 161 (2022), 104130. <https://doi.org/10.1016/j.advwatres.2022.104130>
- [16] Jacopo Dari, Pere Quintana-Seguí, María José Escorihuela, Vivien Stefan, Luca Brocca, and Renato Morbidelli. 2021. Detecting and mapping irrigated areas in a Mediterranean environment by using remote sensing soil moisture and a land surface model. *Journal of Hydrology* 596 (2021), 126129. <https://doi.org/10.1016/j.jhydrol.2021.126129>
- [17] Daniel G. Eshete, Berhanu G. Sinshaw, and Kassaye G. Legese. 2020. Critical review on improving irrigation water use efficiency: Advances, challenges, and opportunities in the Ethiopia context. *Water-Energy Nexus* 3, Int. J. Res. Innov. Earth Sci. 3 5 2016 (2020), 143–154. <https://doi.org/10.1016/j.wen.2020.09.001>
- [18] Humanitarian Data Exchange. 2019. Continent of Africa: High Resolution Population Density Maps. Retrieved April 8, 2019 from <https://data.humdata.org/dataset/highresolutionpopulationdensitymaps>
- [19] Giacomo Falchetta, Shonali Pachauri, Edward Byers, Olha Danylo, and Simon C. Parkinson. 2020. Satellite Observations Reveal Inequalities in the Progress and Effectiveness of Recent Electrification in Sub-Saharan Africa. *One Earth* 2, 4 (2020), 364–379. <https://doi.org/10.1016/j.oneear.2020.03.007>
- [20] Giacomo Falchetta, Nicolò Stevanato, Magda Moner-Girona, Davide Mazzoni, Emanuela Colombo, and Manfred Hafner. 2021. The M-LED platform: advancing electricity demand assessment for communities living in energy poverty. *Environmental Research Letters* 16, 7 (2021), 074038. <https://doi.org/10.1088/1748-9326/ac0cab>
- [21] Michael Fiebig, Andreas Wiartalla, Bastian Holderbaum, and Sebastian Kiesow. 2014. Particulate emissions from diesel engines: correlation between engine technology and emissions. *Journal of Occupational Medicine and Toxicology* 9, 1 (2014), 6. <https://doi.org/10.1186/1745-6673-9-6>
- [22] Simone Fobi, Varun Deshpande, Samson Ondiek, Vijay Modi, and Jay Taneja. 2018. A longitudinal study of electricity consumption growth in Kenya. *Energy Policy* 123 (Dec. 2018), 569–578.
- [23] Qi Gao, Mehrez Zribi, Maria Jose Escorihuela, Nicolas Baghdadi, and Pere Quintana Seguí. 2018. Irrigation Mapping Using Sentinel-1 Time Series at Field Scale. *Remote Sensing* 10, 9 (2018), n/a–n/a. <https://doi.org/10.3390/rs10091495>
- [24] Defne Gencer, Peter Meier, Richard Spencer, and Hung Tie Van. 2011. *State and People, Central and Local, Working Together : The Vietnam Rural Electrification Experience*. Technical Report. World Bank, Washington DC. <https://openknowledge.worldbank.org/handle/10986/27315>
- [25] RCMRD GeoPortal. 2020. Datasets. Retrieved January 28, 2022 from <http://geoportal.rcmrd.org/>
- [26] Aristeidis K Georgoulas, K Folkert Boersma, Jasper van Vliet, Xiumei Zhang, Ronald van der A, Prodromos Zanis, and Jos de Laat. 2020. Detection of NO₂ pollution plumes from individual ships with the TROPOMI/S5P satellite sensor. *Environmental Research Letters* 15, 12 (2020), 124037. <https://doi.org/10.1088/1748-9326/abc445>
- [27] Sachin D Ghude, S Fadnavis, G Beig, S D Polade, and R J van der A. 2008. Detection of surface emission hot spots, trends, and seasonal cycle from satellite-retrieved NO₂ over India. *Journal of Geophysical Research* 113, D20 (2008), n/a–n/a. <https://doi.org/10.1029/2007jd009615>
- [28] GIZ. 2020. *Solar irrigation market analysis in Ethiopia*. Technical Report. GIZ, Addis Ababa, Ethiopia.
- [29] Daniel L Goldberg, Susan C Anenberg, Gaige Hunter Kerr, Arash Mohegh, Zifeng Lu, and David G Streets. 2021. TROPOMI NO₂ in the United States: A Detailed Look at the Annual Averages, Weekly Cycles, Effects of Temperature, and Correlation With Surface NO₂ Concentrations. *Earth's future* 9, 4 (2021), e2020EF001665. <https://doi.org/10.1029/2020ef001665>
- [30] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. KNN Model-Based Approach in Classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Robert Meersman, Zahir Tari, and Douglas C. Schmidt (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 986–996.
- [31] Iolanda Ialongo, Henrik Virta, Henk Eskes, Jari Hovila, and John Douras. 2019. Comparison of TROPOMI/Sentinel-5 Precursor NO₂ observations with ground-based measurements in Helsinki. *Atmospheric Measurement Techniques* 13, 1 (2019), 205–218. <https://doi.org/10.5194/amt-13-205-2020>
- [32] IEA, IRENA, UNSD, World Bank, and WHO. 2021. *Tracking SDG 7: The Energy Progress Report (2021)*. Technical Report. World Bank, Washington DC. <https://www.irena.org/publications/2021/Jun/Tracking-SDG-7-2021>
- [33] Lyatt Jaeglé, Linda Steinberger, Randall V Martin, and Kelly Chance. 2005. Global partitioning of NO_x sources using satellite observations: Relative roles of fossil fuel combustion, biomass burning and soil emissions. *Faraday Discussions* 130, 0 (2005), 407. <https://doi.org/10.1039/b502128f>
- [34] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. 2019. Multivariate LSTM-FCNs for time series classification. *Neural Networks* 116 (2019), 237–245. <https://doi.org/10.1016/j.neunet.2019.04.014>
- [35] Patrick D.M.C. Katoto, Liliane Byamungu, Amanda S. Brand, Jolynne Mokaya, Hans Strijdom, Nandu Goswami, Patrick De Boever, Tim S. Nawrot, and Benoit Nemery. 2019. Ambient air pollution and health in Sub-Saharan Africa: Current evidence, perspectives and a call to action. *Environmental Research* 173 (2019), 174–188. <https://doi.org/10.1016/j.envres.2019.03.029>
- [36] Francis Kemausuor, Edwin Adkins, Isaac Adu-Poku, Abeeku Brew-Hammond, and Vijay Modi. 2014. Electrification planning using Network Planner tool: The case of Ghana. *Energy for Sustainable Development* 19 (2014), 92–101. <https://doi.org/10.1016/j.esd.2013.12.009>
- [37] Masami Kojima and Chris Trimble. 2016. *Making Power Affordable for Africa and Viable for Its Utilities*. Technical Report. World Bank. <https://doi.org/10.1596/25091>
- [38] Alexandros Korkovelos, Babak Khavari, Andreas Sahlberg, Mark Howells, and Christopher Arderne. 2019. The Role of Open Access Data in Geospatial Electrification Planning and the Achievement of SDG7. An OnSSET-Based Case Study for Malawi. *Energies* 12, 7 (2019), 1395. <https://doi.org/10.3390/en12071395>
- [39] Chuanjun Li, Latifur Khan, and Balakrishnan Prabhakaran. 2006. Real-Time Classification of Variable Length Multi-Attribute Motions. *Knowl. Inf. Syst.* 10, 2 (aug 2006), 163–183. <https://doi.org/10.1007/s10115-005-0223-8>
- [40] A Lorente, K F Boersma, H J Eskes, J P Veefkind, J H G M van Geffen, M B de Zeeuw, H A C Denier van der Gon, S Beirle, and M C Krol. 2019. Quantification of nitrogen oxides emissions from build-up of pollution over Paris with TROPOMI. *Scientific reports* 9, 1 (2019), 20033. <https://doi.org/10.1038/s41598-019-56428-5>
- [41] Enes Makalic and Daniel Francis Schmidt. 2011. Review of Modern Logistic Regression Methods with Application to Small and Medium Sample Size Problems. In *AI 2010: Advances in Artificial Intelligence*, Jiuyong Li (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 213–222.
- [42] Randall V. Martin, Daniel J. Jacob, Kelly Chance, Thomas P. Kurosu, Paul I. Palmer, and Matthew J. Evans. 2003. Global inventory of nitrogen oxide emissions constrained by space-based observations of NO₂ columns. *Journal of Geophysical Research* 108, D17 (2003), n/a–n/a. <https://doi.org/10.1029/2003jd003453>
- [43] Dimitrios Mentis, Mark Howells, Holger Rogner, Alexandros Korkovelos, Christopher Arderne, Eduardo Zepeda, Shahid Siyal, Costantinos Taliotis, Morgan Bazilian, Ad de Roo, Yann Tanvez, Alexandre Oudalov, and Ernst Scholtz. 2017. Lighting the World: the first application of an open source, spatial electrification tool (OnSSET) on Sub-Saharan Africa. *Environmental Research Letters* 12, 8 (2017), 085003. <https://doi.org/10.1088/1748-9326/aa7b29>
- [44] Ray Nassar, Timothy G Hill, Chris A McLinden, Debra Wunch, Dylan B A Jones, and David Crisp. 2017. Quantifying CO₂ Emissions From Individual Power Plants From Space. *Geophysical Research Letters* 44, 19 (2017), 10,045–10,053. <https://doi.org/10.1002/2017gl074702>
- [45] Federal Democratic Republic of Ethiopia. 2019. *National Electrification Program 2.0. Integrated Planning for Universal Access*. Technical Report. Federal Democratic Republic of Ethiopia, Addis Ababa.
- [46] Mutlu Ozdogan and Garik Gutman. 2008. A new methodology to map irrigated areas using multi-temporal MODIS and ancillary data: An application example in the continental US. *Remote Sensing of Environment* 112, 9 (2008), 3520–3537. <https://doi.org/10.1016/j.rse.2008.04.010>
- [47] Pascal Prunet, Olivier Lezeaux, Claude Camy-Peyret, and Hervé Thevenon. 2020. Analysis of the NO₂ tropospheric product from S5P TROPOMI for monitoring pollution at city scale. *City and Environment Interactions* 8 (2020), 100051. <https://doi.org/10.1016/j.cacint.2020.100051>
- [48] Payam Refaailzadeh, Lei Tang, and Huan Liu. 2016. Encyclopedia of Database Systems. , 7 pages. https://doi.org/10.1007/978-1-4899-7993-3_565-2
- [49] Juan José Rodríguez, Carlos J. Alonso, and José A. Maestro. 2005. Support vector machines of interval-based features for time series classification. *Knowledge-Based Systems* 18, 4 (2005), 171–178. <https://doi.org/10.1016/j.knosys.2004.10.007>
- [50] Guzman Santafe, Iñaki Inza, and Jose A Lozano. 2015. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review* 44, 4 (2015), 467–508. <https://doi.org/10.1007/s10462-015-9433-y>
- [51] Marciano Saraiva, Eglen Protas, Moisés Salgado, and Carlos Souza. 2020. Automatic Mapping of Center Pivot Irrigation Systems from Satellite Images Using Deep Learning. *Remote Sensing* 12, 3 (2020), n/a–n/a. <https://doi.org/10.3390/rs12030558>
- [52] Gautam Kumar Saw, Sagnik Dey, Hemant Kaushal, and Kanhaiya Lal. 2021. Tracking NO₂ emission from thermal power plants in North India using TROPOMI data. *Atmospheric Environment* 259 (2021), 118514. <https://doi.org/10.1016/j.atmosenv.2021.118514>
- [53] Petra Schmitter, Kefyalew S. Kibret, Nicole Lefore, and Jennie Barron. 2018. Suitability mapping framework for solar photovoltaic pumps for smallholder farmers in sub-Saharan Africa. *Applied Geography* 94 (2018), 41–57. <https://doi.org/10.1016/j.apgeog.2018.02.008>
- [54] scikit-learn developers. 2007. *Support Vector Machines*. scikit-learn developers. Retrieved January 28, 2022 from <https://scikit-learn.org/stable/modules/svm.html>
- [55] Skyler Seto, Wenyu Zhang, and Yichen Zhou. 2015. Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition. *CoRR abs/1512.06747* (2015), n/a–n/a. [arXiv:1512.06747](http://arxiv.org/abs/1512.06747)

- [56] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- [57] Benjamin K. Sovacool, Christopher Cooper, Morgan Bazilian, Katie Johnson, David Zoppo, Shannon Clarke, Jay Eidsness, Meredith Crafton, Thiagarajan Velumail, and Hilal A. Raza. 2012. What moves and works: Broadening the consideration of energy poverty. *Energy Policy* 42 (2012), 715–719. <https://doi.org/10.1016/j.enpol.2011.12.007>
- [58] Evan Thomas, Daniel Wilson, Styvers Kathuni, Anna Libey, Pranav Chintalapati, and Jeremy Coyle. 2021. A contribution to drought resilience in East Africa through groundwater pump monitoring informed by in-situ instrumentation, remote sensing and ensemble machine learning. *Science of The Total Environment* 780 (2021), 146486. <https://doi.org/10.1016/j.scitotenv.2021.146486>
- [59] USGS. 2020. NASA Shuttle Radar Topography Mission Global 1 arc second. Retrieved January 28, 2022 from <https://lpdaac.usgs.gov/products/srtmgl1v003/>
- [60] Lorenzo Vergni, Alessandra Vinci, Francesca Todisco, Francesco Saverio Santaga, and Marco Vizzari. 2021. Comparing Sentinel-1, Sentinel-2, and Landsat-8 data in the early recognition of irrigated areas in central Italy. *Journal of Agricultural Engineering* 52, 4 (2021), n/a–n/a. <https://doi.org/10.4081/jae.2021.1265>
- [61] Tijn Verhoelst, Steven Compernelle, Gaia Pinaridi, Jean-Christopher Lambert, Henk J. Eskes, Kai-Uwe Eichmann, Ann Mari Fjæraa, José Granville, Sander Niemeijer, Alexander Cede, Martin Tiefengraber, François Hendrick, Andrea Pazmiño, Alkiviadis Bais, Ariane Bazureau, K. Folkert Boersma, Kristof Bognar, Angelika Dehn, Sebastian Donner, Aleksandr Elokhov, Manuel Gebetsberger, Florence Goutail, Michel Grutter de la Mora, Aleksandr Gruzdev, Myrto Gratsa, Georg H. Hansen, Hitoshi Irie, Nis Jepsen, Yugo Kanaya, Dimitris Karagkiozidis, Rigel Kivi, Karin Kreher, Pieter F. Levelt, Cheng Liu, Moritz Müller, Monica Navarro Comas, Ankie J. M. Piters, Jean-Pierre Pommereau, Thierry Portafaix, Cristina Prados-Roman, Olga Puertedura, Richard Querel, Julia Remmers, Andreas Richter, John Rimmer, Claudia Rivera Cárdenas, Lidia Saavedra de Miguel, Valery P. Sinyakov, Wolfgang Stremme, Kimberly Strong, Michel Van Roozendaal, J. Pepijn Veefkind, Thomas Wagner, Folkard Wittrock, Margarita Yela González, and Claus Zehner. 2020. Ground-based validation of the Copernicus Sentinel-5P TROPOMI NO₂ measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandora global networks. *Atmospheric Measurement Techniques* 14, 1 (2020), 481–510. <https://doi.org/10.5194/amt-14-481-2021>
- [62] Marco Vriens, Nathan Bosch, Chad Vidden, and Jason Talwar. 2022. Prediction and profitability in market segmentation typing tools. *Journal of Marketing Analytics* n/a (2022), 1–30. <https://doi.org/10.1057/s41270-021-00145-4>
- [63] Chunjiao Wang, Ting Wang, Pucui Wang, and Vadim Rakitin. 2020. Comparison and Validation of TROPOMI and OMI NO₂ Observations over China. *Atmosphere* 11, 6 (2020), 636. <https://doi.org/10.3390/atmos11060636>
- [64] Lin Wang, Zhigang Wang, and Shan Liu. 2016. An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. *Expert Systems with Applications* 43 (2016), 237–249. <https://doi.org/10.1016/j.eswa.2015.08.055>
- [65] Wendan Wang, Danpei Zhao, and Zhiguo Jiang. 2017. Oil Tank Detection via Target-Driven Learning Saliency Model. *2017 4th LAPR Asian Conference on Pattern Recognition (ACPR)* n/a (2017), 156–161. <https://doi.org/10.1109/acpr.2017.70>
- [66] Martijn de Ruyter de Wildt, Henk Eskes, and K Folkert Boersma. 2012. The global economic cycle and satellite-derived NO₂ trends over shipping lanes: ECONOMY CYCLE IN SHIPPING NO₂ FROM SPACE. *Geophysical Research Letters* 39, 1 (2012), n/a–n/a. <https://doi.org/10.1029/2011gl049541>
- [67] Martin Wistuba, Josif Grabocka, and Lars Schmidt-Thieme. 2015. Ultra-Fast Shapelets for Time Series Classification. <https://doi.org/10.48550/ARXIV.1503.05018>
- [68] Yuan Yao, Zhiguo Jiang, Haopeng Zhang, Bowen Cai, Gang Meng, and Deshan Zuo. 2017. Chimney and condensing tower detection based on faster R-CNN in high resolution remote sensing images. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, n/a, 3329–3332. <https://doi.org/10.1109/IGARSS.2017.8127710>
- [69] Yuan Yao, Zhiguo Jiang, Haopeng Zhang, Danpei Zhao, and Bowen Cai. 2017. Ship detection in optical remote sensing images based on deep convolutional neural networks. *Journal of Applied Remote Sensing* 11, 4 (2017), 042611. <https://doi.org/10.1117/1.jrs.11.042611>
- [70] Haopeng Zhang and Qin Deng. 2019. Deep Learning Based Fossil-Fuel Power Plant Monitoring in High Resolution Remote Sensing Images: A Comparative Study. *Remote Sensing* 11, 9 (2019), 1117. <https://doi.org/10.3390/rs11091117>
- [71] Lu Zhang, Zhenwei Shi, and Jun Wu. 2015. A Hierarchical Oil Tank Detector With Deep Surrounding Features for High-Resolution Optical Satellite Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8, 10 (2015), 4895–4909. <https://doi.org/10.1109/jstars.2015.2467377>
- [72] Xiaoyi Zhao, Debora Griffin, Vitali Fioletov, Chris McLinden, Alexander Cede, Martin Tiefengraber, Moritz Müller, Kristof Bognar, Kimberly Strong, Folkert Boersma, Henk Eskes, Jonathan Davies, Akira Ogyu, and Sum Chi Lee. 2020. Assessment of the quality of TROPOMI high-spatial-resolution NO₂ data products in the Greater Toronto Area. *Atmospheric Measurement Techniques* 13, 4 (2020), 2131–2159. <https://doi.org/10.5194/amt-13-2131-2020>
- [73] Naomi Zimmerman, Albert A Presto, Srinivasa P N Kumar, Jason Gu, Aliaksei Haurlyliuk, Ellis S Robinson, Allen L Robinson, and Ramachandran Subramanian. 2017. Closing the gap on lower cost air quality monitoring: machine learning calibration models to improve low-cost sensor performance. *Atmospheric Measurement Techniques Discussions* n/a (2017), 1–36. <https://doi.org/10.5194/amt-2017-260>