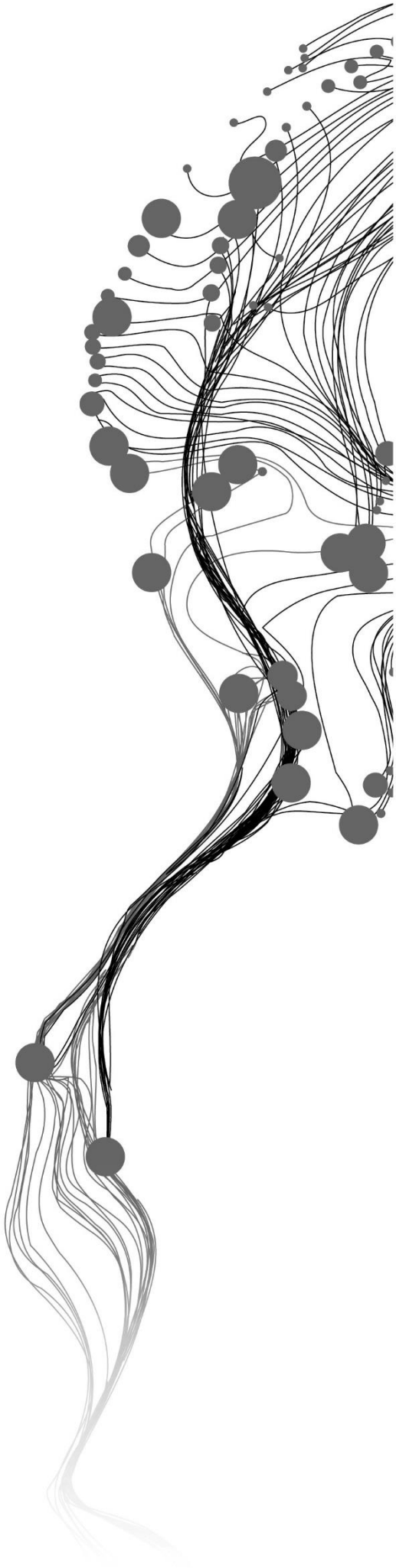
Estimating micronutrient concentrations in maize grains with Sentinel-1 and -2 images

JAYKUMAR HARISHBHAI GOHIL
June, 2023

SUPERVISORS:
dr. M. Belgiu
dr. M.T. Marshall



Estimating micronutrient concentrations in maize grains with Sentinel-1 and -2 images

JAYKUMAR HARISHBHAI GOHIL
Enschede, The Netherlands, June, 2023

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Natural Resource Management

SUPERVISORS:
dr. M. Belgiu
dr. M.T. Marshall

THESIS ASSESSMENT BOARD:
prof.dr.ir. A. STEIN
Prof. Murray Lark (University of Nottingham)

ABSTRACT

Hunger remains a major problem for low and middle-income countries. Moreover, Micronutrient deficiencies(MNDs) or hidden hunger are fatal and prevalent in these regions. Billions, including children, suffer from MNDs. In trace amounts, micronutrients like Selenium, Calcium, Iron, and others are found in the human body. Measurement of these micronutrients is vital, but the current methods available for testing and measuring are time-consuming and expensive. Hence, a method tackling both setbacks from prior techniques is the need of the hour. Several scientists have taken the help of field and crop samples to spatial map these micronutrition concentrations using spatial statistics.

This study aimed to develop a machine learning method comprising remote sensor data from Sentinel 1 and Sentinel 2 and other ancillary information like topographic, climatic, and soil characteristics data for micronutrient concentration in Calcium, Iron, Magnesium, and Zinc. To do this, we take the help of GeoNutrition Surveys as our reference data and use the Random Forest model for building the model. The study area for the research is Malawi. We use the `GEEMAP` library and GEE Python API to fetch all the datasets but the Sentinel 2 L1C, which we used Copernicus hub. L2A Sen2Cor algorithm was used for the atmospheric correction. After filtering many data points with inconsistent data or NaN values. After several combinations of the models, R^2 accuracy for each model was Calcium(0.25), Iron(0.25), Magnesium(0.29), and Zinc(0.23). Finally, partial dependence plots indicated that topographic and climatic features have the highest correlation with micronutrient concentrations. SWIR band depicted most dependency among the spectral features. Cloud cover and the consequences of these missing data were major limitations.

From these findings, this study is a foundation for using spectral and polarimetric features with machine learning techniques for micronutrient concentrations in tropical settings. Moreover, further development in terms of improving models via deep learning could be done. Using UAVs for acquiring spectral features is an experiment for future use.

Keywords: Sentinel-1, Sentinel-2, Worldclim, GeoNutrition, MERIT DEM, spectral, hidden hunger, MNDs, Polarimetric, Random Forest

ACKNOWLEDGEMENTS

I would like to express gratitude to my supervisors, Dr. Mariana Belgiu and Dr. Michael Marshall, for their constant support and guidance throughout the MSc research. Since the very beginning of my research, they have shown faith in me and my skills. I remember several instances where they applauded my efforts and steered my boat in the correct direction when needed. I am thankful to them for setting the bar high to motivate me further. Dr. Mariana demonstrated steadfast faith in my talents and continuous availability and readiness to give help whenever I met challenges. Their understanding, kindness, and approachability made me feel at ease presenting my ideas, asking for guidance, and getting constructive criticism. Their assistance has been vital throughout the research, and I am thankful for their commitment to my academic and personal development. In addition, want to thank Dr. Michael for their intelligent insights, thought-provoking conversations, and rigorous attention to detail. Their knowledge of agricultural remote sensing and willingness to ask probing questions aided in refining my study focus and improving the overall quality of my work. I am grateful to them for their dedication to cultivating an environment of intellectual curiosity and for encouraging me to broaden my views.

I would like to thank my thesis assessment board chair – Dr. Alfred Stein. He helped me better understand the research problem through constructive criticism and questions. I also thank Dr. Lyndon Estes for accepting me as a research intern at Dept. of Geography at Clark University—he and Dr. Michael are the reason why I could live my ten-year-old dream of being in a university in Massachusetts and offering an internship in an exciting branch of remote sensing, i.e., Land cover land use mapping.

This research would have been ongoing if not for the availability of extensive cloud computing support given by the CRIB geospatial computing hub. I thank Dr. Serkan Girgin for this robust and reliable platform. I would also acknowledge the developers providing open-source software and platform like Python and QGIS. They are used for the majority of my work.

Lastly, I thank my family and friends, both here in the Netherlands and back in India, for their support over the span of two years. This would not have been possible without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

The global impact of hunger is severe, with approximately one in every nine individuals facing hunger (FAO, 2020). The World Health Organization (WHO) reports that nearly 45% of child deaths under the age of five are associated with undernutrition, a prevalent malnutrition in low- and middle-income countries (WHO, 2021). Micronutrient deficiencies (MNDs) are one of several types of undernutrition, although the symptoms are not as obvious as those of other types, which is why it is referred to as "Hidden Hunger" (Muthayya et al., 2013). Alarmingly, it is estimated that approximately 1.5 billion individuals worldwide are affected by one or more types of MNDs as Iron (Fe), Zinc (Zn), Iodine (I), Calcium (Ca), or Selenium (Se) deficiencies(FAO, 2018; Saka et al., 2013). These deficiencies commonly occur in regions where grain-based diets dominate, with limited access to nutrient-rich plant and animal foods (Bouis & Saltzman, 2017). Sub-Saharan Africa and Southeast Asia are particularly affected by MNDs (Black et al., 2013), with Sub-Saharan women being disproportionately impacted, subsequently affecting the nutritional well-being of their children as they bear the primary responsibility for caregiving (Conti et al., 2019). Clinical complications can arise, such as anemia resulting from Fe deficiency (Camaschella, 2015) or hypocalcemia due to Ca deficiency (Ross et al., 2011).

MNDs are intricately linked to various health issues. For instance, a deficiency in Zn can result in developmental delays, loss of appetite, impaired immune function, as well as hair loss, diarrhea, delayed sexual development, impotence, hypogonadism in males, and severe cases may even lead to eye and skin diseases. Additionally, weight loss, slowed wound healing, taste abnormalities, and mental lethargy can manifest as potential symptoms (Heyneman, 1996; Prasad, 2004). Inadequate intake of Ca can lead to weakened bones and osteoporosis, characterized by brittle bones and an increased risk of fractures. Ca deficiency may cause rickets in children, while adults may develop other bone diseases. It is worth noting that vitamin D deficiency is more prevalent with Ca deficiency. In children with rickets, the growth cartilage typically fails to mineralize, resulting in irreversible changes to the skeletal structure. Osteomalacia, a condition characterized by poor bone mineralization and softening, can occur in adults and children due to Ca insufficiency (Ross et al., 2011). Chronic latent Magnesium (Mg) deficiency has been associated with atherosclerosis, myocardial infarction, hypertension, malignant tumors, kidney stones, changes in blood lipids, premenstrual syndrome, and psychological conditions (Jahnen-Dechent & Ketteler, 2012).

Micronutrient measurements are typically conducted using biomarkers, which involve assessing micronutrient levels or enzyme activity in human blood and tissues. These biomarkers serve as status indicators and are widely used to evaluate population health (Fairweather-Tait, 2011; Gödecke et al., 2018; King et al., 2015). Establishing sufficient biomarker thresholds poses challenges due to variations in "healthy" ranges among different demographic groups, the influence of physiological buffering, and the immediate impact of infection and inflammation on micronutrient concentrations (Jamison et al., 2006; King et al., 2015).

Biomarker investigations present additional difficulties in terms of logistics and data management. Volunteers face challenges in providing blood or tissue samples, while technical obstacles involve

ensuring stable laboratory hygiene and constant refrigeration, particularly in low-income countries. Alternative approaches for predicting risks of micronutrient deficiencies include analyzing food consumption data. However, these strategies are not only time-consuming but also uncertain due to the regional variations in food quality (Hurst et al., 2013; Joy, 2015; OI Bermudez, 2012).

Although laboratory analyses offer detailed insights into micronutrient deficiencies, they have limitations, as their implementation could be more consistent across different spatial scales. Furthermore, the complexities and challenges associated with these analyses highlight the need for further research and alternative approaches to address micronutrient deficiencies effectively.

Measurement of nutrient concentrations for the consumed crops is usually done using the lab analysis (wet chemistry) on the grains (Huang et al., 2020; Ibrahim et al., 2022; Wang et al., 2013). Other studies used hyperspectral imaging of the grains themselves to assess the nutrients in the grains (Grieco et al., 2021; Herzig et al., 2019 and Mohammadi Moghaddam et al., 2013). The benefit from hyperspectral imaging over the chemical test is the non-destructive assessment of micronutrients. Yet, they both require expensive equipment and trained workers to assess the sample of the grains. Consequently, it is crucial to develop efficient methods for obtaining accurate information on crop nutrient levels and their spatial and temporal variability.

Although laboratory analyses through wet chemistry of the micronutrient concentration of the grains offer detailed insights into potential grain micronutrient deficiencies, this method is expensive and time-consuming and, consequently, cannot be applied across time and large areas. These challenges call for further research and alternative approaches to measure potential micronutrient deficiencies in crops effectively.

Remote sensing data have been successfully used in the past to predict nutrients in the soil, crop canopy, and grains. For example, Kaur et al. (2020) employed combined optical remote sensing data from Landsat-8 and Sentinel-2, along with climate data and ground truth values, to predict soil N, Potassium (K), Phosphorus (P), and Organic Carbon (OC) for select districts in Maharashtra, India. Various linear and non-linear regression models, including Multiple Linear Regression (MLR), Random Forest Regression (RFR), Support Vector Machine for Regression (SVR), and Gradient Boosting (GB), were examined. The findings indicated that RFR and GB outperformed other techniques. Forkuor et al. (2017) focused on mapping the spatial distribution of six soil characteristics including silt, clay, cation exchange capacity (CEC), soil organic carbon (SOC), and N in a 580 km^2 agricultural watershed in south-western Burkina Faso using RapidEye and Landsat images, terrain data, and climatic data. Four statistical prediction models were evaluated and compared, including Multiple Linear Regression (MLR), RFR, SVR, and Stochastic Gradient Boosting (SGB). The results indicated that MLR performed slightly better than the machine learning approaches, with RFR consistently demonstrating the highest accuracy. Similarly, Zhou et al. (2020) conducted a study on assessing different remote sensing sensors such as Synthetic Aperture Radar (SAR) and optical high-resolution imagery to predict SOC and They found the Sentinel-1 and Sentinel-2 based features obtained better R^2 accuracies than others. The results suggest the potential significance of employing Sentinel-1 and Sentinel-2 imagery for estimating nutrient concentrations.

Botoman et al. (2022) conducted a study focusing on the prediction of Zn content in maize grain using soil parameters and environmental factors. They employed a linear mixed model (LMM),

incorporating expert rankings and false discovery rate (FDR) corrections. The predictors included soluble Zn, soil pH, total Zn, potentially present Zn, soil organic carbon, oxalates, effective cation exchange capacity, mean annual precipitation, mean annual temperature, slope, topographic index, and enhanced vegetation index(EVI) from MODIS dataset. EVI is a more sensitive vegetation index to biomass, atmospheric backdrop, and soil condition (Huete et al., 2002a). They found evidence for justifying Similarly, Gashu et al. (2020) conducted a study aiming to spatially predict the concentration of selenium in grains(teff and wheat) within the Amhara region of Ethiopia. They utilized LMM framework, incorporating soil parameters and environmental factors. In addition to these factors, they also included remotely sensed data, specifically bands from Visible region and shortwave infrared regions from the MODIS sensor and the EVI derived from the same. Furthermore, Gashu et al. (2021) revealed a high correlation between soil pH, SOC, and micronutrient concentration for Ca, Fe, Se and Zn in Ethiopia and Malawi. In summary, these studies highlight the utilization of various soil parameters, environmental factors, and remotely sensed data within the LMM framework for predicting micronutrient concentrations in grains. Incorporating these additional predictors like soil parameters and environmental could improve the accuracy and reliability of the predictive models.

Belgiu et al. (2023) used hyperspectral satellite-borne imagery from PRecursore IperSpettrale della Missione Applicativa (PRISMA) to predict micronutrient estimation for Ca, Fe, Mg and Zn for crop wheat, corn, soy and paddy. They predicted the R^2 accuracy for micronutrient concentration between 0.49 and 0.58, emphasizing the potential of hyperspectral remote sensing in predicting micronutrients.

1.2. Problem Statement

Previous studies dedicated to predicting micronutrient concentrations in crop yields rely on medium-resolution satellite data, such as MODIS (Botoman et al., 2020, 2022; Gashu et al., 2021). This research investigates the potential of using RFR on Sentinel 1 and Sentinel 2 data for estimating the micronutrient concentration. Previous studies have primarily used spatial statistics techniques for estimating and mapping micronutrient concentrations (Botoman et al., 2020, 2022; Gashu et al., 2021). In this thesis, RFR will be used due to its capability to predict the non-linear relationship between nutrients and investigated predictors.

Creating detailed maps of MNDs nationally would be particularly beneficial for low- and middle-income countries, which experience a significant burden of MNDs (Black et al., 2013). The study specifically aims to assess the potential of remote sensing and environmental data to estimate nutrients of public health importance, such as Zn, Mg, Ca, and Fe. The analysis will focus on Malawi, where maize is the staple crop grown in the Northern, Central, and Southern regions (Gashu et al., 2021).

1.3. Research Objectives and Questions

Overall Objective: The main aim of this research is to evaluate the potential of remote sensing images and other ancillary geoinformation such as topographic, soil characteristics and climatological data, and to estimate the concentration of micro-nutrients, i.e., Fe, Ca, Mg, and Zn, the grains of maize cultivated in Malawi.

Sub-Objective-1: Evaluate the performance of radar and optical data for estimating micro-nutrient concentrations.

Research Question-1-1: What is the importance of Sentinel-1 and Sentinel-2 spectral and polarimetric features for estimating micro-nutrient concentrations?

Sub-Objective-2: Evaluate the performance of Digital Elevation Model-based derivatives for estimating micro-nutrient concentrations.

Research Question 2-1: What is the importance of topographic features derived from the slope, elevation, and Topographic wetness index relative to Sentinel-1 and Sentinel-2 predictors in estimating micronutrient concentration?

Sub-Objective-3: Evaluate the performance of soil data for predicting micro-nutrient concentrations.

Research Question 3-1: What is the importance of soil parameters relative to Sentinel-1 and Sentinel-2 predictors in estimating micro-nutrient concentration?

Sub-Objective-4: Evaluate the performance of Climatology data like precipitation and temperature as features for estimating micro-nutrient concentrations.

Research Question 4-1: What is the importance of climatological data like temperature and precipitation relative to Sentinel-1 and Sentinel-2 predictors in estimating micronutrient concentration?

2. RELATED WORK

2.1. Spectral and Polarimetric features for (Micro)nutrient Concentration

Many studies have assessed the potential of various remote sensing images for predicting nutrients in the soil. For example, Kaur et al. (2020) used Landsat 8 and Sentinel 2 to predict macronutrients such as N, K, P, and OC in the soils of Maharashtra, India. Remote sensing images have also been successfully used to predict various nutrients in the crop canopy. Rossi et al. (2022) uses Sentinel 2 data for estimating N in rice and maize crops. Studies such as Fu et al. (2020), Li et al. (2014), Ling et al. (2019) and Sharifi (2020) gives evidence for usage of crop canopy in estimating nutrient concentration using remotely sensed inputs.

Other techniques using spectral features are through reflectance spectrometry in the lab or having hyperspectral reflectance measurements using airborne or satellite imagery. Studies like Curran, (1989), Graeff & Claupein, (2003) and Pandey et al. (2017) are an example of using reflectance to measure the amount of nutrients like Nitrogen, Phosphorus, Ca, and Sulphur are detected in various crops like maize an dother plant canopies. Whereas other research by Belgiu et al. (2023), Li et al. (2018), Ling et al. (2019), Liu et al. (2017) and Marang et al. (2021) depicted on use of hyperspectral data for estimating and monitoring nutrients like Mg, N and P, Ca, and K available in various crops.

On the other hand, polarimetric features have a certain major benefit over Multispectral satellite-based sensors as clouds have little effect on radar backscatter readings, which are sensitive to surface roughness and moisture content and can be used to portray the structure of a feature. Because of the thick trunks, leaves, and branches, highly structured elements such as the forest reflect most energy and look brighter. Conversely, barren land has darker characteristics because it reflects the signal away from the antenna in a different direction (ESA, 2020). Studies by Lapaz Olveira et al. (2023) and Munir et al. (2022) are recent developments on how C band SAR data is used to estimate nitrogen from the canopy of maize and oil palm, respectively.

Recently, Belgiu et al. (2023) investigated the potential of hyperspectral data, namely, PRISMA and Sentinel 2 to estimate the composition of macro and micronutrients in crop harvest. The findings of the research indicated that the utilization of remote sensing imagery to estimate the nutritional content of crops has the capacity to provide cost-efficient, timely, and spatially detailed assessment of the nutritional values of crops.

Various studies used remote sensing covariates together with other environmental factors to predict nutrient concentrations in staple crops. For example, Gashu et al., (2020) predicted Se with LMM framework, incorporating similar soil parameters and environmental factors. In addition to these factors, they also included remotely sensed data.

2.2. DEM and its derivatives for Micronutrient Concentration

The region's topographic characteristics considerably influence the spatial distribution of soil qualities (Suleymanov et al., 2021). Farming activities in specific field distribution patterns are may be influenced by the topography features of the geographical area (Husak et al., 2008). Moreover, Botoman et al. (2022) and Gashu et al. (2020) used elevation and its cognates along with other

variables to estimate concentrations of Zn and Se, respectively, and revealed that topography influences the prediction of their respective micronutrient concentration.

2.3. Soil parameters for Micronutrient Concentration

Soil properties such as Soil pH significantly impact soil biogeochemical processes in the natural environment. Soil pH is therefore called the "master soil variable" since it controls many biological, chemical, and physical qualities and processes affecting plant development and biomass output (Neina, 2019). Research by Farwa et al. (2020) where both predict macronutrients like N, P, and K using only soil parameters, whereas studies by Botoman et al. (2022) and Gashu et al. (2020) use soil pH, SOC and oxalates to estimate the concentration of Zn and Se respectively and showed that soil characteristics influenced the prediction of their respective micronutrient concentrations.

2.4. Climatological parameters for Micronutrient Concentration

It is well-established that various environmental elements, including rainfall and average temperature, profoundly shape plants' development and nutritional state as they impact plants’ physiology. These influential factors hold considerable sway in determining the levels of macronutrients present within plants (Köhler et al., 2019). Noteworthy investigations conducted by Botoman et al. (2022) and Gashu et al. (2020) have employed annual downscaled mean temperature and downscaled mean precipitation data for mapping Zn in Malawi and Selenium in Ethiopia, respectively. Both studies have utilized the CHELSA dataset (Karger et al., 2017). The incorporation of climatological parameters for the assessment of micronutrients yields significant insights and implications.

3. STUDY AREA AND DATA USED

3.1. Study Area

The research encompasses the entire country of Malawi as the study area (see Fig. 1). Malawi is located in eastern southern Africa, positioned between latitude 9°22′ and 17°7′ South and longitude 32°40′ and 35°55′ East. It spans a land area of 94,275 square kilometers. Malawi is a landlocked country bordered by Tanzania, Mozambique, and Zambia. The country's topography is highly diverse, with the Great Rift Valley running from north to south and encompassing Lake Malawi. The surrounding regions of the Great Rift valley feature vast plateaus situated at elevations ranging from 800-1,200 meters up to 3,000 meters (Saka et al., 2013).

Malawi experiences a subtropical climate characterized by dry and seasonal conditions. The warm-wet season occurs from November to April, accounting for approximately 95% of the annual precipitation. This period coincides with the maize crop calendar (Botoman et al., 2020).Other crops grown during the same season are cotton, millet, paddy and sorghum. The average annual rainfall varies between 725mm and 2,500mm, with Lilongwe receiving around 900mm, Blantyre receiving 1,127mm, Mzuzu receiving 1,289mm, and Zomba receiving 1,433mm. The lower Shire Valley and certain villages in the Salima and Karonga regions are susceptible to flooding. A cold and dry winter season prevails from May to August, with mean temperatures ranging from 17 to 27 degrees Celsius and occasionally dropping to 4 to 10 degrees Celsius in isolated areas, with frost occurrences during June and July. September through October is characterized by hot and dry weather, with typical temperatures ranging from 25 to 37 degrees Celsius. Humidity levels range from 50% to 87% during the drier months of September/October and the rainy months of January/February (Ministry of Forestry and Natural Resources, n.d.).

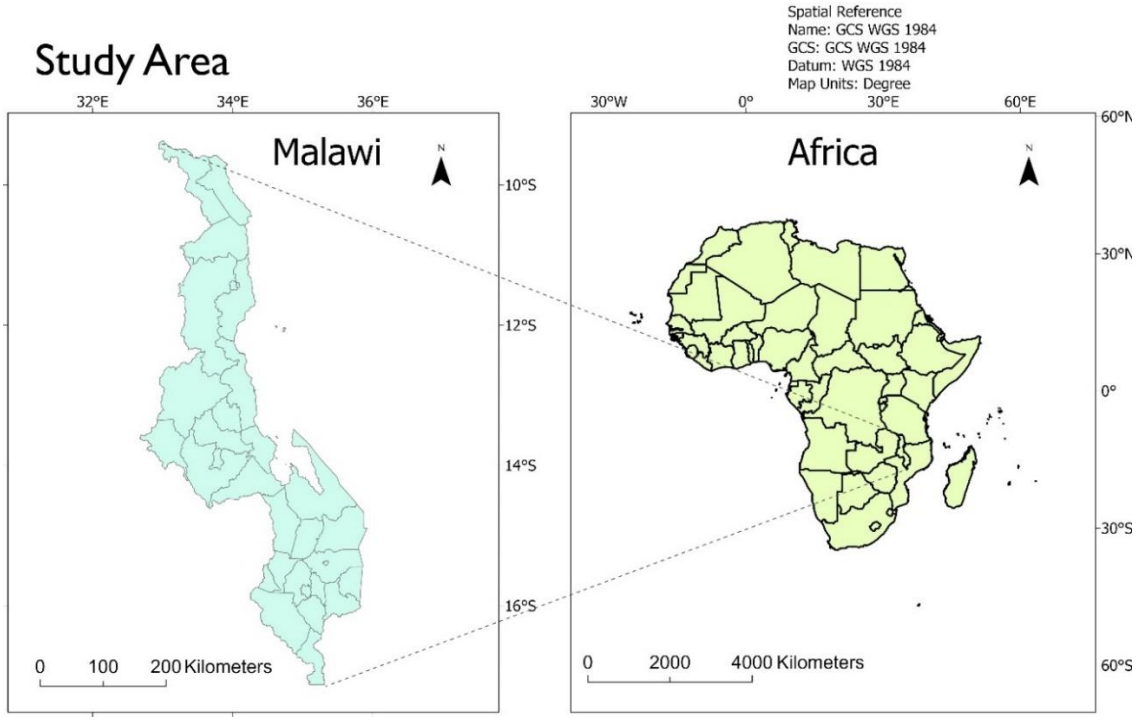


Figure 1 Study area map with district and country boundaries in Malawi and Africa, respectively

As of 2018, the estimated population of Malawi stood at 17.5 million people, with approximately 84% residing in rural areas. The country's economy is vulnerable to external shocks due to its reliance on inadequate irrigation systems and low agricultural productivity. Agriculture plays a central role in Malawi's economy, contributing 30% to the GDP and accounting for more than 80% of national export revenues (JICA, 2022). This sector employs 64% of the country's workforce and plays a crucial role in ensuring food security. Over the past few decades, the percentage of land used for agriculture has steadily increased, with approximately 61% of Malawi's total land area currently dedicated to crop cultivation and livestock grazing. The production of maize and groundnuts, which are the main staple foods, dominates the agricultural sector (FAO, 2018). Agriculture in Malawi heavily relies on rain-fed farming practices, with limited production and consumption of animal products. Consequently, the country faces persistent food shortages at both the national and household levels (JICA, 2022).

3.2. Dataset

The research used two types of datasets. The first type comprises raster imagery obtained from sources such as Sentinel 1, Sentinel 2, MERIT digital elevation model, African Soil Information Service (AfSIS) soil data, and WorldClim climatic data. The second type of data consists of vector data, specifically the Ground reference data collected in the framework of the GeoNutrition project conducted by (Kumssa et al. 2022). The data includes Ca, Zn, Fe, and Mg measurements in various crops, including Maize.

Table 1 The table depicts various datasets, their resolutions, coverage and spectral information, VIS is Visible bands(RGB), NIR(Near Infrared) and SWIR(Short wave infrared)

Dataset & Source	Spatial Resolution	Temporal Resolution	Temporal Coverage (Time Period)	Spectral Information
Sentinel - 2	10m & 20m	5 Days	2017-2018	VIS, NIR (10 m) Red Edge bands (20 m)and SWIR bands (20 m)
Sentinel - 1	5 x 20 m	6 Days	2017-2018	C band IW with VV/VH polarization
MERIT	90m	-	2017	-
WorldClim 2	1000m	-	1970-2000	-
AFSIS Soil Grid	250m	-	2008-2014	-
GeoNutrition Data	-	-	2018	-

3.2.1. Sentinel 1 and Sentinel 2

Sentinel-1a and Sentinel-1b C band Level-1 Ground Range Detected (GRD) SAR images with dual polarization Vertical transmit Vertical receiver (VV), and Vertical transmit Horizontal receiver (VH) (VV+VH) in Interferometric Wide swath (IW) acquisition mode has been used from the year 2017 to 2018. The temporal resolution of the S1a+S1b images is six days with a spatial resolution of 10m. No additional pre-processing has been done on the S1 image. Google Earth Engine (GEE) provides pre-processed S1 GRD images with GRD border noise removal, thermal noise removal, radiometric calibration, and terrain correction. We use the Python API for GEE with an additional library, GEEMAP (Wu, 2020). From Sentinel 1, we use the VV, VH and their polarimetric arithmetic products VV+VH, VV-VV,VV*VV, and VV/VH. Also, we use the Radar vegetation index(RVI) (Kim & Van Zyl, 2009), along with the aforementioned polarimetric independent variables. RVI is calculated as $8 * \frac{VH}{HH+VV+2*VH}$.

Also, RVI is well recognized for being sensitive to vegetation cover and biomass.

Sentinel-2 is a high-resolution, wide-swath imaging mission that supports Copernicus Land Monitoring investigations, such as monitoring plant, soil, and water cover, as well as observing inland rivers and coastal regions. 13 spectral bands are sampled by the Sentinel-2 Multispectral Instrument (MSI), including four bands at a spatial resolution of 10 meters, six bands at 20 meters, and three bands at a resolution of 60 meters. This study used Sentinel-2a and Sentinel-2b with a five-day interval with all bands at 20m (VIS, SWIR, RE) from the years 2017 to 2018 used. Cloud filtering was done using SCL, i.e., Scene Classification Layer, available in Sentinel 2. We performed Sen2Cor, was used for atmospheric correction of the images since only L1C data was available. We fetched data through Copernicus API Hub for large downloads. A total of 13 tiles/granules of Sentinel 2 data cover the entirety of Malawi. From Sentinel 2, we use all the bands available except B01(Ultra blue or Coastal), and deriving indices mentioned in Annexes 8.1.

3.2.2. MERIT Digital Elevation Model

MERIT DEM refers to a high-resolution global Digital Elevation Model (DEM) with a spatial resolution of approximately 3 arcseconds, which is equivalent to about 90 meters at the equator. It was developed by eliminating significant error components from existing DEMs, including the NASA SRTM3 DEM, JAXA AW3D DEM, and Panoramas DEM (Yamazaki et al., 2017). MERIT DEM utilizes multiple satellite datasets and employs various filtering techniques to address absolute bias, tracking noise, speckle noise, and tree height offset. By removing these error components, the accuracy of land area mapping with a vertical accuracy of 2 meters has significantly improved from 39% to 58%. This improvement is particularly notable in flat areas where elevation errors previously exceeded topographic variations. Additionally, the enhanced DEM effectively highlights terrains such as river networks and mountain-valley structures.

This study utilized the slope(degrees), elevation(m), and topographic wetness index (TWI) derived from MERIT DEM. TWI combines information about local upslope contributing area and slope to assess the influence of topography on hydrological processes. The index is calculated based on both the slope and the contributing area per unit width orthogonal to the flow direction.

3.2.3. Climatic data

WorldClim (Hijmans et al., 2005) is a spatially interpolated monthly climate data covering land areas across the globe. It has a spatial resolution of approximately 1 square kilometer. The dataset incorporates various climate variables, including monthly temperature (minimum, maximum, and average), precipitation, solar radiation, vapor pressure, and wind speed. To generate this dataset, data from a large number of weather stations, between 9,000 to 60,000 stations, were collected and

aggregated for the period between 1970 and 2000. The WorldClim dataset provides detailed climate information for different locations worldwide through sophisticated interpolation techniques.
This research had extracted the average temperature and accumulated precipitation over the growing season from the WorldClim dataset. These variables are of particular interest and relevance to the study, as they provide valuable insights into the climatic conditions during the period when crops are cultivated.

Table 2 Climatological variable used for model building

Variable	**Units**	**Source**	**Used in**
Average Temperature(over the season)	Degree Celsius	http://www.worldclim.com/version2	(Botoman et al., 2022; Gashu et al., 2020)
Accumulate**d** Precipitation(over the season)	mm		

3.2.4. AFSIS Soil Data

The AFSIS soil data (Hengl et al. 2015), provides estimates of various soil properties across the African continent. These properties include organic carbon content, pH level, fractions of sand, silt, and clay, presence of coarse debris, bulk density, cation exchange capacity, total N content, exchangeable acidity, and levels of exchangeable bases such as Ca, Mg, potassium, sodium, as well as extractable aluminum. The soil data is derived from an automated mapping framework with random forests. With a spatial resolution of 250 meters, the soil properties are estimated at two or six standard soil depths throughout the entire African continent.
For this research, the focus is on the soil depth of 100 centimeters. This depth is considered significant as it provides the most suitable environment for root growth (Gao et al. 2010). Furthermore, specific variables depicted in the table below will be used in the analysis.

Table 3 Soil variable used for building models

Variable	Units	Used in
Cation Exchange Capacity	Mmol/kg	(Botoman et al., 2022)
Soil Organic Carbon	dg/kg	(Botoman et al., 2022; Gashu et al., 2020)
pH	-	(Botoman et al., 2022; Gashu et al., 2020)
Organic Carbon Stock	t/ha	-
Total Nitrogen	cg/kg	(Gashu et al., 2020)

3.2.5. GeoNutrition survey data

The dataset provided by Kumssa et al. (2022) encompasses primary data from GeoNutrition surveys conducted in Ethiopia and Malawi. This dataset includes information on the concentrations of 29 mineral micronutrients such as Silver, Magnesium, Aluminium, Manganese, Arsenic, Molybdenum, Boron, Nickel, Barium, P, Beryllium, Lead, Ca, Rubidium, Cadmium, Sulfur, Cobalt, Selenium, Chromium, Strontium, Caesium, Thallium, Copper, Uranium, Fe, Vanadium, K, Zn, Lithium in grains and up to 8 soil chemistry characteristics. The concentrations of micronutrients in the grains were determined using inductively coupled plasma-mass spectrometry (ICP-MS) (Kumssa et al., 2022).
Sampling took place across Malawi during the harvest season from April to June 2018 to ensure national coverage. The distribution of concentration values for the micronutrients can be observed in Figure 2, providing insight into their variability. Mg has the highest concentration among all of the micronutrients and Fe has some outliers present. Additionally, Figure 3 illustrates the spatial distribution of the sampling points throughout the region.

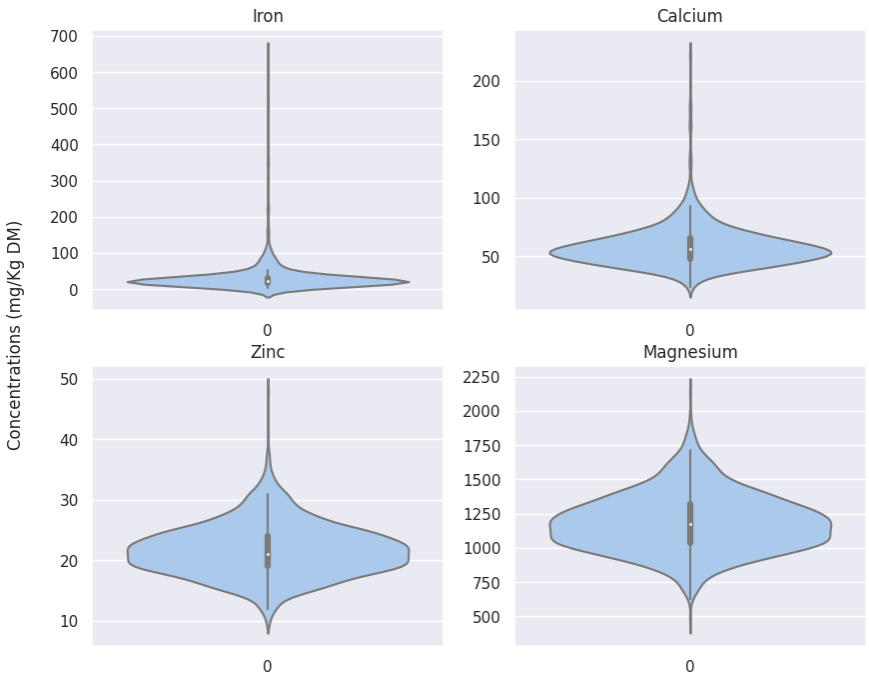


Figure 2 Combined violin and box and whisker plot for the micronutrient concentration in maize grain collected from Malawi

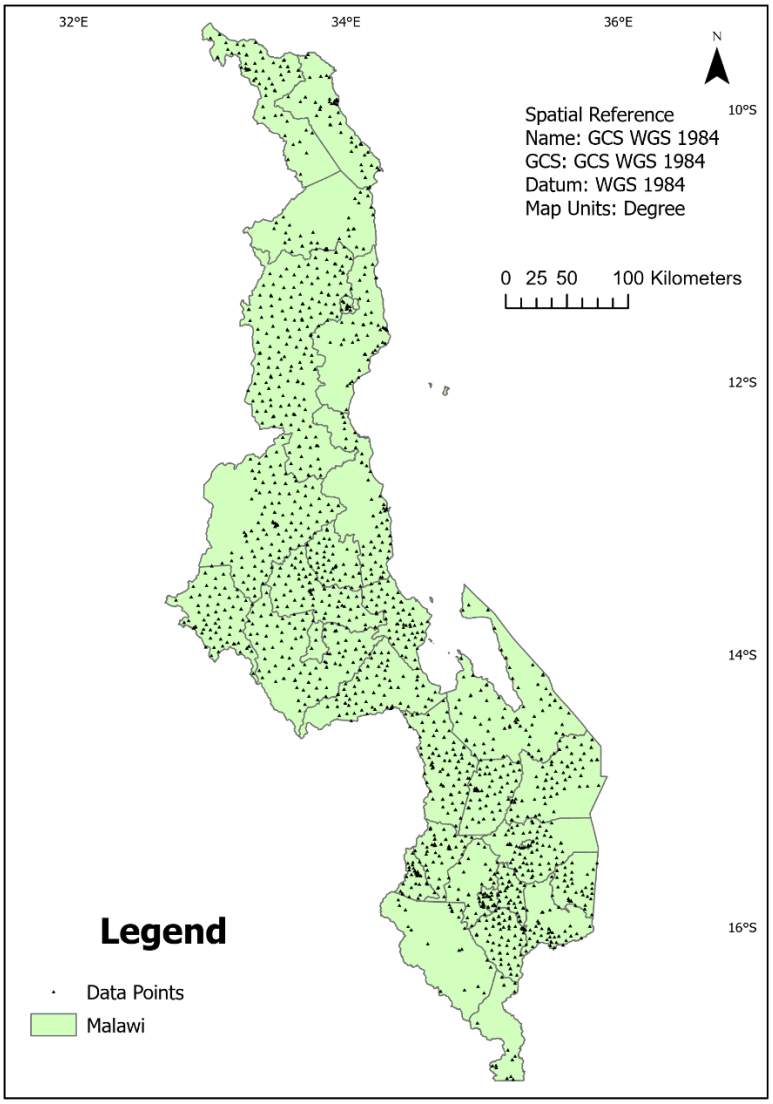


Figure 3 Spatial Distribution of the data points in Malawi representing locations from where grain and soil samples were taken.

4. METHODS

When discussing the methodology, the first step, as illustrated in the figure below, involves preparing the Sentinel 2 and Sentinel 1 data for sampling (extracting raster values). For Sentinel 2 data, atmospheric correction is necessary since only Top-of-Atmosphere reflectance (L1C products) are available for the Malawi region between November 2017 and April 2018. Sen2cor from ESA is employed for this purpose. Cloud masking is applied to the Sentinel 2 data using the SCL bands of the image. Sentinel 1 data is obtained using the GEE Python API (Gorelick et al., 2017). The polarimetric features are averaged over the growing season for each data point. Climatological data is also retrieved using the GEE Python API, specifically WorldClim (Hijmans et al., 2005), which provides average temperature and total rainfall throughout the growing season. The soil data utilized is the AFSIS soil grid with a spatial resolution of 250m (Hengl et al., 2017). Various soil characteristics, such as cation exchange capacity, soil organic carbon, organic carbon stock, nitrogen availability, and soil pH (water-soluble), are analyzed. Topographic features are derived from the MERIT DEM, which has a spatial resolution of 90m at the equator. Elevation, slope, and TWI (Topographic Wetness Index) are extracted from the MERIT DEM. Subsequently, all these data sources are sampled for the available data points. Feature reduction techniques are applied to eliminate excessive independent variables, and a RFR (Breiman, 2001) is trained using the significant independent variables to achieve improved model performance. Hyperparameter tuning uses the Grid Search approach (LaValle et al., 2004). Lastly, Partial Dependency Plots (Friedman, 2001) are used to visualize the influence of an individual or a couple of characteristics on the predicted outcome of the machine learning model.

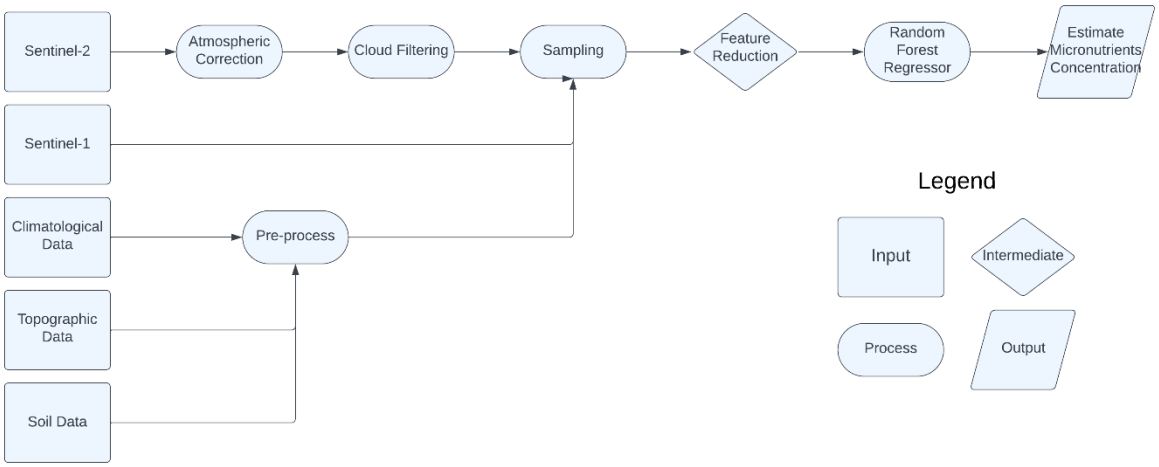


Figure 4 Flowchart depicting the data flow and the generalized methodology

4.1. Sentinel 2

Top of the Atmosphere (TOA) S2 (a+b) imagery captured between November 2017 and April 2018 was obtained through the Copernicus Hub API utilizing the sentinelsat library. Due to the unavailability of Bottom-of-the-Atmosphere(BOA) images, an atmospheric correction was performed on the TOA imagery using Sen2cor L2A, thereby generating BOA images. This correction process involved batch-processing via the terminal. For this study, a consistent spatial resolution of 20m was adopted for all Sentinel 2 bands. To effectively mask out clouds, an intuitive approach leveraging the Rasterio and numpy libraries in Python was employed. The Scene Classification Layer (SCL) excluded everything except vegetation, non-vegetation, and water bodies. It is worth noting that the SCL was initially designed for a 60m resolution; hence, an upscaling factor of 3 was applied to achieve a 20m resolution, which was subsequently utilized in the analysis.

In this study, once the images were ready, we sampled them in two ways, either just taking the reflectance of the particular pixel which coincides with the datapoint and another way, was by creating a buffer of 60m around the datapoint and averaging the reflectance, and finally, averaging them through the growing season; some variations, like dividing the total temporal period into three and two parts, i.e., making three and two-month composites, were also done. Finally, after the images were sampled, we removed the data points with no data, or too few observations i.e., not having at least one cloud free pixel per month. We would only use these data points for further analysis.

Regarding sampling design, as the study uses the sklearn library, which is unable to understand the NaN values as information for the model, especially for the Fe concentration, 439 data points are lost when using the dataset. From 1608 to 884 data points, nearly losing 55% of data was due to missing values and noisy and cloudy data. This raises questions on the generalization of the model. This leads to whether the number of samples is enough to train a robust model. For selection of valid data points, algorithm is used as follows:

the set of all locations as L, and the set of all months as M. We can express the valid locations where the total number of pixel per month is at least 1.

$$\{l \in L \mid (\forall m \in M)(x_{\{lm\}} \geq 1)\}$$

In this notation:

- l ∈ L represents that location l is an element of the set of all locations L.
- m ∈ M indicates that month m is an element of the set of all months M.
- x_{lm} represents the total number of pixel per month m at location l.

Also, we employed averaged bands for calculating spectral indices, which are widely used in vegetation/crop remote sensing shown in Annex 5.1. Therefore we used a total of 20 spectral features(9 raw bands and 11 indices) were used as independent variables for the model

4.2. Sentinel 1

The pre-processing steps, including generating a mean composite for the growing season and extracting the values, were conducted using the `GEEMAP` library and the GEE Python API. The data points obtained from the previous Sentinel 2 processing were utilized in this stage. As for the Sentinel 1 GRD, VV and VH data, which have a resolution of 10m, were resampled to a 20m resolution using

bilinear resampling. The S1 GRD data had already undergone radiometric and terrain correction. Lastly, we sample the VV and VH values for the data points. We also calculate and use polarimetric arithmetic like addition(VV+VH), ratio(VV/VH), and difference(VV-VH). A remark made during this was that VV and VH are logarithmic values, and so for calculating the other derived polarimetric values, this was taken care of.

Furthermore, we use the Radar Vegetation Index (Kim & Van Zyl, 2009) for its ability to be sensitive to crop cover and biomass. Overall, a total of six polarimetric independent variables for the model viz VV,VH,VV/VH,VV+VH,VV-VH, and RVI.

4.3. Climatological data

WorldClim climatological data was used in this research.. Also, as mentioned for Sentinel 1 data for this data too, we use `GEEMAP` with GEE Python API. The point to be noted was that this study used the mean average temperature over the growing season, and we aggregated the precipitation over the temporal period too. Finally, we sample them with the data points. Hence, from climatological data, we get Average temperature and accumulated precipitation as independent variables for the model.

4.4. Soil data

In this research, we use AFSIS soil data which comprises estimations of diverse soil characteristics across Africa. These characteristics encompass the content of organic carbon, pH level, proportions of sand, silt, and clay, presence of coarse debris, bulk density, cation exchange capacity, total nitrogen, exchangeable acidity, and levels of exchangeable bases such as Ca, Mg, potassium, sodium, as well as extractable aluminum. The spatial resolution of AFSIS is 250m. This research uses Cation Exchange Capacity, Soil Organic Carbon, Organic Carbon Content, N, and pH. Again, we use `GEEMAP` with GEE Python API for fetching and sampling. Here, we use only the first four bands of each soil characteristic since the depth of 100cm is reached. Therefore, we use a total of five independent variables for the model.

4.5. Topographic data

MERIT DEM was used in the study, which has a spatial resolution of 90m. We resampled to 20m resolution and used GEEMAP with GEE Python API. Also, using `ee.Terrain` package for computing slope. We use TWI to understand how topography influences hydrological processes. Therefore, we use three independent variables from the topographic data for the model.

4.6. Feature reduction

Feature reduction is crucial in reducing dataset size and eliminating unnecessary variables. Fortunately, several models are available to calculate the relevance of features, enabling us to disregard less useful ones. One such model is the RF proposed by Breiman (2001). In this study, we will utilize the `sklearn` library, developed by Pedregosa et al. (2011), to perform feature selection in Python.

The Recursive Feature Elimination (RFE) algorithm was employed, a feature selection/reduction technique based on the wrapper approach Guyon et al. (2002). This study used a blend of RFE and Cross-Validation(CV) called RFECV. RFECV, an acronym for Recursive Feature Elimination with Cross-Validation, represents a feature selection technique accessible within the scikit-learn (`sklearn`) library. RFECV integrates the concepts of Recursive Feature Elimination (RFE) and

cross-validation to automatically determine the ideal number of features for a given machine learning model.

The RFECV procedure commences by fitting the designated estimator (machine learning model) using the complete set of features. Subsequently, it progressively eliminates features through recursive iterations, considering their importance, and re-fits the model using the reduced feature set. The significance of each feature is assessed based on a predefined metric for feature importance, such as coefficient values for linear models or feature importance scores for tree-based models.

Following each round of feature elimination, RFECV employs cross-validation to assess the model's performance using the reduced feature set. Cross-validation partitions the data into multiple folds, training the model on a subset of the folds and evaluating its performance on the remaining fold. This process is repeated several (Total number of features/step size) times, and an average performance score is computed.

RFECV iterates the recursive elimination process until the specified number of features or a minimum threshold is reached. Throughout each elimination step, it tracks the cross-validated performance scores. This facilitates the identification of the optimal number of features that maximizes the model's performance. The RFECV algorithm provides both a ranking of feature importance and a visualization of the cross-validated performance scores plotted against the number of selected features. These outputs aid in comprehending the trade-off between feature selection and model performance, enabling informed decision-making regarding feature subset selection. In summary, RFECV integrates recursive feature elimination with cross-validation to automatically determine the optimal number of features that yield the highest model performance. It serves as a powerful tool for feature selection in machine learning tasks, providing valuable insights for enhanced decision-making.

For this study, we used three different RFECV models for each micronutrient model, changing one hyperparameter, i.e., the number of trees for the random forest model. The number of trees were 500, 1000, and 1500. At the end of this process, we save the ‘important’ features and build the model with them.

4.7. Random Forest Model

The process’s next and final major step is building a RFR model. In this research, To estimate the concentrations of micronutrients and construct RF models, we employed the Python library sklearn (Pedregosa et al., 2011). In this study, the micronutrient concentration served as the response variable for the regression task. The data was split into a training set, comprising 80% of the data, and a testing set, encompassing the remaining 20%. The testing set was solely utilized for evaluating the performance of the models, while model creation was conducted solely on the training set.

We used GridSearch CV, which suggested a combination of Grid Search and CV. Grid Search is a technique used in Random Forest (RF) models for tuning hyperparameters. It systematically explored a predetermined set of hyperparameter combinations to find the optimal configuration that maximizes model performance. This technique constructed a grid of hyperparameter values and thoroughly assesses the model's performance for each combination.

Grid search requires defining a range or list of values for each hyperparameter of interest. The algorithm then systematically trains and evaluates the model using all possible combinations of these hyperparameter values. Evaluation metrics, such as mean squared error (MSE) or R-squared (R^2)(in

this research), assess each combination's performance. Throughout the grid search process, the algorithm calculates the model's performance for every hyperparameter combination and tracks the one that achieves the best performance based on the evaluation metric. Once the grid search is complete, the optimal hyperparameter combination is identified, and the model can be retrained using these ideal hyperparameters. Grid search streamlines the hyperparameter tuning process by exhaustively exploring diverse combinations, saving time and effort compared to manual tuning. It enables a systematic and comprehensive search across the hyperparameter space, ultimately enhancing model performance.

For this study, we used (i) the number of decision trees, (ii) the depth of the decision tree, (iii) the lowest number of samples necessary to divide an intermediate node, and (iv) the lowest number of samples necessary at the terminal node as hyperparameters to train our model. The configuration for the GridsearchCV is as follows:

```
param_grid = {
      'n_estimators': [500, 1000, 1500],
      'max_depth': [5, 10, 20],
      'min_samples_split': [2, 3, 5],
      'min_samples_leaf': [1, 2, 4]
}
```

Lastly, we take the natural logarithm of the dependent variables, viz. micronutrient concentrations, for linearizing relationships. For each micronutrient concentration, a separate model was introduced. After training the model, it was tested on the testing dataset, where the evaluation of the model was done using the coefficient of determination (R^2), Root Mean Squared Error (RMSE), relative RSME(RRMSE), and Mean Bias error(MBE). Also, we analyze the model using the Partial dependencies plot and SHAP(SHapley Additive exPlanations) (Lundberg et al., 2017) to determine how the target variable (predicted outcome) changes as one or more input features vary while keeping all other features constant and identify the most influential features in your RF model and their relative contributions, respectively. Features with larger absolute SHAP values have a more substantial impact on the predictions, while features with smaller absolute SHAP values have less influence (Lundberg et al., 2017). We analyzed what features are significant for the model. SHAP feature importance method was used to undergo this operation. It is based on the cooperative game theory notion of Shapley values. SHAP feature significance quantifies the influence of each feature on the model's output for a single instance.

5. RESULTS

This chapter is divided into four main sections. The first part focuses on feature selection, which involves exploring three different variants: varying the number of decision trees (500, 1000, 1500) and utilizing RFECV with a k-fold value of ten. The second part involves evaluating the model by presenting its metrics and assessing its performance. The third part involves utilizing SHAP variable importance to identify and describe the significant features within the model. Lastly, the chapter discusses using PDP to analyze how specific features influence the overall model.

Throughout the research analysis, a total of 85 different models containing for all four micronutrients were constructed to find best model for each micronutrient. Among them, the model with the highest accuracy, measured by the R^2 score, was selected to present the research results. For the model, a temporal window of two months was employed. Additionally, features with missing values (NaN) were removed from the analysis, as the sklearn library does not handle such data. As a result, a dataset consisting of 884 data points was used for further analysis and interpretation.

5.1. Variable Selection

RFECV is a valuable technique for selecting the most significant and informative features from a large pool of variables. This process enhances both the performance and interpretability of the model. Hence, the idea behind the strategy of taking multiple options in terms of the number of decision trees as RFECV itself is a random forest regressor. Therefore additional inclusion of various decision trees was pseudo hyperparameter tuning. The result of RFECV for each micronutrient are shown in the Fig. 5.

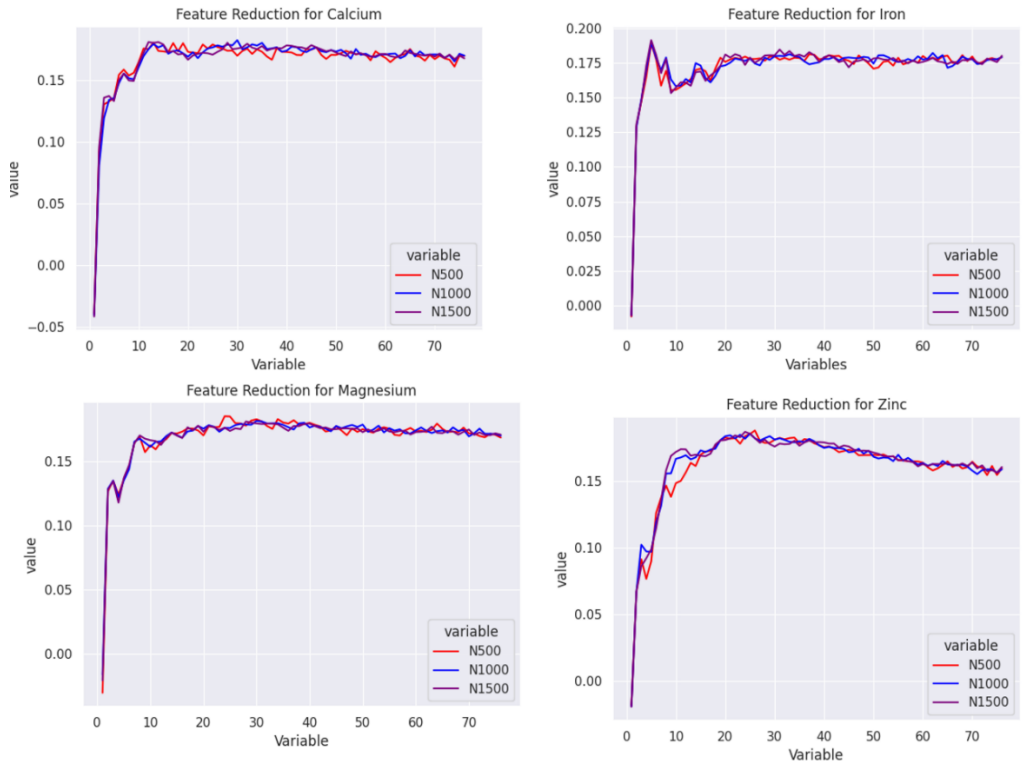


Figure 5 Graphs presenting the feature reduction for each micronutrient

The axis of the charts represents the number of variables and R^2 values for the x and y axis, respectively. Red, blue, and purple lines represent the 500, 1000 and 1500 decision trees, respectively. Lastly, we select the number of variables depending on the highest y-axis value. The features selected were as shown in table below: Note: Please refer to Annex 2 for explanations of variables nomenclature.

Table 4 The table showing all the features selected for the model through RFECV

Model	Selected Features
Ca	'B02_ST1_mean', 'B02_ST3_mean', 'B04_ST1_mean', 'B04_ST3_mean', 'B05_ST1_mean', 'B05_ST3_mean', 'B06_ST2_mean', 'B07_ST1_mean', 'B11_ST1_mean', 'B12_ST2_mean', 'B12_ST3_mean', 'NDRE1_mean_ST3', 'EVI_mean_ST2', 'MSAVI_mean_ST1', 'NDRE3_mean_ST1', 'NDRE2_mean_ST1', 'Prec_acc', 'RVI', 'Temp_avg', 'TWI', 'VH', 'VV', 'DEM', 'SOC', 'CEC', 'pH', 'pol_dif', 'pol_add', 'pol_ratio', 'OCS'
Fe	'Prec_acc', 'Temp_avg', 'VH', 'DEM', 'pol_ratio'
Mg	'B03_ST1_mean', 'B03_ST2_mean', 'B05_ST1_mean', 'B11_ST2_mean', 'B12_ST2_mean', 'NDRE1_mean_ST3', 'MNDWI2_mean_ST2', 'NDWI_mean_ST2', 'EVI_mean_ST1', 'NDRE2_mean_ST1', 'Prec_acc', 'RVI', 'Temp_avg', 'TWI', 'VH', 'Slope', 'DEM', 'SOC', 'CEC', 'pH', 'pol_dif', 'pol_add', 'pol_ratio', 'Nitrogen_mean'
Zn	'B04_ST1_mean', 'B05_ST1_mean', 'B06_ST1_mean', 'B06_ST2_mean', 'B06_ST3_mean', 'B11_ST2_mean', 'MNDWI1_mean_ST3', 'MNDWI2_mean_ST2', 'MNDWI1_mean_ST2', 'SAVI_mean_ST2', 'EVI_mean_ST1', 'NDRE2_mean_ST1', 'Prec_acc', 'Temp_avg', 'TWI', 'VH', 'VV', 'Slope', 'DEM', 'SOC', 'CEC', 'pH', 'pol_add', 'pol_ratio', 'OCS', 'Nitrogen_mean'

5.2. Variable importance

The SHAP values for a feature indicate the feature's average contribution over all conceivable feature combinations. We can determine the relevance of each characteristic in the model's predictions by computing these values. On the x-axis, we have the mean SHAP value, which suggests the impact of the feature on model output and y-axis have the selected features. The variable importance graphs are shown in Fig. 6: Please visit Annexes (b) for abbreviations.

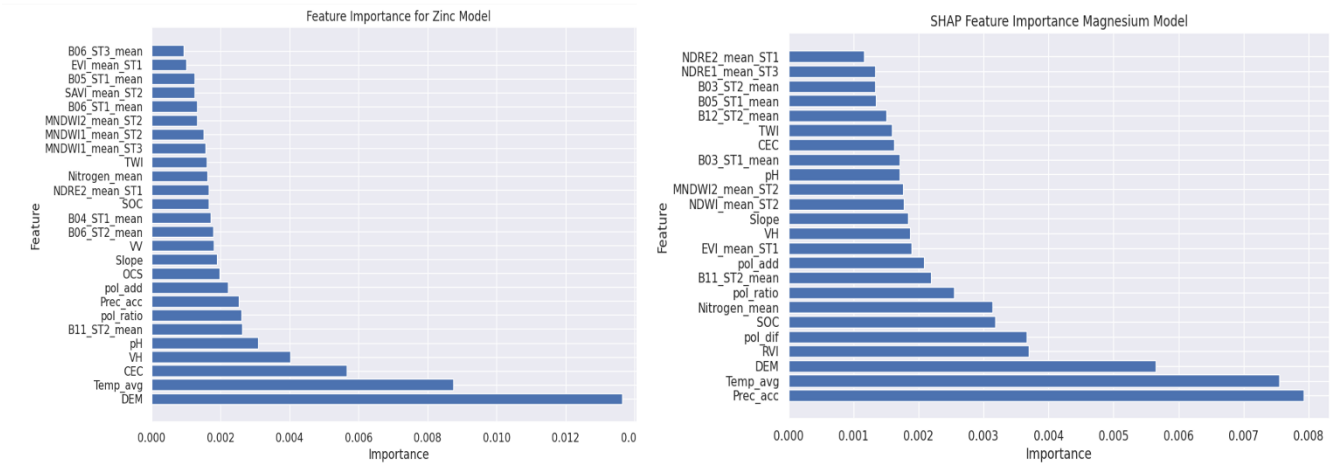
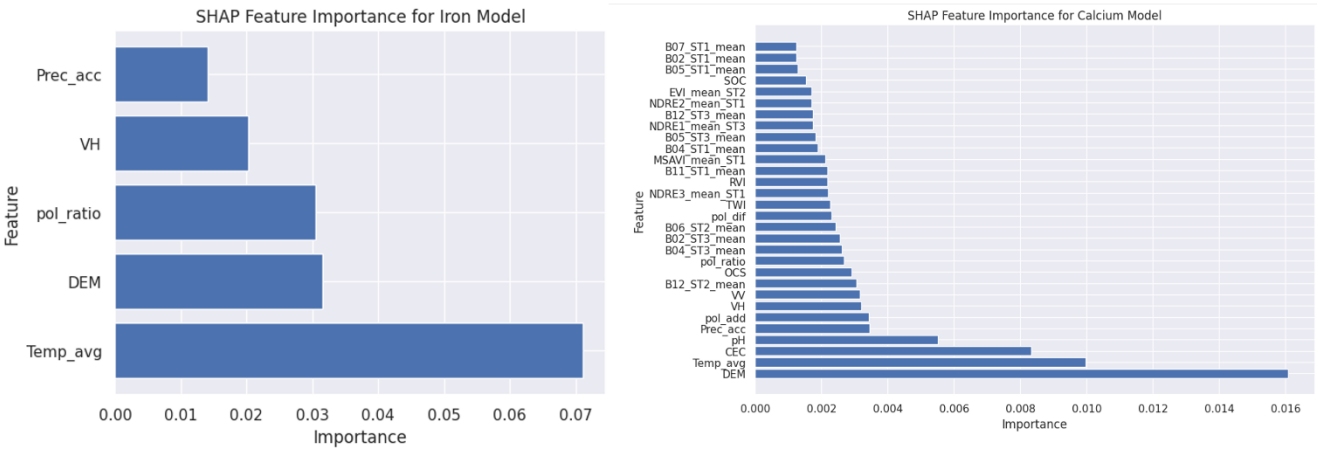


Figure 6 Bar graphs depicting feature importance among selected features



5.3. Model Evaluation

The evaluation results are shown in the Fig. 7 to 10.

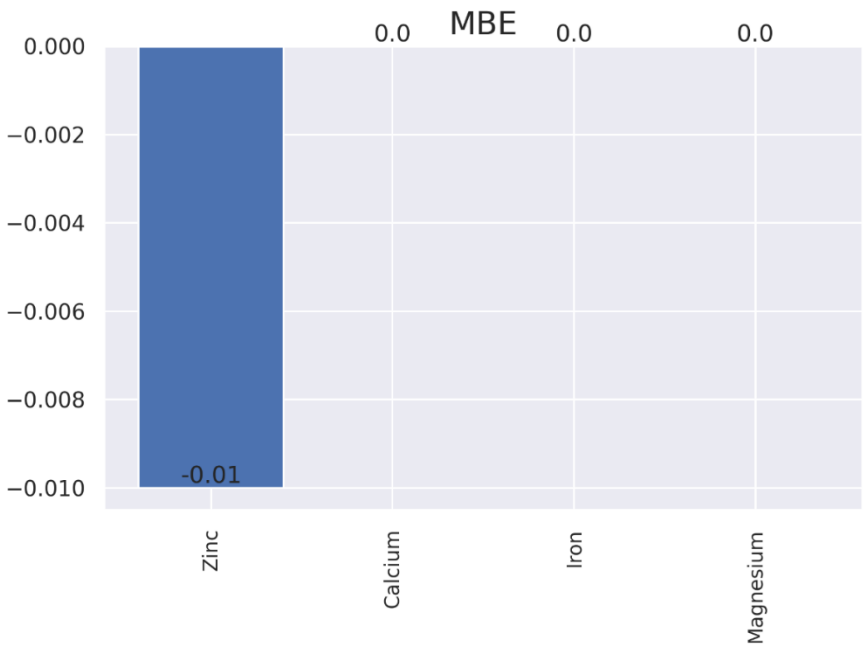


Figure 7 Mean Bias Error for Ca, Zn, Fe, and Mg models

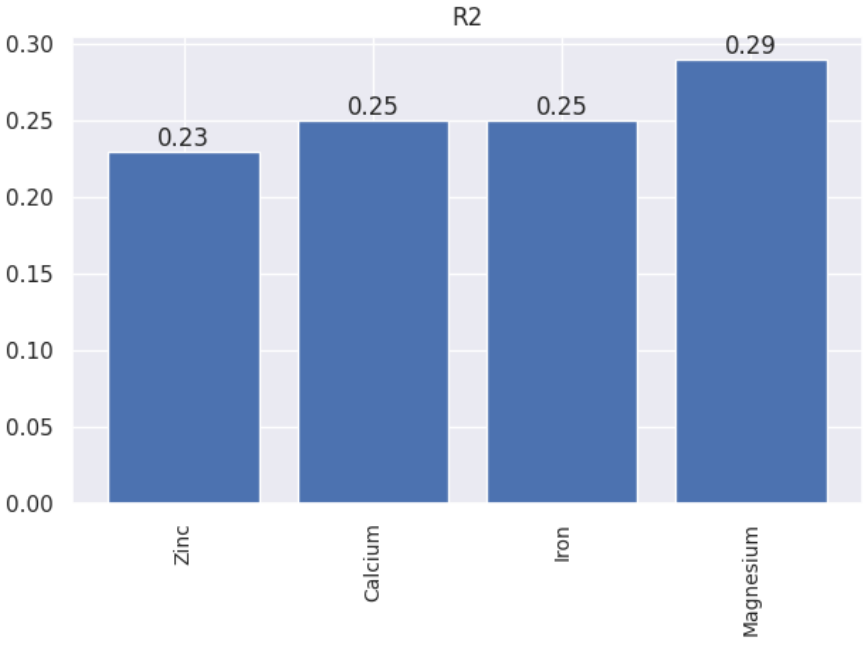


Figure 8 Coefficient of R-square for Ca, Zn, Fe, and Mg models

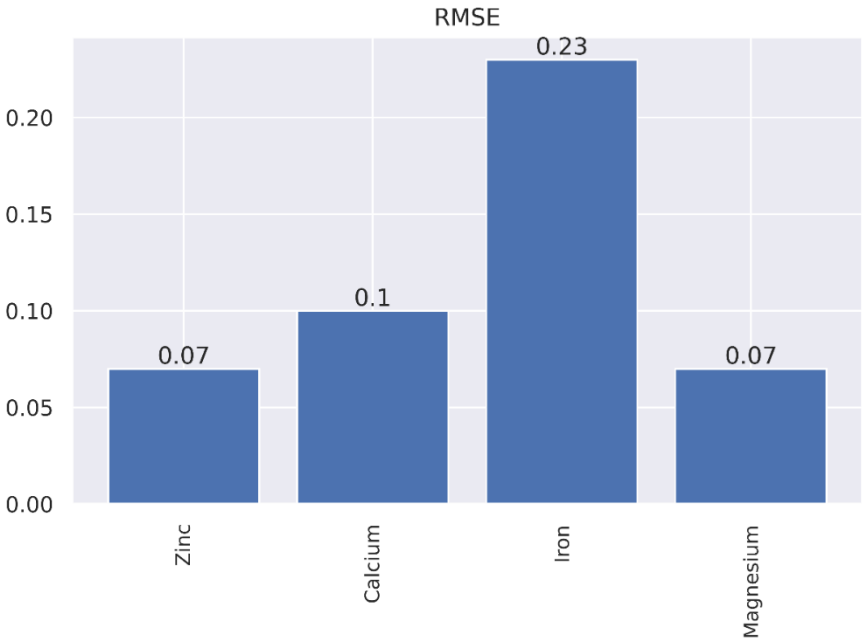


Figure 9 Root mean square error for Ca, Zn, Fe, and Mg models

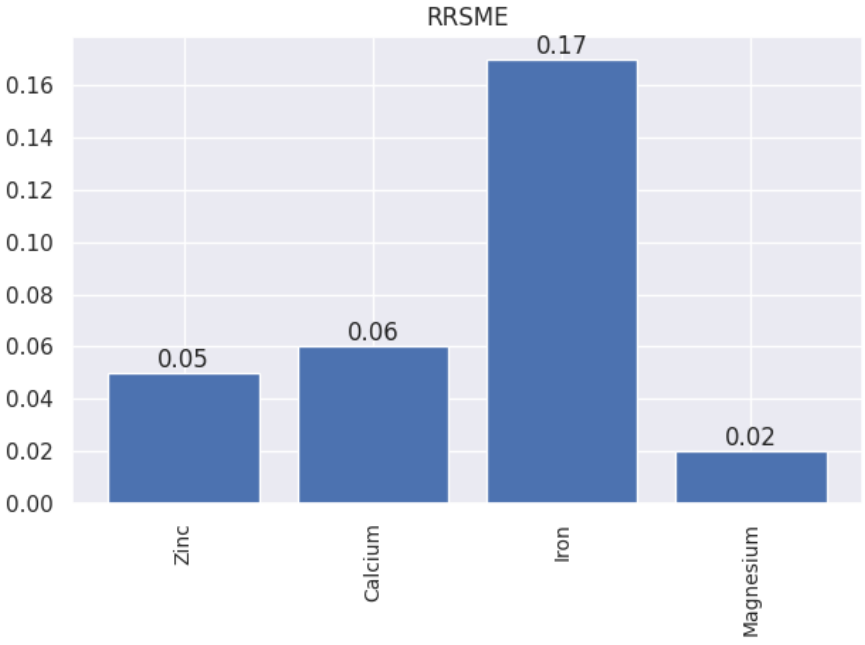


Figure 10 Relative root mean square error for Ca, Zn, Fe and Mg models

It was noticed from the Fig. 7 to 10 collectively, that R^2 highest for the Mg model and lowest for the Zn at 0.29 and 0.23, respectively. MBE was zero for all models except Zn, stating that there is no systematic bias or likelihood for the model's predictions to constantly overestimate or underestimate actual values. For the Fe model, predictions are relatively accurate and have a modest amount of error compared to the variability in the data, as indicated by RMSE and RRMSE values of 0.23 and 0.17, respectively.

5.4. Partial dependence plots

PDPs can be used to describe the link between the input feature and micronutrient concentration estimation, as shown in the Fig. 11 to 14 for few selected features.Across all the model, Elevation(DEM) shows negative relationship to micronutrient concentrations. For the independent features in Fe model(Fig. 11), Elevation(DEM) and polarimetric ratio(VV/VH) shows an negative relationship, while for the mean temperature throughout the growing season show positive relationship. For Zn model(Fig. 12), VH, pH, and CEC and mean temperature depicted positive relationship towards the Zn content, while B11_ST2_mean and DEM showed a negative relationship, inferring that lower the magnitude, more the Zn concentration. We see similar dynamic for Ca model(Fig 13), except for variable 'Prec_acc' which is precipitation aggregated throughout the season, for this variable we see a flat curve stating little-to-less change in Ca content with increase in precipitation. Lastly, for the Mg model(Fig. 14), 'Prec_acc' variable demonstrates a positive correlation, along other variables such as 'RVI', 'Pol_diff' and mean temperature showing direct relationship with Mg concentrations. Meanwhile, DEM, SOC and N content show inversely proportional relationship with the Mg content. The y-axis represents the model's average predicted outcome or response; in this case, the natural log of micronutrient concentration and the x-axis explain the magnitude of input features. Note: Please refer to Annex 8.2 for explanation of abbreviations used.

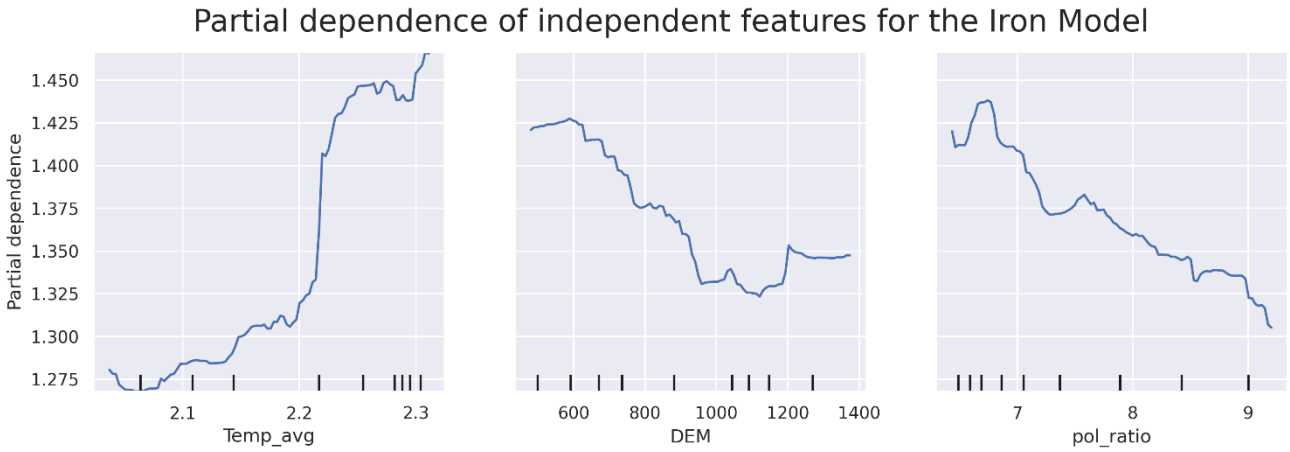


Figure 11 Partial Dependence Plot of selected features for micronutrient Fe per feature

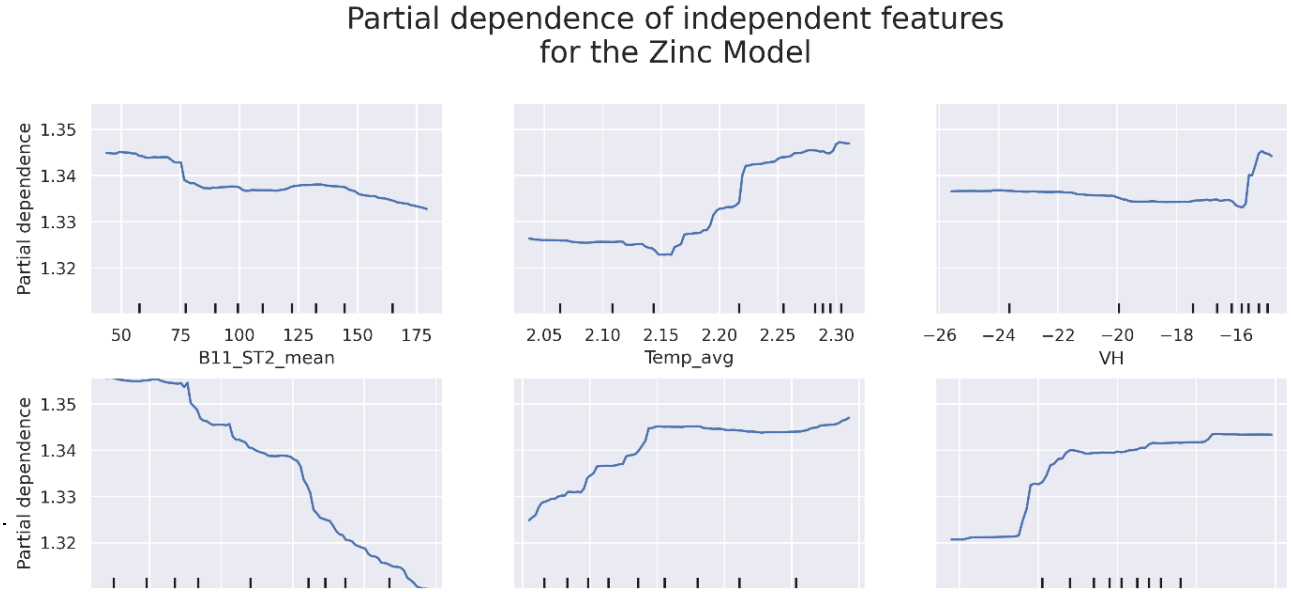


Figure 12 Partial dependence plots of selected features for micronutrient Zn per features

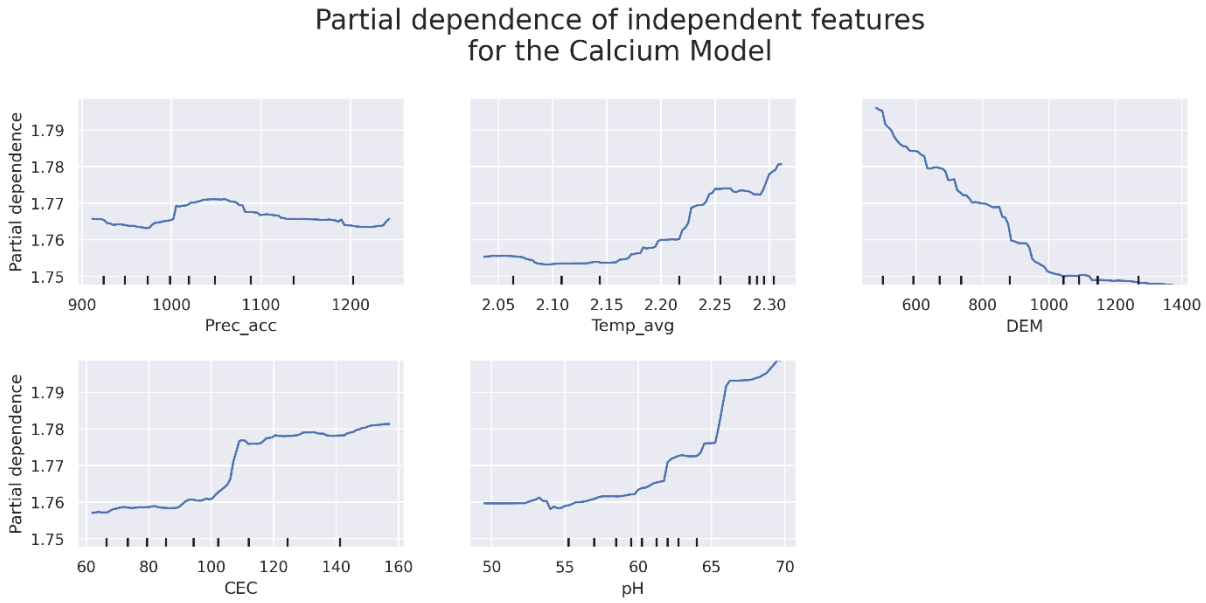


Figure 13 Partial dependence plots of selected features for micronutrient Ca per feature

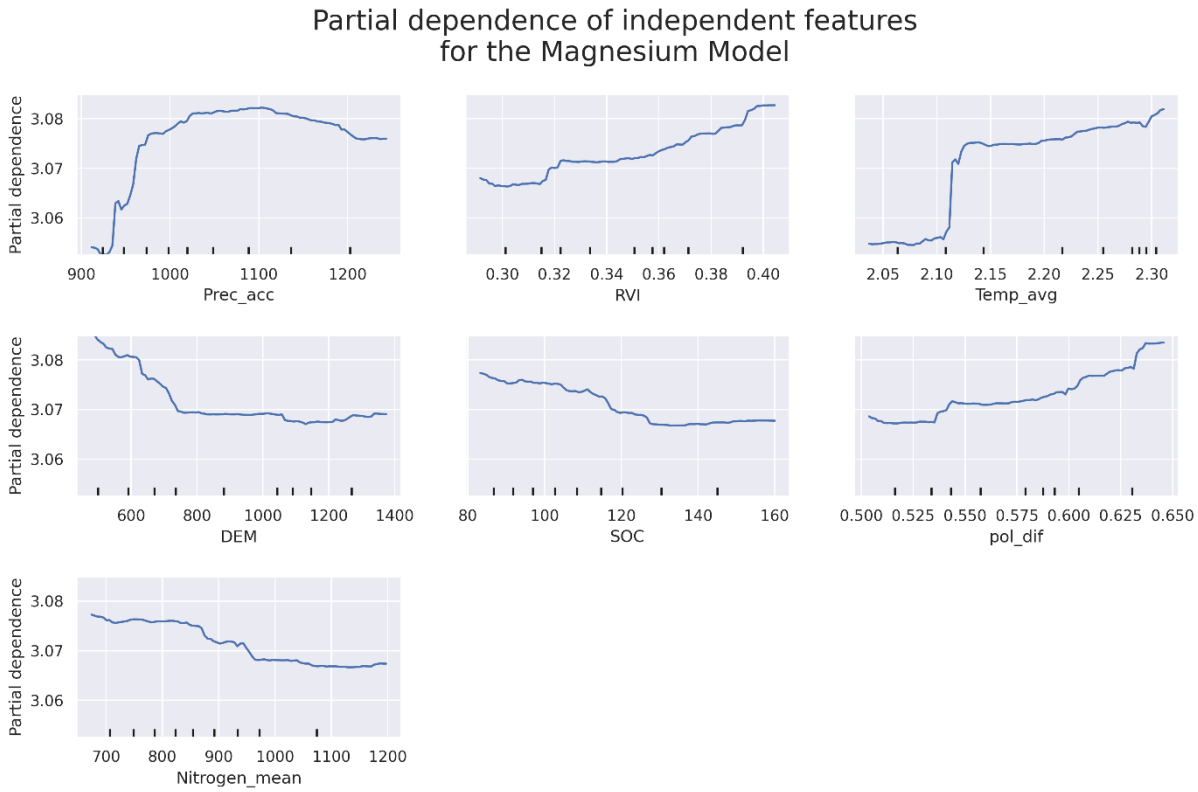


Figure 14 Partial Dependence Plot of selected features for micronutrient Mg per feature

6. DISCUSSION

The present study describes a machine learning based method for predicting micronutrient concentrations utilizing Sentinel 1 and Sentinel 2 satellite imagery and geospatial data as input. This study is a ground-breaking effort to estimate micronutrient concentrations by combining satellite data from Sentinel 1 and Sentinel 2. A notable accuracy level (R^2) of 0.29 was achieved for estimating Mg micronutrient concentration using precipitation accumulated over the season as the most important variable. Soil characteristics extracted from coarser spatial resolution performed better than remotely sensed inputs. Red-edge bands from the Sentinel are also significant in creating a robust model. Lastly, through this study, estimation of micronutrient concentrations was done, returning Mg as most accurately estimated, followed by Fe, Ca, with same R^2 accuracies and Lastly, Zn. The difference between the highest and lowest R^2 accuracies was just 0.06. Major significant features were climatological, topographical, and polarimetric features. Spectral features were polluted by NaN(missing values), noisy and inconsistent data. Change in the type of sensor used for extracting spectral features may have significant implications on the model.

The best prediction model for Fe included none of the remotely sensed inputs or derivatives from the Sentinel 2. The features selected were precipitation throughout the season, average temperature thought out the season, VH polarization, elevation, and polarimetric ratio. Another interesting observation was that average temperature and elevation were always presented as the models' most important features. As a research by Pelegrino et al. (2019) also supported the claim that elevation has association with micronutrient concentration. Among, all four CEC was highly significant among them since CEC plays a vital role in the uptake of macro and micronutrients from the soil to the shoot of the plant and is an indicator of soil fertility (Ulhassan et al., 2022). Also, CEC is significantly correlated with the micronutrient content is the soil (Najafi-Ghiri et al., 2013). Another study by Miner et al. (2018) for the crop maize provided evidence that SOC is contributes significantly for availability of micronutrients. Hence, justifying the correlation between soil parameters and micronutrient concentration. PDPs depicted the type of influence that would be made by the magnitude increase in the input features. For example, PDPs of the Fe model showed that trend in the elevation is linear negative but for average temperature(Temp_avg) is linear positive and similarly for other micronutrients too. The distribution and availability of micronutrients were notably influenced by temperature (Najafi-Ghiri et al., 2013). Furthermore, many approximately flat lines, which helps better understand the low accuracies in all models were shown by PDPs. PDPs for all the models and all their features is shown in Annex 8.3.

Spectral features, including raw bands and indices, were unimportant compared to the features of their counterparts such soil and climatological features. It is understood from the results that SWIR 1 band was the feature that excelled in the group of all spectral features. It's very sensitive to moisture and known for water absorption. Also, Paz-Kagan et al. (2020) proposed that utilizing reflectance spectroscopy in the shortwave infrared (SWIR) spectral range could enable accurate estimation of N content. Evidence from Belgiu et al. (2023), justifies the SWIR bands for estimating micronutrient concentration as they found SWIR important for estimating micronutrient content. Apart from SWIR bands, Band 4 and Band 6, and SAVI and NDVI using the Red-edge showed some significance, yet nowhere came out as the best predictors of the models. A reason behind the failure of Spectral data

could be inconsistent data in terms of non-noisy and regularity in the temporal dimension. This finding co-aligns with known difficulty of using remote sensing in a tropical and humid region (Tseng et al., 2008).

Figures 11-14 showed a strong relationship, either positive or negative among the features, among which climatological, topographic and soil were frontrunners. Studies from Botoman et al. (2022) and Gashu et al (2020) supports their presence as they found soil and landscape features showing influence on the model. According to the PDPs, polarimetric, topographic, climatological, and soil features were among the models' most significant features. Therefore, answering the research questions regarding their significance in estimating micronutrient concentrations. Existing literature highlights the importance of polarimetric(Munir et al., 2022), climatological, topographic, and soil features, stating them as crucial for estimating as well as mapping micronutrient concentration in sub-Saharan Africa (Botoman et al., 2022; Gashu et al., 2020; Miner et al., 2018; Najafi-Ghiri et al., 2013; Paz-Kagan et al., 2020).

A limitation of these models is that they are trained on Malawi dataset only, GeoNutrition surveys also mentions Ethiopia's micronutrition concentration data, as it is still not known that how will these models would perform on Ethiopia dataset. Hence the generalization of model and fine-tuning the model to fit these demands is still pending. Technically, the research was carried out utilizing high-performance computing, which provides efficient processing with time complexity of O(n), and the cloud computing environment made available through the GEE Python API. To improve maintainability and reusability, the code was modularized. Hence, reusability of code is benefits for training on different dataset than Malawi.

The main limitation of the study was the unavailability of cloud-free imagery for Sentinel 2 since the growing season of maize overlaps with the rainy season in Malawi. Using airborne imagery collected by unmanned aerial systems or vehicles (UAV) would offer a viable alternative. In terms of improving spectral features, since already using UAVs, employing hyperspectral sensors would greatly help. The unavailability of Sentinel 2 L2A products on GEE and Copernicus hub added an additional process of atmospherically correcting the data. In terms of computational time and delay, this process goes on for hundreds of hours of processing. An experiment for utilizing maximum of datapoints by removing Sentinel 2 features was done, results from the experiment had no significant change compared with results of models with spectral features. Other experiments with input data as both combing environmental covariates with Sentinel 1 and only using environmental covariates as inputs was also done. In both cases, there is was no significant change in models' performance.

The Random Forest regressor method was used with cross-validation for variable selection/Feature reduction and hyperparameter tuning with ten-fold CV. Hence, creating at least 810 combinations to find the best-fit model for each micronutrient. Also, to increase the accuracy of the models, outlier removal techniques like inter-quartile range and Isolation forest (Tony Liu et al., 2008) were also used, but they had no significant impact. Strategies regarding changing the training and testing ratio were also done i.e., changing the ratio from 80:20 to 60:40, but they produced results with lower accuracies than prior trained models. Other machine learning methods with no hyperparameter tuning were also tested, yet Random Forest was best among them. Scope for implementing deep learning methods like 1D-CNN (Lecun et al., 1998) and LSTM (Hochreiter & Schmidhuber, 1997) known for their usage

and exemplary performance in time-series problems, could produce better results in the estimation of micronutrient concentration. Yet, more importance should be given to preparing the dataset to the highest levels for completeness.

For future directions, UAVs and deep learning methods should be used. Applications of the findings from this research could be combined with findings of malnutrion surveys conducted by the non-profit organization or local government to apply changes in policies of region and implement better schemes for the vulnerable population. Understanding that farmers in Malawi perform subsistence farming (Benson, 2021). Through official channels of extension services, they should be made aware of what their crops and soil lack in terms of micronutrients content.

7. CONCLUSION

7.1. Conclusion

This study aims to estimate micronutrient concentrations by integrating data from Sentinel 1 and Sentinel 2 satellites and GIS data, such as topographic and soil characteristics. The findings reveal that Sentinel 2 features exhibit lower significance than other features due to missing. Topographic and climatological features consistently demonstrate more association with all four models. The research addresses the challenge of identifying hidden hunger in Malawi, specifically micronutrient deficiencies (MNDs). Obstacles such as cloud cover and the unavailability of atmospherically corrected Sentinel 2 data pose challenges. To build the models and optimize their performance, the RF regressor from sklearn is utilized, along with appropriate hyperparameter tuning. Various variations of the RF algorithm are explored, and the models with the highest accuracy for all micronutrients are selected. However, it is crucial to consider the model's adaptability when tested in different settings, as generalizability is paramount. For future studies, developing a new deep-learning model could prove advantageous, and generating a crop mask specific to Malawi from 2017-2018 would be necessary for accurately mapping micronutrient concentrations. Subsequently, the resulting map could be utilized for further investigations into MNDs in Malawi.

7.2. Answers to Research Questions

1. **What is the importance of Sentinel-1 and Sentinel-2 spectral and polarimetric features for estimating micro-nutrient concentrations?**

Among Sentinel-1 and Sentinel-2, Sentinel-1 turns out to be better predictor of micronutrient concentrations.

2. **What is the importance of topographic features derived from the slope, elevation, and Topographic wetness index relative to Sentinel-1 and Sentinel-2 predictors in estimating micronutrient concentration?**

Topographic feature perform better at estimating micronutrient concentration, relative to Sentinel 1 and Sentinel-2. Fig. 6 shows the variable importance which supports the argument.

3. **What is the importance of soil parameters relative to Sentinel-1 and Sentinel-2 predictors in estimating micro-nutrient concentration?**

Soil parameters also show better influencing in estimating micronutrient concentration than Sentinel 1 and Sentinel 2 predictors(Fig. 6).

4. **What is the importance of climatological data like temperature and precipitation relative to Sentinel-1 and Sentinel-2 predictors in estimating micronutrient concentration?**

Climatic data show better association with micronutrient concentration than Sentinel 1 and Sentinel 2 predictors (Fig. 6).

LIST OF REFERENCES

Bannari, A., Morin, D., Bonn, F., & Huete, A. R. (1995). A review of vegetation indices. *Remote Sensing Reviews*, *13*(1–2), 95–120. https://doi.org/10.1080/02757259509532298

Belgiu, M., Marshall, M., Boschetti, M., Pepe, M., Stein, A., & Nelson, A. (2023). PRISMA and Sentinel-2 spectral response to the nutrient composition of grains. *Remote Sensing of Environment*, *292*, 113567. https://doi.org/10.1016/J.RSE.2023.113567

Benson, T. (2021). *Disentangling food security from subsistence agriculture in Malawi*. https://doi.org/10.2499/9780896294059

Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., De Onis, M., Ezzati, M., Grantham-Mcgregor, S., Katz, J., Martorell, R., & Uauy, R. (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *The Lancet*, *382*(9890), 427–451. https://doi.org/10.1016/S0140-6736(13)60937-X

Botoman, L., Chagumaira, C., Mossa, A. W., Amede, T., Ander, E. L., Bailey, E. H., Chimungu, J. G., Gameda, S., Gashu, D., Haefele, S. M., Joy, E. J. M., Kumssa, D. B., Ligowe, I. S., McGrath, S. P., Milne, A. E., Munthali, M., Towett, E., Walsh, M. G., Wilson, L., … Nalivata, P. C. (2022). Soil and landscape factors influence geospatial variation in maize grain zinc concentration in Malawi. *Scientific Reports 2022 12:1*, *12*(1), 1–13. https://doi.org/10.1038/s41598-022-12014-w

Botoman, L., Nalivata, P. C., Chimungu, J. G., Munthali, M. W., Bailey, E. H., Ander, E. L., Lark, R. M., Mossa, A. W., Young, S. D., & Broadley, M. R. (2020). Increasing zinc concentration in maize grown under contrasting soil types in Malawi through agronomic biofortification: Trial protocol for a field experiment to detect small effect sizes. *Plant Direct*, *4*(10), e00277. https://doi.org/10.1002/PLD3.277

Bouis, H. E., & Saltzman, A. (2017). Improving nutrition through biofortification: A review of evidence from HarvestPlus, 2003 through 2016. *Global Food Security*, *12*, 49–58. https://doi.org/10.1016/J.GFS.2017.01.009

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

Camaschella, C. (2015). Iron-Deficiency Anemia. *New England Journal of Medicine*, *372*(19), 1832–1843. https://doi.org/10.1056/NEJMra1401038

Clevers, J. G. P. W., & Gitelson, A. A. (2013). Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. *International Journal of Applied Earth Observation and Geoinformation*, *23*(1), 344–351. https://doi.org/10.1016/j.jag.2012.10.008

Conti, M. V., Campanaro, A., Coccetti, P., De Giuseppe, R., Galimberti, A., Labra, M., & Cena, H. (2019). Potential role of neglected and underutilized plant species in improving women's empowerment and nutrition in areas of sub-Saharan Africa. *Nutrition Reviews*, *77*(11), 817–828. https://doi.org/10.1093/NUTRIT/NUZ038

Curran, P. J. (1989). Remote sensing of foliar chemistry. *Remote Sensing of Environment*, *30*(3), 271–278. https://doi.org/10.1016/0034-4257(89)90069-2

Fairweather-Tait, S. (2011). Selenium in human health and disease. *Antioxid. Redox Signal.*, *14*, 1337–1383.

FAO. (2018). *SMALL FAMILY FARMS COUNTRY FACTSHEET*. www.fao.org/family-farming/data-sources/dataportrait/farm-size/en

FAO. (2020). *The state of food security and nutrition in the world 2020 : transforming food systems for affordable healthy diets*. Food and Agriculture Organization of the United Nations,.

Farwa, U. E., Rehman, A. U., Qasim Khan, S., & Khurram, M. (2020). Prediction of Soil Macronutrients Using Machine Learning Algorithm. *International Journal of Computer (IJC) International Journal of Computer (IJC*, *38*(1), 1–14. http://ijcjournal.org/

Forkuor, G., Hounkpatin, O. K. L., Welp, G., & Thiel, M. (2017). High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLOS ONE*, *12*(1), e0170478. https://doi.org/10.1371/JOURNAL.PONE.0170478

Fu, Y., Yang, G., Li, Z., Li, H., Li, Z., Xu, X., Song, X., Zhang, Y., Duan, D., Zhao, C., & Chen, L. (2020). Progress of hyperspectral data processing and modelling for cereal crop nitrogen monitoring. *Computers and Electronics in Agriculture*, *172*(February), 105321. https://doi.org/10.1016/j.compag.2020.105321

Gao, B. C. (1996). NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, *58*(3), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Gashu, D., Lark, R. M., Milne, A. E., Amede, T., Bailey, E. H., Chagumaira, C., Dunham, S. J., Gameda, S., Kumssa, D. B., Mossa, A. W., Walsh, M. G., Wilson, L., Young, S. D., Ander, E. L., Broadley, M. R., Joy, E. J. M., & McGrath, S. P. (2020). Spatial prediction of the concentration of selenium (Se) in grain across part of Amhara Region, Ethiopia. *Science of The Total Environment*, *733*, 139231. https://doi.org/10.1016/J.SCITOTENV.2020.139231

Gashu, D., Nalivata, P. C., Amede, T., Ander, E. L., Bailey, E. H., Botoman, L., Chagumaira, C., Gameda, S., Haefele, S. M., Hailu, K., Joy, E. J. M., Kalimbira, A. A., Kumssa, D. B., Lark, R. M., Ligowe, I. S., McGrath, S. P., Milne, A. E., Mossa, A. W., Munthali, M., … Broadley, M. R. (2021). The nutritional quality of cereals varies geospatially in Ethiopia and Malawi. *Nature 2021 594:7861*, *594*(7861), 71–76. https://doi.org/10.1038/s41586-021-03559-3

Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS- MODIS. *Remote Sensing of Environment*, *58*(3), 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7

Gödecke, T., Stein, A. J., & Qaim, M. (2018). The global burden of chronic and hidden hunger: Trends and determinants. *Global Food Security*, *17*, 21–29. https://doi.org/10.1016/J.GFS.2018.03.004

Graeff, S., & Claupein, W. (2003). Quantifying nitrogen status of corn (Zea mays L.) in the field by reflectance measurements. *European Journal of Agronomy*, *19*(4), 611–618. https://doi.org/10.1016/S1161-0301(03)00007-8

Grieco, M., Schmidt, M., Warnemünde, S., Backhaus, A., Klück, H.-C., Garibay, A., Antonia, Y., Moya, T., Jozefowicz, A. M., Mock, H.-P., Seiffert, U., Maurer, A., & Pillen, K. (2021). *Dynamics and genetic regulation of leaf nutrient concentration in barley based on hyperspectral imaging and machine learning*. https://doi.org/10.1016/j.plantsci.2021.111123

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning 2002 46:1*, *46*(1), 389–422. https://doi.org/10.1023/A:1012487302797

Herzig, P., Backhaus, A., Seiffert, U., Von Wirén, N., Pillen, K., & Maurer, A. (2019). *Genetic dissection of grain elements predicted by hyperspectral imaging associated with yield-related traits in a wild barley NAM population*. https://doi.org/10.1016/j.plantsci.2019.05.008

Heyneman, C. A. (1996). Zinc Deficiency and Taste Disorders. *Annals of Pharmacotherapy*, *30*(2), 186–187. https://doi.org/10.1177/106002809603000215

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). VERY HIGH RESOLUTION INTERPOLATED CLIMATE SURFACES FOR GLOBAL LAND AREAS.

INTERNATIONAL JOURNAL OF CLIMATOLOGY Int. J. Climatol, *25*, 1965–1978. https://doi.org/10.1002/joc.1276

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Huang, S., Sasaki, A., Yamaji, N., Okada, H., Mitani-Ueno, N., & Ma, J. F. (2020). The ZIP Transporter Family Member OsZIP9 Contributes To Root Zinc Uptake in Rice under Zinc-Limited Conditions. *Plant Physiology*, *183*(3), 1224–1234. https://doi.org/10.1104/PP.20.00125

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002a). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1–2), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002b). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1–2), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Hunt Jr., E. R., Daughtry, C. S. T., Eitel, J. U. H., & Long, D. S. (2011). Remote Sensing Leaf Chlorophyll Content Using a Visible Band Index. In *Agronomy Journal* (Vol. 103, pp. 1090–1099).

Hurst, R., Siyame, E. W. P., Young, S. D., Chilimba, A. D. C., Joy, E. J. M., Black, C. R., Ander, E. L., Watts, M. J., Chilima, B., Gondwe, J., Kang'Ombe, D., Stein, A. J., Fairweather-Tait, S. J., Gibson, R. S., Kalimbira, A. A., & Broadley, M. R. (2013). Soil-type influences human selenium status and underlies widespread selenium deficiency risks in Malawi. *Scientific Reports*, *3*. https://doi.org/10.1038/SREP01425

Husak, G. J., Marshall, M. T., Michaelsen, J., Pedreros, D., Funk, C., & Galu, G. (2008). Crop area estimation using high and medium resolution satellite imagery in areas with complex topography. *Journal of Geophysical Research: Atmospheres*, *113*(D14), 14112. https://doi.org/10.1029/2007JD009175

Ibrahim, S., Saleem, B., Rehman, N., Zafar, S. A., Naeem, M. K., & Khan, M. R. (2022). CRISPR/Cas9 mediated disruption of Inositol Pentakisphosphate 2-Kinase 1 (TaIPK1) reduces phytic acid and improves iron and zinc accumulation in wheat grains. *Journal of Advanced Research*, *37*, 33–41. https://doi.org/10.1016/J.JARE.2021.07.006

Jahnen-Dechent, W., & Ketteler, M. (2012). Magnesium basics. *Clinical Kidney Journal*, *5*(Suppl_1), i3–i14. https://doi.org/10.1093/NDTPLUS/SFR163

Jamison, D. T., Breman, J. G., Measham, A. R., Alleyne, G., Claeson, M., Evans, D. B., Jha, P., Mills, A., & Musgrove, P. (2006). Disease Control Priorities in Developing Countries. *Oxford University Press, New York, Pp 1245–1261*, 1293–1307. https://www.ncbi.nlm.nih.gov/books/NBK11728/

Joy, E. (2015). Dietary mineral supplies in Malawi: spatial and socioeconomic assessment. *BMC Nutrition*, *1*, 42.

Kaur, G., Das, K., & Hazra, J. (2020). Soil Nutrients Prediction Using Remote Sensing Data in Western India: An Evaluation of Machine Learning Models. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 4677–4680. https://doi.org/10.1109/IGARSS39084.2020.9324201

Kim, Y., & Van Zyl, J. J. (2009). A time-series approach to estimate soil moisture using polarimetric radar data. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(8), 2519–2527. https://doi.org/10.1109/TGRS.2009.2014944

King, J. C., Brown, K. H., Gibson, R. S., Krebs, N. F., Lowe, N. M., Siekmann, J. H., & Raiten, D. J. (2015). Biomarkers of Nutrition for Development (BOND)-Zinc Review. *The Journal of Nutrition*, *146*(4), 858S-885S. https://doi.org/10.3945/JN.115.220079

Kriegler, F. ~J., Malila, W. ~A., Nalepka, R. ~F., & Richardson, W. (1969). Preprocessing Transformations and Their Effects on Multispectral Recognition. *Remote Sensing of Environment, VI*, 97.

Kumssa, D. B., Mossa, A. W., Amede, T., Ander, E. L., Bailey, E. H., Botoman, L., Chagumaira, C., Chimungu, J. G., Davis, K., Gameda, S., Haefele, S. M., Hailu, K., Joy, E. J. M., Lark, R. M., Ligowe, I. S., McGrath, S. P., Milne, A., Muleya, P., Munthali, M., … Nalivata, P. C. (2022). Cereal grain mineral micronutrient and soil chemistry data from GeoNutrition surveys in Ethiopia and Malawi. *Scientific Data 2022 9:1*, *9*(1), 1–12. https://doi.org/10.1038/s41597-022-01500-5

Lapaz Olveira, A., Saínz Rozas, H., Castro-Franco, M., Carciochi, W., Nieto, L., Balzarini, M., Ciampitti, I., & Reussi Calvo, N. (2023). Monitoring Corn Nitrogen Concentration from Radar (C-SAR), Optical, and Sensor Satellite Data Fusion. *Remote Sensing 2023, Vol. 15, Page 824*, *15*(3), 824. https://doi.org/10.3390/RS15030824

Lecun, Y., Bottou, E., Bengio, Y., & Haffner, P. (1998). *Gradient-Based Learning Applied to Document Recognition*.

Li, D., Wang, C., Jiang, H., Peng, Z., Yang, J., Su, Y., Song, J., & Chen, S. (2018). Monitoring litchi canopy foliar phosphorus content using hyperspectral data. *Computers and Electronics in Agriculture*, *154*(October 2017), 176–186. https://doi.org/10.1016/j.compag.2018.09.007

Li, F., Mistele, B., Hu, Y., Chen, X., & Schmidhalter, U. (2014). Reflectance estimation of canopy nitrogen content in winter wheat using optimised hyperspectral spectral indices and partial least squares regression. *European Journal of Agronomy*, *52*, 198–209. https://doi.org/10.1016/j.eja.2013.09.006

Ling, B., Goodin, D. G., Raynor, E. J., & Joern, A. (2019). Hyperspectral Analysis of Leaf Pigments and Nutritional Elements in Tallgrass Prairie Vegetation. *Frontiers in Plant Science*, *10*(February), 1–13. https://doi.org/10.3389/fpls.2019.00142

Liu, H., Zhu, H., & Wang, P. (2017). Quantitative modelling for leaf nitrogen content of winter wheat using UAV-based hyperspectral data. *International Journal of Remote Sensing*, *38*(8–10), 2117–2134. https://doi.org/10.1080/01431161.2016.1253899

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://github.com/slundberg/shap

Major, D. J., Baret, F., & Guyot, G. (1990). A ratio vegetation index adjusted for soil brightness. *International Journal of Remote Sensing*, *11*(5), 727–740. https://doi.org/10.1080/01431169008955053

Marang, I. J., Filippi, P., Weaver, T. B., Evans, B. J., Whelan, B. M., Bishop, T. F. A., Murad, M. O. F., Al-Shammari, D., & Roth, G. (2021). Machine Learning Optimised Hyperspectral Remote Sensing Retrieves Cotton Nitrogen Status. *Remote Sensing*, *13*(8), 1428. https://doi.org/10.3390/rs13081428

Miner, G. L., Delgado, J. A., Ippolito, J. A., Barbarick, K. A., Stewart, C. E., Manter, D. K., Del Grosso, S. J., Halvorson, A. D., Floyd, B. A., & D'Adamo, R. E. (2018). Influence of long-term nitrogen fertilization on crop and soil micronutrients in a no-till maize cropping system. *Field Crops Research*, *228*, 170–182. https://doi.org/10.1016/J.FCR.2018.08.017

Mohammadi Moghaddam, T., Razavi, S. M. A., & Taghizadeh, M. (2013). Applications of hyperspectral imaging in grains and nuts quality and safety assessment: A review. *Journal of Food Measurement and Characterization*, *7*(3), 129–140. https://doi.org/10.1007/S11694-013-9148-1/TABLES/2

Munir, S., Seminar, K. B., Sudradjat, Sukoco, H., & Buono, A. (2022). The Use of Random Forest Regression for Estimating Leaf Nitrogen Content of Oil Palm Based on Sentinel 1-A Imagery. *Information 2023, Vol. 14, Page 10*, *14*(1), 10. https://doi.org/10.3390/INFO14010010

Muthayya, S., Rah, J. H., Sugimoto, J. D., Roos, F. F., Kraemer, K., & Black, R. E. (2013). The Global Hidden Hunger Indices and Maps: An Advocacy Tool for Action. *PLOS ONE*, *8*(6), e67860. https://doi.org/10.1371/JOURNAL.PONE.0067860

Najafi-Ghiri, M., Ghasemi-Fasaei, R., & Farrokhnejad, E. (2013). Factors Affecting Micronutrient Availability in Calcareous Soils of Southern Iran.

Http://Dx.Doi.Org/10.1080/15324982.2012.719570, *27*(3), 203–215. https://doi.org/10.1080/15324982.2012.719570

OI Bermudez, K. L. M. S. J. F. (2012). Estimating micronutrient intakes from Household Consumption and Expenditures Surveys (HCES): an example from Bangladesh. *Food Nutr. Bull.*, *33*(Suppl), S208–S213.

Pandey, P., Ge, Y., Stoerger, V., & Schnable, J. C. (2017). High Throughput In vivo Analysis of Plant Leaf Chemical Properties Using Hyperspectral Imaging. *Frontiers in Plant Science*, *8*(August), 1–12. https://doi.org/10.3389/fpls.2017.01348

Paz-Kagan, T., Schmilovitch, ev, Yermiyahu, U., Rapaport, T., & Sperling, O. (2020). *Assessing the nitrogen status of almond trees by visible-to-shortwave infrared reflectance spectroscopy of carbohydrates*. https://doi.org/10.1016/j.compag.2020.105755

Pedregosa, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pelegrino, M. H. P., Weindorf, D. C., Silva, S. H. G., de Menezes, M. D., Poggere, G. C., Guilherme, L. R. G., & Curi, N. (2019). Synthesis of proximal sensing, terrain analysis, and parent material information for available micronutrient prediction in tropical soils. *Precision Agriculture*, *20*(4), 746–766. https://doi.org/10.1007/S11119-018-9608-Z/FIGURES/5

Prasad, A. S. (2004). Zinc deficiency: its characterization and treatment. *Metal Ions in Biological Systems*, *41*, 103–137.

Ross, A., Taylor, C. L., Yaktine, A. L., & Valle, H. B. Del. (2011). Dietary Reference Intakes for Calcium and Vitamin D. In *Dietary Reference Intakes for Calcium and Vitamin D*. National Academies Press (US). https://doi.org/10.17226/13050

Rossi, M., Candiani, G., Nutini, F., Gianinetto, M., & Boschetti, M. (2022). Sentinel-2 estimation of CNC and LAI in rice cropping system through hybrid approach modelling. *Https://Doi.Org/10.1080/22797254.2022.2117651*. https://doi.org/10.1080/22797254.2022.2117651

Saka, J. D. K., Sibale, P., Thomas, T. S., Hachigonta, S., Sibanda, L. M., & others. (2013). Malawi. *IFPRI Book Chapters*, 111–146.

Sharifi, A. (2020). Using Sentinel-2 Data to Predict Nitrogen Uptake in Maize Crop. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 2656–2662. https://doi.org/10.1109/JSTARS.2020.2998638

Suleymanov, A., Abakumov, E., Suleymanov, R., Gabbasova, I., & Komissarov, M. (2021). The Soil Nutrient Digital Mapping for Precision Agriculture Cases in the Trans-Ural Steppe Zone of Russia Using Topographic Attributes. *ISPRS International Journal of Geo-Information 2021, Vol. 10, Page 243*, *10*(4), 243. https://doi.org/10.3390/IJGI10040243

Tony Liu, F., Ming Ting, K., & Zhou, Z.-H. (2008). *Isolation Forest*. https://doi.org/10.1109/ICDM.2008.17

Tseng, D. C., Tseng, H. T., & Chien, C. L. (2008). Automatic cloud removal from multi-temporal SPOT images. *Applied Mathematics and Computation*, *205*(2), 584–600. https://doi.org/10.1016/j.amc.2008.05.050

Ulhassan, Z., Khan, A. R., Hamid, Y., Azhar, W., Hussain, S., Sheteiwy, M. S., Salam, A., Hakeem, K. R., & Zhou, W. (2022). Interaction of nanoparticles with soil–plant system and their usage

in remediation strategies. *Metals and Metalloids in Soil-Plant-Water Systems: Phytophysiology and Remediation Techniques*, 287–308. https://doi.org/10.1016/B978-0-323-91675-2.00024-X

Wang, Y. D., Wang, X., Ngai, S. M., & Wong, Y. S. (2013). Comparative proteomics analysis of selenium responses in selenium-enriched rice grains. *Journal of Proteome Research*, *12*(2), 808–820. https://doi.org/10.1021/PR300878Y/SUPPL_FILE/PR300878Y_SI_002.XLS

WHO. (2021). *Fact sheets - Malnutrition*. World Health Organisation. https://www.who.int/news-room/fact-sheets/detail/malnutrition

Wu, Q. (2020). geemap: A Python package for interactive mapping with Google Earth Engine. *Journal of Open Source Software*, *5*(51), 2305. https://doi.org/10.21105/JOSS.02305

Xu, H. (2007). *International Journal of Remote Sensing Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery*. https://doi.org/10.1080/01431160600589179

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., & Lausch, A. (2020). High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. *Science of The Total Environment*, *729*, 138244. https://doi.org/10.1016/J.SCITOTENV.2020.138244

8. ANNEXES

8.1. Spectral Indices

Table 5 Detailed description of spectral indices

Spectral Indices		Mathematical Formulae	Author
Normalized Difference Vegetation Index	NDVI	(NIR – RED) / (NIR + RED)	(Kriegler et al., 1969)
Red Edge Chlorophyll	RECl	(NIR / RED) – 1	(Clevers & Gitelson, 2013)
Normalized Difference Red-Edge	NDRE	(NIR – RED EDGE) / (NIR + RED EDGE)	(Hunt Jr. et al., 2011)
Modified Soil Adjusted Vegetation Index	MSAVI	(2 * Band 4 + 1 – sqrt ((2 * Band 4 + 1)2 – 8 * (Band 4 – Band 3))) / 2	(Bannari et al., 1995)
Green Normalized Vegetation Index	GNDVI	(NIR – GREEN) / (NIR + GREEN)	(Gitelson et al., 1996)
Normalized Difference Water Index	NDWI	(GREEN – NIR) / (GREEN + NIR)	(Gao, 1996)
Soil Adjusted Vegetation Index	SAVI	((NIR – RED) / (NIR + RED + L)) * (1 + L) where L = 0.5	(Major et al., 1990)
Enhanced Vegetation Index	EVI	2.5 * ((NIR – RED) / ((NIR) + (6 * RED) – (7.5 * BLUE) + 1))	(Huete et al., 2002b)
Modified Normalized Difference Water Index	MNDWI	(Green – SWIR)/(Green+SWIR)	(Xu, 2007)

8.2. Abbreviations

Table 6 Abbreviation used in the thesis

Acronym	Meaning
ST_1,ST_2,ST_3	Represents the three windows of 2 month interval. For Example: 'B11_ST2_mean' indicates mean of SWIR band reflectance over the second temporal window.
B02_mean till B12_mean	Average of raw bands
VV,VH	Polarizations
CEC	Cation Exchange Capacity
SOC	Soil Organic Carbon
pH	pH
OCS	Organic Carbon Stock
Nitrogen	Total Nitrogen
DEM	Elevation
Temp_avg	Average temperature
Prec_acc	Aggregated precipitation throughout the growing season

8.3. PDPs for all selected features

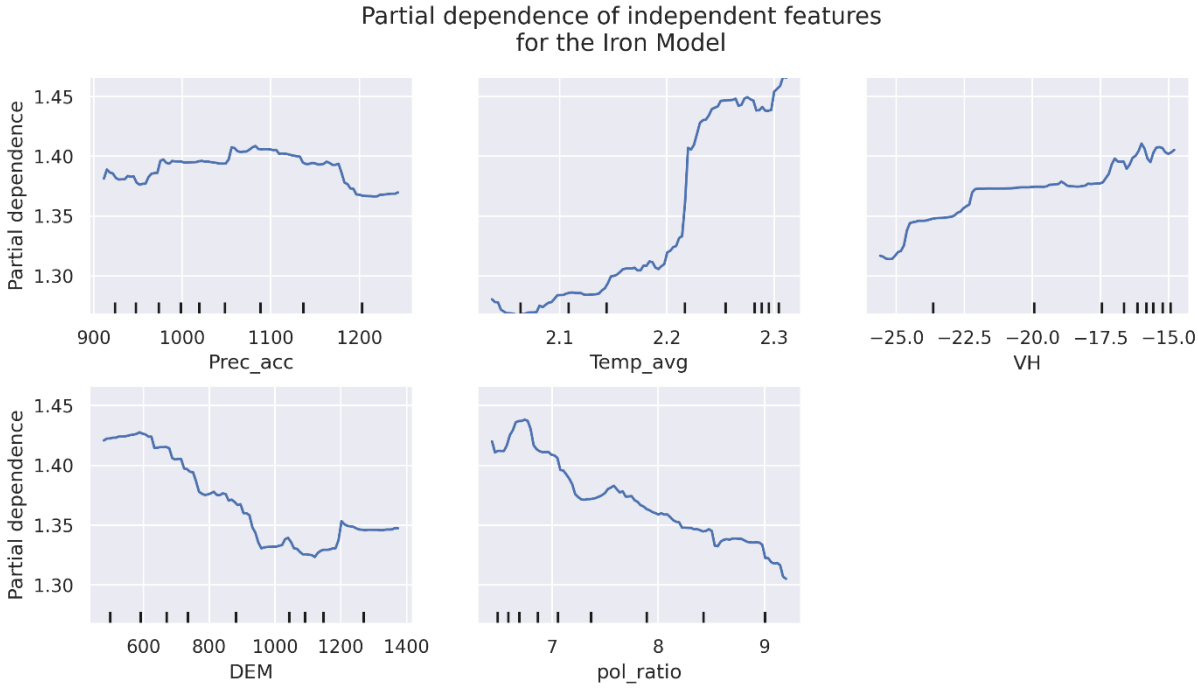


Figure 15 PDPs for all the features used for building Iron model

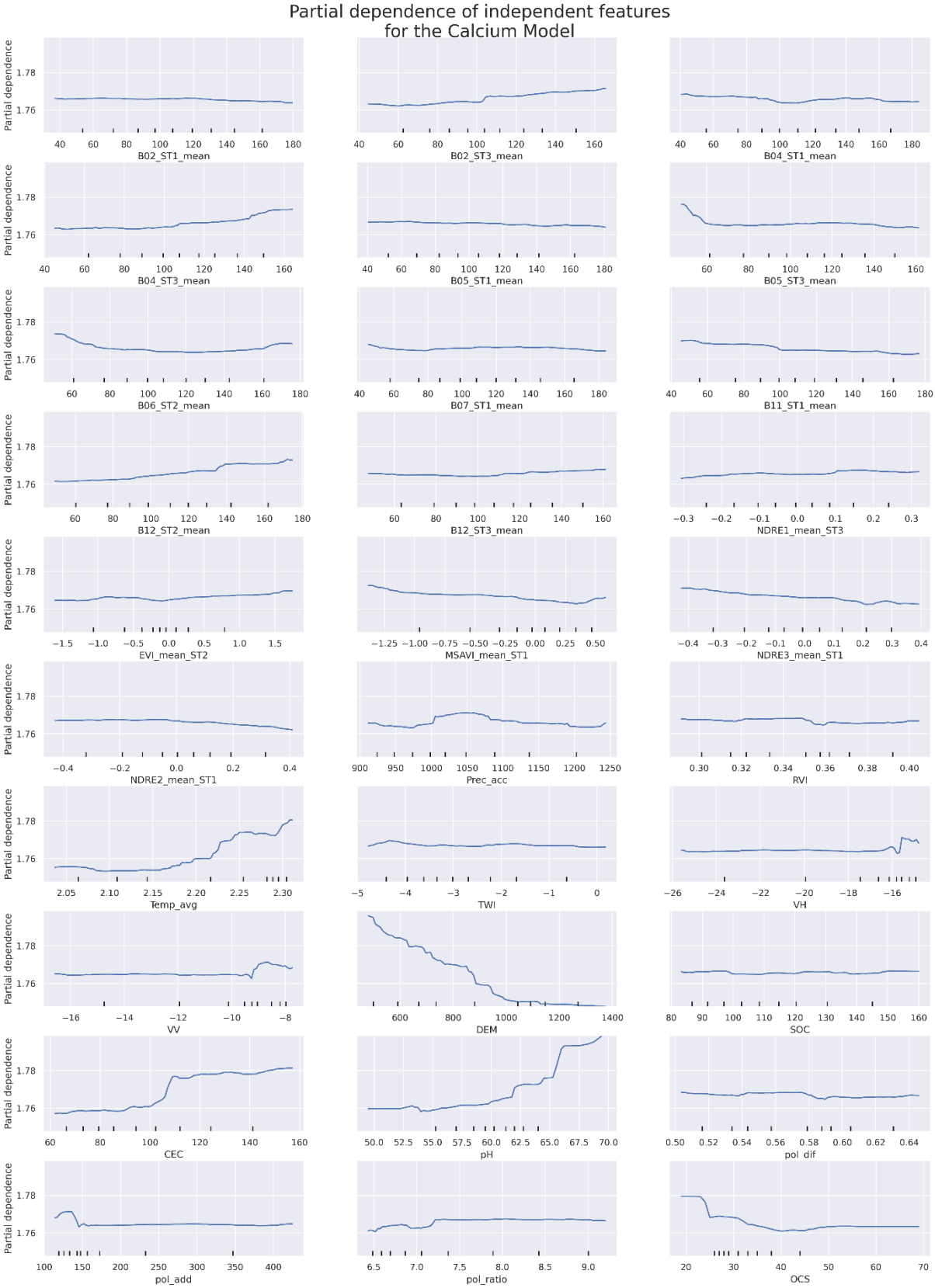


Figure 16 PDPs for all the features used for building Calcium model

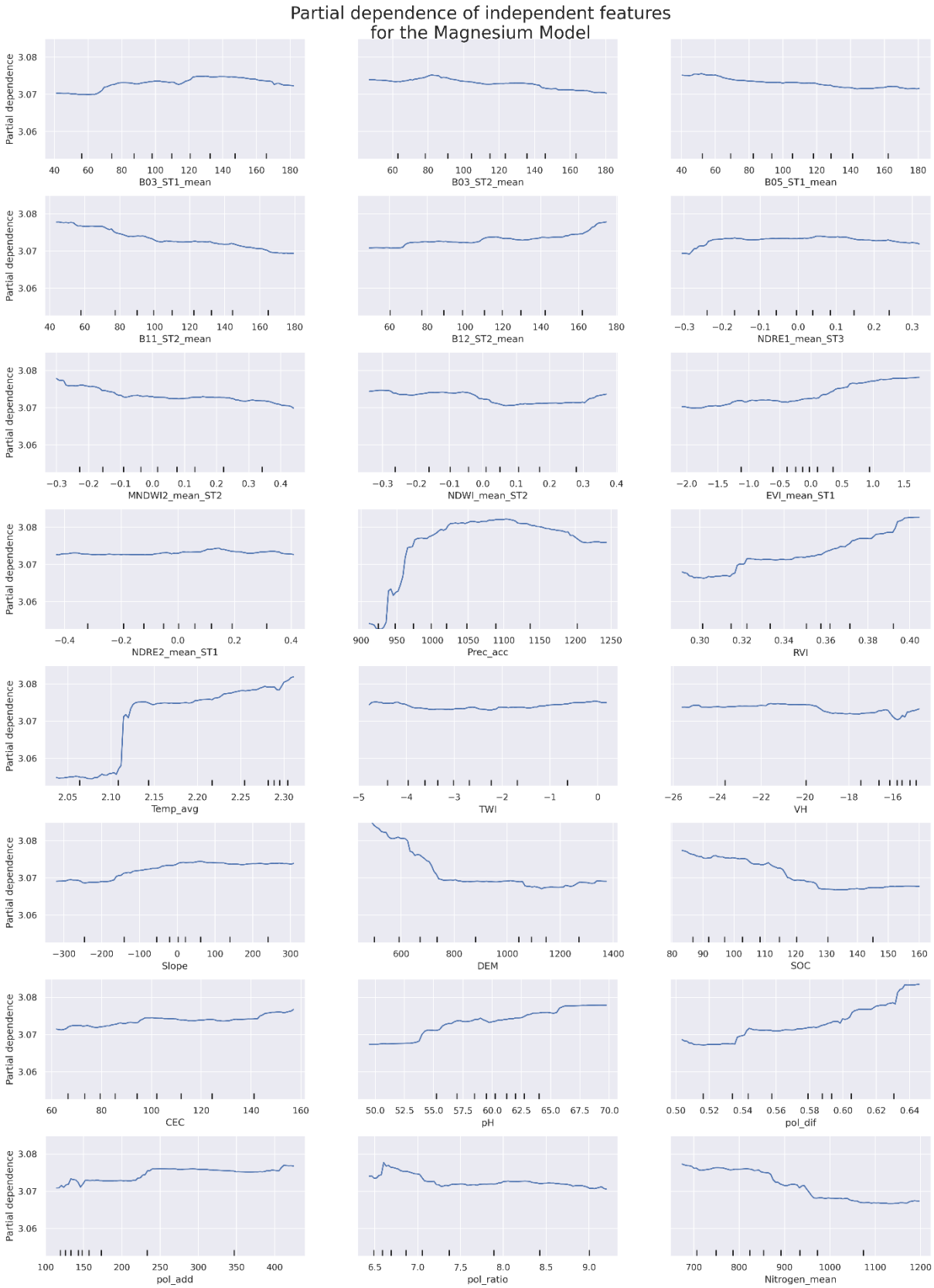


Figure 17 PDPs for all the features used for building Magnesium model

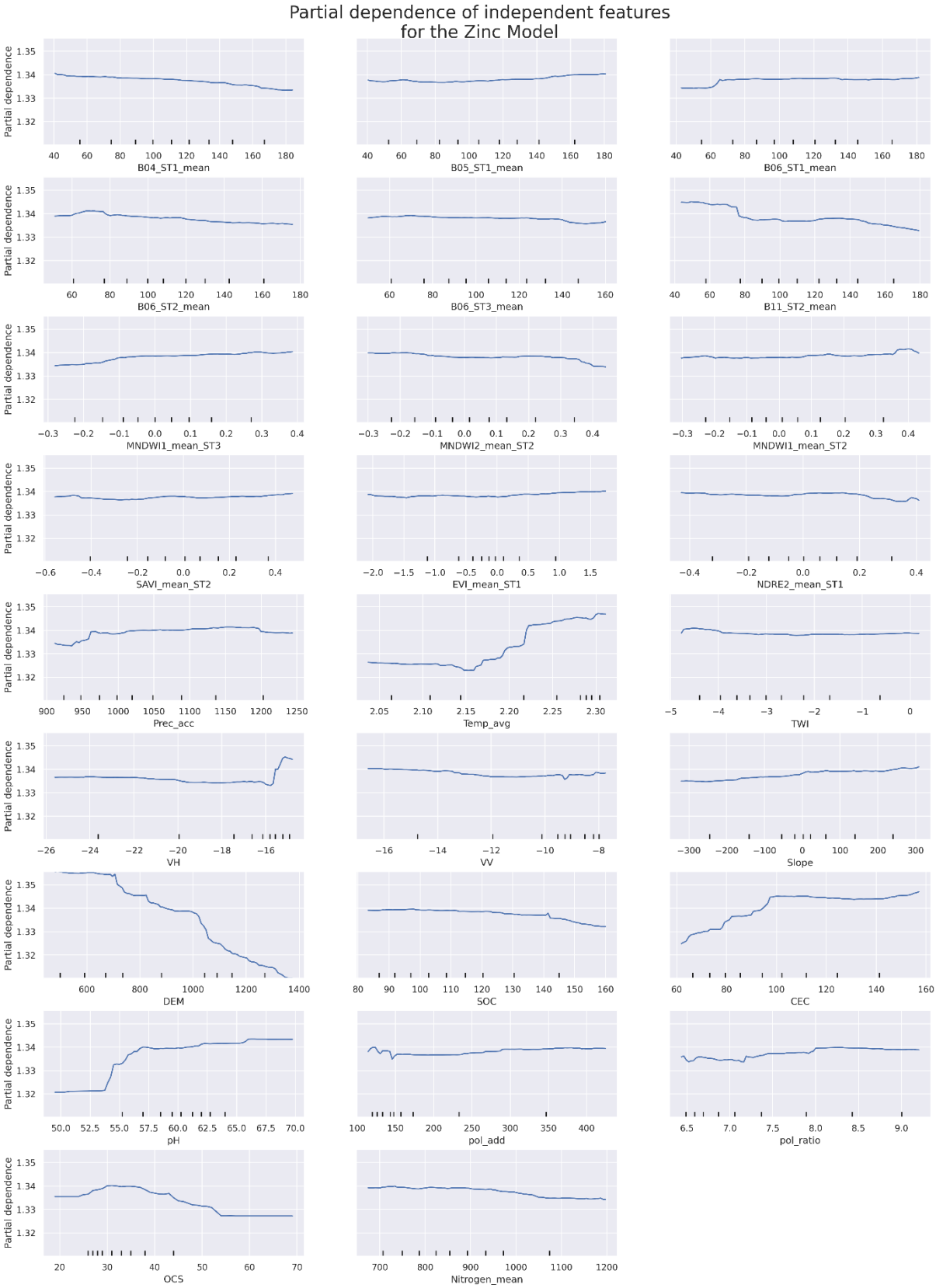


Figure 18 PDPs for all the features used for building Zinc model