FORECASTING OF THE RAIN-FED MAIZE YIELD IN TANZANIA USING MACHINE LEARNING

FORECASTING OF THE RAIN-FED MAIZE YIELD IN TANZANIA USING MACHINE LEARNING

By

Bertha Msuliche Lebalwa

A Dissertation Submitted in Partial Fulfillment of the Requirements for Award of the Degree of Master of Science in Information Technology and Systems (MSc-ITS) of Mzumbe University

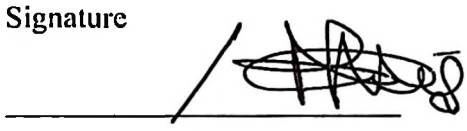
Mzumbe University

2022

CERTIFICATION

We, the undersigned, certify that we have read and hereby recommend for acceptance by the Mzumbe University, a dissertation entitled **Forecasting of the Rain-fed Maize Yield in Tanzania using Machine Learning**, in partial/fulfilment of the requirements for award of the degree of Master of Science in Information Technology and Systems of Mzumbe University.

Signature



Major Supervisor/Dr. Tupokigwe Isagah Mwalukasa

Signature

Internal Examiner/Dr.-Ing. Morice Daudi

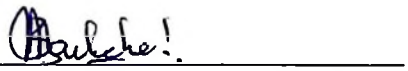
Accepted for the Board of Faculty of Science and Technology

Signature

DEAN, FACULTY OF SCIENCE AND TECHNOLOGY

DECLARATION AND COPYRIGHT

I, Bertha Msuliche Lebalwa, declare that this dissertation is my own original work and that it has not been presented and will not be presented to any other university for a similar or any other degree award.

Signature 

Date 12/10/2022

ACKNOWLEDGMENT

The author of this dissertation owes most sincere gratitude to almighty God for giving her strength, health and everything from the start to the end of this Research Project. The Author would also like to express deep and sincere gratitude to the supervisor, financial supporters, the Department of Computing Science Studies (CSS) under the Faculty of Science and Technology (FST), and colleagues.

First hand, the author appreciates the supervisor, Dr Tupokigwe Isagah Mwalukasa, for accepting to supervise her. Dr Tupokigwe's expertise and experience have been of great help to my research career. Her encouraging directives and suggestions on improving this document have been so valuable.

During this research, the author enrolled as a master's degree candidate in the CSS department. Therefore, the author thanks Dr Tupokigwe Isagah, the Head of the CSS department, for her commitment to organizing various activities regarding this research.

The author would like to thank the Higher Education Student's Loans Board (HESLB) and Sokoine University of Agriculture (SUA) for the financial support provided throughout her master's degree.

This work obtained and used the Ministry of Agriculture (MoA) yield data. The Author would like to appreciate the help and support from the Ministry of Agriculture in Tanzania.

The author would like to give great regard and warmest thanks to colleagues and master's degree candidates, Mr Jovin John, Mr Shadrack Mbilinyi, Mr Maseke Charali, and Mr Basilei Mkude for discussions, comments and encouragement.

Finally, I am grateful to my dear husband, kids, mom, dad and siblings for their enduring hearts and unending prayers encouraging me to pursue my dreams fully. They have provided me with all the love and support I have ever experienced.

LIST OF ABBREVIATION AND ACRONYMS

AgMIP	Agricultural Model Intercomparison and Improvement Project
ANN	Artificial Neural Network
ATDC	Agricultural Technology Demonstration Centre
CFNS	Comprehensive Food and Nutrition Security
ECAW	Enhancing Climate Change Adaptation in Agriculture and Water Resources in the Greater Horn of Africa
FAO	Food and Agriculture Organization
GHG	Green House Gases
GOT	The Government of Tanzania
INDVI	Integrated Normalized Difference Vegetation Index
IT	Information Technology
MAE	Mean Absolute Error
ML	Machine Learning
MOA	Ministry of Agriculture
MSE	Mean Squared Error
NFRA	National Food Reserve Agency
NOAA	National Oceanic and Atmospheric Administration
RF	Random Forest
RMSE	The Root Mean Squared Error
SDGs	Sustainable Development Goals
SHEP	Smallholder Horticulture Empowerment Project
SUA	Sokoine university of Agricultute
UN	United Nation
WRSI	Water Requirement Specification Index

ABSTRACT

Yield monitoring is vital for food security in the country. The uncertainty that may influence fluctuations in yield is impeccable to the nation's food security strategy. Due to climate change around the world, yield fluctuation has dramatically been affected and led to food supply shortage in most developing countries, including Tanzania, since most of the staple food depends on rainfed agriculture, which is hit by high temperatures and variations in rainfall. Bad weather also contributes to diseases, pests and weeds which greatly challenge the growth of crops, particularly the top grain and maize. Most countries have adopted several yield prediction methods to mitigate the effects to resolve the situation. Most countries have shown that emerging techniques such as machine learning provide good predictions to help countries mitigate the problem and secure food security.

Machine learning regression models (linear, AdaBoost, gradient boosting, k-Neighbour, random forest and stacking ensemble) are trained and evaluated using dataset obtained from the online database. Normalized Difference Vegetation Index (NDVI) is used to predict vegetation that are later assumed to locate probability of maize fields. The climate data from those areas is then subjected to training on forecasting maize yield.

Therefore, the results have shown that Machine learning methods like a stacking ensemble, which combine several other models and use Random Forest as the final model achieved low Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of 0.27, 0.12 and 0.34 Ton per Ha for each district respectively.

The results suggest machine learning model like stacking ensemble can be used by policy makers and farmers to mitigate the effects of climate change in yields for a particular season.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

This chapter presents the background of the research problem by outlining the research motivations and problem statement. It highlights the research objectives and questions of the study. Also, the chapter introduces research contributions to researchers and practitioners, and subsequently, a summary of the research report is presented.

1.1 Background

United Nations (UN) assembly set a resolution to attain the 17 sustainable development goals (SDGs) to improve the world by the year 2030[1]. One of the goals (SDG 2)[2] aims to end hunger by enhancing food security, improving nutrition and promoting sustainable agriculture. According to the UN, the world may achieve zero hunger by increasing crop production while conserving the environment (United Nations, 2017). As a result, developing countries are striving to realize the goal by taking various interventions to ensure food security by improving farming technology such as tractors, implements, inputs like seeds, chemicals and fertilizers, creating awareness and providing agro advisory programs and subsidies (Mgendi et al.,2021). For example, Egypt promoted the use of new environmentally friendly technology in Agriculture like biofertilizers, organic mulching and composite application to improve crop yield (Eid et al.,2018). Also, the Ethiopian government has dramatically improved information sharing and communication among local farming societies through smartphones with Global Positioning System (GPS) (Fabregas et al.,2019). Through smartphones, the government can deliver market information, provide relevant weather information, video-based farming advice and

[1] https://sdgs.un.org/goals
[2] https://sdgs.un.org/goals/goal2

pests information in specific areas to promote agriculture productivity (Fabregas et al.,2019). In Tanzania, the Agricultural Technology Demonstration Centre (ATDC) implemented a project called Smallholder Horticulture Empowerment Project (SHEP) that improves the productivity and livelihood of rural households by promoting technology transfer and exchange between Chinese and Japanese (Mgendi et al.,2021).

On the other hand, the world population is expected to increase to nine (9) billion by 2050 while the land remains the same (Underson, 2019). Consequently, in Africa, the population will double up to 2.5 billion (Tian et al.,2021 and Grote et al.,2021). Theoretically, this will imply high competition over natural resources such as water and land for agriculture increases pollution, possibly accelerating climate change (Tian et al., 2021). Climate change affects agriculture and makes crop yield unpredictable due to unprecedented weather calamities (Grote et al., 2021). For example, many crops die due to high temperature that increases evapotranspiration and moisture loss that dries seasonal streams and rivers used for irrigation (Tumbo et al.,2017). Furthermore, climate change causes a very high rate of land degradation due to desertification, thus reducing the quality of soil for crop production (Arora,2019). Arora et al., (2019) further highlighted a high increase in land degradation, a global threat to food security (Arora,2019). According to Aryal et al. (2020), land degradation has globally affected crop production and yield of the four most important food crops, i.e., rice, soybeans, wheat and maize (Aryal et al.,2020). As a result, there is a need to address climate change issues to improve agricultural productivity.

Looking into Tanzania, agriculture is the backbone of the economy since it employs more than 60% of Tanzanians (Society & Shelter, 2010). The industry brings more than 30% of the food exports and 65% of the raw materials used by the domestic industries (Society & Shelter, 2010). Most agricultural activities in Tanzania depend heavily on rain-fed production (Mourice et al., 2014 &Levira, 2009). This has gradually made the industry grow at a low speed because of limited irrigation

schemes and climate change (Doggart et al., 2020 & Newell et al., 2019). For instance, grain production, which dominates the industry by providing staple food, experienced stagnant growth in production for over 50 years due to unpredictable rain seasons and a lack of irrigation schemes to support crop production (Mourice et al., 2014 &Levira, 2009). These challenges accelerate food shortages in most cities in the developing world, including Tanzania (Ojija et al., 2017). According to Rowhani et al. (2011), intra-seasonal rainfall changes of more than 20% may dramatically reduce farm yield by 7.6%, 4.2%, and 7.2% for rice, maize and sorghum, respectively (Rowhani et al., 2011). Tumbo et al., (2017) further projected the decrease in maize (among the famous staple food) yields between 5% and 42% from the baseline due to climate change (Tumbo et al., 2017). Such a decrease due to climate change affects the food security in countries like Tanzania that depend on staple food. Therefore, monitoring and predicting crop yield is imperative to ensure food security in countries vulnerable to climate change (Grote et al., 2021).

Concerning crop yield prediction, Fue et al. (2017) presented a web-based tool called Enhancing Climate Change Adaptation in Agriculture and Water Resources in the Greater Horn of Africa (ECAW). It uses the output from the Agricultural Model Intercomparison and Improvement Project (AgMIP), which links the climate, crop and economic modelling communities with Information Technology (IT) to study the performance of the improved crop cultivar and its economic models that relate directly to the projections of the next generation agriculture (Rosenzweig et al., 2013). The ECAW tool visualizes weather and climate data interactively, making crop modellers understand the climate changes affecting crop growth. However, their tool does not provide interpretation and prediction of crop performance to help plan food security strategies. As a result, currently, Tanzania uses the Water requirement Specification Index (WRSI) algorithm to interpret weather information and predict crop performance. The algorithm incorporates a limited number of factors: temperature and rainfall, and assumes other crop management practices are constant (variety(cultivar), soil, fertiliser use, pests and diseases) (Tarnavsky et al., 2018; Boult et al, 2020). Also, the algorithm does not learn or adjust according to historical

yield information despite the potential of such data to reveal parameters that influence maize production in a particular location (Tarnavsky et al., 2018 &Senay, 2004). These shortcomings reduce the approach's accuracy in predicting crop performance, including yield. This calls for other methods with high prediction accuracy.

With the increasing amount of data in agriculture from geo satellites, the application of Machine Learning (ML) and other modern methods in agriculture to interpret the data accurately and timely is required (Kamilaris et al., 2017). Machine learning is a technique for learning from data and deriving a specific goal or prediction (Kamilaris et al., 2017). Machine learning has been used extensively on big agricultural datasets worldwide to predict yield and estimate crop biomass (Liakos et al., 2018). The datasets may include data that is structured, semi-structured, or unstructured. Unlike WRSI, ML prediction may consist of data with various parameters such as meteorological, crop factor, environmental, economic, and harvest (Liakos et al., 2018). For example, Kung et al., (2016) used meteorological data (air temperature, rainfall, and relative humidity), environmental data (growing area, yield, yield per unit volume and harvested area), and economic data (market prices and production costs) to predict tomato yield to near 77.68% using ML. Khaki et al., (2020) predicted corn and soybean yields using a machine learning method and achieved root mean square error (RMSE) of 9% and 8% of the average yield, respectively. Their model was robust enough to generalize yield prediction to unseen environments without dropping its accuracy. Henceforth, ML can significantly improve crop performance and provide accurate yield prediction. However, there is still a little exploration of ML in the prediction of maize, despite the potential of the approach in agriculture (Kamilaris et al, 2017). Therefore, there is a need to explore the applicability of machine learning in predicting maize yield in Tanzania.

1.2 Statement of the Problem

Climate change has affected rain-fed agriculture, particularly the staple food industry(Mourice et al., 2014). For example, crops like maize, which are popularly

consumed among large populations, have been hit hardly by climate change, leading to low production (Mourice et al., 2014). The Government of Tanzania (GoT), in collaboration with other stakeholders, conducted Comprehensive Food and Nutrition Security (CFNS) vulnerability assessment for crop production season 2016/2017 (URT, 2017). The study revealed that 35,491 tons of food, including staple food such as maize, were needed to feed 1,186,028 people from February to April 2017. 118,603 people (Destitute Population) were identified to require 3,549 tons of subsidized food since they could not access food (URT, 2017). GoT reported that in some regions, especially Geita and the northern zone, crops such as maize and beans have experienced wilting due to a lack of enough soil moisture (URT, 2016). In January 2017, the ministry reported that overall, Mwanza crop conditions were poor, and the neighbouring region, Shinyanga, was kept under watch owing to the delayed start of seasonal rains (URT, 2017). To resolve such problems in Tanzania, Five Year Development Plan II (2016-2020)[3] proposed the adoption of digital technologies to evaluate and monitor the performance of the crops. This initiative is also supported by the current Five-Year Development Plan III (2020-2025)[4], emphasizing modern digital technology such as artificial intelligence and machine learning to promote agricultural production and strengthen food security.

Maize is among the famous and essential crops in Tanzania, which is rain-fed and grown in various regions such as Mbeya, Iringa, Njombe, Morogoro, Ruvuma, Rukwa, and Katavi (Mourice et al., 2014). Unfortunately, there is low production of maize in Tanzania due to climate change, as highlighted by various studies (Mourice et al., 2014). The low production in maize yields is caused mainly by water stress due to unpredictable seasonal rains (Mourice et al., 2014). Though yield can be highly influenced by better crop management practices such as use of fertilizer and quality seed, natural weather conditions still affect maize performance in Tanzania

[3]https://extranet.who.int/nutrition/gina/sites/default/filesstore/FYDP2_II__April%201.pdf
[4]https://www.tro.go.tz/wp-content/uploads/2021/06/FYDP-III-English.pdf

(Mourice et al., 2015). Since most people in Tanzania depends on maize for food (Mourice et al., 2015). So, there is a need to monitor maize performance for high production. However, monitoring alone is not enough, it is important to forecast the yield to ensure community food security. The forecast inference will help propose mitigation strategies to enhance food security in case of low yield or crop failure. Currently, WRSI is in practice to predict maize performance and yield in Tanzania (Tarnavsky et al., 2018 and Boult et al., 2020). However, its accuracy is low (less than 0.61) because it uses only two parameters, temperature and rainfall, to predict the performance of the crop (Tarnavsky et al., 2018 and Boult et al., 2020) while the performance is also affected by many parameters. In addition, the WRSI provides categorical predictions like excellent, very good, good, satisfactory, and poor performance, which is not as important as a quantitative prediction to estimate food security (Tarnavsky et al., 2018, Verdin and Klaver, 2002, Senay and Verdin, 2003 and Boult et al., 2020). As a result, better approaches are needed to predict maize yield in Tanzania.

On the other hand, other parts of the world are using machine learning to manage crop performance and predict quantitative (numerical) yield information with high accuracy (Liakos et al., 2018).For example, Su et al., (2017) developed an ML model to predict rice yield and obtained a relative error of between 21% and 8.5%.Gandhi et al., (2016) used ML and got 78% accuracy on correct predictions. While Patanzi et al., (2016) predicted wheat yield using ML and achieved an accuracy of 78.3%. These performances of the machine learning models demonstrate superiority over the WRSI, which uses limited features.

With regard to maize yield, using the Machine Learning model in prediction increases the prediction efficiency in South Africa (Adisa et al., 2019). Moreover, Wang et al., (2018) developed Baseline and ML Model that showed promising results in predicting soybean and maize crop yields in Argentina using ML techniques for utilizing remotely sensed data, such as satellite imagery. But, all these existing ML models for maize prediction were developed differently from Tanzania

Agricultural practices. For instance, the models were developed based on irrigation agriculture and unimodal seasons, while Tanzania agricultural practices are based on rain-fed agriculture and involves bimodal seasons. In Tanzania, Laudien et al., (2020) attempted to apply statistics using basic machine learning models to predict maize yield to about six weeks before harvest, which is important for a farmer and policymakers to do farm management amendments to mitigate crop failure in case it has been detected earlier. The model is based on linear regression, a basic machine learning model that tries to predict yield based on given independent inputs (climate data) without including time as a parameter. Thus, it is not suitable for time series data like yield data which is crucial in learning crop performance based on historical data. Henceforth, there is a need to explore the applicability of machine learning algorithms suitable for time series data in Tanzania for maize yield forecasting.

1.3 Research Objectives and Research Questions

Generally, the study aimed to forecast rain-fed maize yield in Tanzania in a crop season using Machine Learning. The developed model uses climate and yield data to predict maize yield. The following were the specific objectives of the study.

i. To identify appropriate climate parameters and yield data from the available open data sources.
ii. To develop machine learning models for the maize yield forecasting.
iii. To evaluate machine learning models for the maize yield forecasting.

To achieve the first specific objective, identified climate parameters from the literature and yield data were collected from Copernicus and ministry of agriculture databases, respectively. Then the data were cleaned and arranged in a tabular format, ready for processing. As well, for the second specific objective, ML algorithms for prediction were identified from the literature, trained and tested using collected data by setting experiments. The performance of each model was compared and tuned to achieve the best level so as to ensure a fair comparison of the models. Finally, the

developed models were evaluated using testing datasets to determine their accuracy in crop prediction.

1.3.1 Primary Research Questions

Apart from the stated objectives, this study attempted the following research questions:

Q1: What are the main climate parameters contributing to the prediction of maize crops yield in rain-fed agriculture in Tanzania?

Q2: Can ML models forecast maize yield in Tanzania?

Q3: How accurate is ML model in forecasting maize yield in Tanzania?

1.4 Contribution of the Study

The study intended to predict maize yield in Tanzania using Machine Learning. Through achieving the objectives, findings of this research contribute to the body of knowledge (research) and practitioners as follows.

1.4.1 Contribution to Research

The key output of the research is the ML model that demonstrates its applicability and usefulness in predicting maize yield in the Tanzanian context. The model contributes to the research by supporting the potential of machine learning algorithms in prediction. It ignites the interest of the researchers to explore other machine learning models that have not been covered in the research, beyond black-box forecast and toward interpretability of agricultural mechanism. Also, the study provides a baseline of the next stage to improve the algorithms to predict maize yield.

The study adds knowledge to the literature on using a combination of multiple ML algorithms and various parameters for better accuracy in maize prediction. Also, findings from the demonstration of the model in the Tanzania context (rainfed

agricultural and bimodal seasons) bridge the gap in the literature on lacking such contexts.

1.4.2 Contribution to Practice

Through the demonstrated model, the Tanzanian government will be able to assess and analyze the state of food levels from the prediction and prepare strategies to mitigate food shortage with the help of the National Food Reserve Agency (NFRA). Usually, the government, through the ministry of agriculture's national food security division, has a section called Crop Monitoring and Early Warning that prepares monthly reports on crop production forecasts using manual work. Thus, the developed model, once implemented, will provide the estimation earlier.

For policymakers, the demonstrated model will prompt evidence-based agricultural policy using historical agricultural and inference provided by the machine learning model. These will help the Ministry of Agriculture (MoA) and Local government authorities (TAMISEMI) to prepare enough funds and subsidies to mitigate the problem. Also, knowing the extent of the problem and targeted regions (early prediction) would give them a quick response, especially in supplying food to anticipated communities (low maize production) and delivering alternative advice to farmers to change their farming practice to ensure agricultural productivity.

For farmers, the output of the study will help them understand prior conditions that would lead to crop failure because machine learning can explicitly show features that may influence the yield. As a result, they mitigate their crop management approach at an early stage.

1.5 Outline of the Research Report

This report consists of five (5) chapters. Chapter one presents the background of the study, the problem statement, research objectives and research questions. Also, in this chapter, the contributions to the research and practice are highlighted. Chapter two discusses the literature review on climate change's effect on agriculture and

maize yield. The theory of forecasting, the Machine Learning approach to maize prediction, and different prediction models' strengths and limitations are presented in the chapter. Also, the research gap is identified, and research objectives are justified.

Chapter three explains the methodology adopted to carry out the study. This includes the data collection and cleaning, algorithm selection, model development training and validation. Chapter four presents and discusses results from the developed, trained and validated model, including the accuracy and errors. Chapter five concludes the conducted study by providing a summary of the research objectives, findings, limitations and areas for future research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter reviews previous works on monitoring crop yield, effects of climate change on crops and yield forecasts. It provides a review of climate change's impact on agricultural productivity. Also, the chapter highlights the strength and weakness of the existing approaches in predicting crop yield and establishes a need for using machine learning to forecast maize yield in order to increase food security. Existing initiatives taken to address the challenges are elaborated on in the chapter.

2.2 Agricultural Productivity and Climate Change

Increasing crop production is essential to curb the food requirements of the fast-growing world's population which is projected to reach 9 billion by the year 2050 (Bhakta et al., 2019). In 2018, World Food Programme (WFP) report revealed that the increase in farm yields per hectare is significantly slower when the rates of population rise (Arora, 2019). The increasing demand for food due to the ever-growing population has resulted in intensive agricultural practices affecting the productive land. According to Bhakta et al. (2019), as the population increases, up to 1.3 billion (19%) people will be engaged in agricultural practices to provide enough food (Bhakta et al., 2019). This growth in agricultural practices may temper the quality of land and, as a result, lower crop production. To address this, new advanced techniques like using hybrid seeds, which require a lot of fertilizer, are currently used; however, such practices may detrimentally affect soil fertility, lead to health hazards and increase environmental pollution (Bhakta et al., 2019). Also, aggressive measures to improve productivity may spike irrigation water use in extensive areas and lead to soil erosion and an increase in soil salinity, which declines soil biodiversity, making it unproductive. (Lin et al., 2020). Therefore, population increase directly affects factors that influence crop production, including

plantation area, variations in rainfall and temperature (climate change), and soil quality (Agarwal and Tarar, 2021).

Climate change and agriculture have tricky links; unforeseen changes in climatic conditions have threatened food security globally (Arora, 2019). According to Thornton et al., (2018), climate trends negatively affect agricultural production. For example, the rise in temperature leads to a decrease in rain, affecting fishing ecosystems and crop production. Over the previous 30 years, climate change has reduced wheat and global maize production by 1-5% per decade globally (Loboguerrero et al., 2018). In 2016 Food and Agriculture Organization (FAO) report revealed that if emissions of Green House Gases (GHG) continue, there will be a decrease in significant staple food production from 5-50% for wheat, 20-45% for maize, 20-30% for rice and 5-50% for maize by the year 2100 (Zougmoré et al., 2018). Climate change and variability are significant threats to sub-Saharan Africa's development since most countries' economies rely on agricultural activities based on rainfed agrarian production. Thus, the variation in climate patterns has affected the economy of such countries, including Tanzania leading to poor agricultural productivity. Ojija et at., (2017) revealed that fluctuation in precipitation and temperature patterns in some cities in Tanzania caused food shortages. For instance, a decline in rainfall makes water inadequate for farming activities (Ojoyi et al., 2017) because many seasonal streams and rivers used for irrigation depend on rains. Additionally, many crops die due to heat stress caused by a high temperature that increases evapotranspiration and moisture loss (Ojoyi et al., 2017). In Tanzania, the seasonal temperature increases of about 2°C as predicted by 2050 will dramatically affect the production of sorghum, rice and maize by 8.8%, 7.6% and 13%, respectively (Rowhani et al., 2011). At the same time, a 20% rise in rainfall changes will reduce crop yield by 7.6%, 7.2%, and 4.2% for rice, sorghum, and maize by 2050. This decline in crop yield threatens food security in communities depending on such crops (Rowhani et al., 2011).

According to Grote et al. (2021), having a good amount of agricultural products in societies enhances food security. As a result, it is important to foresee crop yield to ensure food availability in the next year despite the challenges (Grote et al., 2021). Also, crop yield can be improved by detecting and dealing with the factors that affect production, including climate change, diseases and pests related to crop yields in the early stages (Chaudhary and Kausar, 2020). Detecting such factors may need data from weather stations and large-scale meteorological datasets that may be used to improve decision support systems and hence increase crop yield (Burke and Lobell, 2017).

2.3 Crop Yield Predictions

Crop yield prediction is essential in the domestic food supply, farm planning and management, ecosystem sustainability and international food trade (Chaudhary and Kausar, 2020). It is one of the essential aspects of agriculture as it helps in decision-making at global, regional and field levels. For instance, for a country with the largest population and scarce farmland like China, accurate crop yield prediction helps the governments supply enough food by estimating the amount of production in a particular season of planting (Han et al., 2020). Similarly, a country with ample farmland with few people like Australia relies on crop yield prediction to optimize the production of farm crop exports to other needy countries like China (Cai et al., 2019). On the other hand, farmers can utilize crop yield predictions to reduce losses when unfavourable adverse conditions occur (Burke and Lobell, 2017). Also, predictions can be used to maximize crop production since a farmer may decide to change farming management, such as fertilization and irrigation style, to enhance a favourable situation for farming (Burke and Lobell, 2017).

There are several ways of estimating crop yields, like field scouting and surveys that are trusted to produce excellent yield estimates (Burke and Lobell, 2017). Still, these approaches are labour intensive, costly, and can be challenging in large-scale farming (Burke and Lobell, 2017). Alternatively, crop yield can be estimated using crop growth models such as regression growth model and mechanistic growth model

incorporating eco-physiological processes in simulating crop development from planting to harvesting according to farm management practices, soil properties and weather and climate data (Burke and Lobell, 2017). Crop growth models mimic the crop lives by simulating interactions of the plants and their environment. It is possible to predict crop yield that can drive agronomic decisions in crop management (Burke and Lobell, 2017). The models also study the potential impacts of climate change on food security (Burke and Lobell, 2017). Inadequate spatial information on actual conditions of the farm, such as soil, crop and weather information, may hinder the performance of the crop growth models (Burke and Lobell, 2017).

Furthermore, predictions may be presented qualitatively or quantitatively. In the former, results from prediction can be expressed in words or meanings such as 'high, low, medium, good and poor.' The qualitative prediction is subjective according to the opinion of the experts or farmers or algorithms and is mainly made when no historical data is available (Dambacher et al., 2019). Quantitatively, the prediction results are presented in numerical numbers or variables, for example, in specific tonnes per hectare, and it is analyzed on historical data (Basso et al., 2019). Also, quantitative forecasting requires understanding the numerical past data and knowledge if future data will have a relationship with the past data (Basso et al., 2019). It is measured by determining the accuracy, which dictates how the model accurately forecasts the yields (Basso et al., 2019).

There are several methods used to predict crop yield apart from direct methods. This research identified and explored these methods into two categories presented subsequently.

2.3.1 A Review of Crop Yield Predictions using Statistical and Empirical Methods

Statistical prediction methods use historical data to build mathematical models that capture essential trends (Basso et al., 2019). Also, the model uses current data to predict what will happen next based on the captured trends from historical data, and

it can suggest actions to take for best outputs (Kamilaris et al., 2017). Also, the statistical prediction method estimates the relationship between dependent variables, such as yield and one or more independent variables, such as temperature, precipitation, humidity, wind speed, sunlight and soil moisture (Debnath et al., 2018). The dependent variable refers to the variable being tested or measured, while the independent variable is not influenced by other variables measured in an experiment (Evstatiev&Gabrovska-Evstatieva, 2021). In prediction, an investigation is conducted to determine how the independent variables affect the dependent variable, and regression analysis is used to achieve the goal (Evstatiev&Gabrovska-Evstatieva, 2021). Regression analysis is a statistical method in the prediction that can be used to assess the relationship between variables so as to model the future relationship between them (Basso et al., 2019). Regression analysis can be linear, multiple linear and nonlinear, used to determine the relationship between dependent and independent variables (Debnath et al., 2018).

Robert et al., (2018) demonstrated a significant relationship between extreme heat and its effects to crop performance in predicting corn yield from the United States (US) Midwest. The work improved the statistical work Schlenker and Roberts' classical econometric method in 2009. The improved model fused with statistical models (including crop model output within statistical models) used crop models output to parametrize statistical models (Roberts *et al.*,2018). Although the model has shown improvement in prediction, it has some limitations on adding parameters and cannot learn from historical data in prediction. Thus, Adisa et al., (2019) suggested the use of machine learning to improve the performance of the model as it has been proved by other researchers like to perform better than traditional statistics.

Furthermore, Peng (2019) presented statistical modeling for the prediction of rain-fed corn production in the Midwest U.S. Using satellite variables, climate variables, and country-specific fixed effects, and they predicted raw yield after detrending. The model achieved high yield prediction performance, RMSE of 1.04 t/ha (16.6 bu/acre) and R2 of 0.79 from 2003 to 2016 (Peng, 2019). However, it failed to include many

parameters, learn from historical data and cannot suit time series data like yield data well. Adding such parameters might have increased the model's performance.

Apart from purely statistical prediction models, there are empirical prediction models, sometimes called statistical models (Lai et al., 2018). An Empirical model is a mathematical formulation based on literature facts, reaction rates and input parameters (Lai et al., 2018). The empirical model is developed by relating inputs to the respective output using the common established methodology. The model relies on observation rather than theory, meaning that if some particular outcome is observed following some specific circumstance, then the model can reliably predict that outcome in future (Lai et al., 2018).

Lai et al, (2018) predicted wheat yield at a field level using an empirical model in northern Australia. Using Landsat time-integrated Normalized Difference Vegetation Index (iNDVI), they trained a model to predict yield from data obtained from 17 farms from 2001 to 2016. The model showed promising results by reaching an average RMSE of 0.79 Mg/ha. Unfortunately, it failed to learn from historical data, despite being supplied with data for specific fields, including farms and years where production happened. Also, they concluded that the empirical model prediction was unsuitable for time series data like yield data (Lai et al, 2018).

Additionally, Tarnavsky et al., (2018) developed an empirical model known as Water Requirement Specification Index (WRSI) for maize yield prediction in Tanzania that characterized the impact of using different rainfall input datasets like African Rainfall Climatology Version 2 (ARC2), Climate Hazards Center InfraRed Precipitation with Station (CHIRPS), and Tropical Application of Meteorology Using Satellite (TAMSAT), on key WRSI model parameters and obtained very low correlation ($R^2 < 0.61$). Basically, WRSI indicates the crop's performance during the growing season and whether it can manage to maturity or fail before maturity to water stress (McNally et al., 2015). The model used temperature and rainfall parameters and provided categorical predictions like excellent, very good, good, satisfactory, and poor performance (Tarnavsky et al., 2018). This is a challenge

because the numerical prediction of yield is required to estimate the food security of the country (Tarnavsky et al., 2018, Verdin and Klaver, 2002, Senay and Verdin, 2003 and Boult et al., 2020).

To improve those limitations, Peng (2019) recommended the application of the emerging trend of machine learning models for yield prediction. Furthermore, the massive amount of data available in the prediction industry has been among the challenges in prediction accuracy when using traditional statistical Models or WRSI (Tarnavsky et al., 2018, Verdin and Klaver, 2002, Cai et al., 2017; Mathieu and Aires, 2016).

2.3.2 Machine Learning (ML) Approaches in Crop Prediction

Machine Learning (ML) is the fundamental field of artificial intelligence. It makes computers infer the data without being explicitly programmed (Doupe, Faghmous, and Basu, 2019). When these computers are fed with new data, they learn, grow, change, and develop by themselves. Thus, Machine Learning (ML) models reach their goals by determining relationships in patterns and correlations in datasets. The models are subjected to training data and expected to learn from the data based on experience (Adisa et al., 2019). As a result, several fields such as healthcare, construction, transportation, social media, politics, business, security, education, fishing, and mining, apply ML models to improve business operation and service provision (Adisa et al., 2019). For example, hospitals use machine learning to predict diseases and organ failures (Qiu et al., 2019). Such prediction improves the quality of service and reduces the chances of dying due to late diagnosis. ML is used in social media to recommend users about people to follow (e.g., Instagram and Facebook) from the prediction based on tags, likes and trends (Balaji et al., 2021).

According to Adisa et al. (2019), ML can reveal important "hidden" features from big historical datasets and forecast unseen environments better than statistical approaches. However, machine learning methods are developed because their operations are not so understandable, and it can be cumbersome to trace their

operations compared to statistical methods. Adisa et al. (2019) further argued that combining statistical and machine learning models improves performance. Therefore, statistical methods are still used in crop yield prediction problems to enhance interpretability, clarity, and simplicity.

ML can be used to describe or predict depending on the problem or goal of the researcher. In describing knowledge out of the field obtained data, descriptive models are used, while on prediction of the future trends of data, predictive models are used (Alpaydin, 2010). ML models work like human brains by finding patterns in the data to forecast inference of the data or by learning on past data and its pattern and predict future inference learning (Dutta et al.,2020).

Regarding agriculture, ML can be used to manage crop growth and predict yield, which appears to be among the challenging problems in precision agriculture (Xu et al., 2019). This is because crop yield is determined by the quality of inputs parameters available timely while in the growth stage, including soil texture, fertilizer, and weather (Xu et al., 2019). Thus, crop growth and yield prediction management require large data sets from several parameters. Therefore, selecting suitable algorithms to solve problems is vital, and these algorithms must be capable of handling the volume of data (Alpaydin, 2010). For example, Dutta et al., 2020 used several machine learning approaches (random forest, Support Vector Machine and Artificial Neural Network) in studying the influences of socio-economic, management and biophysical events in describing maize yield. The main parameters incorporated in their models include farm details such as farm location, size, soil properties, fertilizer amount and type, labour force and seed rate. Unfortunately, models such as artificial neural networks (ANN) performed miserably by scoring an average of 75% on the classification on validation datasets. On the contrary, the random forest showed robustness in describing the association between maize productivity and farm size (Dutta et al., 2020). However, this study did not integrate crop simulation or empirical growth models to enhance yield estimations.

Crane-Droesch et al. (2018) described an approach to yield modelling using semiparametric neural networks (SNN). They used corn production data from US Midwest. SNN has shown robustness in high-dimensional datasets by revealing complex nonlinear relationships (Crane-Droesch et al.. 2018). Their approach has demonstrated superior performance compared to traditional statistical methods and neural networks in predicting corn yield.

Wang et al., (2018) developed deep learning (an ML model) models for working with satellite imagery data that would provide affordable crop yield prediction. Their work delivered an excellent model that predicts soybean crop yield in Argentina. The inexpensive method demonstrated that transfer learning provides an opportunity for developing countries to utilize remotely sensed data (Wang et al.,2018).

According to Adisa et al.(2019), using machine learning approaches in crop modelling increases the production efficiency. In their study, they used a particular type of ML model called the artificial neural network (ANN) model, which was used to predict maize in the major maize-producing provinces of South Africa. They used the following climate variables: potential evapotranspiration (PET), precipitation (PRE), maximum temperature (TMX), land cultivated (Land) for maize, minimum temperature (TMN), and soil moisture (S.M.). The 28-year data from 1990 to 2017 was divided by 20% and 80% for testing and training datasets respectively. The output showed that the maize yield was more influenced by TMN, S.M., TMX, PRE, Land, and PET in the Free State province (Adisa et al., 2019).

Leroux et al. (2019) researched to predict maize yield using multi-domain remote sensing indices in West Africa. Their study combined Remote sensed data, crop modelling, and statistical method to build a model that overcame the limitation of crop yield estimation due to the Direct techniques based on field surveys. Surface Soil Moisture (SSM), Moderate Resolution Imaging Spectroradiometer (MODIS), Normalized Difference Vegetation Index (NDVI), and SMOS (Soil Moisture and Ocean Salinity) and MODIS Land Surface Temperature (LST) were utilized to determine phenotypic characteristics of the vegetation stress, drought and vigour.

Multiple Linear (MLR) and nonlinear Random Forest (R.F.) models were tested and compared. The Random Forest (R.F) model, a nonparametric algorithm, and Multiple Linear (MLR) models were then calibrated separately for AGB-F and Cstr and the nonlinear model (R.F) has shown the robust capability to predict crop performance which is a highly nonlinear parameter.

Most models mentioned above are combined (ensemble) to form an agricultural yield prediction model, dramatically increasing accuracy (Shahhosseini et al., 2020). Hence, machine learning models have shown outstanding robustness and accuracy in crop yield prediction in many countries. As a result, this research aimed to explore the applicability of ML for crop yield prediction in Tanzania context using maize crops as a case study.

From the review of existing prediction models discussed in sections 0 and 0, Table 1 presents a summary of the strengths and weaknesses of each model.

Table 1: Strengths and Limitations of the Existing Prediction Models

Prediction Model	Strength	Limitation
WRSI (Water requirement specification Index)	• Does not need historical data • It is easy to implement • It incorporates the agricultural methodology of raising crops. For example, the Start of the season, the middle of the season and the final season • It incorporates factors that are specific to a particular crop. For example, how much a specific crop needs water and how much it loses its water	• Its prediction is Linear since it does not include historical data • Its output is qualitative ie., excellent, very good, good, average, poor
Regression Model	• It can predict yields anomalies • It can predict absolute yield for about six weeks • It is easy to implement	• Sometimes, yield outliers can be influenced, especially when the data has many outliers • It is not suitable for time series data like yield data • Its accuracy depends on the linearity of the data, so it is not suitable for nonlinear environmental data
Conventional Machine Leaning, (CML) - (Random Forest (RF), Adaboost, gradient boosting, and kneighbor regression)	• It can embed nonlinear data like agricultural data hence can-do nonlinear decisions • It can provide Quantitative output with acceptable accuracy • Training requires a short time, with a limited amount of data • CML tuning is easy compared to Artificial Neural networks since it is insensitive to randomization of weight	• Providing more data in training does not improve the accuracy • Most of the research has shown conventional machine learning provides low accuracy in prediction compared to Neural networks • CML performance is only comparable to the shallow neural network
Neural Network: Artificial Neural Network, Deep learning Neural Network (Convolutional Neural Network and Recurrent Neural Network)	• Providing more data in training increases the accuracy in prediction of yield • It can embed nonlinear data like yield data hence can-do nonlinear decision • It can provide Quantitative output with acceptable accuracy which is good for yield data • Works best with time series data like yield data especially Recurrent Neural Network) • Most of the current research has shown that deep learning Neural Network provides More accuracy in prediction of yield	• Training takes longer time • It is complex and deep Neural Network • Need more computing resources

From Table 1, the conventional machine learning model does not need many training resources such as time, computing chips and tuning hyperparameters compared to deep learning. It has shown the essential requirement of predicting nonlinear data such as maize yield. Hence, this study explores conventional machine learning methods (such as random forest, adaboost, gradient boosting, etc.) and all available climate data parameters (such as temperature, soil temperature, Evaporation, Precipitation, Leaf area index, Soil water, Wind speed, Runoff, etc.) from the online datasets.

2.4 Maize Yield Prediction Models in Tanzania

Maize is an essential crop in Tanzania, and it grows in all regions as a rain-fed crop. Fortunately, the leading producer of maize in Sub-Saharan Africa is Tanzania. For over forty years, Tanzania has been rated highly among the best 25 maize-producing countries (Mourice et al., 2014 &Levira, 2009). Most of the production is done by smallholder farmers. For example, in 2013/2014 season, Tanzania produced 500 million metric tons of maize, in which 85% was contributed by smallholder farmers (Society & Shelter, 2010). Unfortunately, Tanzania has been experiencing a drop in maize yield for the past three decades due to climate change affecting important factors for maize production like soil water, evaporation, precipitation(rainfall), soil temperature, and air temperature (Kukal et al., 2018). This led to low production of maize crops, affecting communities' food security. Henceforth, it is important to predict maize yield to inform the stakeholders of the performance to ensure food security in the community.

Different studies have discussed several models used to predict maize yields in Tanzania. Liu & Basso, (2020) associated field survey data with crop modelling to improve forecasts for maize yield in Tanzania. They used a crop simulation model called Systems Approach to Land Use Sustainability (SALUS) to integrate field-based survey data. SALUS was run using farming parameters such as soil water inputs, crop and weather inputs, and management inputs. They downloaded temperature and soil data from the Modern-Era Retrospective Analysis (AgMERRA)

for the years 1981 to 2010 (Funk et al. 2015; Ruane et al. 2015). Then, they extracted rainfall data from 0.05°-resolution Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) gridded dataset (Funk et al. 2015; Ruane et al. 2015). In order to have a well-descriptive experiment, three maize varieties were identified for simulations. The varieties represented short, medium, and long-duration cultivars. Then, they utilized four types of soils; extremely fertile, fertile, medium and low fertile soils. Also, on-farm management they considered fertilizer application, planting rates/densities and irrigation amounts as reported from the survey done for Morogoro, Kagera, and Tanga districts. Then, they simulated the data using SALUS to find yield information in Tanzania. Research in this thesis used the real data reported from the ministry and climate reanalyzed data from Copernicus climate store, which is similar to the SALUS model but simulated with a more advanced model. The author of this dissertation opted for data obtained from the Copernicus climate store.

Currently, Tanzanians use Water Requirement Satisfaction Index (WRSI) as a method of crop yield prediction. Tarnavsky et al., (2018) developed an adapted WRSI model for maize yield prediction in Tanzania. The model is based on a procedural model that incorporates input parameters such as rainfall, soil water holding capacity, temperature, the start of the season, length of the growing period, end of the season, and crop factor data. WRSI incorporates a limited number of factors that affect crop performance, only temperature and rainfall. It assumes other factors are constant; hence, it is not very accurate (Tarnavsky et al., 2018). Also, WRSI model does not learn or adjust according to historical yield information. This is a weakness since historical yield may reveal factors influencing maize production a particular location (Tarnavsky et al., 2018 &Senay, 2004). For example, research on the WRSI model on maize production and yield in Tanzania resulted in a very low correlation ($R^2 < 0.61$) of the prediction of maize yield performance (Tarnavsky et al., 2018). Boult et al., (2020) conducted a similar study using WRSI for maize yield prediction in Kenya and got a correlation of about 0.38, which is relatively low. The low accuracy is due to the fact that WRSI is designed as a water-balanced crop

model, thus focusing on water requirement compared to other factors that may contribute to crop failure (Boult et al., 2020). These findings draw a need for using other modern methods such as machine learning, which has been mainly successful applied in agriculture.

Laudien et al. (2020) attempted to apply statistics using basic machine learning models to predict Maize yield. They demonstrated maize yield forecasting per region of Tanzania and predicted yield anomalies and absolute yield about six weeks before harvest. They achieved a median Nash-Sutcliffe efficiency coefficient of 0.72. The model is based on regional regression model, which tries to predict yield based on given independent input values (climatic data). The model is based on Linear regression, a basic machine learning model that tries to predict yield based on given independent inputs (climate data). Since this model was developed using a basic machine learning algorithm unsuitable for time series data like yield data, it cannot be accurately reliable. Also, no studies have explored successful machine learning methods like boosting algorithms, random forests and recurrent neural networks in yield prediction in Tanzania which are suitable for time series data. Henceforth, it is important to demonstrate the use of improved ML algorithms in yield forecasting in Tanzania to improve the current situation and help mitigate and adapt to climate change.

2.5 Findings from the Literature Review

The literature revised various topics that show how climate and other parameters may affect the crop performance(yields) on the farm, specifically, maize yield. Rainfed agriculture is more prone to climate change, and it is threatening crop production and rural livelihood as a result. Several models have been deployed worldwide to understand the yield and ultimately predict it. Models may be based on empirical formulas like WRSI or conventional statistical methods. However, both traditional approaches have shown failure in robust prediction in modern times.

Fortunately, machine learning methods and deep learning have shown more strength in yield prediction worldwide. It is of paramount importance to accurately predict the performance in qualitative like how it is done with WRSI and in quantitative, as it is presented using regression models. AI and ML have shown superiority in this area, but unfortunately, they are not yet deployed in Tanzania. Building from this fact, this thesis is delivering a state-of-the-art model that has been trained using machine learning to demonstrate the accuracy and feasibility of using the models on the national stage.

From the above review, it was decided to use the following parameters; Temperature, Soil temperature, Evaporation, Precipitation, Leaf area index, Soil water, and Wind speed to predict the yield using conventional machine learning models such as random forest, adaboost, linear regression, gradient boosting and stacking model.

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter discusses the research methodology by pinpointing tools used and methods designed and deployed. It explicitly shows how data was obtained and processed technically. It discusses the research design and approach of this thesis. Also, the chapter shows details of all algorithms and models deployed to train and test machine learning models. Procedures taken to conduct machine learning experiments are described in this chapter. Subsequently, the chapter highlights mechanisms used to ensure the responsible design of the demonstrated approach.

3.1 Research Design

The research design used to achieve the study's objectives was based on the methodology described by Bhavsar et al., (2012). According to (Bhavsar et al., 2012), ML projects have mainly four stages. Firstly, problem establishment and project set up where by the problem to be worked on was decided. Secondly, data collection and preprocessing where by data was fetched from online databases and then formatted in a format that would be understandable for an ML model to learn from. Third, Model development training and validation whereby the models were set with optimized parameters and trained well; lastly, model testing and evaluation whereby the trained was tested for its inference to estimate its coefficient of determination and errors. These stages are further explained subsequently.

3.1.1 Problem Establishment and Project Set-up

A thorough literature review was conducted to establish the problem. According to Hart (1998), a literature review aims to examine previous assertions of the research objectives and seek ways of improving them. Thus, the study followed a systematic literature review to explore the effects of climate change in crop yields like hunger due to food shortage; different crop yield prediction methods, including empirical, statistical and machine learning models; factors(parameters) used in crop yield

prediction, and finally strength and weakness of different exiting Machine Learning algorithms used for crop yield prediction.

Research database Google Scholar was used to retrieving scientific publications. Also, the review was conducted by using online search engines such as google.com and bing.com to gather information and non-scientific articles in areas such as crop yield prediction, and the effect of climate change on agriculture and maize production in Tanzania. The quest for information from the mentioned sources was performed by querying various keywords to obtain relevant articles. Keywords used were *'Agriculture Productivity,' 'Impact of climate change on Agricultural productivity,' 'Crop yield prediction,' 'Machine Learning approach in crop yield prediction' and 'maize yield prediction models used in Tanzania,'*. The keywords were searched together and separated using OR statements in which the most relevant were brought up. *So, out of many articles that match, only the first 125 articles were chosen as appropriate.* In out of 125 articles obtained, only 94 articles were used by screening on relevancy to Africa and Tanzania.

To ensure that all relevant information is retrieved, a list of references was visited for each scientific paper to spot pertinent other papers. Likewise, for each paper, the author looked into its citations in google scholar to identify additional relevant articles. These searching approaches are referred to as backward and forward searches(Webster & Watson, 2002; Levy & Ellis, 2006).

Findings from the literature established the need for maize yield prediction in Tanzania. Most of the studies showed that policymakers and farmers are interested in finding out the performance of the crop on the farm in ever-changing weather conditions and predict the yield of rain-fed maize to ensure food security.

3.1.2 Data Collection and Preprocessing

In this study, two kinds of data which are yield and climate (weather) data, were collected from the area of study (Tanzania). This section describes the study area,

how these kinds of data were collected and pre-processed, and how the variable for the ML model trained in this study was selected.

3.1.2.1 Area of Study

Tanzania, which is located in East Africa, is the site where the study was undertaken. Figure 1 shows the location of Tanzania in the African continent (Basalirwa et al., 1999).

Figure 1. Map showing the location of Tanzania in Africa from tanzania.go.tz

Figure 2 shows location of all regions of Tanzania. Rain seasons in Tanzania behave differently in northern and southern regions, with those located in the northern part having two rain maxima (bimodal) and other regions in the southern part have only one long rain seasonal rainfall (unimodal).

Figure 2. A map showing regions of Tanzania from tanzania.go.tz

These unimodal are the high production zones. Most rain seasons start in October and end in May (Basalirwa et al., 1999). With respect to the modelling in this study, the model assumed that the starting of rain accumulation is from October to May of each year. The model determined the maize yield for 2018 season using data from 2005 to 2017. This is because the data obtained from the ministry of agriculture was incomplete and did not have yield information for 2019 to 2022.

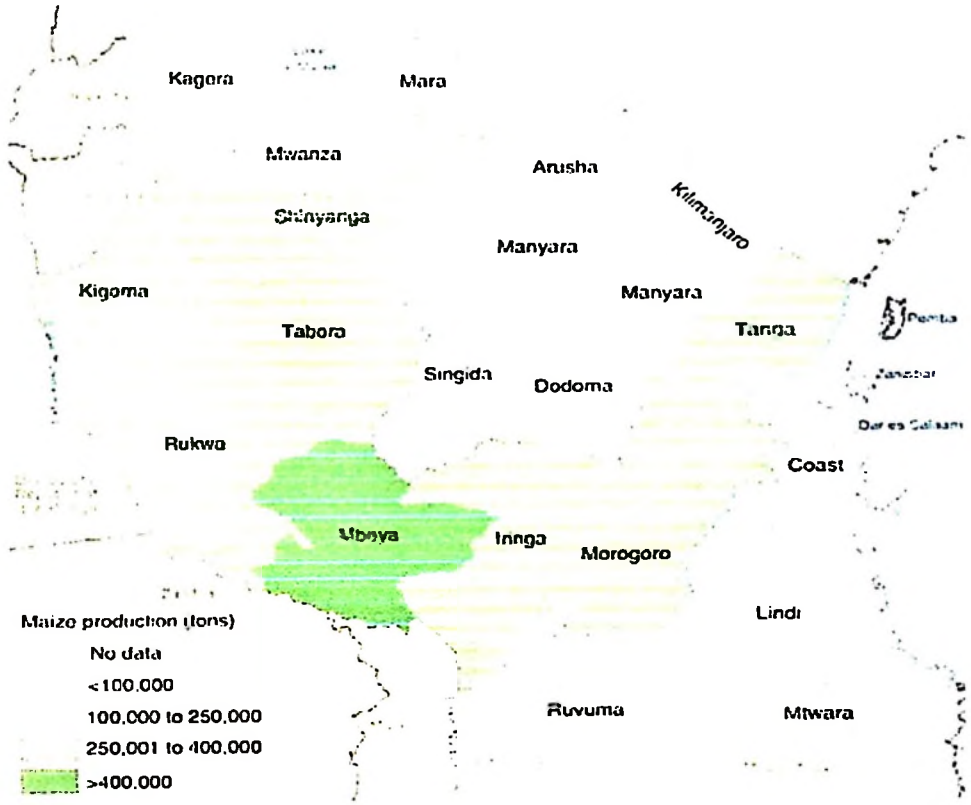


Figure 3: Maize Production in Tanzania (Season 2018/2019) from apps.fas.usda.gov

3.1.2.2 Yield Data

By email, maize yield data from 2005 to 2018 was requested from the National Food Security Division of the Ministry of Agriculture (MoA) of Tanzania. Their contacts are available at www.kilimo.go.tz. The dataset reports maize yield per district. Then, the data file was shared by the MoA to the authors.

Maize accounts for more than 25% of the total cropland in Tanzania, equating to a total harvest area of 3,428,630 ha in 2019 (MoA, 2019). Figure 3 shows 2018/2019 maize yield data for each region. Usually, the MoA reports official data annually by including yield data from all regions of mainland Tanzania except the island Zanzibar. Unfortunately, some of the district's boundaries (such as Kilosa was divided to form new Kilosa and Gairo districts) were changed in 2012 so the weighted average-based analysis was done to estimate the yield distribution between old and new districts.

3.1.2.3 Weather Data

Weather variables had been created based on one climate data source. ERA5-Land reanalysis climate data store (https://cds.climate.copernicus.eu) provided at a spatial resolution of $0.1° \times 0.1°$ was used. European Centre produces the ERA5-Land for Medium-Range Weather Forecasts (ECMWF) by estimating atmospheric, land-surface and sea-state parameters using uncertainty and physics rules. The ERA5-Land Monthly Reanalysis (1950 to present) was downloaded from European Copernicus Climate Data Store [Copernicus is the European Union's Earth observation programme]. ERA5 (and the most advanced ERA5-Land) is the most recent reanalyzed product and outperforms ERA-Interim for East Africa (Gleixner et al., 2020). The weather variables (Temperature, Soil temperature, Evaporation, Precipitation, Leaf area index, Soil water, Wind, Runoff) were created for the period of available yield statistics (2005–2018). The data was obtained by visiting the Copernicus data store (https://cds.climate.copernicus.eu) and then choosing ERA-Land reanalysis data. The portal allows selecting the boundaries of the interesting areas. Tanzania is bounded by latitude (-1^0 to -13^0) and Longitude (27^0 to 41^0). Then, the data date was chosen from January 2005 to December 2018. The data can be downloaded into two formats, either Grib or NetCDS. The data for this study was downloaded as grib file because it is easy to process the data using the opensource library pygrib which is well documented and has a community supporting it. Then,

Copernicus processed the data for 30 minutes and provided a direct download link to get the data.

For rainfall seasons which are bimodal season (vuli) which starts October to December, only data for October and November will be used while unimodal season (masika) which starts from March to May, only data for March and April will be used (Suleiman and Rosentrater, 2015). Only two months each season is enough because the maize cultivars may take three to five months to harvest. So, choosing less than three months makes sense for this study as it aims to forecast yield at least four weeks before harvesting, and there is no data on maize cultivars used in each district because farmers are free to use any cultivar.

3.1.3 Preprocessing of the Data

The input was obtained from the Copernicus website; ERA5-Land data was in grib format. The grib files were compressed climate data files that require a specific tool to open and process the data. In Python language used in this study, pyGrib library was installed to help process the data. The data was then cleaned and formatted. However, in Tanzania, there are no mapped maize planted areas, so it is difficult to estimate for the whole district. Instead, a Normalized Difference Vegetation Index (NDVI) was used to predict vegetation areas. NDVI is a dimensionless vegetation index that describes plant greenness by calculating the difference in reflection on red and near-infrared (NIR) light (Pettorelli et al., 2005). So, to process the data for each district in Tanzania, NDVI for March 2018 was downloaded from National Oceanic and Atmospheric Administration (NOAA). Then, it was used to get areas with NDVI greater than 0.5, indicating vegetation areas. Normally, NDVI greater than 0 indicates vegetation areas of the land (Pettorelli et al., 2005). Figure 4 shows NDVI cumulative values for the whole country. Hence, NDVI greater than 0.5 (Equivalent to 150 in Figure 4) lets our model get data from areas that may have been planted with maize leaving urban areas.

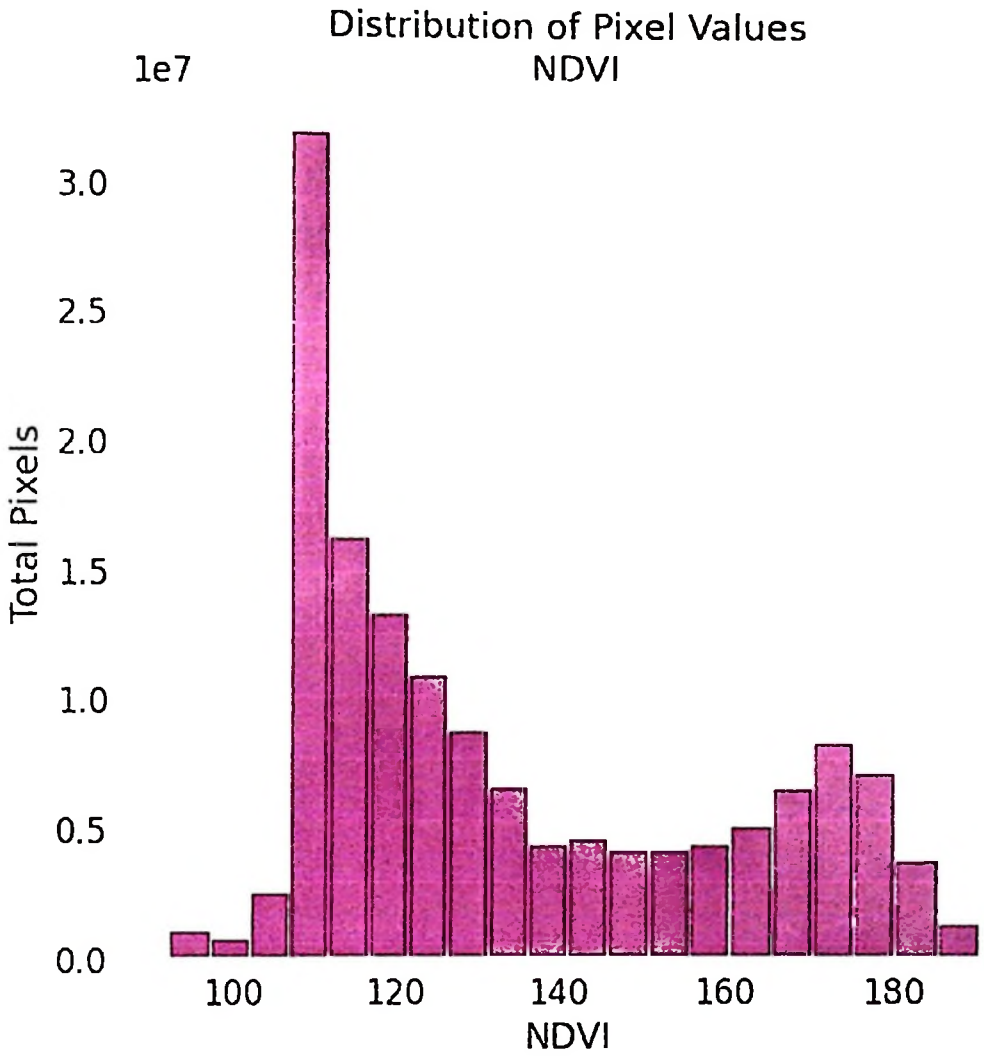


Figure 4. The NDVI pixels in Tanzania. NDVI values 0 to 200 is equivalent NDVI true values from -1 to 1. Which means greater than 100 is vegetation

The data had several missing values which are normally set as 9999, null or zero.

For yield data, variables with zero value were removed consequently, only 94 district yield data remained out of 183 districts. Also, data with variables with Pearson's r >0.1 in correlation to yield were taken on boarding, which led to the reduced number of the training data from 4849 to 318 entries. Only 43 districts remained that were

processed. Also, this data cleaning process removed 2009 and 2013 data since they had a lot of missing data. The yield input data was transformed to logarithmic values. Weather data was then standardized to make the normal distribution, ready to be used by the machine learning model.

The input data was arranged in tabular format in which each weather data for each month was set as a column. The last column is the area of the district. The longitude and latitude of the areas chosen from each district was included. The year column was also included. This makes a total of 127 columns.

3.1.4 Variable Selection

Yield variability in different districts was determined by appropriate variable selection. To achieve clean and appropriate ML model, the following was done:

- Variables that show zero yield were removed as they present false information that a particular district has zero yield, which is unlikely in Tanzania
- The variables with strong collinear and good correlation to yield data were selected. So, to make it easy for the model to learn, only variables with Pearson's $r > 0.1$ were selected. So, it means all the variables that did not show good correlation with yield were removed. Also, it was important to remove such variables so as to reduce the over-fitting of the model
- Also, all the input variables were subjected to standardization to improve data scaling which is important for machine learning models to learn

3.1.5 Machine Learning (ML) Algorithm Selection and Model Development

This research proposed to model conventional supervised models (adaboost, gradient boosting, K-neighbor and random forest) for crop production because they have not been explored well in Tanzania yield calculation and they have shown superiority in regression estimation like yield data prediction (Benos et al., 2021, Geetha et al., 2020, Keerthan Kumar et al., 2019, Yuchi et al., 2019, Liakos et al., 2018, Wu et al.,

2017). The supervised ML models can be trained to give out results for classification and regression problems. The random forest has shown efficacy on prediction using regression analysis for soil and health data that provide quantitative recommendations (Yuchi et al., 2019 and Wu et al., 2017). The models were designed to capture the time dependencies of crop yield over several years. The selection of the models was based on the fact that the conventional ML models are suitable for time series data prediction like climate data, and training. Testing is readily simple and does not require specific Graphics cards. And the models were well designed to predict time-series data (Yuchi et al., 2019 and Wu et al., 2017). So, to improve the ML models' performance, a stacked ensemble regression model was designed to use RF as a final model after getting predictions from other regression models such as ada boosting (is a machine learning technique that fits a regressor on the original dataset and fits additional copies on the same dataset. In this algorithm the weights of the instances are adjusted due to errors of the current prediction), gradient boosting (is an ensemble model of a weak prediction model that are usually decision trees by optimizing differentiable loss function) and Kneighbor (a model that tries to predict based on nearest neighbors of the observed data) Also, linear regression (the model that tries to fit the observed data by means of linear equation) was used for comparison purposes as it has been explored in Tanzania too. Ada boosting model also known as adaptive boosting, is an ensemble model. Stacking regression is the method for forming combination of multiple regression models to improve the accuracy of final model. Furthermore, Table *2* depicts the strengths and challenges of each model to show why the models were chosen for this research explicitly.

Table 2. Strengths and Challenges of the selected models

Models Selected	Strengths	Challenges	Suitability for time-series data
Linear regression	• It works well for linearly trending data • It is very easy to understand and implement training and evaluation • It is fast	• Presumptions that data is linear between various variables are not always applicable • Overfitting on noise data is common	• Not suitable and may need further improvement like Explainable boosted linear regression to work (Ilic et al., 2021)
Adaboost regression	• It works well for weak classifiers by combining to strong classifiers • It combats well overfitting	It is a linear model	• Suitable for nonlinear data and time-series data (Xiao et al., 2019)
Gradient Boosting regression	• It provides vast flexibility by optimizing multiple loss functions • Quite good in processing data with missing variables • No need to differentiate between numerical and categorical values	• May over fit quite easily • It is quite memory and computationally expensive because it processes vast amount of trees • Optimization may need extensive grid search to obtain optimal performance	Suitable for nonlinear data and time series data (Ribeiro and Dos Santos, 2020)
Kneighbor regression	• Fast computation time • Works well for the regression model • Very simple model • Provides accurate prediction	• Scalability is a problem • Heterogenous data might be a problem • Needs definition for the optimal number of neighbors. Grid search may be required to find this • It is quite sensitive to outliers • Missing variable value can be a big problem	Suitable for nonlinear and time-series data (Cristian, 2018)
Stacking/ensemble Random Forest regression	• Combines all the advantages of the above • Provides a chance to choose optimal set of parameters to improve performance • Mostly outperforms the stacked models	It inherits uncertainties of other models	Works well with nonlinear and time series data (Pavlyshenko, 2020)

To achieve, a performing model, the following user-defined parameters for the used data set for each model were set as provided in Table *3*. The user-defined or hyperparameters were obtained by using ski-learn kit library method known as randomisedsearch CV which implements randomized search on hyperparameters by fitting and comparing the score and suggest the best parameters.

Table 3: Optimal hyperparameters for different machine learning algorithms

Algorithm used	Hyperparameters
Linear regression	Default settings
Ada Boost regression	Random_state=0
Gradient Boosting regression	Random_state=0
K neighbors regression	n_neighbors=20, metric=Euclidean
Stacking Random Forest regression	random_state= 1, n_estimators=120, max_features= 'auto', max_depth= 10, bootstrap= True

Fortunately, all the models have been well implemented in Scikit-learn[5], a free machine learning library for predictive data analysis using the Python programming Language (version 3.9).

3.1.6 Machine Learning Model Training and Validation

The model developed based on the yield data from 2005 to 2017 (12 years) was used to provide a completely independent forecast for the harvest years 2016, 2017 and 2018. The model predicts the maize yield for the years 2016, 2017, and 2018 by using weather/climate information obtained from 2005 to 2018. The model is based on district-wise yield prediction. Prediction for year 2016 involves data from 2005 to 2015 for Training (Testset 1). Prediction for Year 2017 involves data from 2005 to 2016 for Training (Testset 2) and Prediction for Year 2018 involves data from 2005 to 2017 for Training (Testset 3).

[5]https://scikit-learn.org/

The yield data for 2018 was not included in the Training and Validation dataset and instead it was used as a final year for evaluation as testing dataset. Training was done by optimizing the loss function by comparing the observed with the predicted data. Cross-Validation which involved splitting the training dataset to assess how the model fits the independent variables was done by splitting in k small sets "folds" and quite suitable for small dataset like the data used on this study. The model used 5-folds (k=5) cross validation. The root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), coefficient of determination (or accuracy) and Pearson's correlation coefficient were reported to benchmark the model performance.

3.1.7 Machine Learning Model Inference Evaluation.

Model inference evaluation was done by statistically calculating coefficient of determination (or accuracy) which is comparing observed data versus true data for the year 2018 (Testset 3). Furthermore, as done and suggested by Khaki et al., (2020), to better improve evaluation of the models, models were trained by using data from 2005 to 2015 and tested on 2016 (Testset 1) data and then trained by using data from 2005 to 2016 and tested on 2017 (Testset 2) data. The stacking random forest model was compared to other regression Machine Models (linear regression, adaboost, gradient boosting, Kneighbor) used in prediction of maize yield that has been done based on Tanzania.

3.1.8 Model for Estimating and Forecasting Yields before Harvest

For Maize, it may take 3 months to harvest or more. For in-season forecasting, it was important to find inference of the yield after two months. The model was be designed to estimate yield 1 month before harvest depending on reproductive phase of each region. The values of yield were assessed such that anomalies are detected and reported as well.

3.2Responsible Design for Maize Yield Forecasting

This research was designed to consider the target groups' ethical, social and economic dimensions. The privacy of the target group data was embraced by ensuring that individual challenges such as data are not shared with the public. Hence, to enforce privacy more, the data is anonymized and generalized to a district to hide the details. The information shared in this publication has been considered fair treatment with the partners to enhance trust and promote good research practice.

The sustainability of the solution proposed in this study has taken a broader dimension on climate change by embracing an alternative prediction methodology that guides policymakers to a better decision that would protect farmers and promote development in rural settings, especially for women farmers who are heavily hit by climate change. It alerts the policymakers earlier to protect the most vulnerable societies in Tanzania that depend on rain-fed maize production by embracing sustainable mitigatory approaches.

To ensure fairness the data was downloaded from trusted sources and preprocessed to remove outliers. The yield data included all districts except those with missing data in some of the years. The climate data was downloaded from Copernicus climate data store which uses world standard algorithms to simulate climate information to all locations even those that do not have weather stations. This makes it possible to assess all the districts on crop production.

3.3 Summary of the Research Methodology Activities

The methodology of this study is presented structurally in four sections: first is how the problem of this study was established, second is Data collection and preprocessing, third section is Model Training and Model Debugging, and the fourth section is Model Testing and Evaluation. Also, the chapter highlights the responsible design of this research where the issues of ethical, explainability, fairness, and privacy were considered. Under the mentioned sections, this methodology comprises three main goals and their activities as summarized in table (Table 4).

Table 4: A broad view of the research methodology

Goals	Activity	Method/Approach/Technique	Data Source & Tools	Deliverables
1.Explore and Identify dataset (climate and yield)	a) Identify sources of data	Literature search	Google scholar &web science	List of Databases
	b) Retrieve clean data from identified data sources	File transfer protocol (FTP) or direct download	Copernicus ERA5 -land, National Oceanic and Atmospheric Administration (NOAA). Ministry of Agriculture	Data sets • ERA5 • Land (climate data) • Normalized Difference Vegetation Index (NDVI) • Yield
2. Develop a prediction Model	a) Develop and Train a Machine learning model	Random forest and loss function minimization	Jupyter Notebook	Prediction Model
	b) Validate the Model	Coefficient of determination and Pearson correlation coefficient score	Jupyter Notebook Inputs from 1 and 2b	Validated Prediction Model
3. Evaluate Prediction Model	Model inference & Measure Performance	Mean absolute error (MAE). Root mean square error (RMSE) and Relative Mean Square Error were used to compare Random Forest with other models (Linear Regression, adaboost, kneighbor and Gradient Boosting)	Inputs from 2. Jupyter Notebook and SK Learn	Model Accuracy

The first goal was to fetch dataset from online databases and then clean and rearrange the data according to our needs. Then, the second and third was training/validating and evaluating the model.

CHAPTER FOUR

RESULTS AND DISCUSSION

This chapter presents the results obtained from training, validation and testing (evaluation) of the models (Linear Regression, Adaboost, Gradient boosting, K-neighbor, Random Forest and Stacking Random Forest). Each model was separately trained and tested on the same data to find direct performance comparisons. Results are discussed in different graphs in this chapter to show how models were learning, scaling and inferring in given dataset during training and testing. The chapter concludes by using stacking algorithm with random forest as the final estimator to boost the performance of the model to an acceptable rate. But also, yield prediction can be well predicted using adaboost, gradient boosting and K-neighbour as the weak estimators to improve the performance of the random forest estimator.

4.1 Assessment of the Machine Learning Models

Five performance evaluation parameters were used to assess the machine learning models in predicting the maize yield. These parameters were Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), coefficient of determination (or accuracy) and Pearson's correlation. They were calculated from two datasets; yield data from 2006 to 2018 and climate data from 2005 to 2018. The results of regression machine learning prediction models of this study were presented by plotting actual and predicted values of maize yield of both training and testing datasets. The model was formatted to predict yield for each district in Tanzania. To determine the model performance, three datasets were arranged;

i. Prediction for the year 2016 maize yield which involves data from 2005 to 2015 for training (Testset 1),
ii. Prediction for the year 2017 maize yield which involves data from 2005 to 2016 for training (Testset 2), and

iii. Prediction for the final year 2018 which involves data from 2005 to 2017 for training (Testset 3).

Then, for each dataset, the results from the regression machine models (linear regression, adaboost, gradient boosting, K-neighbor and Random Forest) were compared with the final regression machine model (Stacking Random Forest Regressor). The inference of the models can well be benchmarked on multiple testing grounds which may bring confidence on the model performance.

4.2 Results from Training and Testing of the Machine Learning Models.

Figure 5 shows the performance of the model's Linear regression, adaboost regression, K-neighbor regression, gradient boosting regression and random forest regression for testset 1.

(a)

(b)

(c)

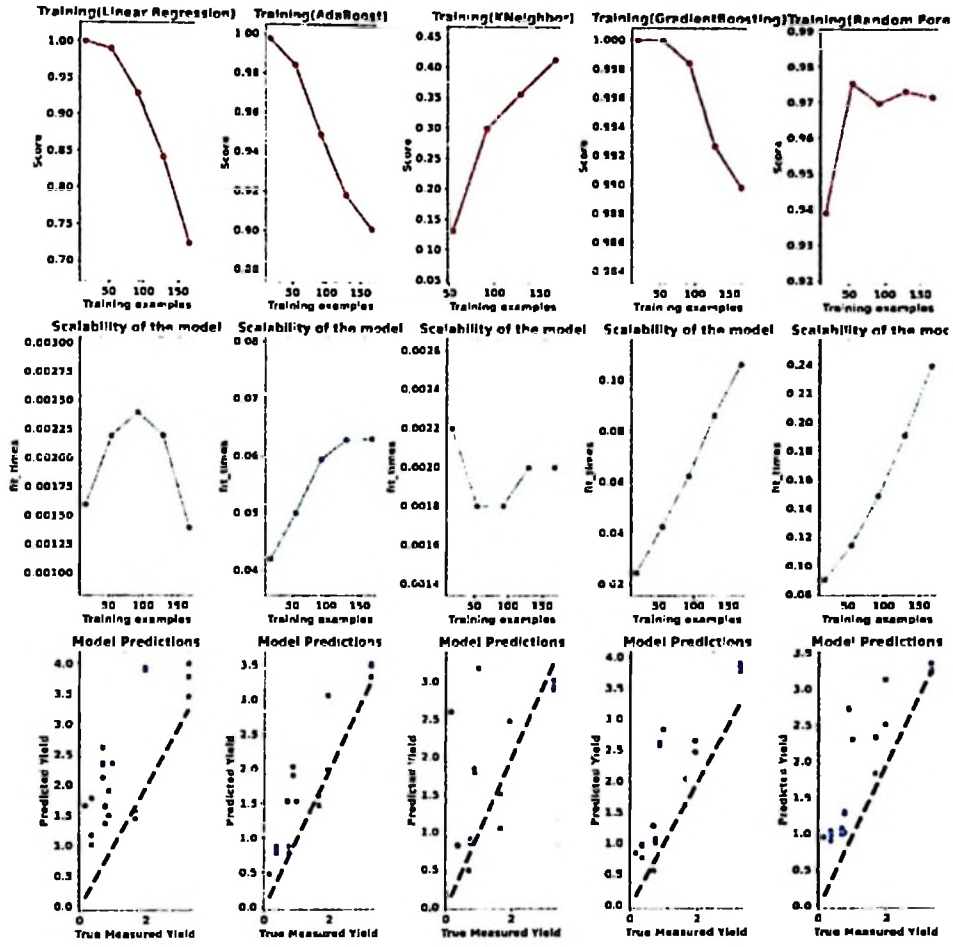


Figure 5: Training and Testing of the all Models for 2016 [Testset 1]

Figure 6 shows the performance of the model's Linear regression, adaboost regression, K-neighbor regression, gradient boosting regression and random forest regression for testset 2.

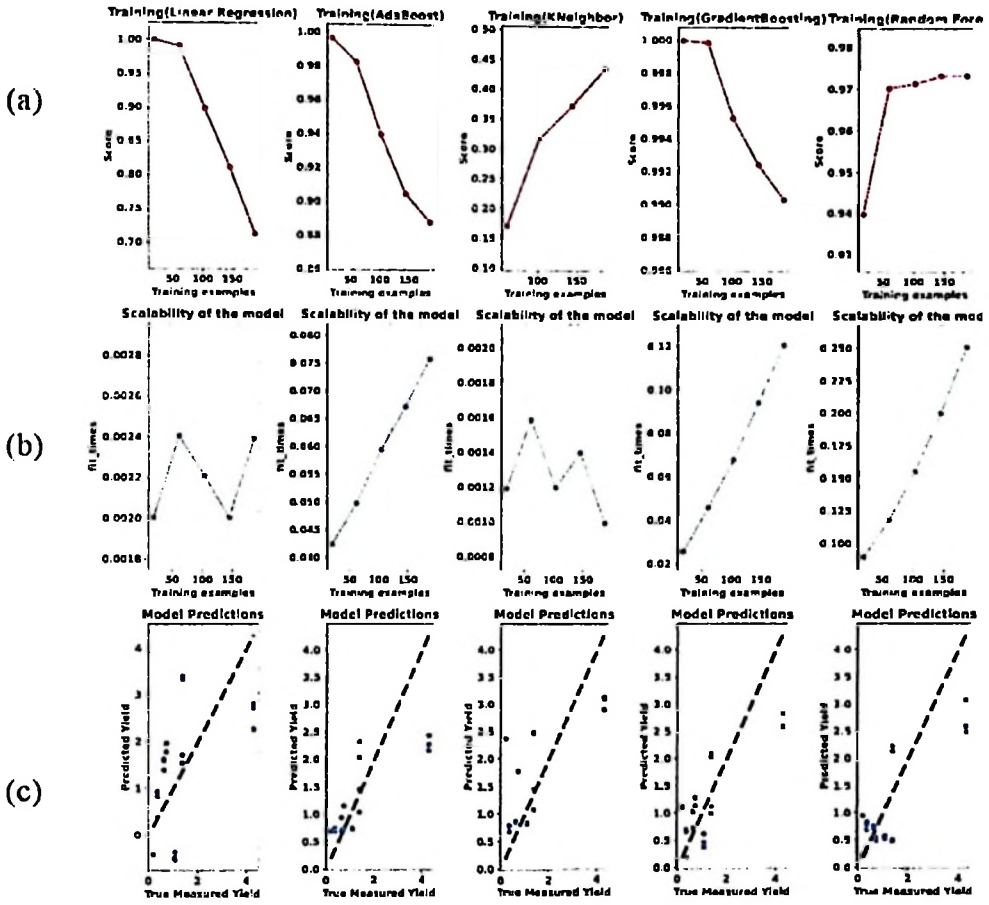


Figure 6. Training and Testing of the all Models for 2017 [Testset2]

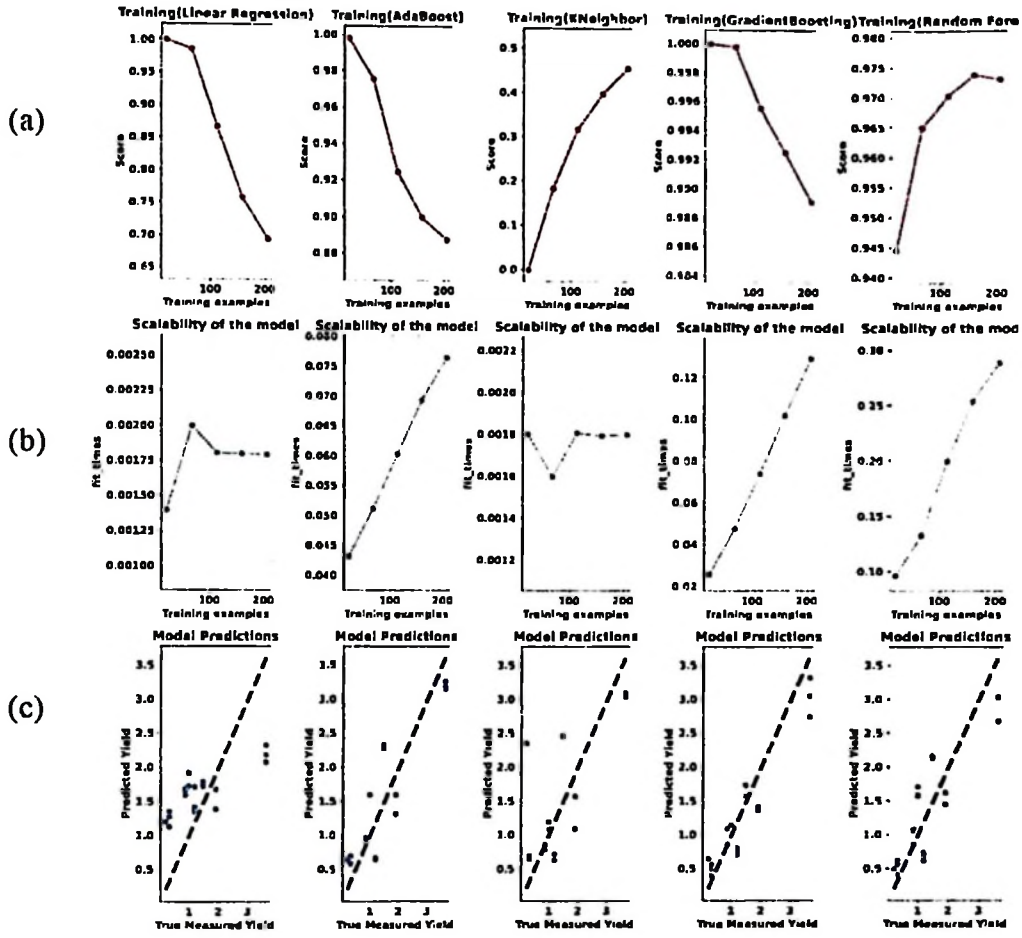


Figure 7. Training and Testing of the all Models for 2018 [Testset 3]

The Figures show how the models were trained, scaled and inferred. All the figures show model performance by measuring how the increase on dataset in training was affecting the scores. The first row Figures5(a), 6(a) and 7(a) show the response of the model scores against increasing number of examples, while the second row Figures 5(b), 6(b) and 7(b) show the response of the model on scalability by measuring how the model was generalizing on the data it was trained, while the third row Figure 5(c), 6(c) and 7(c) show the response of the model on inference of the trained model by comparing the predicted yield data against true yield data.

The decrease in training score illustrated in Figures 5(a), 6(a) and 7(a) show the failure of the models except K-neighbor and Random Forest regression to learning

from the data. It shows that the two models were learning well from the data as the score was increasing when number of examples was increased. However, Figure 5(b), 6(b) and 7(b) shows that K-neighbor was not scaling with the increase or decrease of the number of examples given which means decrease or increase of the number of examples was not important in finding the nearest neighbors for K-neighbor model. Also, Figure 5(b), 6(b) and 7(b) shows K-neighbor, gradient boosting, and random forest have showed good scalability behaviour since they were increasing as the training examples were increased. Therefore, Figure 5(c), 6(c) and 7(c) shows that adaboost, gradient boosting and random forest were able to predict well the yield compared to true measured yield since the dispersion of the points from the prediction line is quite small and distributed fairly equally compared to linear regression and K-neighbour. Also, Linear regression has failed to learn, scale and predict and it cannot be used for maize yield prediction. Therefore, to get advantages from the best other four models, it was good idea to stack the algorithms (K-neighbor, gradient boosting and adaboost) with Random Forest regression to form a robust stacking random forest model.

The stacking models minimize the weakness of the (K-neighbor, gradient boosting and adaboost) and get the greatness out of them and add them to the final model which was random forest model and finally improved the performance of the final model.

4.3 Comparing Results between RF against RF Stacking Regression Models.

Figures8(a), 9(a) and 10(a) show that training of the RF and its stacking estimator (RF stacking) were comparable equal. The two models display the similar behaviour on scalability fitting too as shown in Figures8(b), 9(b) and 10(b). Fortunately, in Figures8(c), 9(c) and 10(c) prediction of the RF stacking was the best on removing the outliers and bringing together the accurate prediction since RF was improving from the strength of its estimators while minimizing the weakness of its estimators.

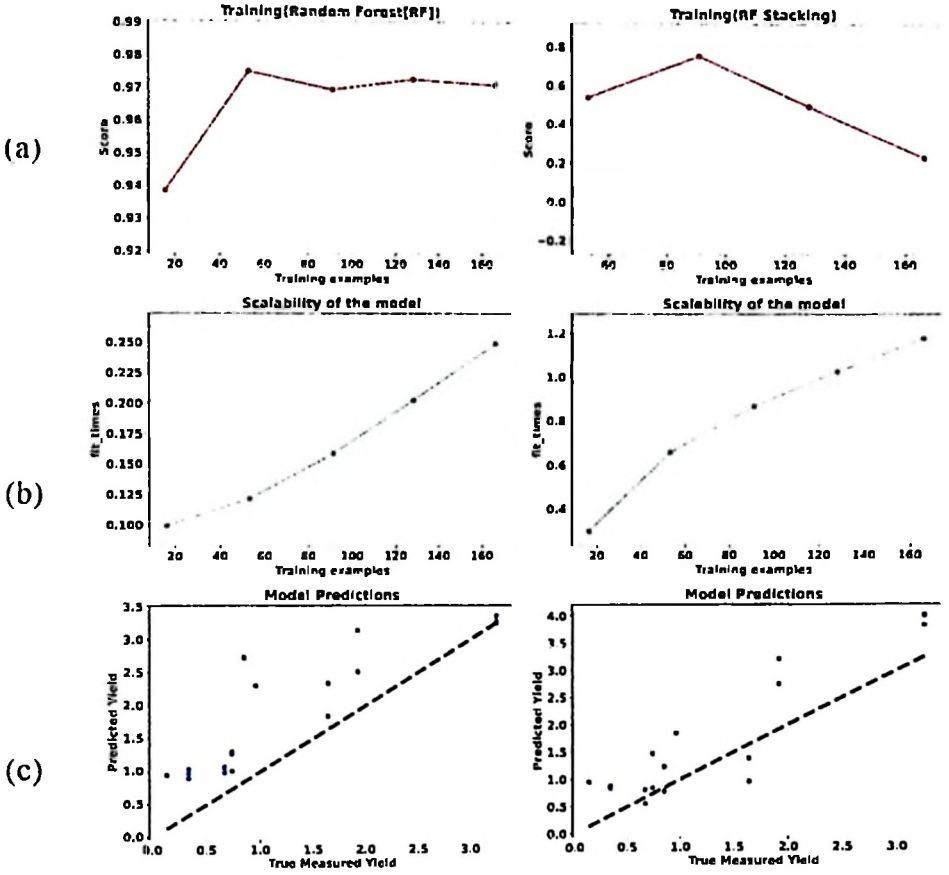


Figure 8. Comparing RF model with Stacking model for 2016 [Testset 1]

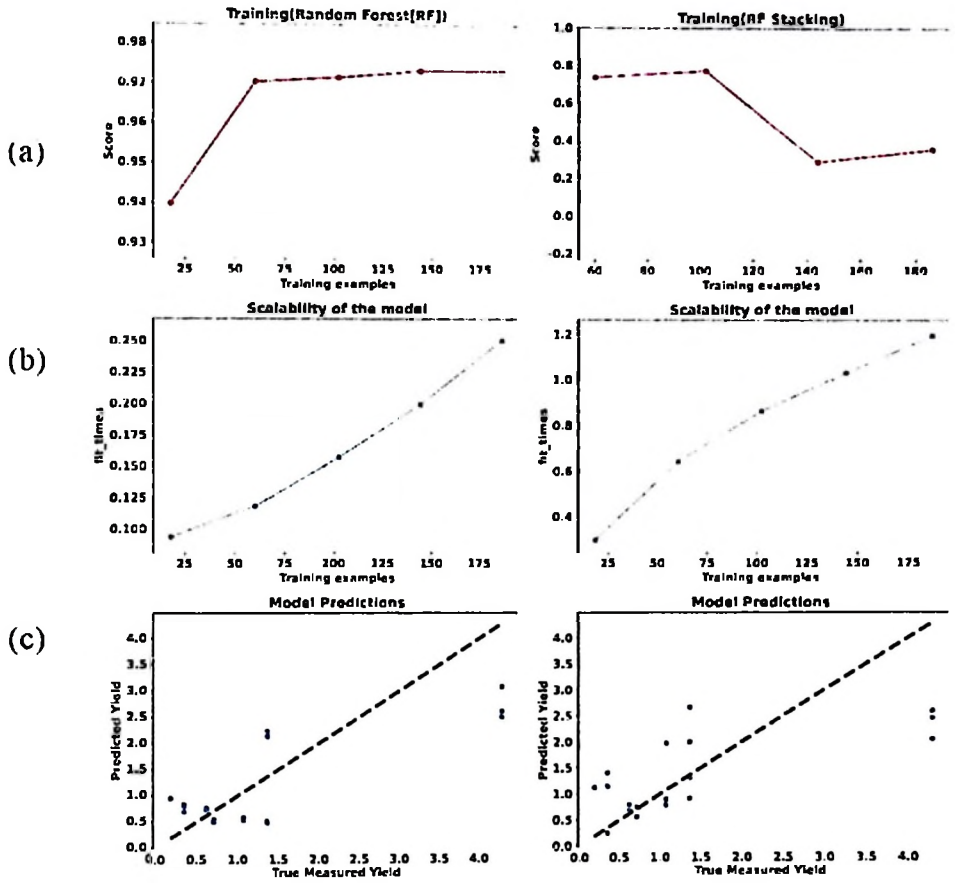


Figure 9: Comparing RF model with Stacking model for 2017 [Testset 2]

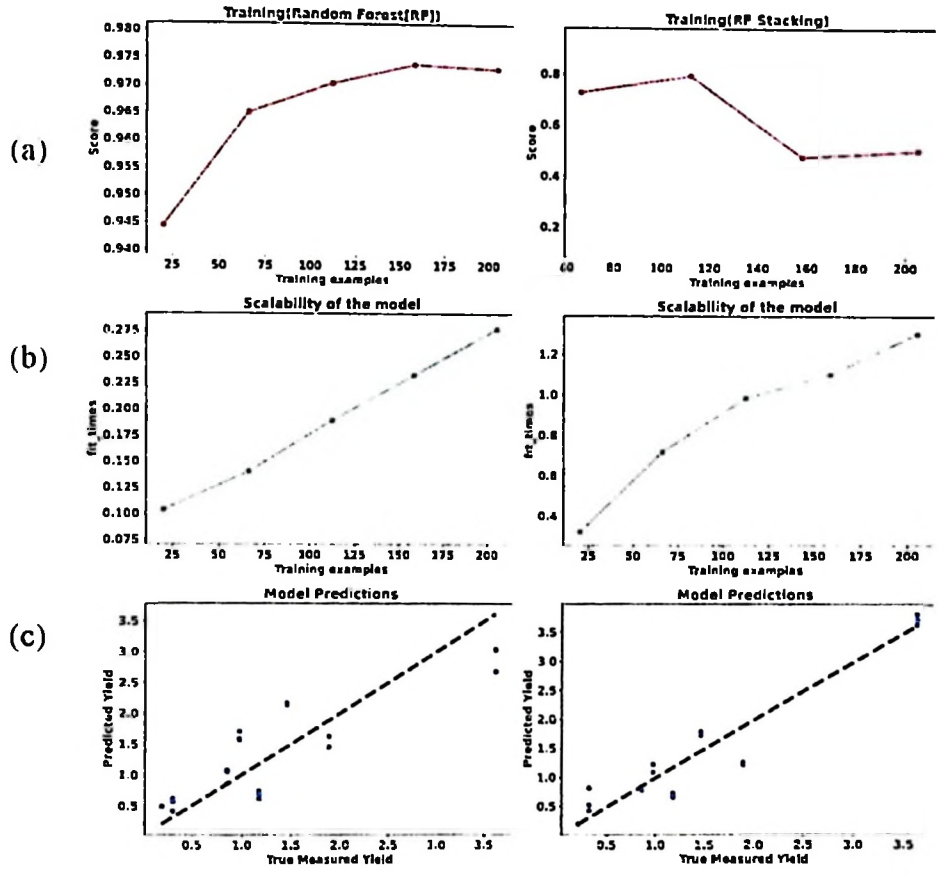


Figure 10: Comparing RF model with Stacking model for 2018 [Testset 3]

The training is congruent with the inference results as presented in Table 5, the prediction dispersion of the stacking RF is close to the regression line indicating decreasing of the standard deviation of the prediction which in turn improved the performance of model's coefficient of determination, Pearson's correlation, MAE, MSE and RMSE in 2016 from 0.18, 0.82, 0.68, 0.78 and 0.88 to 0.55, 0.90, 0.55, 0.12 and 0.42. Unfortunately, in 2017 the performance of the model's coefficient of determination, Pearson's correlation, MAE, MSE and RMSE decreased from 0.62, 0.82, 0.66, 0.64 and 0.80 to 0.48, 0.71, 0.68, 0.88 and 0.94 respectively. The data in year 2017 had discrepancy because the yield data had great outliers which were not well predicted by the estimators leading to decreased performance of the stacking RF compared to RF. In 2018, the performance of the model's coefficient of determination, Pearson's correlation, MAE, MSE and RMSE increased from 0.74,

0.86, 0.48, 0.28 and 0.53 to 0.89, 0.94, 0.27, 0.12 and 0.34 respectively as shown on Table5.

Table 5: Evaluation of the machine learning models

Model	Validation Year	MAE	MSE	RMSE	Coefficient of Determination	Pearson's correlation coefficient
Linear Regression	2016	1.04	1.48	1.21	-0.54	0.76
	2017	1.16	1.75	1.32	-0.01	0.46
	2018	0.74	0.73	0.85	0.34	0.76
AdaBoost Regression	2016	0.52	0.44	0.66	0.53	0.89
	2017	0.62	0.82	0.9	0.52	0.78
	2018	0.46	0.26	0.51	0.76	0.87
Gradient Boosting Regression	2016	0.69	0.75	0.86	0.21	0.86
	2017	0.66	0.63	0.79	0.63	0.83
	2018	0.34	0.16	0.4	0.85	0.94
Neighbor Regression	2016	0.55	0.65	0.81	0.31	0.71
	2017	0.71	0.79	0.88	0.54	0.76
	2018	0.51	0.46	0.68	0.58	0.76
Random Forest Regression	2016	0.68	0.78	0.88	0.18	0.82
	2017	0.66	0.64	0.8	0.62	0.82
	2018	0.48	0.28	0.53	0.74	0.86
Random Forest Stacking Regression	2016	0.55	0.42	0.65	0.55	0.9
	2017	0.68	0.88	0.94	0.48	0.71
	2018	0.27	0.12	0.34	0.89	0.94

In conclusion, the data inconsistences in manual collected information can be detrimental to the conclusion as the discrepancy may mislead the machine learning model on finding the best goal or prediction as it happened in 2017 data. Fortunately, the results show that machine learning models can significantly be important in accurate prediction of the maize yield. The stacking regression model developed can be used by the stakeholders to predict yield and guide stakeholders' understanding of the maize yield in a particular season.

CHAPTER FIVE

SUMMARY, CONCLUSION AND FUTURE WORK

The chapter presents concluding remarks and future open streams. It revisits the context of the study, the proposed solution and existing limitations. Areas for further research are also highlighted in this chapter.

5.1 Research Summary and Conclusion

Recent advancements in satellite technology and machine learning have provided a profound opportunity to learn how agriculture and natural resources are performing by ensuring that large areas are monitored and evaluated easily. The satellite technology and estimation algorithms available can be used to estimate weather and climate measurements through earth observation satellites and advanced simulation databases like the one hosted by the European Copernicus Climate Repository.

These developments have provided an alternative and cheap way to monitor and evaluate rain-fed agriculture in Africa, particularly Tanzania. Copernicus provides climate information to a resolution of 9km, which makes it possible to monitor district-level weather and climate conditions.

Hence, with the emergence of adverse weather conditions and climate change, it is paramount to keep communities and governments aware of the parameters that might influence the performance of rain-fed agriculture to enhance mitigation. Assessment of such parameters can be done using machine learning methodology, which looks at background climate information to predict the performance of the crop.

In this study, appropriate climate parameters and yield data from the Copernicus open data store was identified, and machine learning models to predict maize yield was developed and evaluated. The proposed model was developed to predict maize crop performance at district-level in Tanzania. This study shows the possibility of predicting yield by district-level in Tanzania with the accuracy in coefficient

determination of 78.5% with Pearson correlation of 89%. The performance of the model may be improved by incorporating more advanced algorithms like artificial neural networks, convolutional neural networks and recurrent neural networks. But also, data cleaning showed that most of the data (out of 4849, only 318 which is 6%) was not clean and it was removed. So, it is advisable for Tanzania authorities to improve yield data collection procedures.

5.2 Future Research

5.2.1 On Data

This study was not able to cover all districts of Tanzania because the ministry of agriculture yield dataset had missing spots. So, only 94 districts out of 138 districts were analyzed. Improvement of the yield data capturing is advised to the government so as to improve the performance of the prediction algorithms that would help mitigate climate change effects on crop yield. Also, this research used reanalysis data from European Union's data store which is derived using physics and simulation of the historical climate data obtained from weather stations. Using data that comes from weather station datastore is important to improve the performance of the model as there are very few weather stations deployed in Tanzania. Also, in future, I propose that satellite imagery data be incorporated on analysis to improve prediction. Satellite imagery may present good approach on estimation of the maize yield. However, such improvements may need good models to detect maize fields from the satellite imagery.

5.2.2 On Analysis and model prediction

Furthermore, in order to improve analysis and prediction of the model, more advanced models such as recurrent neural networks (RNN) that provide alternative advanced analysis may be used for future models. The models have shown superior performance in some of the yield prediction models (Liakos et al., 2018). This can involve categorizing the data to multiple groups so as to improve prediction by

classification instead of regression which may be accurate for the current available data and be useful on mitigating climate change in agriculture.

REFERENCES

Adeniyi, P. A. (2016). *Climate Change induced Hunger and Poverty in Africas.* Journal of Global Biosciences, 5(3), 3711-3724.

Adisa, O. M., Botai, J. O., Adeola, A. M., Hassen, A., Botai, C. M., Darkey, D., &Tesfamariam, E. (2019). *Application of artificial neural network for predicting maize production in South Africa.* Sustainability (Switzerland), 11(4), 1–17. https://doi.org/10.3390/su11041145

Agarwal, S. and Tarar, S. (2021). *A Hybrid Approach for Crop Yield Prediction for Crop Yield Prediction using Machine Learning and Deep Learning* Algorithm. Journal of Physics: Conference Series, 1714, p.012012.

Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., &Radiom, S. (2018). *Machine learning regression techniques for the silage maize yield prediction using time-series images of Landsat* 8 OLI. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(12), 4563-4577.

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). *Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage* paper 56. FAO, Rome, 300(9), D05109.

Alpaydin, E. (2010). *Introduction to Machine Learning*, 2nd ed. Retrieved from https:// books.google.nl/books?hl=nl&lr=&id=TtrxCwAAQBAJ&oi=fnd&pg=PR7&dq= introduction+to+machine+learning&ots=T5ejQG_7pZ&sig=0xC_H0agN7mPhYW7oQsWiMVwRnQ#v=onepage&q=introduction to machine learning&f=false.

Arora, N. K. (2019). *Impact of climate change on agriculture production and its sustainable solutions. Environmental Sustainability,* 2(2), 95–96. https://doi.org/10.1007/s42398-019-00078-w

Arora, N.K., (2019). *Impact of climate change on agriculture production and its sustainable solutions. Environmental Sustainability,* 2(2), pp.95-96.

Aryal, J.P., Sapkota, T.B., Khurana, R., Khatri-Chhetri, A., Rahut, D.B. and Jat, M.L., (2020). Climate change and agriculture in South Asia: Adaptation

options in smallholder production systems. Environment, Development and Sustainability, 22(6), pp.5045-5075.

Balaji, T. K., Annavarapu, C. S. R., &Bablani, A. (2021). *Machine learning algorithms for social media analysis:* A survey. Computer Science Review, 40, 100395.

Barker, R., & Hayami, Y. (1976). *Price support versus input subsidy for food self-sufficiency in developing countries.* American Journal of Agricultural Economics, 617-628.

Basalirwa, C. P. K., Odiyo, J. O., Mngodo, R. J., &Mpeta, E. J. (1999). *The climatological regions of Tanzania based on the rainfall characteristics.* International Journal of Climatology, 19(1), 69-80.

Basso, B., & Liu, L. (2019). *Seasonal crop yield forecast: Methods, applications, and accuracies.* advances in agronomy, 154, 201-255.

Basso, B., Cammarano, D., &Carfagna, E. (2013, July*). Review of crop yield forecasting methods and early warning systems.* In Proceedings of the first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics, FAO Headquarters, Rome, Italy (Vol. 41).

Benos, Lefteris, Aristotelis C. Tagarakis, Georgios Dolias, RemigioBerruto, *DimitriosKateris, and DionysisBochtis. "Machine learning in agriculture: A comprehensive updated review."* Sensors 21, no. 11 (2021): 3758.

Bhakta, I., Phadikar, S., & Majumder, K. (2019). *State of the art technologies in precision agriculture:* a systematic review. Journal of the Science of Food and Agriculture, 99(11), 4878-4888.

Bhavsar, H., & Ganatra, A. (2012). *A comparative study of training algorithms for supervised machine learning. International Journal of Soft Computing and Engineering* (IJSCE), 2(4), 2231-2307.

Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. (2012, July*). Machine learning strategies for time series forecasting. In European business intelligence summer school* (pp. 62-77). Springer, Berlin, Heidelberg.

Boult, V. L., Asfaw, D. T., Young, M., Maidment, R., Mwangi, E., Ambani, M., ... & Black, E. (2020). *Evaluation and validation of TAMSAT ALERT soil moisture and WRSI for use in drought anticipatory action.* Meteorological Applications, 27(5), e1959.

Burke, M., Lobell, D.B., (2017). *Satellite-based assessment of yield variation and its determinants in smallholder African* 661 systems. Proc. Natl. Acad. Sci. U. S. A. 114, 2189–2194. doi:10.1073/pnas.1616919114

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., ... & Peng, B. (2019). *Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches.* Agricultural and forest meteorology, 274, 144-159.

Chaudhary, K. and Kausar, F., (2020). *Prediction of Crop Yield using Machine Learning. International Journal of Engineering Applied Sciences and Technology*, 04(09), pp.153-156.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). *Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and electronics in agriculture*, 151, 61-69.

Crane-droesch, A. (2018). *Machine learning methods for crop yield prediction and climate change impact assessment in agriculture Machine learning methods for crop yield prediction and climate change impact assessment in agriculture.*

Cristian, M. (2018). *Average monthly rainfall forecast in Romania by using K-nearest neighbors regression. Annals of'ConstantinBrancusi'*University of Targu-Jiu. Economy Series, (4).

Dawkins, N., Shosho, N., Mamiro, P., & Pace, R. (2015). *Food Insecurity and Maternal Education on Complementary Feeding Practices in a Rural Village in Tanzania.* The FASEB Journal, 29(1 Supplement), 585-25.

Debnath, K. B., &Mourshed, M. (2018). *Forecasting methods in energy planning models.* Renewable and Sustainable Energy Reviews, 88, 297-325.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., &Ceccato, P. (2018). *Validation of the CHIRPS satellite rainfall estimates over eastern Africa. Quarterly Journal of the Royal Meteorological Society*, 144, 292-312.

Doggart, N., Morgan-Brown, T., Lyimo, E., Mbilinyi, B., Meshack, C. K., Sallu, S. M., &Spracklen, D. V. (2020). *Agriculture is the main driver of deforestation in Tanzania. Environmental Research Letters*, 15(3). 034028.

Doupe, P., Faghmous, J. and Basu, S. (2019*). Machine Learning for Health Services Researchers.* Value in Health. 22(7). pp.808-815.

Driessen, P.M., and Konijn, N.T. (1992). *Land-use Systems Analysis. Wageningen Agricultural University*. Wageningen. The Netherlands.

Du-Plessis, J. (2003). *Maize Production. Agricultural Information Services, Department of Agriculture, Pretoria, South Africa.[8] FAO (Food and Agriculture Organization). (1977). Crop water requirements. FAO Irrigation and Drainage* Paper No. 24. by Doorenbos J and W.O. Pruitt. FAO, Rome, Italy.

Dutta, S., Chakraborty, S., Goswami, R., Banerjee, H., Majumdar, K., Li, B. and Jat, M. (2020). *Maize yield in smallholder agriculture system—An approach integrating socio-economic and crop management factors*. PLOS ONE, 15(2), p.e0229100.

Eid, A.R. and Negm, A., (2018). *Improving Agricultural Crop Yield and Water Productivity via Sustainable and Engineering Techniques. Conventional Water Resources and Agriculture in Egypt*, pp.561-591.

Fabregas, R., Kremer, M. and Schilbach, F., (2019). *Realizing the potential of digital development: The case of agricultural advice. Science*, 366(6471), p. eaay3038.

FAO (Food and Agriculture Organization). (1988). *FAO/UNESCO Soil Map of the World: Revised Legend. FAO, Rome.* World Resources Report Number 60.

FAO (Food and Agriculture Organization). (1996). World *Food Summit Plan of Action*. http://www.fao.org/ docrep/003/w3613e/w3613e00.htm. Rome, Italy.

FAO (Food and Agriculture Organization). (1998). *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements. FAO, Rome. Irrigation and Drainage* Paper 56.

Fue, K.G., Sanga C.A., &Tumbo, S.D. (2017). *Ecawsoft: A Web based Climate and Weather Data Visualization for Big Data Analysis. Global Journal of Computer Science and Technology.* 17(1), 33-44.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., &Hoell, A. (2015). *The climate hazards Infrared precipitation with stations—A new environmental record for monitoring extremes.* Scientific Data, 2, 150066

Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016, July). *Rice crop yield prediction in India using support vector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering* (JCSSE) (pp. 1-5). IEEE

Geetha, V., Punitha, A., Abarna, M., Akshaya, M., Illakiya, S., & Janani, A. P. (2020, July). *An effective crop prediction using random forest algorithm. In 2020 International Conference on System, Computation, Automation and Networking* (ICSCAN) (pp. 1-5). IEEE.

Geoffrey, O., Wietse, F., Ronald, H., Supit, I., and Omondi, P., (2018). *Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts. Agriculture and forest meteorology* 250:243-261

Gleixner, S., Demissie, T., &Diro, G. T. (2020). *Did ERA5 improve temperature and precipitation reanalysis over East Africa. Atmosphere*, 11(9), 996.

González Sánchez, A., Frausto Solís, J., & Ojeda Bustamante, W. (2014). *Predictive ability of machine learning methods for massive crop yield prediction.*

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., & Zhang, J. (2020). *Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sensing.* 12(2), 236.

Idrobo, A. M. (2015). *Extension of the Geospatial Data Abstraction Library (GDAL/OGR) for OpenDRIVE Support in GIS Applications for Visualisation*

and Data Accumulation for Driving Simulators (Doctoral dissertation, Technische Universität München).

Ilic, I., Görgülü, B., Cevik, M., &Baydoğan, M. G. (2021). *Explainable boosted linear regression for time series forecasting. Pattern Recognition*, 120, 108144.

Kamilaris, A., Kartakoullis, A., &Prenafeta-Boldú, F. X. (2017). *A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, 143*, 23-37.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., & Anderson, M. (2020). *Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest.* Environmental Research Letters, 15(6), 064005.

Keerthan Kumar, T. G., Shubha, C., & Sushma, S. A. (2019). *Random forest algorithm for soil fertility prediction and grading using machine learning.* Int J Innov Technol ExplorEng, 9(1), 1301-1304.

Keynes, J. M. (1933). *National self-sufficiency.* Studies: An Irish Quarterly Review, 177-193.

Khaki, S., Wang, L., &Archontoulis, S. V. (2020). *A cnn-rnn framework for crop yield prediction.* Frontiers in Plant Science, 10, 1750.

Kukal, M. S., & Irmak, S. (2018). *Climate-driven crop yield and yield variability and climate change impacts on the US Great Plains agricultural production.* Scientific reports, 8(1), 1-18.

Kulwa, K. B., Mamiro, P. S., Kimanya, M. E., Mziray, R., &Kolsteren, P. W. (2015). *Feeding practices and nutrient content of complementary meals in rural central Tanzania:* implications for dietary adequacy and nutritional status. BMC pediatrics, 15(1), 171.

Kung, H. Y., Kuo, T. H., Chen, C. H., & Tsai, P. Y. (2016). *Accuracy analysis mechanism for agriculture data using the ensemble neural network method. Sustainability, 8(8), 735.*

Lai. Y., Pringle, M. J., Kopittke, P. M., Menzies, N. W., Orton, T. G., & Dang, Y. P. (2018). *An empirical model for prediction of wheat yield, using time-integrated Landsat NDVI. International journal of applied earth observation and geoinformation*, 72, 99-108.

Laudien, R., Schauberger, B., Makowski, D. et al. *Robustly forecasting maize yields in Tanzania based on climatic predictors.* Sci Rep 10, 19650 (2020). https://doi.org/10.1038/s41598-020-76315-8

Leroux, L., Castets, M., Baron, C., Escorihuela, M. J., Bégué, A., & Lo Seen, D. (2019). *Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices.* In European Journal of Agronomy (Vol. 108, pp. 11–26). https://doi.org/10.1016/j.eja.2019.04.007

Levira, P. W. (2009). *Climate change impact in agriculture sector in Tanzania and its mitigation measure. In IOP Conference Series: Earth and Environmental Science (*Vol. 6, No. 37, p. 372049). IOP Publishing.

Li, C., Wang, Y., Ma, C., Chen, W., Li, Y., Li, J., ... & Xiao, Z. (2021). *Improvement of Wheat Grain Yield Prediction Model Performance Based on Stacking Technique.* Applied Sciences, 11(24), 12164.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., &Bochtis, D. (2018*). Machine learning in agriculture: A review*. Sensors, 18(8), 2674.

Lin, J., Guan, Q., Tian, J., Wang, Q., Tan, Z., Li, Z., & Wang, N. (2020). *Assessing temporal trends of soil erosion and sediment redistribution in the Hexi Corridor region using the integrated RUSLE-TLSD model. Catena*, 195, 104756.

Liu, L., & Basso, B. (2020). *Linking field survey with crop modeling to forecast maize yield in smallholder farmers' fields in Tanzania. Food Security*, 12(3), 537–548. https://doi.org/10.1007/s12571-020-01020-3

Loboguerrero, A.M., Birch, J., Thornton, P., Meza, L., Sunga, I., Bong, B.B., Rabbinge, R., Reddy, M., Dinesh, D., Korner, J., Martinez-Baron, D., Millan, A., Hansen, J., Huyer, S., & Campbell, B. (2018). *Feeding the World in a Changing Climate: An Adaptation Roadmap for Agriculture. The Global*

Commission on Adaptation. Rotterdam and Washington, DC. 20 p. Available online at www.gca.org

McNally, A., Husak, G. J., Brown, M., Carroll, M., Funk, C., Yatheendradas, S., ... & Verdin, J. P. (2015). *Calculating crop water requirement satisfaction in the West Africa Sahel with remotely sensed soil moisture*. Journal of Hydrometeorology, 16(1), 295-305.

Mgendi, G., Mao, S., &Qiao, F. (2021). *Is a Training Program Sufficient to Improve the Smallholder Farmers' Productivity in Africa? Empirical Evidence from a Chinese Agricultural Technology Demonstration Center in Tanzania. Sustainability*, 13(3), 1527.

Mgendi, G., Shiping, M. and Xiang, C., 2019. *A review of agricultural technology transfer in Africa: Lessons from Japan and China case projects in Tanzania and Kenya. Sustainability*, 11(23), p.6598.

Minot, N., &Pelijor, N. (2010). *Food security and food self-sufficiency in Bhutan. Washington, D.C: International Food Policy Research Institute (IFPRI) and Ministry of Agriculture and Forests (MoAF).* (agris.fao.org accessed in 04/25/2017)

Morgen, S. (2001). *The Agency of Welfare Workers: Negotiating Devolution, Privatization, and the Meaning of SelfSufficiency. American Anthropologist*, 103(3), 747-761.

Mourice, S. K., Tumbo, S. D., Nyambilila, A., &Rweyemamu, C. L. (2015). *Modeling potential rain-fed maize productivity and yield gaps in the Wami River sub-basin, Tanzania. Acta Agriculturae Scandinavica, Section* B—Soil & Plant Science, 65(2), 132-140.

Muthoni, F. K., Odongo, V. O., Ochieng, J., Mugalavai, E. M., Mourice, S. K., Hoesche-Zeledon, I., ... &Bekunda, M. (2019). *Long-term spatial-temporal trends and variability of rainfall over Eastern and Southern Africa. Theoretical and Applied Climatology*, 137(3), 1869-1882.

Ojija, F., Abihudi, S., Mwendwa, B., Leweri, C. M., & Chisanga, K. (2017). T*he Impact of Climate Change on Agriculture and Health Sectors in Tanzania: A*

review. International Journal of Environment, Agriculture and Biotechnology, 2(4), 1758–1766. https://doi.org/10.22161/ijeab/2.4.37

Ojoyi, M., Mutanga, O., Mwenge Kahinda, J., Odindi, J., & Abdel-Rahman, E. M. (2017). *Scenario-based approach in dealing with climate change impacts in Central Tanzania. Futures,* 85, 30–41. https://doi.org/10.1016/j.futures.2016.11.007

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., &Mouazen, A. M. (2016). *Wheat yield prediction using machine learning and advanced sensing techniques. Computers and Electronics in Agriculture,* 121, 57-65.

Pavlyshenko, B. (2020). *Bayesian regression approach for building and stacking predictive models in time series analytics. In International Conference on Data Stream Mining and Processing* (pp. 486-500). Springer, Cham.

Peng, B. (2019). *Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U. S Field Crops Research Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the March.* https://doi.org/10.1016/j.fcr.2019.02.005

Pettorelli, N., Vik, J. O., Mysterud, A., Gaillard, J. M., Tucker, C. J., &Stenseth, N. C. (2005). *Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends in ecology & evolution, 20(9), 503-510.*

Qiu, Q., Nian, Y. J., Guo, Y., Tang, L., Lu, N., Wen, L. Z., ... & Liu, K. J. (2019). *Development and validation of three machine-learning models for predicting multiple organ failure in moderately severe and severe acute pancreatitis. BMC gastroenterology,* 19(1), 1-9.

Ramos, P. J., Prieto, F. A., Montoya, E. C., & Oliveros, C. E. (2017). *Automatic fruit count on coffee branches using computer vision. Computers and Electronics in Agriculture, 137, 9-22.*

Ribeiro, M. H. D. M., & dos Santos Coelho, L. (2020). *Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Applied Soft Computing, 86, 105837.*

Robert, C. P., Elvira, V., Tawn, N., & Wu, C. (2018). *Accelerating MCMC algorithms. Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5), e1435.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S. &Asseng, S. (2013). *The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. Agricultural and Forest Meteorology, 170, 166-182.*

Rowhani, P., Lobell, D. B., Linderman, M., &Ramankutty, N. (2011). *Climate variability and crop production in Tanzania. Agricultural and Forest Meteorology, 151(4), 449-460.*

Senay, G. (2004). *Crop Water Requirement Satisfaction Index (WRSI) Model Description.* http://iridl.ldeo.columbia.edu/documentation/usgs/adds/wrsi/WRSI_readme.pdf [Originally developed by FAO]

Senay, G. B., & Verdin, J. (2003). *Characterization of yield reduction in Ethiopia using a GIS-based crop water balance model. Canadian Journal of Remote Sensing, 29(6), 687-692.*

Sengupta, S., & Lee, W. S. (2014). *Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions. Biosystems Engineering, 117, 51-61.*

Shahhosseini, M., Hu, G., &Archontoulis, S. (2020). *Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science*, 11, 1120.

Shahhosseini, M., Martinez-Feria, R. A., Hu, G., &Archontoulis, S. V. (2019). *Maize yield and nitrate loss prediction with machine learning algorithms. Environmental Research Letters*, 14(12), 124026.

Su, Y. X., Xu, H., & Yan, L. J. (2017). *Support vector machine-based open crop model (SBOCM): Case of rice production in China. Saudi journal of biological sciences*, 24(3), 537-547.

Suleiman, R. A., & Kurt, R. A. (2015). *Current maize production, postharvest losses and the risk of mycotoxins contamination in Tanzania. In 2015 ASABE Annual International Meeting* (p. 1). American Society of Agricultural and Biological Engineers.

Sun, J., Di, L., Sun, Z., Shen, Y., & Lai, Z. (2019). *County-level soybean yield prediction using deep CNN-LSTM model. Sensors,* 19(20), 4363.

Tarnavsky, E., Chavez, E., & Boogaard, H. (2018). *Agro-meteorological risks to maize production in Tanzania: Sensitivity of an adapted Water Requirements Satisfaction Index (WRSI) model to rainfall. International Journal of Applied Earth Observation and Geoinformation,* 73, 77-87.

Thakur, D., & Biswas, S. (2022). *Machine Learning in Sustainable Healthcare. In Advanced Computational Techniques for Sustainable Computing* (pp. 79-91). Chapman and Hall/CRC.

Thornton, P., Dinesh, D., Cramer, L., Loboguerrero, A. M., & Campbell, B. (2018). *Agriculture in a changing climate: Keeping our cool in the face of the hothouse.* 47(4), 283–290. https://doi.org/10.1177/0030727018815332

Tumbo, S., Sanga, C., Sumari, N., &Kahimba, F. (2015). *Assessing the impacts of climate variability and change on agricultural systems in Eastern Africa while enhancing the region's capacity to undertake integrated assessment of vulnerabilities to future changes in climate-Tanzania. Columbia University.*

United Nations (2017) *Resolution adopted by the General Assembly on 6 July 2017, Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development* (A/RES/71/313 Archived 28 November 2020 at the Wayback Machine)

URT. (2017). *National Food Security Bulletin. Ministry of Agriculture, Livestock and Fisheries. December 2016, Dodoma, Tanzania.*

URT. (2017). *National Food Security Bulletin. Ministry of Agriculture, Livestock and Fisheries. January 2017, Dodoma, Tanzania.*

Verdin, J., &Klaver, R. (2002). *Grid cell based crop water accounting for the famine early warning system. Hydrological Processes,*16(8), 1617-1630.Wang, A.

X., Lobell, D., &Ermon, S. (2015). Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data.

Wu, H., Cai, Y., Wu, Y., Zhong, R., Li, Q., Zheng, J., ... & Li, Y. (2017). *Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression. Bioscience trends.*

Xiao, C., Chen, N., Hu, C., Wang, K., Gong, J., & Chen, Z. (2019). *Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach. Remote Sensing of Environment,* 233, 111358.

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., ... & Allen, R. W. (2019). *Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environmental pollution,* 245, 746-753.

Zougmoré, R. B., Partey, S. T., Ouédraogo, M., Torquebiau, E., & Campbell, B. M. (2018). *Facing climate variability in sub-saharanafrica: Analysis of climate-smart agriculture opportunities to manage climate-related risks.* Cahiers Agricultures, 27(3). https://doi.org/10.1051/cagri/2018019

APPENDICES

Appendix 1: Research Clearance Letter

UNITED REPUBLIC OF TANZANIA
MINISTRY OF EDUCATION, SCIENCE AND TECHNOLOGY

MZUMBE UNIVERSITY
OFFICE OF THE DEPUTY VICE CHANCELLOR (A)

[illegible]

Ref No. MU/EPCS/INT/38/Vol IV/152 Date 03rd December 2021

TO WHOM IT MAY CONCERN

RE: INTRODUCTION OF MS. [illegible] V. LEBALWA

1. The bearer of this letter Ms. Bertha V. Lebalwa whose registration number is 1941200121 is a postgraduate student at our University (Mzumbe University) pursuing Master of Science in Information Technology (MSc. ITS). As part of requirements for completion of her studies, she is collecting data on "FORECASTING OF THE RAIN FED MAIZE YIELD IN TANZANIA USING MACHINE LEARNING".

2. This letter seeks to address three purposes. Firstly, to introduce her to you. Secondly, to request you to grant her permission to undertake data collection at your organization. And thirdly, to request you to facilitate any form of assistance she might need in order to successfully pursue this noble exercise at your organization. We can assure you that this activity is entirely for academic purpose and will never be used for any other purposes.

3. We trust that you will provide the student with necessary assistance.

4. Sincerely yours,

Dr. Edward Pakaya (PhD)
FOR DEPUTY VICE CHANCELLOR (Academic Affairs)

Dissertation

ORIGINALITY REPORT

15%	11%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Mzumbe University Student Paper	1%
2	scholar.mzumbe.ac.tz Internet Source	1%
3	link.springer.com Internet Source	1%
4	www.nature.com Internet Source	<1%
5	media.neliti.com Internet Source	<1%
6	www.mdpi.com Internet Source	<1%
7	Balraj Singh, Parveen Sihag, Karan Singh, Sanjeev Kumar. "ESTIMATION OF TRAPPING EFFICIENCY OF VORTEX TUBE SILT EJECTOR", International Journal of River Basin Management, 2018 Publication	<1%

UNITED REPUBLIC OF TANZANIA

MINISTRY OF EDUCATION, SCIENCE AND TECHNOLOGY

MZUMBE UNIVERSITY
FACULTY OF SCIENCE AND TECHNOLOGY

Tel: +255 23 2931220/21/22
Fax: +255 23 2931216
E-Mail: fst@mzumbe.ac.tz

P. O. Box 87,
Mzumbe,
TANZANIA

11/10/2022

Head of Department,
Computing Science Studies
Mzumbe University

RE: ERROR FREE CERTIFICATE FOR MS. BERTHA MSULICHE LEBALWA

Please refer to the captioned matter above.

Ms Bertha Msuliche Lebalwa successfully defended her dissertation and declared passed *"with minor correction"* on October 04, 20202. The title of her dissertation is ***Forecasting of the Rain-Fed Maize Yield in Tanzania using Machine Learning.***

This letter confirms that Ms Bertha Msuliche Lebalwa has incorporated all raised comments from the examiners and oral panellists as per MU guidelines. The submitted matrix stipulates how the candidate addressed comments from the internal and external examiners.

Also, I submitted the document to Turnitin software to check for plagiarism and similarity before sending it to a proofreader. The result was 15%, and the breakdown is attached to the thesis document.

Therefore, I confirm that the submitted thesis document is ***Error-Free,*** and it can proceed to the appropriate organ for further steps.

Best regards,

Dr Tupokigwe Isagah

Lecturer-CSS Department

UNITED REPUBLIC OF TANZANIA

MINISTRY OF EDUCATION, SCIENCE AND TECHNOLOGY

MZUMBE UNIVERSITY
FACULTY OF SCIENCE AND TECHNOLOGY

Tel: +255 23 2931220/21/22
Fax: +255 23 2931216
E-Mail: fst@mzumbe.ac.tz

P. O. Box 87,
Mzumbe,
TANZANIA

11/10/2022

Head of Department,
Computing Science Studies
Mzumbe University

RE: ERROR FREE CERTIFICATE FOR MS. BERTHA MSULICHE LEBALWA

Please refer to the captioned matter above.

Ms Bertha Msuliche Lebalwa successfully defended her dissertation and declared passed *"with minor correction"* on October 04, 20202. The title of her dissertation is ***Forecasting of the Rain-Fed Maize Yield in Tanzania using Machine Learning***.

This letter confirms that Ms Bertha Msuliche Lebalwa has incorporated all raised comments from the examiners and oral panellists as per MU guidelines. The submitted matrix stipulates how the candidate addressed comments from the internal and external examiners.

Also, I submitted the document to Turnitin software to check for plagiarism and similarity before sending it to a proofreader. The result was 15%, and the breakdown is attached to the thesis document.

Therefore, I confirm that the submitted thesis document is ***Error-Free,*** and it can proceed to the appropriate organ for further steps.

Best regards,

Dr Tupokigwe Isagah

Lecturer-CSS Department

MOROGORO, TANZANIA

Cel. 075709911

E-mail mohhashim@gmail.com,

mohhashim@sua.ac.tz

06th October 2022

TO WHOM IT MAY CONCERN

RE: CERTIFICATION OF LANGUAGE EDITING OF A DISSERTATION WRITTEN BY BERTHA MSULICHE LEBALWA

This is to certify that I, the undersigned, have proofread and edited a Dissertation written by **Bertha Msuliche Lebalwa**, *Forecasting of the Rain Fed Maize Yield in Tanzania Using Machine Learning.*

Proofreading and editing focused on grammar, meaning, vocabulary selection, structuring of sentences at clause and discourse levels and all other matters related to language.

I wish to confirm that the author has made all the corrections and the recommended improvement to my satisfaction.

Dr Hashim Issa Mohamed

Language Editor