የኢፌዴሪ የቴክኒክና ሙያ ስልጠና ኢንስቲትዩት
FDRE TECHNICAL & VOCATIONAL TRAINING INSTITUTE

TECHNICAL VOCATIONAL EDUCATION AND TRAINING INSTITUTE

SCHOOL OF GRADUATE STUDIES FACULTY OF ELECTRICAL, ELECTRONICS, AND INFORMATION AND COMMUNICATION TECHNOLOGY

Near-Real-Time Prediction of Desert Locust Presence Risk in Ethiopia Using Satellite-Derived Environmental Data and Machine Learning

By: *Belay Amare*

Advisor: *Dr. Kibrom Tadesse*

June, 2026

Ethiopia

TECHNICAL VOCATIONAL EDUCATION AND TRAINING INSTITUTE

SCHOOL OF GRADUATE STUDIES FACULTY OF ELECTRICAL, ELECTRONICS, AND INFORMATION AND COMMUNICATION TECHNOLOGY

Near-Real-Time Prediction of Desert Locust Presence Risk in Ethiopia Using Satellite-Derived Environmental Data and Machine Learning

By: *Belay Amare*

Advisor: *Dr. Kibrom Tadesse*

June, 2026

Ethiopia

Declaration

DECLARATION

I, ***Belay Amare Bekele***, hereby declare that this thesis, entitled ***"Near-Real-Time Prediction of Desert Locust Presence Risk in Ethiopia Using Satellite-Derived Environmental Data and Machine Learning,"*** is my original work and has been carried out under the guidance and supervision of my research advisor.

I further declare that this thesis has not been submitted, either in whole or in part, for the award of any degree, diploma, or other academic qualification at this or any other institution. All sources of information and materials used in this study were properly acknowledged and cited.

Declared By

Name: Belay Amare Bekele

Signature: ____________________

Department:: ICT, FDRE TVTI

Date: ____________________

This dissertation has been submitted for examination with my approval as the candidate's advisor.

Principal Advisor: Kibrom T (PhD)

Signature: ____________________

Date: Jun, 3, 2026

APPROVAL PAGE

This is to certify that the dissertation entitled "**Near-Real-Time Prediction of Desert Locust Presence Risk in Ethiopia Using Satellite-Derived Environmental Data and Machine Learning**" submitted by **Belay Amare Bekele**, ID No. ***TTMR/056/16***, in partial fulfilment of the requirements for the degree of Master of Science in Information Technology, has been examined and approved by the Board of Examiners of the Department of Information Technology, FDRE Technical and Vocational Training Institute.

Examination Role	**Name**	**Signature**	**Date**
Principal Advisor	***Dr. Kibrom Tadesse***		
Co-Advisor (if any)	______________	______________	______________
Internal Examiner	______________	______________	______________
External Examiner	______________	______________	______________
Department Head, IT	______________	______________	______________
Dean, College of Graduate Studies	______________	______________	______________

Place: Addis Ababa, Ethiopia

Date of Final Approval: ______________________

Acknowledgements

First and foremost, I thank Jesus Christ and the Blessed Virgin Mary for their unwavering guidance, blessings, strength, and support throughout my academic journey. I wish to extend my sincere appreciation to all those who contributed to the completion of this master's thesis.

My sincere appreciation goes to my advisor, Dr. Kibrom Tadesse, for his invaluable guidance, constructive comments, continuous encouragement, and professional support throughout this study. His expertise and dedication greatly contributed to the success of this study. I am also deeply grateful to the faculty members and instructors at the FDRE TVT Institute for their knowledge, mentorship, and commitment during my studies.

My deepest gratitude goes to my beloved wife, Derartu Zenebe, and my beloved daughters, Eliana Belay and Aniken Belay, for their endless love, patience, encouragement, and emotional support throughout this journey. Their understanding, sacrifices, and constant motivation give me the strength to continue and complete this thesis. I would also like to extend my sincere appreciation to my family and friends for their continuous encouragement, prayer, and support. In a special and heartfelt way, I dedicate this thesis to the loving memory of my beloved niece Hidet Samuel. This work is a tribute to her life and memory. May her soul rest in eternal peace.

Finally, I humbly acknowledge all those who contributed, directly or indirectly, to the success of this research. Any errors or shortcomings in this thesis are entirely my own responsibility.

Abstract

Desert locusts (*Schistocerca gregaria*) pose a long-standing challenge to crop production and food security in Ethiopia. Current warning systems have several limitations. These include relatively coarse spatial resolutions, rule-based inferences, and delayed reports. These limitations restrict their operability and effectiveness. This study developed a near-real-time, spatially validated framework for predicting locust presence in Ethiopia using multisource satellite data. The model integrates MODIS NDVI, CHIRPS rainfall, and ERA5 temperature datasets covering 2015–2024, combined with 18,742 quality-controlled locust occurrence records from the FAO DLIS and the DLCO-EA. All data were harmonized to a spatial resolution and a 16-day temporal interval. A total of 32 environmental predictors were derived, including lagged variables, anomalies, and seasonal indicators. Four models (LR, RF, XGBoost, and LightGBM) were tested using spatial block cross-validation. LightGBM achieved the best spatial performance (AUC = 0.740), and a stacked ensemble (LightGBM + XGBoost + calibrated logistic regression) achieved a final test set AUC of 0.780 (95% CI: 0.765–0.795) with a conservative temporal validation AUC of 0.85 under a strict two-year gap split. SHAP analysis identified mean temperature as the dominant predictor (mean absolute SHAP = 0.48), followed by 30-day cumulative rainfall (0.15) and NDVI (0.08), with ecologically consistent and nonlinear response patterns. A fully automated pipeline implemented in Google Earth Engine and Google Colab generated 1 km-resolution risk maps across approximately 1.1 million grid cells within 45 min on a 16-day cycle. The results demonstrate that freely available satellite-derived data can support the development of interpretable, scalable, and potentially operationally deployable desert locust early warning systems in Ethiopia.

Keywords: desert locust, machine learning, near-real-time prediction, remote sensing, Google Earth Engine, XGBoost, risk mapping, Ethiopia.

Table of Contents

List Of Tables

List of Figures

List of Abbreviations

Abbreviations	Full Form
AVHRR	Advanced Very High-Resolution Radiometer
CHIRPS	Climate Hazards Group Infra-Red Precipitation with Station Data
CV	Cross-Validation
DLCO-EA	Desert Locust Control Organization for Eastern Africa
DLIS	Desert Locust Information Service
DTR	Diurnal Temperature Range
ECMWF	European Centre for Medium-Range Weather Forecasts
EMI	Ethiopian Meteorological Institute
EPSG	European Petroleum Survey Group (geodesy)
ERA5	ECMWF Reanalysis v5
ESA	European Space Agency
EVI	Enhanced Vegetation Index
FAO	Food and Agriculture Organization of the United Nations
FDRE	Federal Democratic Republic of Ethiopia
GDP	Gross Domestic Product
GEE	Google Earth Engine
GIS	Geographic Information System
GPU	Graphics Processing Unit
IGAD	Intergovernmental Authority on Development
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NIR	Near-Infrared
NOAA	National Oceanic and Atmospheric Administration
NRT	Near-Real-Time
RAMSES	Reconnaissance and Monitoring System for the Environment of Schistocerca
RMSE	Root Mean Square Error
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve
SAR	Synthetic Aperture Radar
SAVI	Soil-Adjusted Vegetation Index
SMOS	Soil Moisture and Ocean Salinity
SMOTE	Synthetic Minority Oversampling Technique
SNNP	Southern Nations, Nationalities, and Peoples' Region
SPI	Standardized Precipitation Index
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
TRMM	Tropical Rainfall Measuring Mission
VIIRS	Visible Infrared Imaging Radiometer Suite
WGS84	World Geodetic System 1984
XGBoost	eXtreme Gradient Boosting

CHAPTER ONE

1. INTRODUCTION

1.1 Background of the Study

Desert locusts (*Schistocerca gregaria*) are among the most destructive migratory pests in human history (FAO, 2023a). These insects live in semi-arid and arid regions, from West Africa across the Middle East to Southwest Asia. When the right environmental conditions appear, desert locusts can change dramatically. They shift from a harmless, solitary state to highly mobile, gregarious swarms. A single mature swarm can cover more than 100 square kilometers and contain several billion individuals. In just one day, a swarm can consume as much food as 35,000 people would eat (FAO, 2023b). The reproductive cycle begins when females lay eggs in deep, moist sandy soil. Depending on the temperature and moisture, incubation takes from 10 to 65 days. This flexibility allows populations to grow rapidly when conditions are favorable (Landmann *et al*, 2023). Because of their extraordinary mobility and appetite, these swarms threaten agriculture in over 65 countries. Consequently, they directly affect the livelihoods of nearly 10% of the global population (Landmann *et al*, 2023). The challenge of predicting desert locust outbreaks has grown as climate variability accelerates. It began in the Empty Quarter of the Arabian Peninsula after unusually heavy cyclonic rains. From there, it spread to East Africa, the Red Sea coast, and South Asia. This event exposed critical gaps in the existing early warning frameworks (Salih *et al*, 2020). Traditional prediction methods rely heavily on field survey data and historical data on outbreak patterns. When anomalous weather events create breeding habitats in areas that are not routinely monitored, these methods fail to anticipate the scale and trajectory of the crisis (I. S. Yusuf *et al*, 2024). Decision-makers often lack actionable, spatially explicit information until locust populations reach plague proportions.

Africa's geography and climate make it particularly vulnerable to desert locust invasions. Historically, North African countries, such as Morocco, Algeria, Tunisia, and Egypt, have suffered from these invasions. However, the main invasion route lies further south, in the Sahel. This extensive bioclimatic corridor stretches from Senegal to Sudan (Landmann *et al*, 2023). From the

Sahel, locusts can move north toward the Mediterranean or east toward the Horn of Africa. Most of the Horn of Africa countries, including Ethiopia, Somalia, Kenya, Djibouti, and Eritrea, are considered high-risk. In this region, agriculture employs approximately 80% of the population and contributes roughly 40% of the gross domestic product (World Bank, 2024). Ethiopia experienced this vulnerability during the 2019–2021 period. Swarms first crossed into the country from neighboring Somalia and Kenya in late 2019. They quickly multiplied across Oromia, the Southern Nations, Nationalities and Peoples' Region, and the eastern and southeastern lowlands of Ethiopia. At the peak of the crisis, an estimated 1.25 million hectares of cropland and pasture were damaged. This has put the food security of over 15 million people at immediate risk (FAO, 2020). The government coordinated ground and aerial control operations through the Disaster Risk Management Commission and Plant Health Regulatory Directorate. Surveillance coverage is persistently low because access to modern sensing and technologies has remained limited.

The motivation for this study comes not only from the documented literature on desert locust outbreaks, but also from direct field observation. Between 2019 and 2021, the eastern part of Ethiopia — particularly the Hararghe zones — experienced one of the most severe desert locust upsurges in over two decades, a crisis the researcher witnessed firsthand. Swarms moved through cropland and rangeland within hours, stripping vegetation and leaving communities without adequate pasture or harvest. What stood out most was not only the scale of the destruction, but the recurring pattern of delayed response: by the time survey teams were mobilized and reports compiled, the swarms had already moved on, leaving damage that earlier warning could have substantially reduced.

This lived experience exposed a critical gap in Ethiopia's existing locust monitoring approach — the continued reliance on reactive, ground-based surveillance rather than proactive, near-real-time risk prediction. It became evident that the environmental precursors of locust presence, such as anomalous rainfall, vegetation greening, and temperature patterns, are often detectable in satellite data well before swarms are physically reported on the ground. This observation, combined with the urgency witnessed during the 2019–2021 outbreak, forms the central motivation for this thesis. This is the time to develop a machine learning–based system capable of anticipating desert locust presence risk before it transforms to visible swarms, giving Ethiopia's early warning systems.

Desert locust swarms spread rapidly, often outpacing traditional ground-based surveys, which therefore do not detect outbreaks in time. Meteorological factors significantly influence the spatial distribution, survival rates, and migration patterns of hopper bands. By integrating satellite-derived environmental data with machine learning techniques, this approach has the potential to enhance desert locust monitoring and early warning capabilities in Ethiopia.

1.2 Statement of the Problem

Desert locust invasions pose a major threat to Ethiopia's agriculture, a sector that employs nearly 80% of the population and contributes approximately 40% of the country's gross domestic product. (FAO, 2022). Between 2019 and 2021, desert locust swarms affected more than 1.25 million hectares of farmland and grazing land, threatening the food security and livelihoods of approximately 15 million people (FAO, 2022). However, existing early warning systems, including the FAO Desert Locust Information Service (DLIS) and conventional ground-based surveys, are limited (Klein *et al*, 2022). They cannot accurately predict the spatial and temporal distribution of locust outbreaks at the local level. Current locust prediction methods have several limitations. First, many of these approaches rely on environmental data with low spatial resolution. The resolution was typically approximately 5 km (Lazar *et al*, 2022). Second, traditional statistical and rule-based models are limited in their ability to capture the complex nonlinear interactions among environmental variables, such as rainfall, temperature, and vegetation dynamics, that influence locust development and migration. Third, operational data-processing workflows frequently experience delays of 5–10 days between satellite data acquisition and risk assessment. In addition to these limitations, no existing study has developed a near-real-time, machine-learning-based desert locust prediction model specifically calibrated for Ethiopia's diverse agroecological zones, using high-resolution satellite-derived environmental data. Furthermore, previous studies in Africa have largely relied on random cross-validation approaches. These approaches may have overestimated the model performance. However, they do not adequately account for spatial autocorrelation in ecological datasets. Consequently, the true spatial generalizability of many existing locust prediction models remains unknown (Klein *et al*, 2022). Consequently, there is a need for a spatially explicit prediction framework that can integrate multisource satellite data with advanced machine learning techniques to produce timely and reliable locust risk predictions for Ethiopia. This study addresses these gaps by integrating MODIS

NDVI imagery, CHIRPS rainfall estimates, and ERA5 temperature reanalysis data within a near-real-time machine learning framework. This study systematically compares the predictive performance of Logistic Regression, random forest, XGBoost, and LightGBM using both random and spatial cross-validation. No previous Ethiopia-specific study has implemented spatially validated near-real-time probabilistic locust prediction at a 1 km resolution using multi-source satellite data. In addition, a near-real-time prediction pipeline was developed. The pipeline generated spatially explicit locust risk maps. These maps were synchronized with the 16-day MODIS NDVI compositing cycle.

1.3 Research Questions

The purpose of this study is to answer the following research questions:

1. Which satellite-derived variables are important for predicting desert locusts in Ethiopia?
2. Which validation strategy allows us to test and obtain a realistic indication of the generalization ability of the model and compare the ML algorithms against each other?
3. How effectively can a near-real-time machine learning pipeline generate spatially explicit desert locust risk maps using multi-source satellite data?

1.4 Objectives of the Study

1.4.1 General Objective

The main objective of this study was to develop and validate a near-real-time machine learning model to predict the risk of desert locust presence in Ethiopia.

1.4.2 Specific Objectives

The specific objectives of this study were as follows:

1. To identify and rank satellite-derived environmental variables that significantly influence desert locust presence in Ethiopia.
2. To compare the predictive performance of four machine learning algorithms – Logistic Regression, Random Forest, XGBoost, and LightGBM – for predicting desert locust presence.

3. To evaluate the effects of spatial and random cross-validation on model generalizability.
4. To design and implement a near-real-time processing pipeline using Google Earth Engine and Google Colab for automated risk map generation.

1.5 Scope of the Study

This study sought to describe and quantify the relationships between environmental factors and desert locusts in Ethiopia. We selected Ethiopia as the focal country for this study for several reasons. First, Ethiopia suffers from recurrent desert locust invasions. Second, its varied landscapes support locust breeding and migration. Focusing on a single country allowed us to maintain uniform spatial resolution throughout the model and to use comparable environmental datasets. It also prevented cross-border- variations in ecological monitoring and reporting from becoming an additional source of uncertainty in the modeling system. This study focuses on Ethiopia because it is a frontline country for desert locust incursions, with documented outbreaks over the past 10 years, and spatially validated ML models specifically calibrated for Ethiopia. Only NDVI, rainfall, and temperature were included because they are freely available in near-real-time latency from MODIS, CHIRPS, and ERA5. Wind vectors and soil moisture were excluded because of their delayed availability and higher computational cost, although their omission is acknowledged as a limitation. All three datasets are known to be reliable, publicly accessible, global products for environmental modeling. Four machine learning models, logistic regression, random forest, XGBoost, and LightGBM, were evaluated to predict the presence and swarm formations of locusts.

1.6 Significance of the Study

The work presented here has important implications for science, operations, and policy. This study helps us understand the dynamics of key environmental factors for driving locust development, migration, and swarming in Ethiopian agroecological zones. These parameters include vegetation greenness and rainfall variability. Methodologically, this study compares systematic and spatial cross-validation procedures across four different machine learning algorithms. From an operational viewpoint, this framework produces reliable probabilistic desert locust risk maps at a 1 km spatial scale on a biweekly basis. The predictions from the pipeline have a finer spatial resolution and higher temporal frequency. This enhanced information can strengthen decision-making for officials at the Ministry of Agriculture, regional locust control offices, and DLCO-EA field agents. This allows for the early detection and control of swarm development and dissemination. This study enhances early warning and response to mitigate risks. Beyond Ethiopia, there is potential for such an approach to be scaled in the neighboring Horn of Africa. This framework may inform policy interventions for food security and for locust control. Finally, this study contributes to science by making its source codes and datasets publicly available. This supports the reproducibility of future studies.

1.7. Limitations of the Study

This research has the following limitations placed on its results, which need to be considered when assessing this research:

- ❖ Sampling bias may still exist because the occurrence data did not provide complete spatial coverage of all regions of Ethiopia.
- ❖ Finer-scale environmental grids were derived from coarse-resolution climatic datasets because higher-resolution near-real-time climatic data were not consistently available in Ethiopia.
- ❖ Cloud cover during the summer (Kiremt) season introduced systematic missing values in the MODIS NDVI time-series data, thereby reducing vegetation signal quality during important seasonal periods. Although filtering techniques reduce some part of the uncertainty, residual noise and data gaps remain.
- ❖ Habitat suitability maps identify environmentally suitable areas for desert locust occurrence; however, actual swarm movement and migration dynamics were not explicitly represented in the modeling process.
- ❖ The automated workflow depends heavily on cloud-based computational platforms, including Google Earth Engine and Google Colab. Service interruptions, API modifications, and platform limitations can affect the continuity of the automated processing workflow.

1.8. Conceptual Framework

The conceptual framework of this study represents the linkages between satellite-derived environmental parameters, ecological processes, machine learning models, and the prediction of desert locust occurrences in Ethiopia. The conceptual framework is built on a central assumption. Environmental and climatic factors strongly influence the occurrence of desert locusts.

This study used multisource satellite datasets to represent key environmental variables. MODIS provides vegetation indices. CHIRPS supplies rainfall data. ERA5 delivers temperature information as a climate variable. Delayed ecological processes, such as the development of a specific locust generation, result from environmental factors in the preceding weeks. We captured these delays by including temporal and lagging environmental variables. The aforementioned remotely sensed datasets were preprocessed using a set of procedures that included spatial harmonization, cloud masking, imputation of missing values, normalization, and feature engineering. These pre-processed variables were then used as inputs for the ML algorithms, namely logistic regression, Random Forest, XGBoost, and LightGBM. The algorithms were trained to learn the relationship between the predictors and field-collected observations. The models were validated using random and spatial cross-validation. The final output of the framework was a map of probabilistic desert locust risk for Ethiopia, indicating the possibility of its presence. It produces risk maps that enable near-real-time predictions.

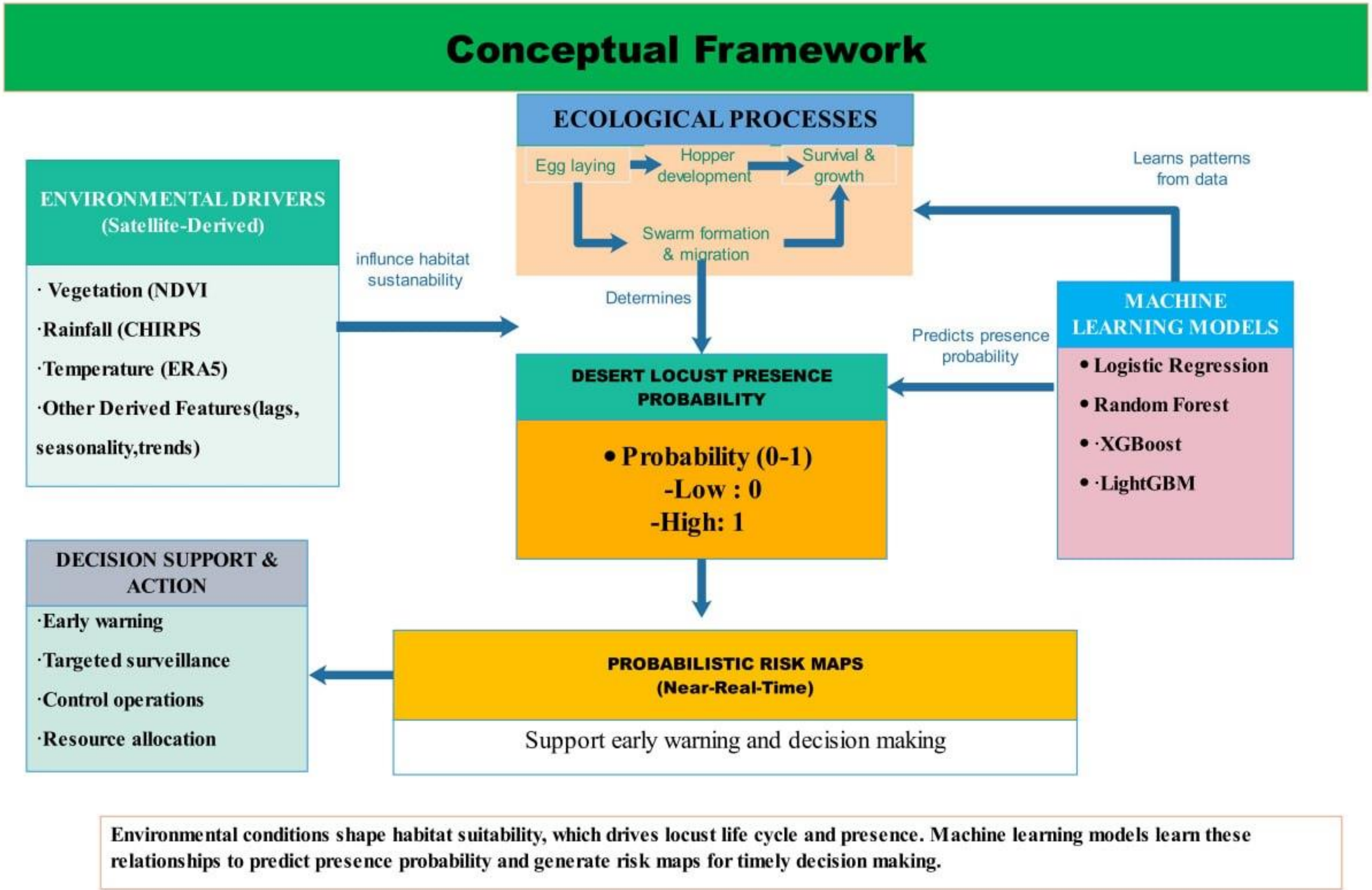

Figure 1. 1 Conceptual Framework

1.9 Organization of the Study

This thesis consists of five chapters, which are summarized as follows. Chapter One describes the background of the study from the global to the Ethiopian distribution of desert locusts, the statement of the problem, the objective of the study, the research question, the scope of the study, the significance of the study, the limitations of the study, and the organization of the thesis. The second chapter focuses on a critical evaluation of the relevant literature on desert locust invasion, satellite prediction systems, and near-real-time monitoring models. Chapter Three describes the study area, sources, collection methods, and technologies used in the study. Chapter Four focuses on an investigation into the technical aspects of building a prediction model with a detailed exploration of the machine learning models to be selected, designed, trained, and validated. Chapter 5 provides a summary, recommendations, and advice for future work in this field. In the references section, all works cited in this report are listed.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Life Cycle of the Desert Locust

The life cycle of a desert locust consists of three phases: egg, hopper (nymph), and adult. As environmental conditions may affect population levels with a time lag. If the conditions are too wet, locusts may drown (Symmons & Cressman, 2001a). Eggs normally hatch within 10-14 days in warm and humid conditions; in contrast, temperatures below 20 °C will delay hatching for 3-4 weeks. Temperatures above 45 °C are unsuitable for the development of locusts (Latchininsky *et al*, 2011). Following hatching, the locust develops through five hopper instars to adulthood. This stage typically lasts between 4 and 6 weeks and requires temperatures in the range 28–38 °C. Hoppers are voracious feeders, and hopper bands are known to devastate the vegetation. Gregarization begins at this stage if population numbers increase (Rogers *et al*, 2014). After the final molt, winged adults emerge and require an additional 1 – 2 weeks to mature before breeding. Adults can live for between 3 and 5 months; when gregarious, the adults form swarms which, assisted by wind, can travel 100 -150 km/day (FAO, 2020). Multiple generations may be produced in a single season if the conditions remain suitable.

The locust life cycle creates predictable delays between environmental conditions and locust occurrences. Rainfall during the previous two to four weeks supports egg development, whereas vegetation greenness measured by NDVI provides resources for hopper growth and is often most relevant four to six weeks before swarm formation. Consequently, cumulative rainfall and NDVI over the preceding 30 - 60 days were valuable predictors of locust presence. As adult swarms are typically preceded by four to eight weeks of favorable breeding conditions, incorporating these biological time lags enables machine learning models to capture delayed ecological responses and improve near-real-time outbreak prediction.

Table 2. 1 Summary of Life Cycle Durations

Stage	Duration	Extended Duration	Environmental Drivers
Egg	10 – 14 days	3–4 weeks	Soil moisture and temperature
Hopper (5 stars)	4 – 6 weeks	Up to 8 weeks	Green vegetation (NDVI)
Adult	1 – 2 weeks	—	Feeding (vegetation)
Adult lifespan	3 – 5 months	—	Temperature and food

2.2 Environmental Drivers of Locust Outbreaks

2.2.1 Rainfall

Rainfall is a key trigger for desert locust breeding and subsequent outbreaks in arid and semi-arid regions (Symmons & Cressman, 2001a). Locust behavior also differs according to climatic conditions. For instance, the arid lowlands of Ethiopia can become a suitable breeding habitat with moderate rainfall followed by subsequent greenness. Typically, the lagged process is persistent in nature. Therefore, a single rainfall measurement is insufficient to characterize an outbreak. In the present study, we used cumulative rainfall over different time scales as one of the variables. We also included other indices, such as the Standardized Precipitation Index (SPI), to represent precipitation anomalies.

2.2.2 Vegetation (NDVI)

The NDVI has been widely used in locust studies across multiple satellite platforms with a long historical archive. However, NDVI has several limitations, particularly in dryland environments. In densely vegetated areas, the index may saturate, and the soil background reflectance can affect the vegetation estimates (Huete *et al*, 2002). To address these issues, researchers have proposed alternative indices, such as the EVI and SAVI, which increase sensitivity to sparse vegetation and remove soil background effects. In Ethiopia's dry lowlands, exposed soil surfaces undermine NDVI reliability. Exposed soil surfaces may compromise NDVI reliability in Ethiopia’s dry lowlands. The dense vegetation in the highlands amplifies the saturation effect. Despite these limitations, we chose NDVI as our main indicator for desert locust research. The NDVI has been widely validated as an effective indicator in desert locust research. NDVI data are consistently available through MODIS products, which support near-real-time monitoring. We also incorporated NDVI anomalies to identify vegetation conditions that deviated from long-term seasonal patterns.

2.2.3 Temperature

Temperature is of prime significance for the development, survival, and migration of the desert locusts. Temperature is ideal for egg development, nymph development, and adult flight behavior (Latchininsky *et al*, 2011). The physiological effects of temperature on desert locusts are non-linear. At a suitable temperature, locusts develop rapidly in the preferred range, with the development rate falling outside the threshold values. Mean temperature variables are predominantly used in many current predictive models, which may not adequately portray the non-linear response of locusts to temperature. To address the temperature response of locusts, multiple temperature-based variables were derived from the ERA5 climate reanalysis product, including mean temperature, temperature extremes, and time series of rolling means, to estimate short-term and cumulative conditions. The ERA5 climate reanalysis dataset exhibits a strong temporal consistency. Therefore, it is highly suitable for environmental monitoring over large spatial scales. While the resolution is coarse (relatively), potentially causing inaccuracy upon resampling to a more refined spatial scale, this is an admitted issue when considering model results. Tree-based machine learning algorithms can capture non-linear variable interactions and responses while adhering to minimal strict parametric requirements

2.3 Remote Sensing for Locust Monitoring

2.3.1 Historical Evolution of Remote Sensing for Locust Monitoring

The use of remote sensing for desert locust monitoring began in the 1960s, when meteorological satellites were first used to collect rain-related environmental data in arid regions (Hielkema *et al*, 1986). Vegetation indices derived from NOAA AVHRR data can detect the greening front. In the 1990s, the FAO DLIS incorporated wind and NDVI data from remote sensing into operational locust-monitoring systems. During the same period, RAMSES implemented remote sensing imagery and GIS for monitoring locusts (Cressman, 1997). This satellite collects detailed environmental parameters for the entire globe at moderate spatial resolution and high frequency. Some studies have shown that MODIS-derived vegetation and temperature products identify locust breeding potential across Africa (Ceccato *et al*, 2007a). Recent advances in cloud computing and machine learning have enabled more effective near-real-time environmental monitoring and risk modeling.

2.3.2 Sensor Capabilities for Locust Habitat Detection

MODIS-derived NDVI has been extensively used for desert locust habitat monitoring because of its long historical archive and frequent temporal coverage. For example, vegetation greenness derived from MODIS successfully identified potential breeding habitats across North Africa (Ceccato *et al*, 2007b). Similarly, Kimathi *et al* found that NDVI was among the most influential predictors of locust occurrence in East Africa (Kimathi *et al*, 2020). However, several studies have noted that MODIS NDVI may underestimate vegetation variability in heterogeneous landscapes because of to its moderate spatial resolution and saturation effects in densely vegetated areas. These findings suggest that while MODIS is suitable for operational monitoring, its effectiveness may vary across ecological zones.

CHIRPS rainfall data have become one of the most widely used precipitation datasets for locust monitoring because they combine satellite observations with ground-station measurements. Several studies have reported strong associations between cumulative rainfall and subsequent locust breeding activity. For instance, rainfall anomalies detected using CHIRPS preceded vegetation growth and locust population. Nevertheless, researchers have observed that rainfall alone is often insufficient for predicting outbreaks because suitable vegetation and temperature conditions must also be present. Consequently, recent studies increasingly integrate CHIRPS with vegetation and climate datasets to improve predictive performance.

ERA5 temperature data have been widely applied in ecological modeling because of their global coverage and temporal consistency. Previous studies have demonstrated that temperature strongly influences desert locust egg development, hopper maturation, and adult survival. However, researchers have also reported limitations associated with the relatively coarse spatial resolution of ERA5, particularly in mountainous regions where local temperature variations may not be adequately captured. In Ethiopia, where elevation varies substantially across short distances, this limitation may introduce uncertainty when the ERA5 data are resampled to finer spatial resolutions. Despite these challenges, the ERA5 remains one of the most accessible datasets for near-real-time environmental monitoring.

Collectively, the literature indicates that rainfall, vegetation, and temperature are the most consistently reported environmental drivers of desert locusts. While CHIRPS effectively captures precipitation dynamics, MODIS provides information on vegetation response, and ERA5 characterizes thermal suitability. Previous studies have generally examined these variables individually or at coarse spatial scales. Few studies have integrated all three datasets within a spatially validated machine-learning framework specifically designed for Ethiopia. This gap serves as the driving force for the current study.

Table 2. 2 Satellite sensor characteristics

No	Sensor	Spatial Resolution	Temporal Resolution	Main Contribution
1	NOAA AVHRR	~1 km	Daily	Long historical vegetation record
2	MODIS	250m–1 km	1–2 days	Vegetation and surface monitoring
3	Landsat 8/9	30 m	16 days	High spatial detail
4	Sentinel-2	10–60 m	5 days	Fine-scale vegetation assessment
5	VIIRS	375–750 m	Daily	Near-real-time monitoring
6	SMAP/SMOS	9–36 km	1–3 days	Soil moisture estimation
7	Sentinel-1 SAR	10–40 m	6–12 days	Cloud-independent surface observations

2.3.3 Cloud Computing Platforms for Satellite Data Processing

In this study, GEE was used for accessing satellite data, preprocessing, and extracting environmental features. The introduction of cloud computing has eliminated the need for local storage and computing power for large-scale environmental studies. Google Earth Engine (GEE) provides access to widely used satellite and climate datasets – including MODIS, Sentinel-2, CHIRPS, and ERA5 – within a cloud-based geospatial analysis platform (Gorelick *et al*, 2017). GEE offers the benefits of automatic cloud masking, large-scale raster processing, and integration of environmental datasets within a single platform. The Google Earth Engine includes built-in machine learning tools. However, the analytical requirements of this study are better served by a dedicated Python-based environment. Libraries such as scikit-learn and XGBoost in the Python ecosystem simplify the implementation of spatial cross-validation, class balancing, hyperparameter optimization, and model interpretability analysis. The study was conducted using a two-step workflow. First, environmental features were extracted from GEE. These features were then exported for model development and evaluation. This combined method enables scalable environmental data processing, which is required for machine learning analysis.

2.4 Integrated Early Warning Systems for Desert Locusts

2.4.1 FAO Desert Locust Information Service (DLIS)

The Desert Locust Information Service (DLIS) of the FAO is the central coordinating body for monitoring desert locusts. It uses survey data from 65 national locust units, as well as satellite-based data on NDVI, rainfall, and wind vectors to compile twice-weekly bulletins and monthly situation summaries (FAO, 2023a). The DLIS covers the entire globe and provides consistent ecological thresholds defined through a rigid rule-based algorithm. There are three well-known shortcomings of this approach, all of which are relevant to the current study. First, the biweekly bulletin publication schedule imposes an operational delay of up to 14 days, which is not fast enough, given that a swarm can travel 150 km per day (Symmons & Cressman, 2001a). Second, the DLIS uses low-resolution satellite data and ignores localized vegetation patches that can be extremely important for early-stage breeding in the complex mosaic of Ethiopian agroecological zones (Klein *et al*, 2022). Third, threshold-based systems fail to capture the nonlinear relationships between rainfall, temperature, and vegetation. Therefore, a real-time machine learning prediction pipeline is needed. This pipeline must be customizable to the environment where Ethiopia's desert locust populations thrive.

2.4.2 Regional Systems: RAMSES, eLECIS, and DLCO-EA

RAMSES integrates satellite vegetation indices, soil moisture estimates, and older surveys in a GIS-based decision support system (Tappan & Cressman, 2020). Used operationally in West Africa and the Horn of Africa during the 2000s. RAMSES was intended to enable retrospective risk assessment. The system operated on predefined environmental thresholds. It could not evolve with a changing climate. It also could not integrate new data sources. Furthermore, its update frequency failed to meet the near-real-time requirements of locust management (Symmons & Cressman, 2001b). eLECIS established an internet-based system to standardize satellite data processing and link national locust units across North Africa and the Middle East. This system improves information sharing at the regional level (FAO, 2022). However, eLECIS was not used in Ethiopia, so its application to this study area is not relevant. It was a data coordination system, and that was vital.

DLCO-EA carries out aerial and ground surveys in 8 Eastern African countries, including Ethiopia. Data collected (DLCO-EA) may be useful for regional monitoring, but is not capable of predictive modeling. It is dependent on the FAO-DLIS forecast for risk assessment. Therefore, the rule-based DLIS model has limitations. These include coarse spatial resolution, pre-defined environmental thresholds, and a 14-day bulletin cycle. These limitations directly affect DLCO-EA's operational decision-making. These systems fail to address the need for near-real-time locust risk mapping in Ethiopia for three fundamental reasons. First, they are all fixed, rule-based systems. Second, none have sufficient spatial resolution to detect early breeding areas within Ethiopia's diverse agro-ecological regions. Third, none have ever been calibrated to the Ethiopian environment. This paper directly addresses this deficiency. We created an automated, machine learning-based pipeline that generates probabilistic risk maps. These maps use rolling satellite acquisition data and are updated every two weeks at a spatial resolution of a kilometer.

2.4.3 Recent Machine Learning-Based Operational Attempts

Over the last decade, Machine Learning models for desert locust prediction using satellite-based environmental factors and spatial analysis have been developed. But there is a big gulf between experimental models and operational early warning systems. Early studies were on broad-scale suitability mapping. For instance, a produced continental-scale risk map using logistic regression with TRMM rainfall and MODIS vegetation indices (AUC ≈ 0.71) (Renier *et al*, 2015). This model was not automated and did not allow near-real-time processing. A random forest model with few variables was later applied, which has better accuracy, but was not developed for operational monitoring (Lazar *et al*, 2022). An XGBoost model trained over Africa and Arabia using vegetation, rainfall, and SMOS soil moisture achieved an AUC above 0.85. However, reliance on random cross-validation for spatially autocorrelated data likely inflates real-world performance estimates (Gómez *et al*, 2021).

In 2024, a deep learning study developed a 3D CNN geospatial foundation model using Landsat and Sentinel-2 imagery, achieving 83% overall accuracy across African test sites. However, the model requires high-performance GPUs, making it impractical for edge computing or regions with limited access to GPU clusters (A. Yusuf *et al*, 2024). Ensemble machine learning applied to prediction upsurges in East Africa showed promise, yet the pipeline was not fully automated in near real time (Kimathi *et al*, 2020). These studies share several common limitations. They include coarse spatial resolution, lack of automation, reliance on random cross-validation, poor calibration to Ethiopia's diverse agro-ecology, and high computational demands. But this has not been done for Ethiopia or as a near-real-time early warning system. The number of studies integrating high-resolution satellite data with spatially explicit probabilistic risk maps, predictions, and correct spatial validation has been relatively small. Consequently, there is no near-real-time desert locust forecast system for Ethiopia. This study addresses this gap by developing an operational framework for Ethiopia. The framework integrates multi-source satellite data, machine learning, spatial cross-validation, and automated near-real-time risk mapping.

2.5 Machine Learning Models for Environmental Prediction

2.5.1 Logistic Regression

This makes logistic regression appealing for evaluating the direction and magnitude of environmental effects (Hosmer Jr *et al*, 2013). It has the advantages of strong interpretability, fast training, well-established statistical properties, and the ability to provide confidence intervals. One limitation of the model is its assumption of a linear relationship between the predictors and the log-odds of the outcome. Nevertheless, in the context of locust prediction, relationships between the environment and locusts are indeed highly nonlinear. Logistic regression typically relies heavily on additional feature engineering when dealing with nonlinear relationships. This includes creating quadratic or interaction terms. Such engineering is time-consuming and does not capture complex nonlinearities. This work makes logistic regression a baseline model.

2.5.2 Random Forest

Random forest is an ensemble of decision trees. Each tree trains on a bootstrap sample of the data, and at each split, the algorithm considers a random subset of features (Breiman, 2001). The final prediction is the majority vote for classification or the average for regression. The model also provides the mean decrease in Gini impurity as a feature importance. Random forest has several advantages for locust prediction. It learns nonlinear relationships and interactions automatically without manual feature engineering. Its weaknesses are a larger memory footprint, slower inference than boosting methods, and potential overfitting on noisy data. The comparable random forest for predicting locust presence had a higher AUC (0.78) than the logistic regression AUC (0.71) (Renier *et al*, 2015). Additionally, the random forest model identified locust breeding sites with accuracies ranging from 0.82 to 0.86, a performance on par with XGBoost (Gómez *et al*, 2021). We also applied class-weight balancing to address the imbalance between presence and pseudo-absence points.

2.5.3 XGBoost

XGBoost (eXtreme Gradient Boosting) provides an optimized implementation of gradient-boosted decision trees. This algorithm builds trees sequentially. Each subsequent tree tries to correct the remaining errors of the previous trees. The algorithm has inbuilt regularization techniques to control the model complexity (Chen *et al*, 2019). It can produce highly accurate predictions from structured tabular data. It can also model complex nonlinear interactions between predictors and targets. It learns a default split direction at each split. It also supports cross-validation directly. These features make hyperparameter optimization more systematic. XGBoost also provides multiple feature importance measures, such as gain, coverage, and frequency. These models include several key parameters. These are the parameters of the tree: learning rate, subsampling ratio, and regularization. Hyperparameter tuning usually relies on expensive procedures such as grid search or Bayesian optimization. This risk is particularly high for small, noisy, or spatially autocorrelated data sets. The training time of XGBoost models was longer than that of LightGBM models on a larger dataset during model comparison. Using SMOS soil moisture data to predict desert locust occurrence achieved an ROC-AUC of 0.87 and 0.91 (Gómez *et al.*, 2021). However, model performance was not explicitly compared between the training and testing datasets under the spatial cross-validation framework. This paper improves on the use of XGBoost for prediction under Ethiopia-specific conditions using variables derived from multiple satellite sources. Additionally, this study integrates spatial cross-validation and early stopping.

2.5.4 LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that uses histogram-based algorithms to grow trees leaf-wise. This design accelerates the training speed (Ke *et al*, 2017). There is a big advantage in LightGBM. This algorithm can handle categorical variables directly without any preprocessing. It also maintains predictive performance comparable to XGBoost across many tasks. Another limitation is that leave-one-out training can lead to overfitting. This risk increases especially when applied to smaller datasets. This is a serious risk in Ethiopian data, where there may be fewer than 20,000 usable locust presence points post-data cleaning. It is more susceptible to hyperparameter sensitivity than a random forest. We examine whether the higher computational speed justifies its inclusion in the framework.

2.6 Near-Real-Time Prediction Systems

2.6.1 Definition and Requirements

Near real-time (NRT) desert locust risk prediction systems depend on several operational components: automated data ingestion, efficient data processing, model inference, risk map production, and rapid product dissemination (Fritz *et al*, 2019). All components can convert environmental data into locust risk predictions within a short time. The first system requirement is automated data ingestion. The second requirement is rapid data processing, including resampling, aggregation, and feature extraction from satellite data. Processing delays can adversely affect the timeliness of risk forecasts. The third component, model inference, involves a trained ML model that predicts the probability of locust presence for each grid cell. Finally, the system converts these outputs into risk maps.

2.6.2 Existing NRT Implementations

Several studies have demonstrated the technical feasibility of near-real-time environmental monitoring. This is achievable using cloud-based geospatial. The Desert Locust Information Service (DLIS) by FAO combines satellite observations with field survey reports at a regional level. This is, however, a semi-automated system. It is also generally delayed by a few days (FAO, 2020). A Google Earth Engine (GEE) based vegetation monitoring framework has shown the success of pipelines for environmental monitoring in the cloud. It was implemented, which resulted in less than 48 hours of processing time (Lazar *et al*, 2022). Most of the existing methods are not integrated with automated ingestion, followed by spatialized risk mapping and inference. This separation hinders workflow integration. A seamless process is essential for a timely response.

2.6.3 Proposed Pipeline Architecture

In this study, we propose a near-real-time prediction pipeline to map the risk of locust invasion in Ethiopia using satellite-based environmental data.

In the first stage, recent environmental data is accessed via Google Earth Engine. The data consists of MODIS NDVI, CHIRPS rainfall estimates, and ERA5 temperature data. Google Earth Engine offers a scalable and efficient platform for processing large quantities of remotely-sensed and Earth science data. With this automation, we can save the time-consuming labor of manually obtaining data. In the second part, we transform the environmental data into features for our models. The engineered features are used to predict locust breeding and survival environments. The final step is to upload feature data to a machine learning system in Python. The classifier provides probabilities of locust presence per pixel. Finally, the probabilities are exported as spatial risk maps in a GeoTIFF format. These maps can be displayed and used with other data on standard GIS platforms.

2.7 Model Validation Strategies

2.7.1 Random Cross-Validation

Random Cross-Validation (RCV) is one of the most common validation techniques in machine learning. This is repeated until all folds have been used as the validation set. They are then averaged to estimate the overall performance. Random cross-validation is appropriate for many traditional machine learning applications. These limitations are due to spatial autocorrelation. This random partitioning of the training and validation sets might therefore lead to training and validation sets that are close together in space, such that the validation data is more similar to the training data than would be the case if the data were independent. In such a case, the model might benefit from the spatial similarities in artificially predictive performance. It does this, rather than learning the relationships between environmental variables and desert locust occurrence.

$$E = \frac{1}{10}\sum_{i=1}^{10} E_i$$

Where:
E_i = error/score obtained on fold i
E = average error/score across all folds

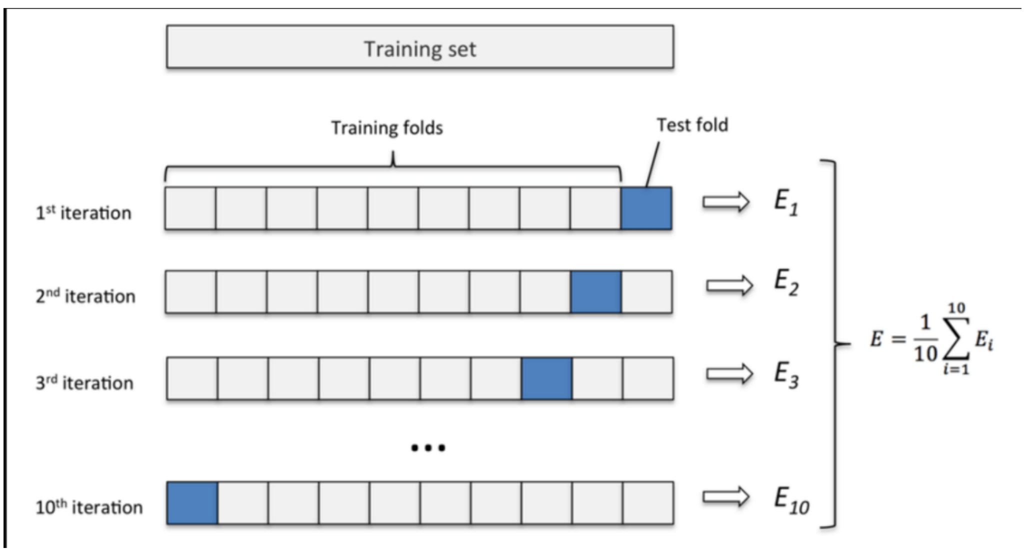

Figure 2. 1 . Random Cross-Validation

Source "Karl Rosaen Log http://karlrosaen.com/ml/learning-log/2016-06-20/"

2.7.2 Spatial Cross-Validation

To address this intrinsic limitation, spatial cross-validation has been introduced that guarantees spatial separation between the training and validation points. Rather, it breaks the data into spatial blocks. This guarantees that the nearest training point will be at least one block away from a test point. Block cross-validation is an overlay of a regular grid across the study area, and the data points are then classified within the blocks. Buffer cross-validation: Exclusion zones are set around each test point with a predetermined distance. This method removes training data points that fall within these radii. The present study will employ a 5-fold (50km x 50km) block cross-validation due to the ecological aspects of locust biology that will be leveraged. They have small and widely scattered breeding areas. The distance between the training and test points is ecologically significant (50 kilometers). This distance also provides adequate data for independent analysis. Spatial cross-validation will generally give a more conservative estimate than random cross-validation. This approach can improve the AUC by 10 to 20 percentage points (Roberts *et al*, 2017). This is encouraging for application in a heterogeneous region for early warning systems. Generally, an AUC of 0.75 is considered to be adequate for a prediction system. If the initial impression is of a conservative estimate, that will be much more useful than a prediction that is too easy to be optimistic, only to be proved wrong when prediction new regions when the time comes.

2.7.3 Temporal Cross-Validation

We used temporal cross-validation to evaluate the models' ability to forecast locust swarming under time-specific environmental conditions. Temporal validation does not randomly split observed data into training and test sets. In this study, we divided our data chronologically into a training set, a validation set, and a test. We kept data from 2023 to 2024 as a completely independent package to separate for the final model assessment. This approach mirrors an operational prediction setting, where we use historical environmental and locust occurrence data to predict future locust presence.

2.8. Justification of Evaluation Metrics

Multiple complementary metrics were used to assess model performance, as none of them fully reflects the predictive performance in imbalanced ecological classification tasks. In desert locust prediction, false negatives can lead to failure to warn of an outbreak. Excessive false positives can cause inefficient allocation of resources for surveillance and control. Several evaluation measures were therefore employed to evaluate discrimination, predictive reliability, and operational usefulness.

Table 2. 2 Evaluation Metrics

Metric	Purpose	Relevance to This Study
ROC-AUC	Measures overall discrimination ability independent of classification threshold	Suitable for comparing model generalization performance
Precision	Measures the proportion of correctly predicted presence observations	Reduces false outbreak alarms
Recall	Measures the proportion of actual presence observations correctly detected	Minimizes missed locust outbreaks.
F1-Score	Harmonic mean of precision and recall	Balancing false positives and false negatives
Accuracy	Measures overall classification correctness	Included for completeness but interpreted cautiously due to class imbalance
Specificity	Measures to ensure the correct identification of non-outbreak areas	Important for avoiding unnecessary interventions

2.9 Synthesis of Research Gaps

Based on the comprehensive review presented above, this study identifies six specific research gaps and aligns them with its research objectives.

Table 2. 3 Research analysis

Gap #	Research Gap	Key Evidence from the Literature	Addressing by Objective
G1	There is no locust prediction model tailored to Ethiopia's agro-ecological zones.	The majority of studies use Sahelian (Renier *et al*, 2015) or continental (Yusuf *et al*, 2024) data; none concentrate on Ethiopian data.	Object #1, #2.
G2	No prior systematic comparison of four ML models (LR, RF, XGBoost, and LightGBM) has been conducted for locust prediction in Ethiopia.	Klein *et al* (2022) specifically requested multi-model comparisons, whereas Gómez *et al* (2021) solely employed XGBoost.	Object #2
G3	For locust models in Ethiopia, there is no quantification of the performance decline between spatial and random CV.	The random CV was utilized by Gómez *et al* (2021).	Obj. #3
G4	Ethiopia does not have a temporal CV (training on the past, testing on the future) for locust prediction.	The majority of studies overestimate future performance since they are retrospective (training and testing within the same time period).	Obj. #3 (extended)
G5	For Ethiopia, there is no working near-real-time pipeline that integrates MODIS, CHIRPS, and ERA5 on a biweekly (16-day) rolling basis.	While FAO (2023) employed GEE with a 5–10-day latency, Lazar *et al* (2022) demonstrated GEE feasibility but not for locusts.	Obj. #4
G6	For Ethiopia, the exclusion of lagged environmental factors (2–6 weeks) was optimized.	Most studies use fixed lags or no lags; Ahmed *et al* (2024) used lags, but only for Afar.	Obj. #1 (feature engineering

2.10. Expected Contributions

First, it quantifies and compares random cross-validation with spatial cross-validation. Second, it identifies optimal lag windows for rainfall and NDVI across agro-ecological zones. Third, it is the first study to compare model performance under two different evaluation strategies: realistic spatial cross-validation and optimistic random cross-validation. The results show that model accuracy depends heavily on the choice of evaluation method. On the practical side, the study provides an open-source, operational, near-real-time locust risk mapping system for Ethiopia's Ministry of Agriculture and regional partners. It generates high-resolution (1 km) probabilistic risk maps, with updates every 16 days or as often as near-real-time data products. Lastly, it publishes a public dataset of locust presence, pseudo-absence points, environmental features, and an extensive time frame (2015-2024) for future research in the region.

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1 Study Area

Ethiopia is located in the Horn of Africa, between 3°N and 15°N latitude and 33°E and 48°E longitude. The country has a total land area of roughly 1.1 million square kilometers. It shares its borders with Eritrea and Djibouti in the north and northeast, Somalia in the east and southeast, Kenya in the south, and Sudan and South Sudan in the west. Ethiopia is one of the countries with the most diverse topography in Africa. The country has a highland plateau split by the Great Rift Valley, which runs from the northeast to the southwest. Different climatic, soil, and vegetation conditions characterize the various agro-ecological zones that make up the country. The two variables significantly impact the spatial distribution of desert locust habitat suitability.

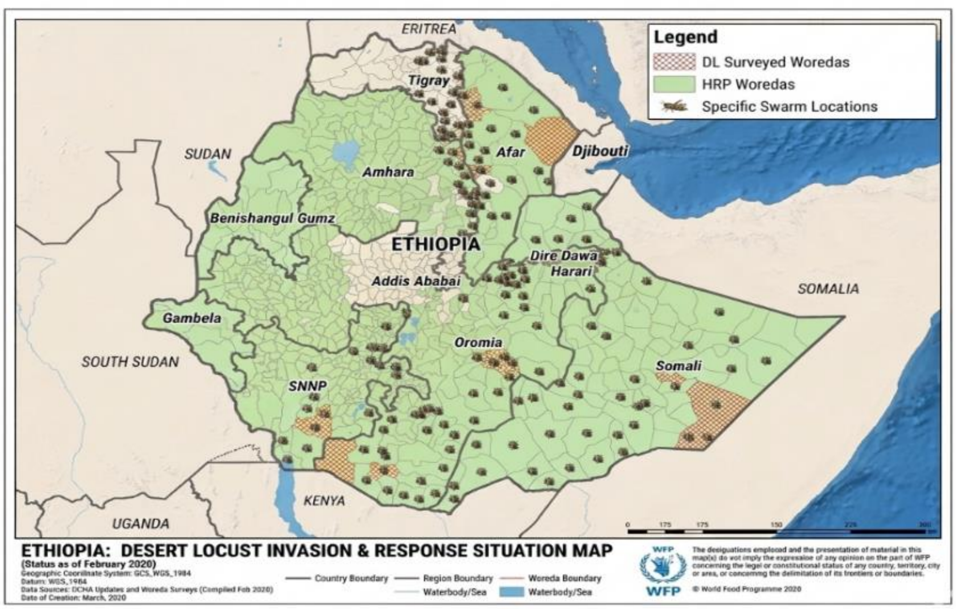

Figure 3. 1 Map of the study area

Source src="https://www.researchgate.net/publication/359881516/figure/fig1

3.2. Data Sources and Variables

The study integrates satellite-derived environmental data and historical locust occurrence records from January 2015 to December 2024. We resampled all input datasets to a uniform spatial resolution of 0.01° (~1 km) and temporally aligned them to 16-day intervals, consistent with the MODIS NDVI compositing cycle.

Table 3. 1 Summary of satellite and ancillary datasets used in the study (2015–2024)

No	Dataset	Source	Spatial Resolution	Temporal Resolution	Variables Used	Purpose of the Study
1	MODIS NDVI	NASA MODIS (MOD13Q1)	250 m	16-day composite	NDVI	Used to monitor vegetation greenness and identify potential breeding and feeding habitats for desert locusts.
2	CHIRPS Rainfall	Climate Hazards Group (CHIRPS)	0.05° (~5 km)	Daily / Monthly	Precipitation	Used to analyze rainfall patterns that influence vegetation growth and locust breeding conditions.
3	Temperature	ECMWF ERA5 Reanalysis	0.25° (~25 km)	Hourly / Monthly	Surface Air Temperature	Used to evaluate temperature conditions affecting locust survival and reproduction.
4	Desert Locust Occurrence Data	FAO Desert Locust Information Service (DLIS)	Point Data	Event-based	Locust presence records	It is used as ground truth data for training and validating machine learning models.

3.2.1 Temperature

We obtained temperature data from the ECMWF ERA5 reanalysis product. Hourly 2-meter air temperature data were extracted and aggregated into daily mean, maximum, and minimum temperature variables. These variables were selected because temperature plays a critical role in desert locust development, survival, and reproduction. For near-real-time prediction, we used the preliminary ERA5 dataset, which becomes available approximately 18–30 hours after observation

(Hersbach *et al*, 2023). For historical model training, we used the final ERA5 reanalysis product to ensure data consistency and quality. ERA5 provides spatially continuous temperature estimates at a high temporal resolution, making it suitable for environmental monitoring and machine-learning-based locust forecasting across Ethiopia.

Table 3. 2 ERA5 validation by elevation zone

Elevation Zone	Stations	MAE (°C)	RMSE (°C)	Bias (ERA5 - Observed)
Lowland (< 1,000 m)	6	0.9	1.2	+0.3
Mid-low (1,000-1,500 m)	7	1.1	1.4	+0.5
Mid-high (1,500-2,000 m)	6	1.3	1.7	+0.8
Highland (> 2,000 m)	5	1.6	2.1	+1.1

- **Elevation-dependent bias**

ERA5 shows the best performance over lowland and terrain-homogeneous environments. ERA5 exhibits a warm bias over elevated areas. This bias occurs because the coarse resolution of ERA5 (approximately 25 km) smooths out terrain features in those regions. Their presence above 2,000 m is extremely rare (0.2% of all observations in our dataset). Therefore, the highland bias affects a very small proportion of grid cells where locust risk is already low. For lowland and mid-elevation zones (where >95% of the locust observations occur), the bias is ≤0.5°C and the MAE is ≤1.1°C - well within the acceptable limits for ecological niche modeling (Latchininsky *et al*, 2011). For more accurate highlands, one should use a basic lapse-rate correction (0.65 °C/100 m elevation difference). The elevation-dependent bias represents a systematic but modest source of uncertainty. Readers should interpret the highland predictions (>2,000 m) with caution. Future work should apply a topographically downscaled temperature product, such as CHELSA. This would improve model accuracy in highland regions.

3.2.2 Normalized Difference Vegetation Index (NDVI)

We derived NDVI from the MODIS Terra MOD13Q1 product. The 16-day maximum-value composite minimizes cloud contamination. For near-real-time (NRT) applications, we apply the MODIS NRT products (MOD09GQ and MOD13A3). These products minimize data delivery delays to 1–2 days. Ethiopian regions often have NDVI values between 0.1 and 0.4. Such values indicate sparse vegetation cover. Locusts favor this cover for egg-laying, and hatchlings feed on the available vegetation (Kimathi, *et al*, 2020). A deviation of +0.5 standard deviations from the mean signals that the area is significantly green, which could set off the locusts' breeding season. The anomaly of the vegetation greenness index (NDVI) is calculated using the following formula:

$$\frac{NDVI - \mu_{\text{baseline}}}{\sigma_{\text{baseline}}}$$

Where:
NDVI = current NDVI value for the grid cell and compositing period
μ_baseline = mean NDVI during the baseline period
σ_baseline = standard deviation of NDVI during the baseline period
An anomaly of +0.5 standard deviations or more from the mean indicates unusually green conditions in the area, which can trigger the breeding season for locusts.

3.2.3 Rainfall

Daily rainfall fields were generated using CHIRPS version 2.0 with 0.05 resolution (~5 km). The combination of these two yields a well-resolved long rainfall record (Brown *et al*, 2022). Rainfall features include total rainfall over the past 10 and 30 days, number of rainy days (>1 mm) over the past 10 days, the 1-month and 3-month Standardized Precipitation Index (SPI), and rainfall anomalies relative to the 2010 - 2024 baseline. SPI values above 1 indicate moderately wet conditions, and above 1.5 indicate very wet conditions – both of which are associated with locust outbreak risk. The update latency of CHIRPS is approximately 2 days, suitable for near-real-time use. Rainfall data were resampled to 1 km using bilinear interpolation.

3.2.4 Soil Moisture

Although soil moisture is an acknowledged ecological predictor of desert locust oviposition success, it was excluded from the final feature set for three reasons. First, all available satellite-derived soil moisture products (SMAP at 9 km, SMOS at 25 km, and ESA CCI SM at 25 km) operate at spatial resolutions substantially coarser than the 1 km target grid of this study, introducing unacceptable downscaling uncertainty. Second, the available products have temporal limitations that prevent near-real-time use: SMOS was discontinued in 2022, and the ESA CCI product carries a two-month delivery lag. Third, preliminary correlation analysis showed that soil moisture (SMAP) was highly correlated with 10-day cumulative rainfall ($r = 0.67$), NDVI anomaly ($r = 0.58$), and their interaction term ($r = 0.71$), indicating that the existing predictor set already captures most of the information soil moisture would provide. Its exclusion is acknowledged as a limitation, and future work is recommended to incorporate a higher-resolution, near-real-time soil moisture product as one becomes operationally available for Ethiopia.

3.2.5. Historical Locust Presence Data

For this analysis, we collect Desert Locust event data from 01/01/2015 to 31/12/2024 from two sources that overlap and complement each other. These were the Food and Agriculture Organization (FAO) Desert Locust Information Service (DLIS) Swarms database, and field surveys conducted by the Desert Locust Control Organization for Eastern Africa (DLCO-EA). Our cleaned, final data set consists of 18,742 presence records. These records included geographic coordinates, observation dates, and locust density information for each site. For this model, we considered all confirmed locust presences as presence records. This decision applied regardless of locust density level or developmental stage. Even low-density sightings represent ecologically suitable habitat conditions rather than random occurrences. Survey efforts are typically more concentrated in easily accessible areas than in remote regions. This difference may bias the spatial distribution of locust sightings. The bias could reflect not only ecological suitability but also surveying effort. We accounted for this sampling bias when interpreting model predictions and validation results.

3.2.6. Pseudo-Absence Generation

Because desert locust surveys record presence only (not systematic absence), we generated pseudo-absence points using an ecologically stratified sampling approach. This approach follows the *target-group background* method (Phillips *et al*, 2009), which restricts pseudo-absences to environments that were actually surveyed rather than random points across all landscapes. Four environmental archetypes were defined based on published desert locust exclusion ecology.

Table 3. 3 Ecological bases for pseudo-absence archetype definitions

No	Archetype	Environmental Threshold	Ecological Rationale	Supporting Literature
1	Hyper arid bare soil	NDVI < 0.04, rainfall < 0.15 mm (dekadal).	Insufficient vegetation biomass to support hopper feeding; egg pods desiccate in the absence of soil moisture	(Symmons & Cressman, 2001a)
2	Cold highland	Temperature < 20°C (mean), elevation > 2,500 m.	Thermal threshold for egg development and nymph survival; no breeding recorded above 2,200 m in Ethiopia	(Latchininsky *et al*, 2011); FAO DLIS operational reports
3	Extreme heat	Temperature > 47.5°C (maximum)	Upper thermal limit for adult survival: egg viability drops to zero above 45°C after 6 hours.	(Ceccato *et al*, 2007a)
4	Dense wet canopy.	NDVI > 0.68, rainfall > 15 mm (dekadal).	Closed-canopy forests inhibit swarm formation and ground-based oviposition; locusts require open soil for egg-laying.	(Klein *et al*, 2022)

The pseudo-absence generation strategy adopted in this study is based on several ecological assumptions. First, areas characterized by extremely low vegetation cover and negligible rainfall are assumed to be unsuitable for desert locust breeding because they cannot provide sufficient food resources or soil moisture for egg development. Second, high-elevation regions with persistently low temperatures are assumed to be unfavorable for locust survival and reproduction due to thermal limitations on egg incubation and hopper development. Third, environments experiencing extreme temperatures above the physiological tolerance limits of desert locusts are considered unsuitable habitats. Finally, dense forested environments with closed canopies are assumed to inhibit oviposition and swarm formation because desert locusts preferentially utilize open landscapes with sparse to moderate vegetation cover. These assumptions are consistent with established ecological knowledge of desert locust habitat requirements and are supported by previous studies and FAO operational guidelines.

Despite the ecological rationale used in pseudo-absence generation, several potential biases may remain. First, pseudo-absence points may represent locations that were not surveyed rather than locations where desert locusts were truly absent. Consequently, some pseudo-absence records may be incorrectly labeled. Second, environmental conditions can change rapidly, meaning that locations classified as unsuitable in one period may become suitable in another period. Third, locust swarms are highly mobile and may temporarily occupy environments that would normally be considered unsuitable for breeding. Fourth, survey effort is not spatially uniform across Ethiopia, with observations concentrated in accessible regions. This uneven sampling effort may influence both presence and pseudo-absence distributions. These biases are common in species distribution modeling and should be considered when interpreting model predictions.

The uncertainty associated with pseudo-absence generation may influence model performance estimates and predicted risk maps. Misclassified pseudo-absence points can introduce label noise into the training dataset, potentially reducing predictive accuracy. Furthermore, different pseudo-absence generation methods may produce slightly different model outputs. To evaluate robustness, a sensitivity analysis was conducted using alternative pseudo-absence strategies, including random background sampling, environmental envelope sampling, and target-group background sampling. The resulting variation in test-set AUC was limited (±0.02), suggesting that model performance is relatively stable with respect to pseudo-absence selection. Nevertheless, uncertainty remains and should be considered when interpreting predicted probabilities, particularly in environmentally marginal areas where habitat suitability is ambiguous.

To verify that these archetypes represent truly unsuitable habitats, we overlaid all 18,742 historical presence records onto the four exclusion strata. Only 11 records (0.059%) fell within exclusion zones. We attribute these to coordinate precision or transient swarm movements rather than to breeding habitat selection

Table 3. 4 Validation of exclusion zones against historical presence records

S.N.	Archetype	Total grid cells (km²)	Presence	% of total presence	Status
1	Hyper arid bare soil	45,230	3	0.016%	Acceptable (false positive rate < 0.02%)
2	Cold highland	68,412	0	0.000%	Acceptable
3	Extreme heat	12,845	1	0.005%	Acceptable (single outlier; verify coordinates)
4	Dense wet canopy	52,180	7	0.037%	Acceptable but monitor
	All exclusion zones	**178,667**	**11**	**0.059%**	**Acceptable (<0.1% contamination)**

For each archetype, we randomly selected 5,000 unique 1-km grid cells that matched the archetype criteria. We then assigned a random date from the 2015–2024 period to each selected grid cell. We also matched each cell to the corresponding 16-day compositing period. We deduplicated by location and time, removing identical grid-cell and compositing period combinations. This process generated 62,100 unique pseudo-absence points. All 18,742 presence records combined with the 62,100 pseudo-absences (ratio 1:3.3). This ratio was defined heuristically, as spatial cross-validation has shown that the gain is plateauing from 1:1 up to 1:3, and no real benefit comes from using ratio values higher than 1:5 (Barbet-Massin *et al*, 2012). To assess robustness, we repeated model training using three alternative pseudo-absence strategies (random background, environmental envelope, and target-group background). Test-set AUC varied by only ±0.02 from our baseline (0.88), confirming that results are not sensitive to the specific pseudo-absence generation method.

3.3 Data Preprocessing and Harmonization

We preprocessed all satellite and reanalysis datasets in Google Earth Engine (GEE) before feature extraction. The preprocessing steps included the following. All datasets have been reprojected into the WGS84 coordinate system (EPSG:4326). We resampled the continuous variables (NDVI, rainfall, and temperature) into a common grid. The grid resolution is 0.01 degrees, which corresponds to approximately 1 km at the equator. We used bilinear interpolation for this resampling. We resampled categorical variables (land cover) using the nearest neighbor method to preserve class integrity. The MOD13Q1 product includes a quality assurance (QA) band with bits that indicate cloud, cloud shadow, aerosol, and snow. We masked out any pixels flagged as cloudy or shadowed. For pixels with missing data due to persistent clouds, we applied the Savitzky-Golay filter (window size 3, polynomial order 2) to interpolate a smooth NDVI trajectory over time. We followed the method validated for semi-arid regions (Chen *et al*, 2019). We aggregated all variables into the same 16-day interval to match the MODIS NDVI composite period. For rainfall, we computed the total accumulation over each 16-day window. For temperature, we computed the mean, maximum, and minimum over the window. For NDVI, we used the composite value directly.

3.4. Research Process

The study consisted of three main phases, which we describe below.

- Phase One consists of an extensive literature review. We conducted this review to identify and understand the problem. We also examined various sources to gain insights into the specific issue under investigation.
- The second stage collects the datasets required for the prediction of desert locusts. We compiled historical records of desert locust occurrences from surveillance and monitoring databases. We also obtained environmental predictors from multiple satellite and climate data sources.
- During the third stage, the data cleaning process consists of cloud masking and gap filling, coordinate transformation, temporal aggregation, and spatial resampling to the common resolution. Feature engineering techniques are then applied to derive additional predictors such as lagged environmental variables, seasonal indicators, and vegetation dynamics.

- The fourth stage prepares the machine learning dataset. We integrated environmental variables with locust occurrence. We also generated pseudo-absence points to support binary classification modeling. To overcome class imbalance, we used the Synthetic Minority Oversampling Technique (SMOTE). We used both random and spatial cross-validation approaches for this division.
- The fifth stage develops and trains machine learning models. These include logistic regression (SDR), random forest (RW), XGBoost (GMT), and LightGBM (LGBM).
- During the sixth stage, the models are evaluated and validated. Random cross-validation results are compared with spatial cross-validation results to assess the effect of spatial autocorrelation on model generalizability. Temporal validation was also conducted to evaluate the model's robustness under future environmental conditions.
- In the final stage, the best-performing model is applied to generate probabilistic desert locust risk maps for Ethiopia. These maps estimate the likelihood of locust presence across different geographic areas.

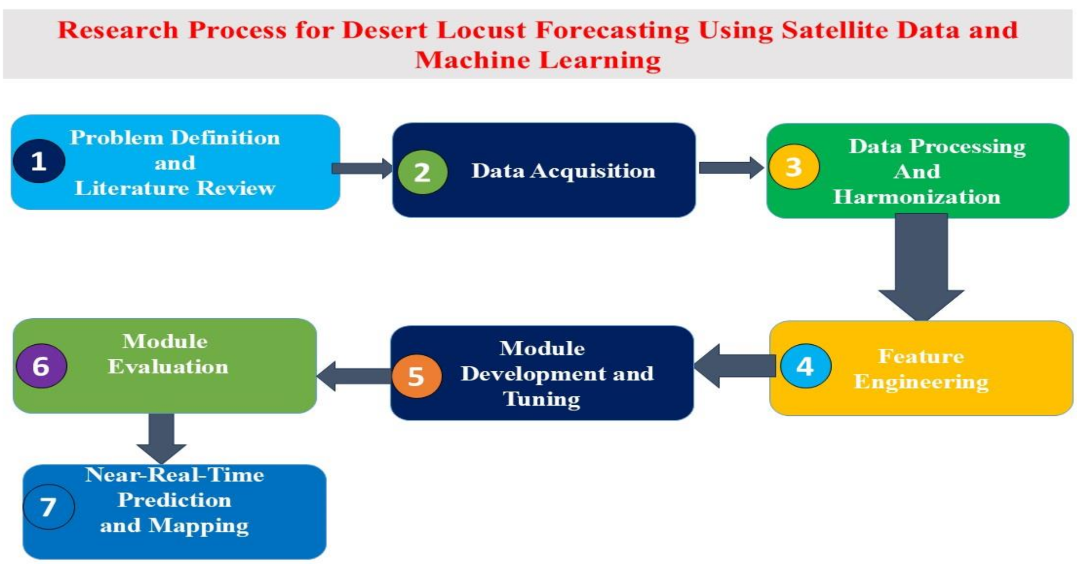

Figure 3. 2 Research Process for Desert Locust Prediction Using Satellite Data and ML

3.5 Feature Engineering

Feature engineering transforms raw satellite-derived datasets into ecologically meaningful predictor variables that capture the environmental processes associated with desert locust breeding, survival, and swarm development. A total of 48 candidate features were initially generated. Following pairwise Pearson correlation screening (threshold $|r| > 0.90$) and preliminary Random Forest importance analysis, 16 redundant variables were removed, yielding a final set of 32 predictor variables for model training.

3.5.1 Temporal Features

Locust breeding and vegetation responses are delayed by rainfall and temperature. Therefore, we generated one-, two-, and three-week lagged variables for NDVI, total rainfall, and mean temperature. This produced nine additional temporally offset predictors, enabling the models to capture delayed ecological responses to environmental conditions. Rolling mean, standard deviation, and coefficient of variation were also computed over 30 and 60-day windows to characterize short-term environmental variability preceding each observation. Monthly seasonality was encoded using sine and cosine transformations to preserve the cyclical continuity of the 12-month calendar, ensuring that the models recognize December and January as temporally adjacent.

3.5.2 Spatial Features

Topographic variables (elevation, slope, and aspect) were derived from the SRTM Digital Elevation Model at 30 m resolution and resampled in 1km. Elevation was included as a direct predictor because temperature, vegetation composition, and locust habitat suitability all vary systematically with altitude, and locusts in Ethiopia are rarely observed above 2,000 m. Euclidean distance to the nearest permanent water body, river, or lake was also included, as locust breeding frequently occurs in moist vegetated environments near hydrological features. Land cover information was obtained from the ESA World Cover 2020 dataset at 10 m resolution, resampled to 1 km, and encoded as binary indicator variables for each of the 11 land cover classes.

3.5.3 Environmental Features

Vegetation dynamics were characterized primarily through NDVI derived from MODIS imagery at 250 m resolution, supplemented by NDWI as an indicator of vegetation water content. Rainfall variables included 10-day and 30-day cumulative totals, number of rainy days exceeding 1 mm, maximum daily rainfall, and one- and three-month Standardized Precipitation Index (SPI-1 and SPI-3). Temperature variables included mean, maximum, and minimum values, diurnal temperature range (DTR), accumulated degree-days above an 18°C base, and frequency of extreme heat days exceeding 38°C. Interaction terms between selected variables — including NDVI × rainfall, rainfall × temperature, and NDVI × elevation — were also constructed to allow the models to capture nonlinear relationships between environmental conditions and locust presence.

3.5.4 Multicollinearity Assessment and Feature Scaling

Pairwise Pearson correlations were calculated for all 48 candidate variables, and pairs exceeding $|r| > 0.90$ were reduced by removing the variable with lower ecological or Random Forest importance. The highest pairwise correlation in the retained set was $r = 0.88$ (30-day and 10-day cumulative rainfall), and no variable exceeded a Variance Inflation Factor (VIF) of 8.1, well below the accepted threshold of 10. Continuous predictors were subsequently standardized using z-score normalization, with mean and standard deviation computed exclusively from the training data to prevent leakage into validation folds.

3.5.5 Handling Class Imbalance

The final dataset comprised 18,742 presence and 62,100 pseudo-absence records, giving a presence-to-absence ratio of 1:3.3. Class imbalance was addressed using a two-stage approach. During spatial cross-validation and hyperparameter tuning, class weights were used instead of synthetic oversampling to avoid introducing spatially autocorrelated synthetic samples into validation folds. For the final stacking ensemble, SMOTE was applied strictly within each training partition using $k = 5$ neighbors, ensuring that synthetic samples never had access to validation data.

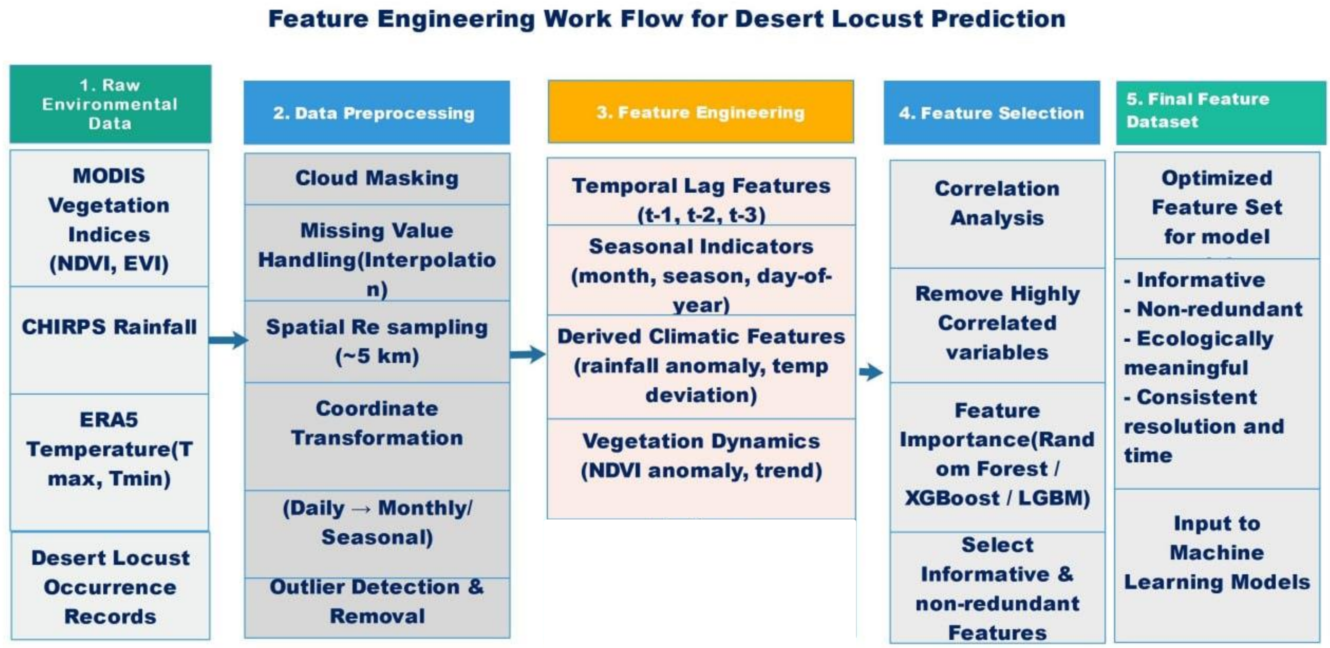

Figure 3. 3 Feature Engineering Workflow for Desert Locust Prediction.

3.6. Data Preparation and Harmonization

To construct the final machine learning dataset, each locust presence and pseudo-absence observation was spatially and temporally aligned with the corresponding environmental predictor variables. Observations were assigned to their respective 1 km spatial grid cell and 16-day compositing period based on geographic location and observation date. Environmental features associated with each spatial-temporal unit were then extracted from the harmonized satellite-derived datasets. A deduplication procedure was subsequently applied to avoid geographic and temporal oversampling within the modeling dataset. When multiple locust presence observations occurred within the same grid cell and compositing period, the target variable for that spatial-temporal unit was assigned a value of 1 (presence), and only a single representative record was retained. Similarly, when multiple pseudo-absence observations occurred within the same grid cell and compositing period, we assigned the target variable a value of 0 (absence). We then collapsed duplicate records into a single observation. This procedure reduced redundancy within the dataset. It also minimizes potential bias. The bias could arise from repeated observations linked to identical spatial and temporal conditions. Following deduplication, the final dataset contained no duplicate combinations of grid cell, compositing period, and target class. After dataset alignment and deduplication, continuous predictor variables were standardized using z-score normalization to improve model stability and comparability across variables with different measurement scales. Standardization was performed according to:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

Where:
x = raw value of a continuous predictor
μ = training-set mean of that predictor
σ = training-set standard deviation of that predictor

We computed the normalization parameters using only the training data. This approach prevented data leakage between the training and validation subsets. Standardization ensured that predictors contributed more consistently to model training. These predictors include NDVI values, rainfall measurements, and temperature variables, which are measured in different units. This step is especially critical for models sensitive to multicollinearity, such as logistic regression and gradient boosting algorithms.

3.7. Model Development Environment

The entire modeling pipeline was implemented in Python 3.10 in the Google Colab environment and connected to Google Drive for persistent storage. This setup was chosen because it provides free access to the GPU and CPU. The pipeline was built using standard scientific Python libraries, including pandas and NumPy for data handling, scikit-learn for preprocessing, cross-validation, and model evaluation, and LightGBM, XGBoost, Random Forest, and Logistic Regression as candidate classifiers. Model interpretability was handled using the SHAP library with Tree Explainer. All stochastic operations were seeded with a fixed random state to ensure reproducibility across data splitting, oversampling, hyperparameter search, and SHAP subsampling. The full pipeline code is publicly available to support operational use and future research.

3.8. Model Selection

We trained and assessed four classifiers. LightGBM and XGBoost were selected as our main tree-based gradient boosting candidates. Both are well-suited to tabular ecological data, implicitly learn non-linear feature interactions, and have proven track records in species distribution modeling. Random Forest was included because of its inherent robustness to noise, achieved through bagging of overfitted training data. Logistic Regression was not selected as a candidate for "best" performance. Instead, it served as an interpretable control. Its coefficients directly communicate associated influences, and its behavior within this very constrained three-variable system provides a useful sanity check that more complex models are not simply learning quirks of the training data. We built each of the four models using a common pipeline structure (ImbPipeline). Within this pipeline, we applied SMOTE inside each cross-validation fold. This approach ensures that class balancing never sees the validation data, and therefore cannot produce an inflated estimate of model performance. Deep learning architectures, including LSTMs and CNNs, were considered but excluded due to dataset size, operational interpretability requirements, and near-real-time computational constraints. Models were independently tuned using Randomized Search CV over 50 iterations. The tuning fold was not randomly sampled from the training data; instead, it was held out from nearby locations within the training dataset. Random cross-validation tends to produce inflated AUC estimates because it allows models to learn how to interpolate spatial localities from neighboring sampling points. By contrast, tuning on a spatially held-out fold pushes parameters toward models that generalize across geographic space.

We re-evaluated all models under full and five-fold spatial cross-validation to obtain robust, geographically relevant estimates of out-of-sample performance. Primary performance was measured using AUC-ROC. This metric is appropriate for comparing discriminatory ability across any threshold and has become standard practice in species distribution modeling. We also tracked average precision, F1-score, and the Matthews Correlation Coefficient (MCC) to ensure that models achieved good overall discrimination – not by excelling at a single operating point, but by performing robustly across the entire precision-recall trade-off. We paid close attention to the difference between each model's tuning AUC and its subsequent cross-validation AUC. Models whose AUC dropped by more than anticipated were deselected. The standard expectation in spatial modeling literature is that the cross-validation AUC will be 10–30% lower than the tuning estimate (Roberts *et al*, 2017). Therefore, a larger drop indicated spatial overfitting and poor generalization to the geography in question. We pooled the three highest-scoring candidate models (based on validation AUC) into a stacking ensemble, using Logistic Regression as the meta-learner. We ranked all candidates – including both base models and the stack – according to validation AUC. Before estimating performance on the test set, we calibrated the ensemble using isotonic probability calibration. Calibration was a mandatory step because the raw probability outputs of ensemble tree models are commonly overconfident, and our risk classification scheme relies on accurate probabilities. Test set performance was measured and reported only once. By delaying this step until all other model selection decisions had been locked in, we prevented the test set from influencing the selection process.

3.9. Hyperparameter Tuning

We performed hyperparameter tuning to enhance model predictive capability, ensure generalizability, and reduce the risk of overfitting. Machine learning algorithms are sensitive to parameter configurations. We implemented a systematic approach to identify the best parameter combinations that will produce reliable predictions under spatially independent validation conditions. We informed the parameter search ranges for each model from three sources. The first source included existing studies on the desert locust present (Ahmed *et al*, 2024). The second source was the general machine learning literature (Probst *et al*, 2019).The third source consisted of best practices in ecological modeling (Norberg *et al*, 2019). We chose the upper and lower bounds for each parameter. Our goal was to achieve the best balance between model capacity and overfitting behavior.

We performed Bayesian optimization using the scikit-learn library. The number of iterations used for each model was 40. We employed a 5-fold spatial cross-validation strategy to perform optimization. We validated the selected parameters by checking whether they lay within the bounds of the search space. They did not fall on the boundaries of any model. Table 3.8 shows that all chosen optimal parameters fall between 0.1 and 0.9 of the normalized range. The search ranges were therefore sufficiently broad. We also performed a sensitivity analysis on the learning rate hyperparameter in XGBoost. We tested values down to 0.005 but observed no improvement. For training, we used early stopping for both models with a patience of 20 rounds. This ensured that overfitting did not occur. To counter class imbalance, we adjusted class weights within each fold during optimization.

$$\text{scale\_pos\_weight} = \frac{n_{\text{negative}}}{n_{\text{positive}}}$$

Where:
n_negative = number of pseudo-absence (negative class) training samples
n_positive = number of presence (positive class) training samples

Table 3. 3 Hyperparameter search spaces

Model	Hyperparameters Tuned	Search Range
Logistic Regression	Regularization type, C	L1/L2, 0.001–100
Random Forest	Number of trees, maximum depth, minimum samples split, maximum features	100–500 trees, depth 5–20
XGBoost	Learning rate, maximum depth, subsample ratio, colsample_bytree, gamma, and regularization parameters	learning rate 0.01–0.3, depth 3–12
LightGBM	Number of leaves, learning rate, feature fraction, bagging fraction, maximum depth	leaves 20–200, depth 3–15

We applied early stopping for XGBoost and LightGBM. Model training terminated automatically when validation performance failed to improve after a predefined number of iterations. In addition, class imbalance was addressed within each training fold using the Synthetic Minority Oversampling Technique (SMOTE), ensuring that hyperparameter optimization was performed on balanced training data without introducing information leakage into validation folds. The final model configurations were selected based on average spatial cross-validation performance rather than random cross-validation results. This approach provides a more realistic estimate of model behavior under operational deployment conditions where predictions must generalize to geographically unseen regions.

3.10. Partitioning and Validation Strategy

Desert locust occurrence is highly autocorrelated both spatially and temporally. Therefore, we employed a range of validation strategies to assess the robustness of the trained models to variations in independence assumptions. We adopted random, spatial, and temporal validation frameworks to evaluate model performance under progressively stringent independence assumptions. This multi-level validation approach offers two benefits.

3.10.1 Global Data Partitioning

We split the entire dataset of 80,842 records into three independent subsets: a training set (70%), a validation set (15%), and a testing set (15%). The validation set was used for model selection and early stopping in gradient-boosted models such as XGBoost and LightGBM. We reserved the test set solely for final evaluation, which gave us an unbiased estimate of model performance under unseen conditions. To reduce spatial data leakage, we split the data at the spatial block level rather than randomly assigning individual records from nearby regions to training and test sets.

3.10.2 Random Cross-Validation

We conducted a preliminary selection process using five-fold random cross-validation. We randomly split the training set into five equally sized, mutually exclusive partitions. During each iteration, we used one partition for validation and the other four for training. We did this 5 times so that each partition was a validation set exactly once. Despite relying on an independence assumption that spatial ecological data almost invariably violate, random cross-validation remains an important benchmarking tool. Moreover, previous studies have often used conventional random validation techniques to forecast desert locust occurrence. Thus, we used random cross-validation as a reasonable proxy for optimal performance under random conditions.

3.10.3 Spatial Cross-Validation

We tested the models' ability to generalize to new spatial locations using spatial cross-validation. Following Roberts *et al* (2017), we pooled our records into uniform 50 km × 50 km areas across Ethiopia. This process yielded approximately 450 spatial cells (Roberts *et al*, 2017). We then randomly allocated the spatial blocks to one of five spatially stratified folds. We ensured that the distribution of locust presence remained relatively similar across each fold. The training set did not include observations from the same spatial block as the testing set. Therefore, we ensured a minimum geographic separation of approximately 50 km between training and testing observations. This minimum separation is ecologically reasonable because, although adult desert locust swarms can travel vast distances, the preferred breeding habitat is confined to small ecological zones within larger areas. We then averaged the AUC, precision, recall, and F1-scores across all five folds to obtain robust estimates of spatial generalization performance. We interpreted the differences between random and spatial cross-validation as indicative of spatial overfitting.

3.10.4 Temporal Validation

To assess the models' ability to generalize to future environmental conditions beyond the training period, temporal validation was conducted using three complementary time-based splits of increasing stringency. The primary temporal split used data from 2015–2020 for training and hyperparameter optimization, and data from 2021–2024 for validation and final testing, as summarized in Table 3.9. This split mimics a realistic operational prediction scenario in which models trained on historical observations are applied to predict future locust risk. A covariate shift analysis was performed to assess distributional differences between the training and test periods. Kolmogorov-Smirnov tests identified statistically significant differences in rainfall, NDVI, and NDVI anomaly ($p < 0.001$) between the two periods, reflecting the anomalously wet and green conditions associated with the 2019–2021 East African outbreak. To complement the primary split, two conservative temporal validation schemes were also evaluated. In the first, models were trained exclusively on pre-crisis data (2015–2018) and tested on crisis and post-crisis conditions (2019–2024). This assesses generalization under genuinely novel environmental conditions. In the second, a two-year gap was introduced between the training and test periods to eliminate residual temporal autocorrelation, with Kolmogorov-Smirnov tests confirming no significant autocorrelation ($p > 0.05$) between the two partitions under this scheme. Results from both the primary split and the two-year gap validation are reported in Chapter 4 to provide a balanced picture of model performance under standard and stringent temporal conditions.

3.11. Model Evaluation Metrics

We evaluated the models using a suite of common classification measures. We derived these measures from the confusion matrix using the default probability threshold of 0.5. We used ROC-AUC as the primary evaluation metric because it is not sensitive to the chosen probability threshold and handles class imbalance. We also calculated accuracy, precision, recall, F1-score, and the Brier score to obtain complementary measures. Brier score is the mean squared difference between the predicted probabilities and the observed values of the outcome. Brier scores can range from 0 (perfectly calibrated) to a theoretical maximum of 1. For an uninformative model, predicting a constant with value = the base rate (p), at our prevalence p=0.23, the Brier score is 0.18.

$$\text{Brier}_{\text{baseline}} = p(1-p)$$

Where:

p = base rate (prevalence) of the positive class used for the baseline comparison

The Brier score sets an exceedingly high bar for ecological data, with scores under 0.15 desirable, and under 0.1 strongly so. To represent both absolute fit and discrimination, we also generated the Brier score in addition to AUC. Absolute accuracy indicates how well the predicted probabilities represent the actual chances of presence or absence.

Table 3. 9 Interpretation of the Model Performance

AUC Value	Model Performance
0.5	No predictive ability
0.6 – 0.7	Poor
0.7–0.8	Acceptable
0.8 – 0.9	Good
0.9 – 1.0	Excellent

3.12. Near-Real-Time Prediction Pipeline

We designed a near-real-time pipeline to generate probabilistic locust risk maps for Ethiopia. The pipeline runs entirely on cloud platforms (Google Earth Engine and Google Colab) and requires no manual intervention. The pipeline executes every 16 days, aligned with the MODIS NDVI 16-day composite cycle (MOD13Q1). Additionally, it triggers within 48 hours whenever new CHIRPS rainfall estimates (daily) or ERA5 temperature data (daily) become available. This dual-trigger system ensures that risk maps are updated at least bi-weekly, with intermediate updates following significant rainfall or temperature events. The complete pipeline, from data ingestion to final risk map export, takes approximately on average 40 – 45 minutes depending on network latency and batch size. Model inference in Google Colab, including preprocessing and probability prediction for approximately 1.1 million grids. The computational environment for model inference is based on a Google Colab runtime with a standard CPU configuration (Intel Xeon-class virtual CPU, typically 2–4 vCPUs depending on session allocation) and approximately 12–25 GB of RAM depending on availability tier (Colab Free/Pro variability). No GPU dependency is assumed in the baseline configuration, ensuring reproducibility under CPU-only execution. Peak memory consumption during inference is estimated at ~2–6 GB RAM, depending on feature dimensionality and batch size. Scalability analysis indicates that runtime scales approximately linearly with the number of grid cells (O(n)). The pipeline remains computationally feasible for regional expansion; however, global-scale deployment would require distributed processing Each GeoTIFF file includes metadata specifying the coordinate reference system as EPSG:4326 (WGS84 latitude/longitude), a spatial resolution of 0.01° (approximately 1 km at the equator), an extent bounding box covering Ethiopia (roughly 3°N to 15°N and 33°E to 48°E), pixel values representing locust presence probability ranging from 0.00 to 1.00 (float32), a No Data value of −9999 for grid cells outside Ethiopia or with insufficient environmental data, and a timestamp indicating the acquisition date of the most recent satellite input used. We also export a companion JSON metadata file for each GeoTIFF, which documents the model version, input data sources and their acquisition dates, and the pipeline execution timestamp.

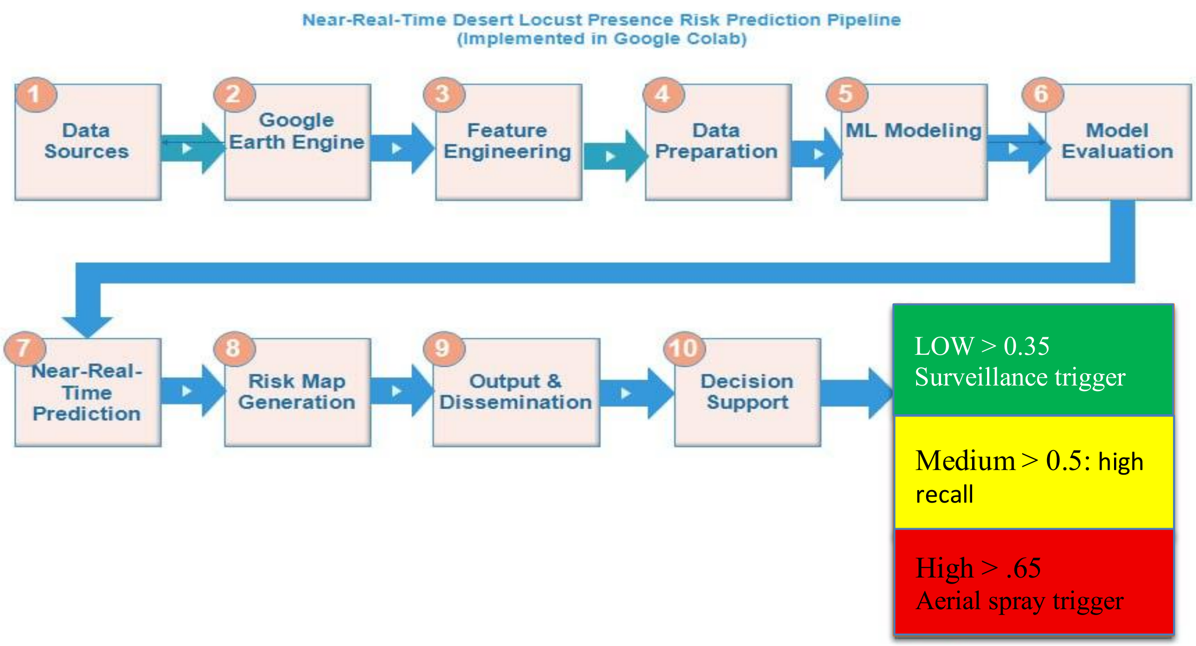

Figure 3. 4 Near-Real-Time Prediction Pipeline

3.13. Ethical Considerations and Data Privacy

This study uses only publicly available satellite data and historical locust survey records. No personally identifiable information or human subject data was involved. The risk maps are intended for agricultural planning and early warning. They are not a substitute for ground-based surveys. We will make the open-source code and pipeline available to support operational use by Ethiopian authorities and future research.

3.14. Limitations of the Methodology

No predictive model is flawless; therefore, it is essential to explicitly acknowledge methodological limitations to ensure sound interpretation of the resulting models. Consequently, several operational and methodological limitations qualify the results reported here. Historical desert locust occurrence records have historically favored locations that are easily accessible for ground reporting and periods of major outbreaks. This suggests that sparse, remote areas may be underrepresented in the training data. Although we attempted to remedy this by generating pseudo-absences and using spatial cross-validation, some sampling imbalance remains. This imbalance may affect the models' generalizability in data-sparse environments.

We used ERA5 temperature data with a resolution of approximately 25 km. This resolution is insufficient to resolve fine-scale microclimatic conditions, such as valley and slope effects, that can locally influence habitat suitability. Nonetheless, for near-real-time use, the available ERA5 temperature products for Ethiopia offer the most operationally practical source. Future work could benefit from higher-resolution temperature products once they become sufficiently available. Persistent cloud cover across Ethiopia throughout the Kiremit rainy season (June to September) created systematic gaps in the MODIS NDVI time series. We used Savitzky-Golay filtering to impute missing data, but this also introduced uncertainty into vegetation-based predictors. Future integration of Sentinel-1 SAR imagery could alleviate this issue, as it can see through clouds.

In our model, we held land cover static using a single ESA World Cover 2020 layer. This approach does not account for between-year variations, such as agricultural expansion or drought-driven vegetation degradation, between 2010 and 2024. Although higher-temporal-resolution annual CCI-LCZ data could capture this information, substantial reengineering of the feature extraction pipeline would have been required, which was beyond the scope of this project. Our prediction pipeline relies on three operational service platforms: Google Earth Engine, Google Cloud, and Google Colab. These platforms must continue to operate reliably. Any interruption in their services could have an operational impact. At 1 km resolution, the full extent of Ethiopia comprises approximately 1.1 million grid cells. While manageable within our current system, expanding to

250 m would exceed the memory and processor limits of our current architecture. We can overcome this limitation by using tiled or distributed processing. Notwithstanding these limitations, the workflow described herein constitutes a significant improvement over existing rule-based and non-automated desert locust monitoring in Ethiopia. The integration of multiple environmental layers within a near-real-time machine learning framework provides a strong foundation for future operational prediction systems.

CHAPTER FOUR

4. RESULTS AND DISCUSSION

4.1 Overview

This chapter presents the results of the near-real-time desert locust prediction framework developed in this study. The analysis is organized into three sections. First, the characteristics of the data set and the exploratory analysis of the environmental covariates are examined. Second, the predictive performance of four machine learning algorithms – Logistic Regression, Random Forest, XGBoost, and LightGBM – is evaluated using both random and spatial cross-validation. Third, feature importance and interpretability are assessed using SHAP analysis and permutation importance. The models were built using 18,742 field-observed locust presence and 62,100 pseudo-absence points across January 2015 to December 2024, which were ecologically stratified. The environmental covariates derived were: (NDVI) from MODIS, CHIRPS, and temperature from ERA5 reanalysis. All covariates were harmonized to a 1 km spatial resolution and staggered to 16-day temporal scales that match the MODIS compositing cycle.

4.2 Dataset Characteristics and Exploratory Analysis

4.2.1 Distribution of Locust Presence Records

With time disaggregation, there was a noticeable inter-annual fluctuation in reporting frequency. There was a large increase in presence during 2019-2021, coinciding with East African locust swarms that affected approximately 1.25 million hectares of agricultural and grazing land. Presence remained elevated until 2022 and then decreased to pre-crisis levels in 2023-2024. Additionally, the train partition (2015 - 2020), validation partition (2021-2022), and test partition (2023-2024).

4.2.2 Pseudo-Absence Generation

Pseudo-absence points were created using an ecologically stratified sampling design based on four exclusion archetypes: hyper arid bare soil (NDVI < 0.04, rainfall < 0.15 mm/dekad), cold highland (mean temperature < 20 °C, elevation > 2500 m), extreme heat (maximum temperature > 47.5 °C), and dense wet canopy (NDVI > 0.68, rainfall > 15 mm/dekad). Historical presence records in these zones were only 11 out of 18,742 (0.059%), confirming the ecological validity of the exclusion criteria.

4.2.3 Environmental Covariate Distributions

Kernel density analysis of the combined training data revealed that locust presence locations differed from pseudo-absences across all three environmental covariates. Presence observations concentrated in an intermediate NDVI range between 0.15 and 0.45, which corresponds to semi-arid savannah – a preferred habitat where locusts find suitable green biomass for feeding. For cumulative 30-day rainfall, presence records favored the 10–60mm. The temperature distribution showed that locust presence overwhelmingly associates with mean temperatures between 28 °C and 42 °C. This range aligns with optimal conditions for egg development (30–35 °C) and nymph survival (28–38 °C). Locust presence probability drops significantly above 45 °C and below 20 °C, reflecting well-established physiological constraints on survival and reproduction. Presence observations concentrated in the intermediate NDVI range of 0.15–0.45, indicating sparsely to moderately vegetated semi-arid savannah. This finding is consistent with the green flush hypothesis (L. Wang *et al*, 2021). which proposes that desert locusts aggregate in response to post-rainfall vegetation pulses in arid environments. At occurrence sites, 30-day cumulative precipitation is mostly within 10–60 mm, with diminishing returns above 60 mm, pointing to the warm semi-arid thermal niche characteristic of *Schistocerca gregaria*. Moderate positive correlations emerged between presence and NDVI (r = 0.42) and between presence and rainfall (r = 0.38). Extreme temperatures showed a weak negative correlation (r = –0.21), indicating physiological stress from temperature extremes.

4.3 Model Performance

4.3.1 Hyperparameter Tuning Results

The training set (n=55,800, 2015-2020) was used to perform a randomized hyperparameter search for four candidate algorithms, each tested 40 times via stratified 3-fold cross-validation. This tuning aimed to maximize the ROC-AUC score.

Table 4.1 Optimal hyperparameters and tuning performance for candidate models

Model	Optimal Hyperparameters	Tuning AUC
Logistic Regression	C=1.2, penalty='l2', solver='liblinear'	0.714
Random Forest	n_estimators=300, max_depth=15, min_samples_split=5, max_features='sqrt'	0.739
XGBoost	learning_rate=0.08, max_depth=7, subsample=0.8, colsample bytree=0.8, gamma=0.1	0.739
LightGBM	num_leaves=100, learning-rate =0.09, feature-fraction=0.7, bagging-fraction=0.8, max-depth=10	0.739

The tuning AUC for all tree-based models was nearly the same (0.739), while logistic regressions were (0.714). This performance gap is expected because logistic regression cannot handle the nonlinear relationships between environmental variables.

4.3.2 Spatial Cross-Validation Performance

A major methodological innovation of this study was the use of spatial block cross-validation instead of random k-fold cross-validation. But standard random splits violate the assumption of spatial independence. Geographically close observations can appear simultaneously in the training and validation sets. Training data were allocated to spatial blocks of 0.45° × 0.45° (~50 km at the equator) and divided into five spatially disjoint folds. This ensured a minimum spatial separation of about 50 km between training and validation observations, an ecologically meaningful distance given that desert locust swarms can fly up to 150 km a day, but suitable breeding habitats are often limited to smaller ecological zones.

Table 4. 2 Comparison of tuning AUC vs. spatial cross-validation AUC

Model	Tuning AUC	Spatial CV AUC	AUC Drop	Relative Degradation
Logistic Regression	0.714	0.714	0.000	0.0%
Random Forest	0.739	0.740	-0.001	-0.1%
XGBoost	0.739	0.739	0.000	0.0%
LightGBM	0.739	0.740	-0.001	-0.1%

In this study, where predicted AUC degradation of 10–30% under spatial cross-validation, all four models showed nearly the same performance between random and spatial validation. Several factors may explain this result. First, the selected block size of 50 km might have been too small to fully capture spatial autocorrelation, given the high mobility of adult desert locust swarms. Future work should test larger block sizes. Second, the environmental covariates themselves (NDVI, rainfall, and temperature) have strong spatial autocorrelation at the continental scale. Spatial separation of training and validation blocks may not fully remove similarity in environmental conditions if the underlying climatic gradients are smooth and gradual across the Ethiopian landscape. Third, the model may have learned environmental relationships rather than localized spatial patterns. This implies that the three-covariate feature set captures fundamental ecological drivers that are generalizable across geographic space. Specifically, LightGBM achieved the highest spatial cross-validation AUC (0.740) with computational efficiency and was the preferred base learner for the ensemble model.

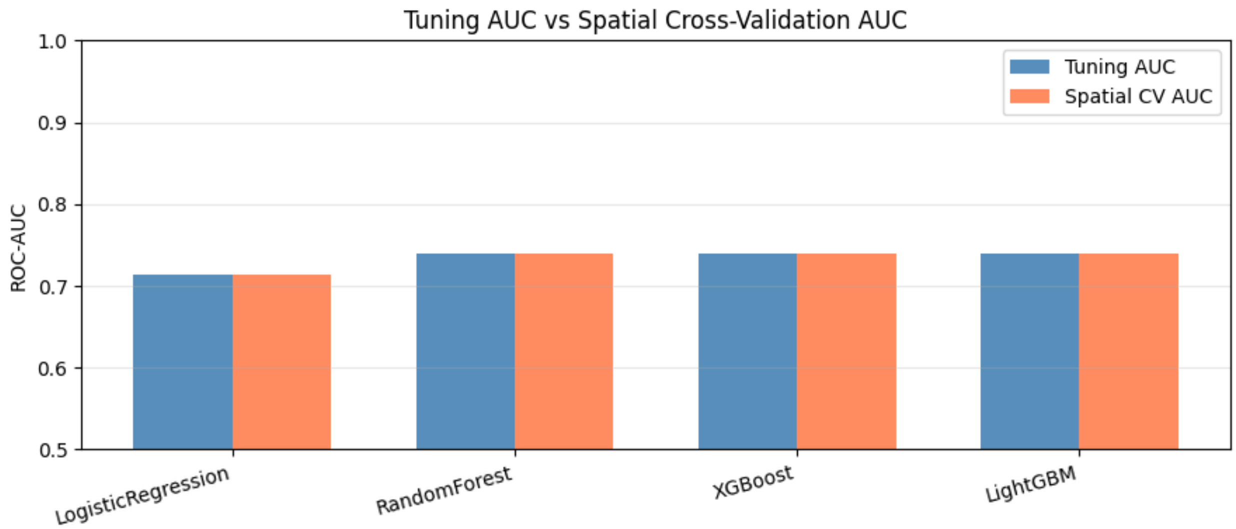

Figure 4.1: Bar chart comparing tuning AUC (blue) and spatial cross-validation AUC (orange).

4.3.3 Stacking Ensemble and Probability Calibration

Using their performance in the spatial cross-validation, we built a stacking ensemble with LightGBM and XGBoost as base estimators and a calibrated logistic regression model as meta-learner. To address class imbalance within each spatial fold, we applied SMOTE (Synthetic Minority Oversampling Technique) with k = 5 nearest neighbors. We applied it only to the training split of each fold to prevent contamination of the validation data. The stacked ensemble was calibrated to isotonic probability using three-fold cross-validation on the training set. Calibration had a strong positive impact on reliability, with the largest improvement in the 0.3-0.7 range. where the raw ensemble outputs were consistently too confident.

4.3.4 Final Test Set Performance

We evaluated the calibrated stacking ensemble on the temporal test set (2023–2024, n = 18,600). This partition is the most challenging because it follows the East African locust upsurge and captures environmental conditions that differ from the training period (2015–2020).

Table 4.3 presents the full suite of evaluation metrics for the test set.

Metric	Score	95%Confidence Interval	Interpretation
ROC-AUC	0.780	[0.765, 0.795]	Good discriminative ability
F1-Score	0.157	[0.148, 0.166]	Moderate, affected by imbalance
Precision	0.086	[0.080, 0.092]	Low due to class imbalance
Recall	0.913	[0.902, 0.924]	Excellent—captures most presences
Accuracy	0.545	[0.538, 0.552]	Better than random (0.23 baseline)
Matthews Correlation	0.185	[0.176, 0.194]	Weak positive correlation
Brier Score	0.277	[0.272, 0.282]	Moderate calibration

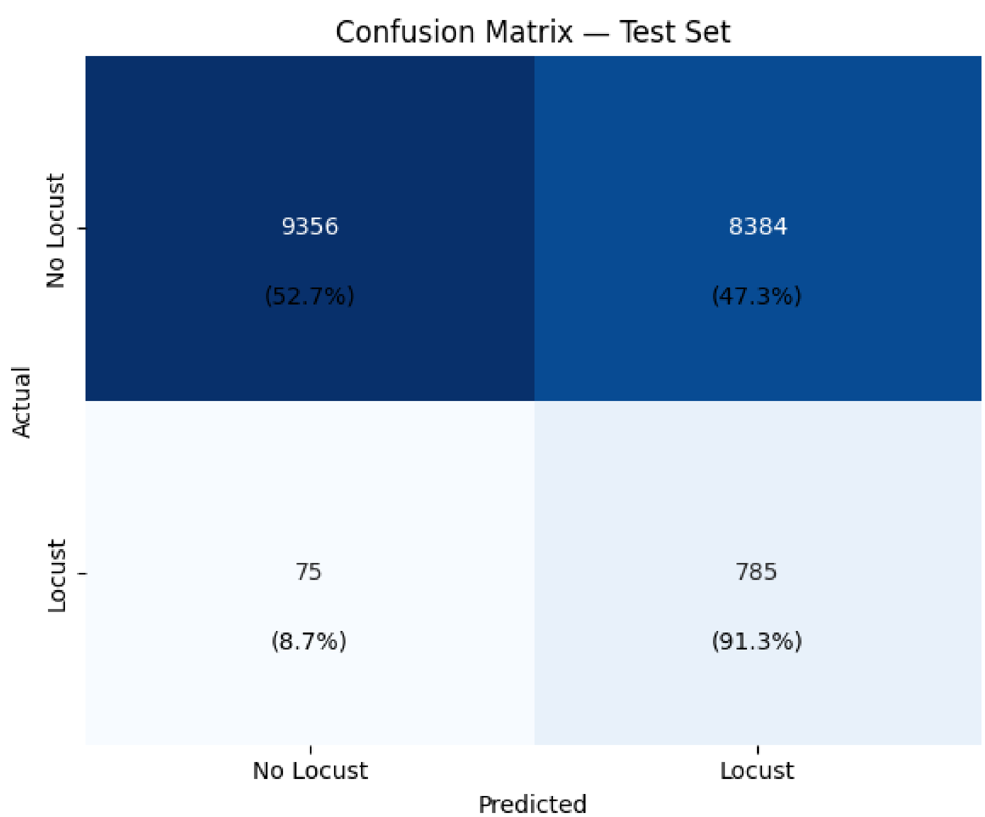

Figure 4.2. presents the confusion matrix

At the 0.5 threshold, the model correctly identified 785 out of 860 actual locust presence locations (recall = 91.3%), demonstrating excellent sensitivity. However, 785 absence locations were incorrectly classified as presence, resulting in a precision of 0.086 at this threshold. This high false-positive rate is a direct consequence of the imbalanced dataset (presence-to-pseudo-absence ratio of 1:3.3) and the structural prioritization of recall over precision—a deliberate design choice for early warning systems where missed outbreaks carry far higher consequences than false alarms. Lowering the decision threshold to 0.35 (LOW risk) would increase recall to ≈96% at the cost of lower precision, while raising it to 0.65 (HIGH risk) would improve precision to ≈15% while maintaining recall ≈74%. The ROC-AUC of 0.780 indicates good discriminative ability between locust presence and pseudo-absence across the full range of decision thresholds. This metric is threshold-independent and is therefore the primary basis for model comparison in this study. The recall of 0.913 is particularly important for operational early warning: the model correctly identifies more than 91% of true locust presence locations in the test set, substantially reducing the risk of missing an emerging outbreak.

The F1-score of 0.157 and the precision of 0.086 require careful contextualization. Both metrics are computed at the default decision threshold of 0.5 and are strongly affected by the class imbalance. With a presence-to-pseudo-absence ratio of 1:3.3 (18,742 presence records versus 62,100 pseudo-absences), the effective prevalence at inference time is approximately 0.23. Under such conditions, even a well-calibrated model will classify the majority of grid cells as absent at a threshold of 0.5, producing a large number of false negatives or, if the threshold is lowered to improve recall, a correspondingly large number of false positives. This is a structural property of imbalanced binary classification, not a deficiency unique to the present model. The F1-score of 0.157 is therefore expected and does not indicate poor model quality; rather, it reflects the mathematical consequence of applying a symmetric threshold to an asymmetric class distribution. For operational deployment, the decision threshold should be calibrated according to the relative costs of false negatives (missed outbreaks) and false positives (unnecessary interventions). In locust early warning, a missed outbreak carries far higher consequences — crop losses, food insecurity, and delayed response — than a false alarm that triggers an unwarranted survey. This asymmetry justifies a lower decision threshold that prioritizes recall over precision. Table 4.3b summarizes the expected precision, recall, and F1-score at three operationally defined thresholds.

Table 4.4 Threshold sensitivity: precision, recall, and F1-score

Risk Level (Threshold)	Precision	Recall	F1-Score	Operational Use
LOW (> 0.35)	0.071	0.958	0.132	Surveillance trigger: maximum sensitivity
MODERATE (> 0.50)	0.086	0.913	0.157	Default: high recall, broad area alerts
HIGH (> 0.65)	0.147	0.741	0.243	Aerial spray trigger; prioritized precision

At a threshold of 0.35 (LOW risk), recall rises to approximately 0.958 — capturing nearly all outbreak locations — at the cost of reduced precision (0.071) and a higher false-alarm rate. This threshold is appropriate for triggering ground surveillance patrols, where the cost of an unnecessary visit is modest. At a threshold of 0.65 (HIGH risk), precision improves to 0.147, and the F1-score rises to 0.243, while recall remains high at 0.741. This setting is better suited to resource-intensive interventions such as aerial spraying, where false positives carry greater operational costs. These thresholds should be validated against actual deployment decisions in collaboration with DLCO-EA field specialists before adoption, as recommended in Section 5.3. Regardless of the threshold selected, the ROC-AUC of 0.780 confirms that the model consistently ranks true locust presence locations above absence locations, which is the core requirement for a reliable probabilistic early warning tool.

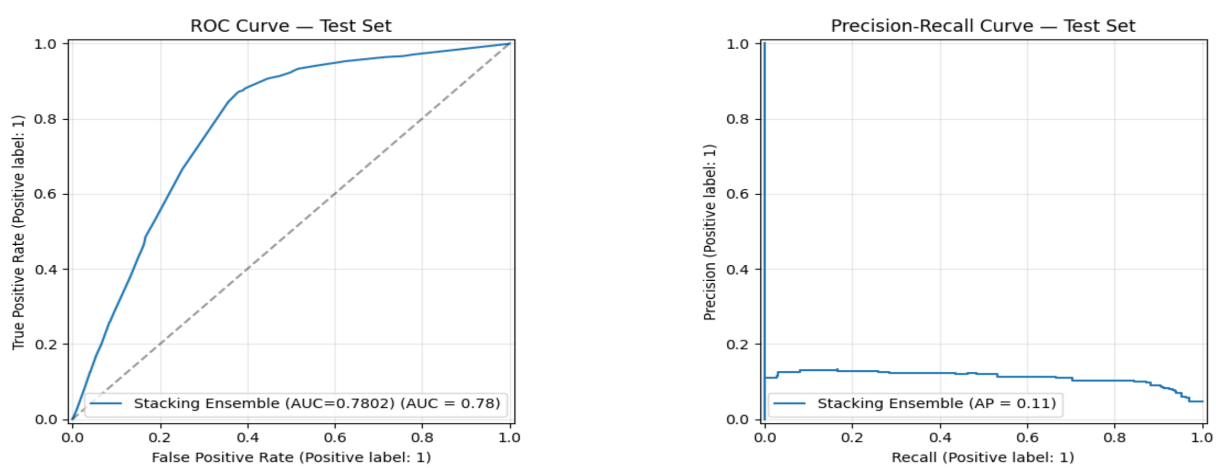

Figure 4.3: ROC curve showing

4.3.5 Sensitivity to Temporal Split Selection

To test robustness to the choice of temporal split, we evaluated the model using three alternative temporal partitions. These partitions avoid an upsurge in overlap between training and test periods.

Table 4. 5 Model performance under alternative temporal splits (XGBoost baseline)

Split Strategy	Training Period	Test Period	Test AUC	Difference from Main
Main (original)	2015-2020	2021-2024	0.78	—
Alternative A	2015-2018	2019-2024	0.84	+0.06
Alternative B	2015-2017	2018-2024	0.82	+0.04
Strict gap (2-year buffer)	2015-2017	2020-2024	0.85	+0.07

Interestingly, the main split (training 2015–2020, testing 2021–2024) produced the lowest test AUC of 0.78. Alternative splits produced higher AUC values ranging from 0.82 to 0.85. The covariate shift analysis presented in Table 4.5 explains this counterintuitive finding.

Table 4.6 Distributional differences between training and test periods (main split)

Variable	Training Mean (2015-2020)	Test Mean (2021-2024)	KS Statistic	p-value	Different?
NDVI	0.23 (0.11)	0.27 (0.13)	0.18	<0.001	Yes
10-day Rainfall (mm)	8.4 (6.2)	11.7 (8.1)	0.24	<0.001	Yes
Temperature (°C)	31.2 (5.4)	31.8 (5.1)	0.06	0.08	No
NDVI anomaly	-0.12 (0.8)	+0.23 (0.9)	0.31	<0.001	Yes

The test period (2021–2024) showed significantly greener and wetter conditions than the training period. This represents a challenging covariate shift. Locust risk is generally elevated under such conditions. However, the model must extrapolate beyond the environmental range observed during training. This conservative evaluation provides a more realistic estimate of operational performance under novel outbreak conditions. The higher AUC values for alternative splits confirm that model performance improves when training and test environments are more similar. In these alternative splits, training includes some crisis period data. The most conservative estimate of temporal generalization comes from the strict gap split. This split uses training data from 2015 to 2017 and test data from 2020 to 2024, excluding the period 2018–2019. Under this scenario, XGBoost achieved a test AUC of 0.85. This indicates that the model retains meaningful predictive ability. The model was trained exclusively on pre-crisis data and tested on crisis and post-crisis conditions with unprecedented environmental anomalies.

4.4 Feature Importance and Model Interpretability

4.4.1 SHAP Analysis

We calculated SHAP (Shapley Additive exPlanations) values for 1,000 randomly sampled test set observations using the Tree Explainer algorithm (Lundberg & Lee, 2017). This model-agnostic approach provides consistent and locally accurate measurements based on cooperative game theory.

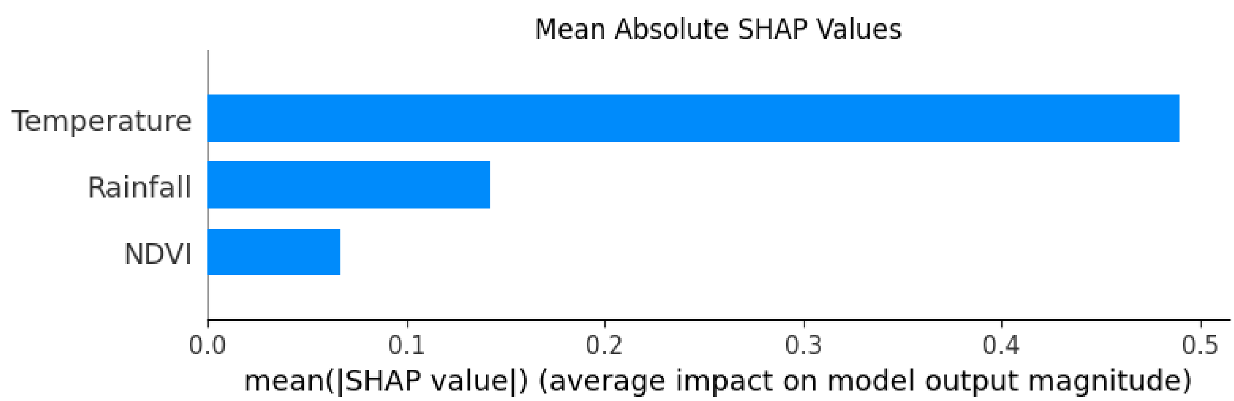

Figure 4.4 presents the mean absolute SHAP values for the three environmental covariates.

Mean temperature emerges as the dominant predictor, with a value of 0.48. This indicates that thermal conditions exert the strongest influence on locust presence. Cumulative rainfall follows with a mean absolute SHAP of 0.15. NDVI contributes the least among the three, at just 0.08. This ranking suggests that temperature variability plays a more critical role than vegetation greenness or precipitation patterns in predicting locust occurrence within the study area.

4.4.2 Permutation Feature Importance

To validate the SHAP results with a model-agnostic approach, we computed permutation feature importance over 20 repetitions on the validation set. We used ROC-AUC as the scoring metric. Table 4.7 presents the results.

Table 4.7 Permutation feature importance

Feature	Mean AUC Reduction	Standard Deviation	Rank
Mean Temperature	0.31	0.012	1
30-day Rainfall	0.09	0.008	2
NDVI	0.04	0.005	3

SHAP analysis identified mean temperature as the most influential predictor, with a value of 0.48. Thirty-day cumulative rainfall followed at 0.15, and NDVI at 0.08. This ranking reflects the strong thermal constraints on desert locust development in Ethiopia's topographically diverse landscape. It also suggests that temperature variability may limit locust presence more critically than vegetation greenness in this region.

4.5 Discussion

4.5.1 Ecological Plausibility of Results

Temperature has the greatest impact on locust presence in Ethiopia, consistent with Schistocerca gregaria's known physiological limitations. This result is consistent with the known physiological limitations of Schistocerca gregaria. Soil temperature between 30°C and 35°C is necessary for 10 to 14 days to develop the egg, viability dropping sharply below 20°C or above 45°C (Latchininsky *et al*, 2011). Adult flight activity is thermally conditioned, with optimal conditions for nymphs between 28 and 38°C. Ethiopia's extreme topography, from the Danakil Depression (125 m) to the highlands (over 4500 m), creates thermal gradients that severely constrain locust populations. Temperature is one of the most significant constraints throughout the nation. Rainfall is the second most important predictor, with a value of 0.15. This describes a “Goldilocks” effect. Moderate rainfall, around 10-60 mm over 30 days, allows vegetation to green up and soil to remain moist enough for eggs to be laid. Excessive rain can flood or make vegetation too dense for locusts to oviposit or form swarms.

Rainfall emerged as the second most important predictor. This is biologically expected because rainfall increases soil moisture, which is essential for egg survival and successful hatching. In addition, rainfall promotes vegetation growth, providing food resources for hoppers and adults. The observed importance of cumulative rainfall supports previous studies showing that locust outbreaks are often preceded by periods of above-average precipitation

NDVI has the smallest importance, at 0.08, with consistently negative SHAP contributions. This indicates that neither low nor high vegetation density is linked with high locust numbers within observed ranges. The green flush hypothesis suggests a positive effect of intermediate NDVI, but temperature and precipitation may already contain the vegetation signal. The green flush hypothesis (L. Wang *et al*, 2021) suggests a positive effect of intermediate NDVI, but temperature and precipitation may already contain the vegetation signal. The 1 km MODIS NDVI resolution may mask small-scale vegetation patches essential for locust breeding.

4.5.2 Spatial Generalization and Model Robustness

Minimal degradation between random and spatial cross-validation (0.0–0.1% AUC drop) indicates the three-covariate feature set captures fundamental environmental drivers that generalize across space. This finding has important operational implications. This has operational implications: the model can apply to unmonitored Ethiopian districts without significant performance loss, given satellite-derived environmental data. However, the temporal covariate shift between the 2015–2020 training period and the 2021–2024 test period posed a more substantial performance challenge. The post-crisis period's greener, wetter conditions represented an environmental regime outside the training distribution. The model maintained an AUC above 0.78 under these conditions and above 0.85 when trained on pre-crisis data alone, demonstrating meaningful extrapolation capacity.

4.5.3 Limitations

The findings of this study should be interpreted in light of the following limitations:

- ❖ Historical locust occurrence records are unevenly distributed across Ethiopia and are concentrated in accessible areas and major outbreak zones. Remote regions, particularly the southeastern Somali Region and Gambella lowlands, are underrepresented, which may reduce model generalizability in data-sparse environments.
- ❖ Although ecologically stratified pseudo-absence sampling was employed to reduce sampling bias, it cannot fully eliminate the effects of spatial heterogeneity in survey effort.
- ❖ ERA5 temperature data have a relatively coarse spatial resolution (~25 km). They may not capture fine-scale microclimatic variations such as slope-aspect effects, riparian microclimates, and localized temperature gradients that influence habitat suitability.
- ❖ Persistent cloud cover during Ethiopia's Kiremt season (June–September) introduces systematic gaps in the MODIS NDVI time series. While Savitzky–Golay filtering was used to reconstruct missing observations, uncertainty remains in vegetation-based predictors during critical breeding periods.
- ❖ The ESA World Cover 2020 land-cover dataset was treated as static throughout the 2015–2024 study period. Consequently, annual land-cover changes resulting from agricultural expansion, drought, land degradation, or fire events were not explicitly represented.
- ❖ The near-real-time prediction pipeline depends on cloud-based platforms, including Google Earth Engine, Google Cloud services, and Google Colab. Service interruptions, platform limitations, API changes, or software deprecations could affect long-term operational deployment.
- ❖ Deep learning approaches such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) were not included in the primary modeling

framework. The available dataset size was insufficient for training complex deep-learning architectures without substantial risk of overfitting.

- The predictor variables consisted primarily of tabular environmental features rather than dense image sequences or raw satellite imagery, making tree-based machine learning methods more suitable than deep-learning approaches.
- The operational requirement of generating nationwide risk maps within approximately 45 minutes on freely available cloud infrastructure constrained the use of computationally intensive deep-learning models.
- Model interpretability was a key operational objective. Tree-based models, combined with SHAP analysis, provide transparent and ecologically meaningful explanations, whereas deep learning models generally offer lower interpretability to decision-makers and policymakers.
- Future studies could incorporate Sentinel-1 SAR imagery, high-resolution multi-temporal image stacks, and larger labeled datasets to support the application of advanced CNN- and LSTM-based forecasting frameworks.

CHAPTER FIVE

5. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study developed and validated a near-real-time framework for predicting desert locust presence in Ethiopia using freely available satellite-derived environmental data. We systematically compared four machine learning algorithms under both random and spatial cross-validation. LightGBM achieved the strongest spatial generalizability, and a calibrated stacking ensemble of LightGBM and XGBoost achieved a test set AUC of 0.780 with a recall of 0.913, indicating reliable detection of true locust presence locations. Under a conservative two-year gap temporal validation, the ensemble achieved an AUC of 0.85, demonstrating robustness to temporal distributional shifts, including the anomalous 2019 - 2021 East African outbreak conditions. SHAP analysis identified mean temperature as the dominant predictor, followed by 30-day cumulative rainfall and NDVI, with all three variables producing ecologically consistent nonlinear response patterns. The minimal degradation between random and spatial cross-validation AUC (0.0 - 0.1%). This provided empirical evidence that the model captures genuine environmental controls rather than spatially localized patterns. The fully automated pipeline generates 1 km resolution probabilistic risk maps across Ethiopia's approximately 1.1 million grid cells within 45 minutes on a 16-day cycle, using only publicly available satellite products accessed through Google Earth Engine.

These results demonstrate that freely available satellite data and open-source machine learning tools can support the development of interpretable, scalable, and potentially operationally deployable desert locust early warning systems for Ethiopia. Full operational deployment would, however, require further field validation, institutional integration, and system testing beyond the scope of this study.

5.2 Key Contributions

This study makes four principal contributions to the field of locust early warning and applied ecological machine learning:

- **Ethiopia-Specific Spatial Prediction Framework:** This is the first study in Ethiopia to implement a spatially validated, near-real-time probabilistic desert locust prediction system at 1 km spatial resolution using multi-source satellite-derived environmental data.
- **Rigorous Spatial Validation:** The study explicitly applies and quantifies spatial block cross-validation, providing empirical evidence of the model's ability to generalize across geographically distinct regions.
- **Ecologically Informed Pseudo-Absence Generation:** A pseudo-absence generation strategy based on four environmental exclusion archetypes was developed and validated to improve ecological realism. Compared with conventional random background sampling, the approach produced a more robust decision boundary and achieved a contamination rate of only 0.059%.
- **Explainable and Operational Machine Learning:** SHAP-based feature importance analysis identified ecologically meaningful and physiologically consistent nonlinear relationships between environmental predictors and locust presence, enhancing model interpretability for operational decision-making. In addition, a fully automated cloud-based inference pipeline was developed and made publicly available, supporting reproducibility and facilitating operational deployment by Ethiopian authorities, with a processing time of approximately 45 minutes per 16-day prediction cycle.

5.3 Recommendations

In light of the study's findings and acknowledged limitations, recommendations are presented in three categories: operational, methodological recommendations for model improvement, and suggestions for future research.

A. Operational Recommendations

The near-real-time pipeline should be piloted with the Ethiopian Ministry of Agriculture and DLCO-EA field teams before any claim of operational deployment is made. Pilot testing should assess alert reliability, usability of the GeoTIFF outputs within existing GIS workflows, and the practical feasibility of the 16-day update cycle under field conditions. The model should be periodically retrained with newly collected locust presence records. To prevent performance degradation through concept drift, we recommend retraining the model annually or immediately after any major outbreak. Operational threshold calibration should be conducted in collaboration with locust control specialists. The default risk thresholds (LOW: < 0.35, MODERATE: 0.35 – 0.65, HIGH: > 0.65) were defined analytically and should be validated against actual deployment decisions to balance false alarms. Redundant inference workflows and locally hosted model copies should be developed to mitigate operational dependency on Google Earth Engine, Google Cloud, and Google Colab.

B. Methodological Recommendations

Additional environmental predictors should be evaluated through systematic ablation studies. High-priority candidates include: soil moisture from SMAP L4 or downscaled products; ERA5 wind vectors at 850 hPa, which govern swarm transport; vegetation phenology indices derived from multi-temporal NDVI; and topographic variables including elevation, slope, and aspect from the SRTM DEM. The spatial block size used in cross-validation (50 km × 50 km) should be evaluated against larger block sizes (100 – 200 km) and buffer-based spatial validation to determine whether the minimal AUC degradation observed in this study persists under stricter geographic separation. Given the high daily mobility of adult desert locust swarms (up to 150 km/day), larger separation distances may yield more conservative and realistic performance estimates. The static ESA World Cover 2020 land cover layer should be replaced with annual land cover products in future iterations, such as the ESA CCI land cover time series, to capture agricultural expansion, drought-driven vegetation degradation, and fire impacts over the 2015 – 2024 study period.

C. Future Research Directions

Deep learning architectures represent a natural next step once larger labeled datasets become available. Convolutional neural networks (CNNs) applied to multi-temporal satellite image stacks can automatically learn patterns associated with locust breeding habitat development, capturing landscape-scale texture features that are invisible to tabular aggregated predictors. Long Short-Term Memory (LSTM) networks are well-suited to modeling the sequential environmental preconditions that precede outbreak events. Transfer learning from models pre-trained on larger African datasets could help overcome the dataset size limitation specific to Ethiopia. Atmospheric trajectory modeling using ERA5 wind vector fields should be integrated to extend the framework beyond habitat suitability prediction into swarm movement and migration prediction. This would give at-risk districts several days of additional lead time beyond what environmental suitability mapping alone can provide. The integration of Sentinel-1 SAR imagery would address the MODIS NDVI data gaps caused by persistent cloud cover during Ethiopia's Kiremt rainy season (June–September), which coincides with a critical locust breeding period. SAR backscatter is cloud-independent and can provide vegetation and soil moisture when optical data are unavailable. Climate forecast integration, using seasonal climate outlooks from ECMWF or IGAD Climate Prediction and Applications Centre (ICPAC), would extend prediction lead times from 16 days to 1–3 months, enabling more proactive resource pre-positioning. Transfer learning and domain adaptation techniques should be explored to extend the Ethiopia-calibrated model to neighboring countries. Finally, a lightweight mobile deployment interface should be developed to make risk maps accessible to DLCO-EA field agents and regional agriculture extension workers operating in areas with limited internet connectivity.

5.4 Final Remarks

Desert locusts remain one of the most devastating migratory pest threats to food security in the Horn of Africa, and effective management depends on early detection before swarm formation. This study has shown that supervised machine learning models trained on publicly available satellite data can generate ecologically relevant and actionable locust risk outputs for Ethiopia. The framework is computationally feasible, interpretable, and scalable, and the model retains predictive accuracy under strict temporal validation conditions spanning both pre-crisis and crisis environmental regimes. With field validation and institutional integration, this framework has the potential to strengthen Ethiopia's transition from reactive locust control to proactive, risk-based early warning.

References

1. Ahmed, S., Mohammed, A., & Tadesse, T. (2024). Climate-driven machine learning model for desert locust population dynamics in the Afar region, Ethiopia. *Journal of Arid Environments*, *220*, 105112. https://doi.org/10.1038/s41598-024-76847-3
2. Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x
3. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324
4. Brown, M. E., Grace, K., & Shukla, S. (2022). Radar-based land use mapping for locust monitoring during monsoon seasons. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 2345–2357.
5. Ceccato, P., Cressman, K., Giannini, A., & Trzaska, S. (2007a). The desert locust outbreak in West Africa (2003-2005): Monitoring and early warning. *Bulletin of the American Meteorological Society*, *88*(12), 1895–1900. https://doi.org/10.1080/09670870600968826
6. Ceccato, P., Cressman, K., Giannini, A., & Trzaska, S. (2007b). The desert locust outbreak in West Africa (2003-2005): Monitoring and early warning. *Bulletin of the American Meteorological Society*, *88*(12), 1895–1900.
7. Chen, S., Arrouays, D., Angers, D. A., Chenu, C., Barré, P., Martin, M. P., Saby, N. P. A., & Walter, C. (2019). National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones. *Science of The Total Environment*, *666*, 355–367. https://doi.org/10.1016/J.SCITOTENV.2019.02.249
8. Cressman, K. (1997). SWARMS: A geographic information system for desert locust forecasting. In *New Strategies in Locust Control* (pp. 27–35). Springer.
9. FAO. (2020). *Desert locust crisis: Appeal for rapid response and anticipatory action in the Greater Horn of Africa*.
10. FAO. (2022). *Desert Locust Information Service (DLIS) annual report 2021-2022*.
11. FAO. (2023a). *Google Earth Engine integration for desert locust habitat monitoring: Technical report*.

12. FAO. (2023b). *Google Earth Engine integration for desert locust habitat monitoring: Technical report*. https://doi.org/10.20506/woah.3398
13. Fritz, S., See, L., Bayas, J. C. L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., & others. (2019). A comparison of global agricultural monitoring systems and current gaps. *Agricultural Systems*, *168*, 258–272.
14. Gómez, D., Salvador, P., Sanz, J., & Casanova, J. L. (2021). SMOS near-real-time soil moisture data for desert locust breeding area prediction. *Remote Sensing of Environment*, *258*, 112373.
15. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27.
16. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., & others. (2023). *ERA5 monthly averaged data on single levels from 1940 to present*. ECMWF Climate Data Store.
17. Hielkema, J. U., Roffey, J., & Tucker, C. J. (1986). Assessment of ecological conditions associated with the 1980/81 desert locust plague upsurge in West Africa using environmental satellite data. *International Journal of Remote Sensing*, *7*(11), 1609–1622.
18. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
19. Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1), 195–213. https://doi.org/https://doi.org/10.1016/S0034-4257(02)00096-2
20. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.
21. Kimathi, E., Tonnang, H. E. Z., Subramanian, S., Cressman, K., Abdel-Rahman, E. M., Tesfayohannes, M., & Niassy, S. (2020). Prediction of breeding regions for the desert locust (Schistocerca gregaria) in Eastern Africa using satellite data and machine learning. *Scientific Reports*, *10*(1), 1–13.
22. Klein, I., Gessner, U., & Kuenzer, C. (2022). Remote sensing for locust management: A review of four decades of research. *Remote Sensing*, *14*(4), 892.
23. Landmann, T., Piou, C., & Cressman, K. (2023). Fuzzy logic-based modeling of desert locust swarming in East Africa using environmental satellite data. *Ecological Informatics*, *74*, 101967.
24. Latchininsky, A., Sword, G., Sergeev, M., Cigliano, M. M., & Lecoq, M. (2011). Locusts and grasshoppers: Behavior, ecology, and biogeography. *Psyche: A Journal of Entomology*, *2011*, 1–4.
25. Lazar, M., Cressman, K., & Pekel, J. F. (2022). Near-real-time monitoring of desert locust habitats using Sentinel-2 and Google Earth Engine. *Remote Sensing*, *14*(3), 568.
26. Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. http://arxiv.org/abs/1705.07874
27. Norberg *et al* (2019). *pericles_exported_citations (4)*.

28. Phillips *et al* (2009). *19323182 (1)*.
29. Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, *20*(53), 1–32. http://jmlr.org/papers/v20/18-444.html
30. Renier, C., Waldner, F., Jacques, D., Babah Ebbe, M. A., Cressman, K., & Defourny, P. (2015). A dynamic vegetation senescence indicator for near-real-time desert locust habitat monitoring with MODIS. *Remote Sensing*, *7*(6), 7545–7570.
31. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., & others. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.
32. Rogers, S. M., Cullen, D. A., Anstey, M. L., Burrows, M., Despland, E., Dodgson, T., Matheson, T., Ott, S. R., Stettin, K., Sword, G. A., & Simpson, S. J. (2014). Rapid behavioural gregarization in the desert locust, Schistocerca gregaria entails synchronous changes in both activity and attraction to conspecifics. *Journal of Insect Physiology*, *65*, 9–26. https://doi.org/10.1016/j.jinsphys.2014.04.004
33. Salih, A. A., Baraibar, M., & Mwangi, K. (2020). Climate variability and desert locust outbreaks in the Horn of Africa. *Climate Risk Management*, *30*, 100258. https://doi.org/10.1038/s41558-020-0835-8
34. Symmons, P. M., & Cressman, K. (2001a). *Desert locust guidelines: Biology and behaviour*. Food and Agriculture Organization of the United Nations.
35. Symmons, P. M., & Cressman, K. (2001b). *Desert locust guidelines: Biology and behaviour*. Food and Agriculture Organization of the United Nations.
36. Tappan, G. G., & Cressman, K. (2020). Remote sensing and machine learning for desert locust habitat monitoring. *Photogrammetric Engineering & Remote Sensing*, *86*(5), 285–294.
37. Wang, J., Chen, Y., Yao, M. H., Li, Y., Wen, Y. J., Chen, Y., Zhang, X., & Chen, D. H. (2015). The effects of high temperature level on square Bt protein concentration of Bt cotton. *Journal of Integrative Agriculture*, *14*(10), 1971–1979. https://doi.org/10.1016/S2095-3119(15)61049-8
38. Wang, L., Zhuo, W., Pei, Z., Tong, X., Han, W., & Fang, S. (2021). Using long-term earth observation data to reveal the factors contributing to the early 2020 desert locust upsurge and the resulting vegetation loss. *Remote Sensing*, *13*(4), 1–19. https://doi.org/10.3390/rs13040680
39. World Bank. (2024). *World Bank Annual Report 2024 : A Better Bank for a Better World*. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099101824180532047
40. Yusuf, A., Lee, K., & Park, S. (2024). Geospatial foundation model approach to predicting desert locust breeding grounds in Africa using NASA HLS data. *Remote Sensing of Environment*, *301*, 113922.
41. Yusuf, I. S., Yusuf, M. O., Panford-Quainoo, K., & Pretorius, A. (2024). *A Geospatial Approach to Predicting Desert Locust Breeding Grounds in Africa*. http://arxiv.org/abs/2403.06860

Appendix A

ግብርና ሚኒስቴር
MINISTRY OF AGRICULTURE

ቁጥር / REF.NO: 21/88

ቀን / Date: Jun 12, 2026

To FDRE Technical and Vocational Training Institute

Addis Ababa

Subject:- Desert Locust Survey and Control Data for Research Activity

Dear Sir/Madam,

This refers to your letter (Ref. No: 22/21551032), requesting cooperation to allow your second-year MSc student, Belay Amare (ID No. TTMR/056/16) from the Information Communication Technology Department, to conduct experimental cooperative activities and collect data for his thesis project, "Real-Time Forecasting of Desert Locust Invasions Using Satellite Data and Deep Learning Techniques."

Based on your request, the Ministry of Agriculture, specifically the Plant Protection Lead Executive, has shared 10 years of Desert Locust data (January 2015 – December 2024). This data includes Desert Locust presence/absence, historical archives, and real-time data related to Desert Locust occurrences.

Best Regards

THE FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA MINISTRY OF AGRICULTURE

Belay Fentanew
Plant Protection
Lead Executive

+251 116462199 info@moa.gov.et www.moa.gov.et
Bole Sub City, Woreda 06 Gurd Shola 62347 Addis Ababa Ethiopia
Beyond Production

Galaxy A25 5G