



WALLAGA UNIVERSITY

SCHOOL OF GRATUATE STUDIES

COLLEGE OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

**PREDICTING LAND SUITABILITY FOR CEREAL CROPS USING
MACHINE LEARNING APPROACHES**

MSc Thesis

By. Yassin Umeta Kumera

May,2025

Nekemte, Ethiopia



WALLAGA UNIVERSITY
SCHOOL OF GRATUATE STUDIES
COLLEGE OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE

PREDICTING LAND SUITABILITY FOR CEREAL CROPS USING
MACHINE LEARNING APPROACHES

Thesis submitted to the College of Engineering and
Technology in partial fulfillment of the requirements for the
degree of Master of Science in Computer Science.

By. Yassin Umeta Kumera

Advisor: Mr.Kamal Mohammed(Ass.Prof)

May, 2025
Nekemte, Ethiopia



WALLAGGA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
P.O. BOX: 395, NEKEMTE, ETHIOPIA

APPROVAL SHEET FOR SUBMITTING OF FINAL THESIS

As members of the Final MSc thesis defense examining board, we hereby confirm that we have thoroughly reviewed and evaluated the thesis entitled “Predicting Land Suitability for Cereal Crops using Machine Learning Approaches” by Yassin Umata Kumera. We recommend its approval as it meets the academic standards and requirements for the Master of Science degree in Computer Science.

_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
Naol Bakala		
_____	_____	_____
External Examiner	Signature	Date

Final Approval and Acceptance
Thesis Approved by

_____	_____	_____
Department PGC	Signature	Date
_____	_____	_____
Dean of College	Signature	Date
_____	_____	_____
Dean of SGS	Signature	Date

STATEMENT OF THE AUTHOR

I, Mr. Yassin Umeta Kumera, declare that this thesis, titled “Predicting Land Suitability for Cereal Crops Using Machine Learning Approaches,” is my original work, conducted under the supervision of Mr. Kamal Mohammed (Ass.Prof). I have adhered to academic integrity and ethical standards throughout its preparation, including data collection, analysis, and writing. All sources have been properly cited, and no part of this work has been fabricated, falsified, or plagiarized. "This thesis is presented as part of the requirements for the MSc in Computer Science at Wallaga University and has not been submitted to any other institution for academic recognition."

I understand that any breach of these principles may lead to disciplinary or legal consequences.

Name: Yassin Umeta Kumera

Signature: _____

Date _____

Department of Computer Science

College of Engineering and Technology

Declaration

I hereby affirm that the thesis prepared by Yassin Umata Kumera, entitled “Predicting Land Suitability for Cereal Crops Using Machine Learning Approaches,” submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science, complies with the university’s academic regulations and meets the required standards of originality and scholarly quality.

Mr. Kamal Mohammed (Ass. Prof)

Name of Advisor

Signature

Date

Acknowledgment

Above all, I sincerely thank the Almighty Allah for blessing me with the strength, guidance, and perseverance needed to successfully undertake and complete this research. I also extend my deep appreciation to my advisor, Mr. Kamal Mohammed(Ass.Prof), for his expert guidance and continuous support throughout the thesis work.

My sincere thanks also go to the Oromia Engineering Corporation for their valuable assistance in supplying the necessary datasets, including soil, climate, and topographic information, which were fundamental to this study.

I am especially thankful to Special thanks to Dr. Dejene Abera (Ph.D.) for his guidance on the dataset, support with labeling, and provision of academic resources.

Finally, I would like to express my sincere gratitude to my family and friends, especially Bilisumma Dassalegn, for their unwavering encouragement and support throughout this journey.

Table of Contents

STATEMENT OF THE AUTHOR	iv
Acknowledgment	vi
Table of Contents	vii
LIST OF FIGURES	xi
LIST OF TABLES	xii
LIST OF EQUATIONS	xiii
ABSTRACTS	xv
CHAPTER ONE	1
1. INTRODUCTION	1
1.1. Background of the Study	1
1.2. Motivation.....	2
1.3. Statement of the Problem.....	2
1.4. Research Questions.....	4
1.5. Objective	4
1.5.1. General Objective	4
1.5.2. Specific objectives	4
1.6. Scope and Limitations.....	5
1.6.1. Scope of the Study	5
1.6.2. Limitations of the Study.....	5
1.7. Significance of the Study	5
CHAPTER TWO	6
2. LITERATURE REVIEW AND RELATED WORK	6
2.1. Definition of Land and Its Resources	6
2.2. Suitability of Land Evaluation	6
2.3. Existing Approaches to Land Suitability Evaluation.....	7

2.4. Land Characteristics.....	7
2.4.1. Characteristics of Soil	8
2.4.2. Climatic Parameters	9
2.4.3. Topographic parameter	9
2.5. FAO-Based Land Suitability Classification.....	9
2.6. Requirements of major cereal crops	10
2.7. Machine Learning	12
2.7.1. Supervised Machine Learning	13
2.7.2. Unsupervised machine learning	13
2.7.3. Semi-Supervised Machine Learning.....	13
2.7.4. Reinforcement Learning	13
2.8. Application of ML in Predicting Agricultural Land Suitability	14
2.9. Feature Selection.....	14
2.10. Related work	16
CHAPTER THREE	19
3. RESEARCH METHODOLOGY	19
3.1. Introduction.....	19
3.2. Research Design.....	19
3.3. Study Area	20
3.4. Data Collection and Sources	20
3.5. Data Annotation and Labeling	21
3.6. Data Preprocessing.....	21
3.6.1. Selection Methods Feature	22
3.6.2. Machine Learning Models	23
3.7. Predictive Model Evaluations	23
3.7.1. Performance Evaluation Metrics for Predictive Modeling	24

CHAPTER FOUR.....	26
4. PROPOSED SOLUTION	26
4.1. Introduction.....	26
4.2. System Architecture for Land Suitability Prediction.....	26
4.3. Data Preprocessing.....	27
4.3.1. Data Cleaning and Handling Missing Values	28
4.3.2. Handling Categorical Data.....	28
4.3.3. Feature Selection.....	28
4.4. Machine Learning Model Building.....	29
4.4.1. Hyperparameter Tuning	31
4.5. Model Evaluation and Prototyping	32
4.5.1. Prototyping.....	32
CHAPTER FIVE	33
5. RESULT AND DISCUSSION	33
5.1. Overview	33
5.2. Working Environment	33
5.3. Deployment Environment.....	34
5.4. Dataset Description.....	34
5.5. Dataset Class Distributions Result.....	36
5.6. Preprocessing Implementations	37
5.7. Results of Feature Selection.....	37
5.8. Model Performance and Evaluation Results.....	38
5.8.1. Evaluation of Model Performance with Feature Selection	38
5.8.2. Performance of GB Model with F S Methods	38
5.9. Results of Hyperparameter Tuning	40
5.10. Comparison	41

5.11. Result and Discussion	43
CHAPTER SIX	45
6. CONCLUSION AND FUTURE WORK	45
6.1. Conclusion	45
6.2. Recommendation	46
6.3. Contribution of the Study.....	47
REFERENCES	48
APPENDICES	i
Appendix A: Selected features by each feature selection methods	i
Appendix B: Result.....	ii
Appendix B:1 Stratified 10-fold CV result.....	ii
Appendix B:2 Classification Reports.....	iii
Appendix C: Sample Code.....	iv
Appendix C:1 Sample code for Preprocessing Implementations	iv
Appendix C:2 Sample code for Hyperparameter tuning Implementations.....	iv

LIST OF FIGURES

Figure 2.1 Feature Selection	15
Figure 2.2 Filter Method Flowchart.....	15
Figure 2.3 Wrapper Method flowchart	16
Figure 2.4 Embedded Method Flowchart	16
Figure 3.1 Workflow Diagram for Predicting Land Suitability.....	19
Figure 3.2 Location of study area.	20
Figure 4.1 Architecture of proposed system	27
Figure 4.2 Steps of Dataset Preprocessing.....	28
Figure 4.3: Feature Selection Workflow and Model Development.....	29
Figure 4.4 Model Building Flow Diagram.	30
Figure 4.5 Land suitability prediction Deployment Architecture.....	32
Figure 5.1 distribution of dataset in each class for cereal crops	37
Figure 5.5 Performance of GB Model with UFS without and with Normalization...	39
Figure 5.6 Performance of GB Model with RFECV without and with Normalization	39
Figure 5.7 Performance of GB Model with SFS without and with Normalization ...	39
Figure 5.2 GUI of the Land Suitability Prediction Prototype.....	44

LIST OF TABLES

Table 2.1 Classification of Land suitability based on FAO.....	9
Table 2.2 Suitability criteria for wheat crop	10
Table 2.3 Suitability criteria for Barley	11
Table 2.4 Suitability criteria for teff crop	11
Table 2.5 Suitability criteria for Maize crop.....	12
Table 2.6 Suitability criteria for Sorghum crop	12
Table 2.7 Summary of Related Works on Land Suitability Prediction	17
Table 3.3 Five class classifications Confusion matrix.....	25
Table 5.1: Tools and Libraries Used in the Study	33
Table 5.2 Description of dataset features.....	35
Table 5.4 Feature Selection Outcomes for Each Machine Learning Model.....	37
Table 5.5 summarized performance of all models with and without feature selection	40
Table 5.6: Selected Hyperparameters for RF,GB and KNN.....	41
Table 5.7 Comparison Between the Proposed Approach and Prior Research Studies	43
Table A.1 List of Parameter selected by each feature Selection methods	i

LIST OF EQUATIONS

Equation 3.124

Equation 3.224

Equation 3.324

Equation 3.425

Equation 4.131

Equation 4.231

Abbreviations

ACC · Accuracy	KNN · K-nearest Neighbor
AHP · Analytic Hierarchy Process	LCD · Liquid Crystal Display
AI · Artificial Intelligence	LDA · Linear Discriminant Analysis
ANP · Analytic Network Process	LR · Linear Regression
ARFF · Attribute Relation File Format	ML · Machine Learning
CaCO ₃ · Calcium Carbonate	NPK · Nitrogen, Phosphorous, Potassium
CSV · Common Separated Value	OM · Organic Matter
CV · Cross Validation	PH · Power of Hydrogen
DT · Decision Tree	RF · Random Forest
EC · Electronic Conductivity	RFE · Recursive Feature Elimination
ECO · Engineering Corporation of Oromia	RFECV · Recursive Feature Elimination with cross validation
FAO · Food and Agricultural Organization	RL · Reinforcement Learning
FS · Feature Selection	RMSE · Root Mean Square Error
GB · Gradient Boosting	SFS · Sequential Forward Selection
GDP · Gross Domestic Product	SMOTE · Synthetic Minority Over Sampling Technique
GIS · Geographic Information System	SVM · Support Vector Machine
GNB · Gaussian Naïve Bayes	UFS · Univariate Feature Selection
GPS · Global Positioning System	

ABSTRACTS

Assessing land suitability is a critical agricultural practice aimed at improving crop productivity. Nevertheless, ensuring food security remains a significant challenge, especially in developing nations like Ethiopia, where population growth intensifies the demand for agricultural output. Traditional land suitability assessment methods are often labor-intensive, time-consuming, costly, and lack the precision and speed needed for timely decision-making. To overcome these challenges, this research utilizes machine learning methods to forecast land suitability for five key cereal crops cultivated in Ethiopia namely Wheat, Barley, Maize, Sorghum, and Teff based on datasets acquired from the Engineering Corporation of Oromia. The dataset underwent extensive preprocessing to prepare it for model training. Three machine learning algorithms Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbors (KNN) were employed alongside feature selection methods including Univariate Feature Selection (UFS), Recursive Feature Elimination with Cross-Validation (RFE CV), and Sequential Forward Selection (SFS). Model performance was optimized using randomized search with cross-validation for hyperparameter tuning. Each model was evaluated under stratified 10-fold cross-validation using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The findings indicated that the Gradient Boosting algorithm, when integrated with Sequential Forward Selection, delivered superior performance achieving an accuracy of 94.41%, a precision of 94.37%, a recall of 94.34%, and an F1-score of 94.35%. Therefore, the GB model paired with SFS is recommended as the most effective approach for accurately and efficiently predicting land suitability for cereal crop cultivation. In addition to developing and evaluating machine learning models, this study suggests that future research could explore deep learning approaches with larger datasets to further enhance prediction accuracy.

Keywords: Land Suitability, Cereal Crops, Machine Learning, Gradient Boosting.

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the Study

Agriculture plays a vital role in ensuring food security and generating income. However, decisions regarding what to cultivate, where, and when remain persistent challenges for farmers and land managers. Land suitability prediction refers to estimating how appropriate a particular area of land is for cultivating various cereal crops. In Ethiopia, a large proportion of the population still lives below the international poverty line, and despite ongoing economic diversification, agriculture remains the backbone of the economy, employing approximately 80–85% of the population (Gotor et al., 2021).

However, many agricultural lands in the country are facing various problems such as erosion, landslides, deforestation, groundwater depletion, annual, and seasonal variations, which are harmful to crop yields. Therefore, to get the optimal benefit from the land, land suitability evaluation which is a process of assessing land performance based on its attributes can help to establish plans to improve agricultural productivity (Tadesse & Negese, 2020).

Ethiopia's agricultural sector is a cornerstone of its economy, with cereal crops such as wheat, maize, teff, barley, and sorghum being central to its agricultural practices. These cereals account for about 75% of the cultivated land area, contribute approximately 29% to the agricultural GDP, and provide 64% of the calorie intake in the Ethiopian diet (IFPRI, 2023). Therefore, where, and which crops to cultivate are the key issues that farmers and land managers face every day (Mokarram et al., 2015).

According to United Nations elaboration, the present population of Ethiopia as of Friday, April 10, 2025 is 129,239,634 (Worldometer, 2024). As the population increases, land use and land covers are also growing and similarly reducing food security (Mokarram et al., 2015) which results in food shortage. To resolve these problems and to ensure food security, land suitability evaluation is a powerful tool to support decision-making in land-use planning (Bagherzadeh et al., 2016) along with machine learning technology.

Machine learning has emerged as a powerful tool in agriculture, offering solutions across a wide range of domains including species identification, soil analysis, land suitability assessment, crop breeding, irrigation management, crop supervision, yield forecasting,

quality assessment, disease identification, weed classification, and animal husbandry management.

Recognizing both the challenges in traditional agricultural practices and the growing potential of machine learning technologies, this study focuses specifically on leveraging machine learning methods to predict land suitability. The ultimate goal is to enhance the productivity of cereal crops by providing data-driven insights for better land use planning.

Therefore, machine learning algorithm and the physico-chemical soil data which were collected from laboratory test results and climates like rainfall, temperature, and topographic data like slope (Tadesse & Negese, 2020) along with crop requirements to identify land suitability classes that show degrees of suitability for the selected crops will be used. According to the FAO method (Hagos, 2022) land suitability classes are highly (S1), moderately (S2), marginally (S3) suitable, and currently (N1) as well as permanently not suitable (N2) for various crops.

1.2. Motivation

Agriculture remains the foundation of Ethiopia's economy, yet traditional methods for assessing land suitability are often inefficient and lack predictive capacity. Although various studies have attempted to improve agricultural productivity, most rely on conventional techniques that are time-consuming and unable to accurately match land attributes with crop requirements (Welteji, 2018; Kahsay et al., 2018; Taghizadeh-Mehrjardi et al., 2020). Consequently, many farmers cultivate crops without understanding the suitability of their land, leading to poor yields and persistent poverty. To address these limitations, this study proposes the use of machine learning techniques for automated and accurate land suitability prediction, aiming to enhance cereal crop productivity and support better decision-making among farmers and agricultural experts.

1.3. Statement of the Problem

Land is the foundation of life for all living organisms, including humans, animals, and plants. It supports agricultural activities by providing essential nutrients through soil, making it a vital resource that must be protected. However, increasing global population and expanding human needs are placing growing pressure on land resources (Abebe & Sewnet, 2020). According to the International Organization of Supreme Audit Institutions Working Group on Environmental Auditing, the world faces severe environmental, social,

and economic challenges that continue to affect agricultural productivity and the sustainability of land use. These challenges are particularly critical for land, a resource that is becoming increasingly imbalanced.

The continuous population growth in developing nations like Ethiopia creates an increasing demand for food production to meet the needs of the rising population (Taghizadeh-Mehrjardi, 2007). Despite this demand, the average farmland size per household in Ethiopia remains under one hectare, posing a serious limitation. Moreover, the number of livestock surpasses the carrying capacity of the available arable land, leading to overexploitation and making it difficult to sustain sufficient labor output for effective land cultivation (Agidew & Singh, 2017).

In Ethiopia, farmers often face challenges in selecting appropriate cereal crops due to a lack of accurate and timely information on land suitability. Traditional assessment methods such as those based on GIS and AHP frameworks are resource-intensive, time-consuming, and lack predictive capacity (Taghizadeh-Mehrjardi et al., 2020). Furthermore, existing studies, has applied machine learning to land suitability for specific crops like wheat and barley, these studies suffer from limitations such as narrow crop focus, inadequate feature selection techniques, lack of end-user deployment tools, and minimal integration of local agro-environmental data.

Moreover, many studies do not clearly address how datasets are labeled according to agronomic requirements, nor do they leverage multiple feature selection strategies to optimize model performance. The absence of a scalable, user-accessible system also limits the practical value of these approaches for real-world agricultural planning.

To address these challenges, this study proposes a machine learning-based system that predicts land suitability for five major cereal crops wheat, barley, maize, teff, and sorghum using an integrated approach that combines soil, topographic, and climatic data. The goal is to develop a robust, automated, and accessible land evaluation tool that improves crop planning decisions and supports food security efforts in Ethiopia.

Accordingly, this study addresses the following key problems:

- The mismatch between the actual requirements, and what is implemented on a land failed to identify the level of land suitability for cereal crops quickly and automatically.

- Performing land suitability assessment is a long process and time-consuming task for both agricultural professionals and land evaluators.

1.4. Research Questions

This study is undertaken to examine and provide answers to the following research questions:

1. What combination of environmental factors (soil, climate, and topographic attributes) are most relevant for accurately predicting land suitability for major cereal crops.
2. How can multiple feature selection methods improve the predictive performance and interpretability of machine learning models in land suitability classification?
3. How do machine learning models (RF, GB, KNN) perform in predicting land suitability, and which model-feature selection combination gives the highest classification accuracy?

1.5. Objective

1.5.1. General Objective

The primary objective of this study is to predict land suitability for cereal crops using machine learning techniques to enhance cereal crops productivity.

1.5.2. Specific objectives

To achieve the main goal of predicting land suitability for cereal crop production using machine learning, the following specific objectives were outlined:

- To Conduct a detailed review of literature on machine learning applications in land suitability assessment for cereal crops.
- To gather and preprocess soil, climate, and topographic data for model training.
- To refine the parameters of machine learning models to improve the accuracy of land suitability predictions
- To evaluate and compare the performance of different land suitability prediction models.
- To design and implement a user-friendly prototype for practical land suitability assessment.

1.6. Scope and Limitations

1.6.1. Scope of the Study

This study focuses on developing an automated land suitability classification system aimed at enhancing cereal crop productivity through machine learning techniques. The analysis is based on soil, climatic, and topographic data collected over one year from the Engineering Corporation of Oromia. The research is geographically limited to the Kellem Wollega Zone in the Oromia Region. It employs three machine learning techniques Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbors (KNN) in combination with three feature selection strategies: Univariate Feature Selection (UFS), Recursive Feature Elimination with Cross-Validation (RFECV), and Sequential Forward Selection (SFS). To ensure accurate and balanced model evaluation, stratified k-fold cross-validation was implemented.

1.6.2. Limitations of the Study

The study does not cover other crop types and may face challenges related to data quality and availability, particularly in remote areas. It also confined to a specific region, the Kellem Wollega Zone Oromia region. This geographic limitation means that the results may not be generalizable to other regions with different soil, climate, and topographic conditions.

1.7. Significance of the Study

The implementation of the proposed land suitability prediction model and its accompanying prototype provides practical benefits to agricultural experts, land evaluators, and farmers. It enables accurate and timely identification of land suitability classes, helping stakeholders determine which cereal crop is best suited to a specific land unit prior to cultivation. By utilizing soil test results along with climatic and topographic data, farmers can make informed crop selection decisions. This approach also reduces uncertainty about land potential, thereby supporting improved crop yields and promoting data-driven agricultural practice.

CHAPTER TWO

2. LITERATURE REVIEW AND RELATED WORK

A literature review serves as a structured examination of scholarly articles, research papers, and other academic sources relevant to the study area. In this context, it focuses on agricultural practices in general, with a particular emphasis on land suitability assessment for cereal crop production using machine learning techniques. This chapter aims to deepen understanding of the subject matter by providing a comprehensive overview, synthesizing past studies, identifying research gaps, and critically evaluating existing methodologies related to automated land suitability prediction. The review will highlight how machine learning has been applied in agriculture and assess its effectiveness in improving land evaluation processes for cereal crops.

2.1. Definition of Land and Its Resources

Land is recognized as a vital yet finite natural asset, fundamental to sustaining human life and driving economic progress. It comprises more than just the Earth's surface it includes elements such as climate conditions, soil properties, topography, vegetation, wildlife, human populations, and the cumulative effects of both historical and ongoing human activities. As highlighted by Neameh (2003) and the Food and Agriculture Organization (FAO) (Ng et al., 1992), land should be viewed as a multidimensional entity that integrates both environmental and human-influenced factors.

2.2. Suitability of Land Evaluation

Building on the understanding of land and its components, evaluating land suitability becomes essential in determining how appropriate a specific land unit is for a given purpose. Land suitability assessment is the process of analyzing how effectively a land area can support particular uses, such as farming, infrastructure development, or housing (Hagos, 2022). In the agricultural sector, such evaluations typically involve comparing the physical and chemical characteristics of land like soil structure, depth, and drainage capacity with the ecological requirements of target crops. Numerous studies have been undertaken both globally and within Ethiopia to assess agricultural land suitability as a means of improving crop.

2.3. Existing Approaches to Land Suitability Evaluation

Given the importance of agriculture as both a means of ensuring food security and generating income especially in developing nations numerous studies have explored land suitability assessment using diverse technological approaches. For instance, Yohannes and Soromessa (2018) conducted a GIS-based multi-criteria evaluation in the Andit Tid watershed of Ethiopia to assess the suitability of land for wheat and barley cultivation. This study incorporated topographical, soil, and climatic variables and aligned them with FAO crop requirement standards. The suitability classification was achieved by integrating weighted and ranked criteria through the use of GIS and Analytic Network Process (ANP) software, resulting in a spatial map delineating suitability zones for the two cereal crops.

In a similar vein, Bozdağ et al. (2016) examined land suitability in the Cihanbeyli region of Turkey by employing a combination of the Analytic Hierarchy Process (AHP) and GIS. This study focused on categorizing land based on soil characteristics, terrain, and climatic data to determine its appropriateness for agricultural use. Likewise, Tashayo et al. (2020) investigated the potential for maize cultivation in Iran's semi-arid regions using GIS and AHP techniques, classifying land areas into different suitability levels for maize production.

2.4. Land Characteristics

Land characteristics refer to the measurable attributes of a specific land unit that can be analyzed through field assessments or soil surveys conducted by agricultural experts and land evaluators. These attributes include factors such as slope, annual rainfall, soil texture, available water capacity, and vegetation biomass (Neameh, 2003). Such characteristics are essential in determining the suitability of land for various uses, particularly in agriculture. According to FAO guidelines, land quality can be assessed either through direct field observations or by evaluating key land attributes against crop-specific requirements (Alemayehu, 2023). The most critical features that influence land suitability classification include the land's physical and chemical soil properties, climatic variables, and topographic features.

2.4.1. Characteristics of Soil

Soil characteristics are essential indicators of land quality and play a vital role in determining agricultural productivity. These properties influence various factors that affect plant development, such as water retention, nutrient availability, and root penetration. Accurate assessment of these attributes is fundamental to land suitability classification.

Soil depth is a crucial role in root development, water retention, and nutrient availability. Deeper soils generally support better plant development, while shallow soils can limit crop performance (FAO, 2007).

Soil texture: refers to the relative composition of mineral particles that make up the soil. This physical attribute is essential for determining water retention, aeration, and nutrient availability. Based on the proportions of these components, soils are typically classified into categories such as sandy, silty, clayey, or loamy; for instance, a balanced mixture of clay and loam is referred to as clay loam.

Soil PH: A chemical attribute indicating soil acidity or alkalinity. It affects nutrient availability and microbial activity, making it critical for crop selection and productivity.

Electronic Conductivity (E.C): EC measures soil salinity, which is closely linked to various physical and chemical soil conditions. High EC values can inhibit plant growth by affecting the soil's water balance (Mylavarapu et al., 2020).

Total Nitrogen and Average Phosphorus: These are essential macronutrients that significantly influence soil fertility. Their deficiency, common in degraded soils, is a major constraint for crop growth (Morgan, 2013).

Potassium (K): Another vital macronutrient, potassium is essential for plant development. Its deficiency, particularly in sandy soils, leads to symptoms such as leaf chlorosis and stunted growth (Morgan, 2013).

Soil Organic Matter (OM): Soil organic matter includes decomposed plant and animal materials, contributing to soil fertility, structure, and biological activity (Lal, 2005).

2.4.2. Climatic Parameters

Temperature: Mean minimum and maximum temperatures influence the thermal conditions required for crop development. In highland regions, temperature variations often correlate with elevation (Staff Soil Survey, 2016).

Rainfall: Rain-fed agriculture dominates in Ethiopia and other developing regions, making rainfall a critical parameter. It determines water availability for crops and thus plays a major role in land suitability assessment (Ritung et al., 2008).

2.4.3. Topographic parameter

Topography influences land suitability through parameters like slope and elevation. Slope affects water runoff, erosion risk, and ease of cultivation, while elevation influences temperature and solar radiation both of which are important for crop growth (Ritung et al., 2008).

2.5. FAO-Based Land Suitability Classification

The FAO land suitability classification system categorizes land into different levels of suitability based on its potential to support sustainable use. The classification is divided into two major orders: Suitable (S) and Not Suitable (N), which are further subdivided into classes as shown in the table below.

Table 2.1 Classification of Land suitability based on FAO

Order	FAO Symbols	Suitability Classes	Description
S	S1	Highly Suitable	Land with ideal soil, climate, and topography that supports optimal crop growth with no minimal limitations.
S	S2	Moderately Suitable	This classification refers to land that is generally appropriate for cultivation but has certain limitations.
S	S3	Marginally Suitable	Such lands are characterized by serious limitations that challenge sustainable agricultural use.
N	N1	Currently Not	These areas are currently unfit for farming because of

		Suitable	significant constraints. However, future advancements in technology or management practices may make their use feasible.
N	N2	Permanently Not Suitable	These land areas possess irreversible or extreme constraints that make them unsuitable for sustainable agricultural use under any foreseeable circumstances.

Source: Adapted from FAO Land Evaluation Guidelines (FAO, 1976).

2.6. Requirements of major cereal crops

Cereal crops namely barley, teff, maize, wheat, and sorghum are foundational to Ethiopian agriculture and play a crucial role in supporting the nation's food security and economy. These crops account for approximately 75% of the total cultivated land, highlighting their dominance in the agricultural landscape. Since 2000, there has been substantial progress in the cultivation of cereals, reflected in expanded acreage, increased productivity, and enhanced overall yields (Seyoum Taffesse et al., 2013). For continued improvement in cereal crop outputs, it is essential to classify land according to its suitability based on the specific agronomic requirements of each crop. Such classification supports informed decision-making and efficient land use (Fekadu & Negese, 2020).

Table 2.2 Suitability criteria for wheat crop

Feature	S ₁	S ₂	S ₃	N ₁	N ₂
Soil Depth (cm)	> 60	37 – 60	26 – 36	16 – 26	< 16
Soil Texture	L, CL, SiL	SCL, SiC	C, SL	LS	S
Soil pH	6.0 – 7.5	5.5 – 6.0 or 7.5 – 8.0	5.2 – 5.5 or 8.0 – 8.3	5.0 – 5.2 or 8.3 – 8.6	< 5.0 or > 8.6
Organic Matter (%)	> 3.0	2.5 – 3.0	2.0 – 2.5	1.0 – 2.0	< 1.0
Slope (%)	0 – 8	8 – 15	15 – 30	30 – 45	> 45
Temperature (°C)	15 – 20	13 – 15 or 20 – 22	12 – 13 or 22 – 24	10 – 12 or 24 – 26	< 10 or > 26
Rainfall (mm)	500 – 750	400 – 500 or 750 – 850	300 – 400 or 850 – 950	200 – 300 or 950 – 1100	< 200 or > 1100
Nitrogen(%)	> 0.22	0.15 – 0.22	0.10 – 0.15	> 0.10	Very Low (v)
Available Phosphorus (mg/kg)	> 10	5 – 10	2 – 5	< 2	Nil
Potassium (cmol/kg)	> 1.3	0.7 – 1.3	0.3 – 0.7	0.1 – 0.3	< 0.1
EC	< 4.5	4.5 – 8.5	–	8.5- 12	> 12.0

(mmho/cm)				.0	
-----------	--	--	--	----	--

Table 2.3 Suitability criteria for Barley

Feature	S ₁	S ₂	S ₃	N ₁	N ₂
Soil Depth (cm)	> 60	35 – 60	25 – 35	15 – 25	< 15
Soil Texture	Sil-C to C	SiCL, SC	L to SCL	-	S, SL
Soil pH	5.8 – 7.2	5.4 – 5.8 or 7.2 – 7.8	5.1 – 5.4 or 7.8 – 8.0	4.8 – 5.1 or 8.0 – 8.5	< 4.8 or > 8.5
Organic Matter (%)	> 3.2	2.7 – 3.2	2.0 – 2.7	1.0 – 2.0	< 1.0
Slope (%)	0 – 10	10 – 18	18 – 32	32 – 55	> 55
Temperature (°C)	14.5 – 20.5	13.5 – 14.5 or 20.5 – 21.5	12.5 – 13.5 or 21.5 – 22.5	11.5 – 12.5 or 22.5 – 24.5	< 11.5 or > 24.5
Rainfall (mm)	450 – 600	350 – 450 or 600 – 800	250 – 350 or 800 – 1100	150 – 250 or 1100 – 1300	< 150 or > 1300
Nitrogen (%)	> 0.28	0.18 – 0.28	0.10 – 0.18	> 0.10	Very Low (v)
Available Phosphorus (mg/kg)	> 10	5 – 10	< 5	-	-
Potassium (cmol/kg)	> 1.3	0.7 – 1.3	0.3 – 0.7	0.1 – 0.3	< 0.1
EC (mmho/cm)	< 3.5	3.5 – 6.5	-	6.5 – 10	> 10

Table 2.4 Suitability criteria for teff crop

Feature	S ₁	S ₂	S ₃	N ₁	N ₂
Soil Depth (cm)	> 65	45 – 65	30 – 45	20 – 30	< 20
Soil Texture	C, Si, SiC, SiL, SiCL	L, CL	SCL	SL	LS, S
Soil pH	5.6 – 7.8	5.3 – 5.6 or 7.8 – 8.1	5.1 – 5.3 or 8.1 – 8.4	4.9 – 5.1	< 4.9 or > 8.4
Organic Matter (%)	> 3.2	2.6 – 3.2	2.0 – 2.6	1.0 – 2.0	< 1.0
Slope (%)	0 – 10	10 – 18	18 – 30	30 – 60	> 60
Temperature (°C)	15.0 – 18.0	14.5 – 15.0 or 18.0 – 19.5	13.5 – 14.5 or 19.5 – 21.0	12.5 – 13.5 or 21.0 – 22.0	< 12.5 or > 22.0
Rainfall (mm)	480 – 680	380 – 480 or 680 – 880	300 – 380 or 880 – 1100	200 – 300 or 1100 – 1300	< 200 or > 1300
Nitrogen(%)	> 0.26	0.16 – 0.26	0.10 – 0.16	> 0.10	Very Low (v)
Available Phosphorus (mg/kg)	> 16	9 – 16	4 – 9	< 4	Nil
Potassium (cmol/kg)	> 1.3	0.7 – 1.3	0.3 – 0.7	0.1 – 0.3	< 0.1
EC (mmho/cm)	< 3.5	3.5 – 7	-	7 – 10	> 10

Table 2.5 Suitability criteria for Maize crop

Feature	S ₁	S ₂	S ₃	N ₁	N ₂
Soil Depth (cm)	> 105	60 – 105	45 – 60	30 – 45	< 30
Soil Texture	L, SC	SiC	C	SL, LS	S, HC
Soil pH	5.6- 6.8	5.3- 5.6 or 6.8 – 7.2	5.1 – 5.3 or 7.2 – 7.5	4.8 – 5.1 or 7.5 – 8.1	< 4.8 or > 8.1
Organic Matter (%)	> 3.0	2.5 – 3.0	2.0 – 2.5	1.0 – 2.0	< 1.0
Slope (%)	0 – 7	7 – 14	14 – 28	28 – 45	> 45
Temperature (°C)	17.5- 22.0	16.5 – 17.5 or 22.0- 23.0	15.5 – 16.5 or 23.0- 24.5	14.5 – 15.5 or 24.5 -25.5	< 14.5 or > 25.5
Rainfall (mm)	600 -800	500 – 600	400 – 500	300 – 400	< 300
Total Nitrogen (%)	> 0.22	0.16 – 0.22	0.10 – 0.16	> 0.10	Very Low (v)
Available Phosphorus (mg/kg)	> 16	9 – 16	4 – 9	< 4	-
Potassium (cmol/kg)	> 1.3	0.7 – 1.3	0.3 – 0.7	0.1 – 0.3	< 0.1
EC (mmho/cm)	< 3.5	3.5 – 7	-	7 – 10	> 10

Table 2.6 Suitability criteria for Sorghum crop

Feature	S ₁	S ₂	S ₃	N ₁	N ₂
Soil Depth (cm)	> 105	60 – 105	40 – 60	25 – 40	< 25
Soil Texture	L, SC	SiL, SiCL	SiC	SL, C	S, LS
Soil pH	5.6 – 7.2	5.3 – 5.6 or 7.2 – 7.4	5.1- 5.3 or 7.4 – 7.9	4.9 – 5.1 or 7.9 – 8.4	< 4.9 or > 8.4
Organic Matter (%)	> 3.0	2.5 – 3.0	2.0 – 2.5	1.0 – 2.0	< 1.0
Slope (%)	0 – 8	8 – 16	16 – 30	30 – 55	> 55
Temperature (°C)	11 – 24	9 – 11 or 24-27	7- 9 or 27- 29	5- 7 or 29- 30	< 5 or > 30
Rainfall (mm)	500- 750	350 – 500 or 750 – 850	250 – 350 or 850 – 1000	150 – 250 or 1000 – 1250	< 150 or > 1250
Total Nitrogen (%)	> 0.22	0.15 – 0.22	0.10 – 0.15	> 0.10	Very Low (v)
Available Phosphorus (mg/kg)	> 16	9 – 16	4 – 9	< 4	Nil
Potassium (cmol/kg)	> 1.3	0.7 – 1.3	0.3 – 0.7	0.1 – 0.3	< 0.1
EC (mmho/cm)	< 11.5	11.5 – 15.5	-	> 15.5	-

2.7. Machine Learning

Machine Learning (ML), a fundamental branch of Artificial Intelligence (AI), involves creating algorithms that enable systems to automatically identify patterns in data (referred

to as experience "E") and progressively improve their performance ("P") on defined tasks ("T") without requiring explicit programming instructions (Perner, 2015). It plays a vital role in computer science and engineering fields by enabling systems to analyze data and generate meaningful predictions or decisions. The primary categories of machine learning include supervised, unsupervised, semi-supervised, and reinforcement learning, each serving distinct purposes.

2.7.1. Supervised Machine Learning

Supervised learning trains models using datasets with input-output pairs, labeling each input with the appropriate output. Algorithms use the relationship between input attributes and output labels to predict previously unseen data. Alpaydin (2014) emphasizes the importance of creating a model that maps input features to class labels as a crucial aspect of supervised learning. There are two major categories of issues in supervised learning

Classification: This entails estimating a category or class label based on supplied data.

Regression: This deals with predicting continuous numerical values. The model learns the relationship between inputs and a continuous output variable, such as forecasting temperature or crop yield.

2.7.2. Unsupervised machine learning

Unsupervised machine learning is a method where algorithms analyze data without labeled outcomes, aiming to uncover hidden patterns or groupings within the dataset (Kotsiantis, Zaharakis, & Pintelas, 2007).

2.7.3. Semi-Supervised Machine Learning

Semi-supervised machine learning is an approach that integrates a limited set of labeled data with a larger volume of unlabeled data to train models more effectively. It helps improve learning performance when labeled data is scarce or costly to obtain (Chapelle, Scholkopf, & Zien, 2006).

2.7.4. Reinforcement Learning

Reinforcement learning is a machine learning approach in which an agent improves its decision-making by interacting with its environment, receiving rewards for favorable actions and penalties for unfavorable ones (Sutton & Barto, 2018). Through trial and error, it gradually learns the optimal strategy to maximize cumulative rewards over time (Perner, 2015).

2.8. Application of ML in Predicting Agricultural Land Suitability

Land suitability classification is a core component of agricultural planning. Following FAO guidelines (Ng et al., 1992), machine learning is increasingly applied to automate this process by integrating diverse datasets including soil, climate, and topography. Key machine learning algorithms used in land suitability assessment include:

Random Forest (RF): Random forest is a robust ensemble learning technique that generates multiple decision trees by randomly selecting subsets of features, and then aggregates their outputs for more accurate and stable predictions (Breiman, 2001).

Support Vector Machine (SVM): SVM is a powerful supervised learning technique that constructs an optimal hyperplane to distinguish between different classes. It is particularly effective in land suitability analysis, ranking just below Random Forest in predictive strength (Taghizadeh-Mehrjardi et al., 2020).

Decision Tree (DT): is a supervised learning algorithm commonly used for both classification and regression, where it learns decision rules from the training data in a tree-like model (Chauhan, 2021).

K-Nearest Neighbor (K-NN): K-NN is a straightforward algorithm that makes predictions by identifying the closest training examples to a given input, typically using distance measures such as Euclidean or Manhattan distance.

Gradient Boosting (GB): Gradient Boosting is a machine learning technique that builds models iteratively, with each new model correcting errors made by the previous one, often used for predictive tasks (Friedman, 2001).

2.9. Feature Selection

Feature selection is a crucial step in the machine learning pipeline aimed at reducing the number of input variables by retaining only the most relevant features for the task at hand. This process helps eliminate noisy or irrelevant data, which in turn enhances model performance, reduces computational overhead, and improves interpretability (Miao & Niu, 2016). The objective is to retain important features while discarding those that add little or no value, without altering the original data itself.

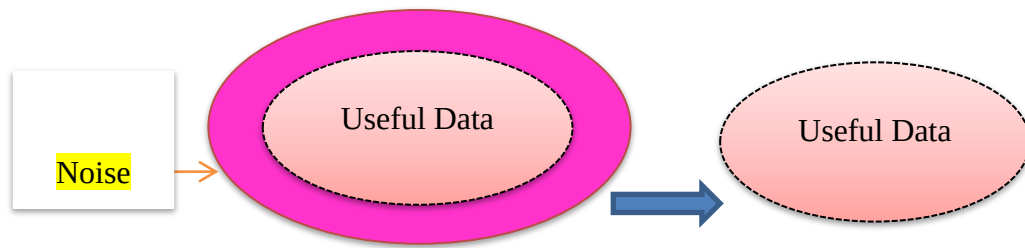


Figure 2.1 Feature Selection¹

Feature Extraction: Feature extraction is a dimensionality reduction technique that transforms the original input data into a smaller set of new features. Rather than choosing from existing variables, it creates meaningful combinations or representations that retain essential information. This approach is particularly useful for simplifying large datasets while maintaining their most relevant characteristics.

Feature Selection Methods

Filter Methods: These techniques serve as a preprocessing step and work independently of any machine learning model. Features are ranked based on statistical measures, and the least relevant ones are discarded. These methods are computationally efficient, scalable, and easy to implement (Wah et al., 2018).

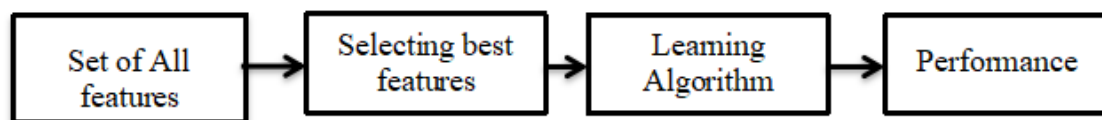


Figure 2.2 Filter Method Flowchart

Wrapper Methods: Unlike filter methods, wrapper methods evaluate subsets of features using a specific machine learning algorithm. Although they often provide better performance by tailoring the selection to a specific model, they are generally more computationally demanding, particularly with high-dimensional datasets (Wah et al., 2018).

¹<https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning>

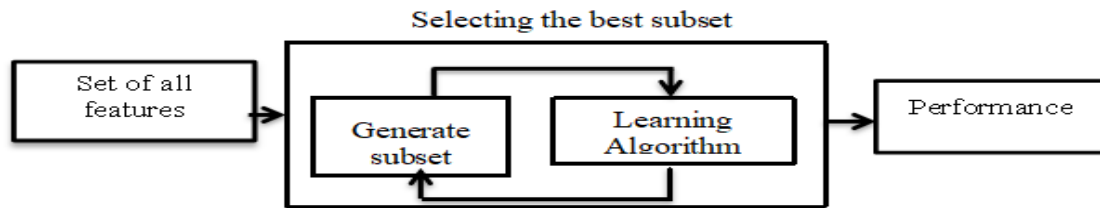


Figure 2.3 Wrapper Method flowchart

Embedded method: These methods are built into specific algorithms that evaluate the importance of each feature while learning the model structure. Common examples include Lasso regression and tree-based models like Random Forest, which inherently assess feature relevance (Guyon & Elisseeff, 2003).

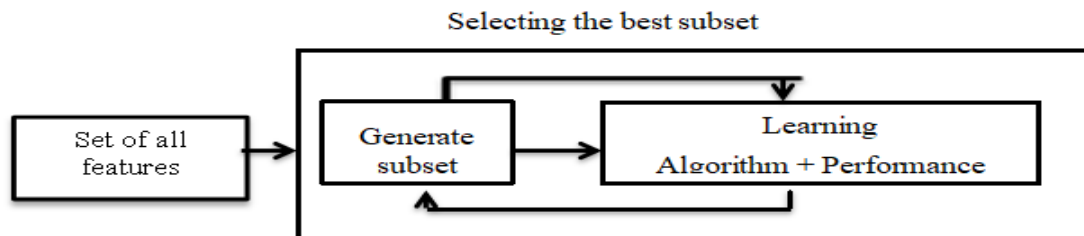


Figure 2.4 Embedded Method Flowchart

2.10. Related work

The absence of reliable land suitability assessments for identifying optimal land units for specific crops remains a significant challenge. Farmers often struggle to align crop types with land characteristics, hindering productivity. In recent years, machine learning techniques have been increasingly applied to tackle these issues and enhance agricultural output. Several studies have explored this topic, as outlined below.

Bikila Abebe (2021) conducted a study on land suitability prediction using machine learning techniques, focusing specifically on wheat and barley cultivation. While the study demonstrated the effectiveness of models such as Random Forest in classifying land suitability, it was constrained by several key limitations. The research was limited to only two cereal crops, which restricts its generalizability to broader agricultural contexts where multiple crop recommendations are essential. Additionally, the study did not explore the comparative benefits of diverse feature selection techniques, which are critical for enhancing model performance and reducing dimensionality.

Komolafe et al. (2019) successfully assessed cassava land suitability using SVM and Decision Trees on a small dataset of 252 records, achieving 87.5% accuracy after 10-fold cross-validation. However, their analysis was unclear on how the dataset was categorized based on crop requirements.

Senagi et al. (2017) used a parallel Random Forest model to assess land suitability for sorghum cultivation, incorporating eight features such as drainage, rainfall, rooting depth, salinity, sodicity, moisture retention, slope, and temperature. They compared several algorithms, including RF, LR, LDA, KNN, GNB, and SVM, with Random Forest achieving the highest accuracy of 90% using 10-fold cross-validation. Notably, key soil fertility indicators like pH, nitrogen, and phosphorus were not included in the study.

Sarmadian et al. (2014) conducted a land suitability assessment for rainfed wheat in Kouhin, Iran, using SVM. The study utilized 32 soil profiles and 10 land variables covering climate (precipitation, temperature), topography (slope, relief), and soil properties (texture, organic carbon, CaCO_3 , gypsum, coarse fragments, and pH). A non-linear two-class SVM model was developed using MATLAB 8.2 for land suitability classification. The model showed good accuracy, with an error rate of 3.72 and an 84% match between predicted results and actual outcomes. However, the study did not specify the dataset source or total record count, although the low RMSE indicated a good model fit.

Bhimanpallewar and Narasinagrao (2017) proposed a decision tree-based model to assess land suitability for Jowar (sorghum) in small-scale farms. Input variables included temperature, rainfall, slope, pH, EC, N, P, K, and soil moisture, aligned with the FAO suitability framework. The predicted land suitability classes included: highly suitable, moderately suitable, marginally suitable, currently unsuitable, and permanently unsuitable. Real-time data collection was mentioned, but the method (e.g., sensor-based or manual) was not clearly explained. The model achieved 98.9% accuracy with only a 1.1% error rate.

Table 2.7 Summary of Related Works on Land Suitability Prediction

Authors & Year	Title	Method & Findings	Identified Gaps
Bikila Abebe (2021)	Predicting land suitability for wheat and barley crop.	The study employed RF, GB and KNN achieving	Focused exclusively on two crops and did not incorporate a

		99.1% accuracy	diverse range of cereal types.
Komolafe et al. (2019)	Predictive Modeling for Land Suitability Assessment for Cassava Cultivation	Applied SVM and Decision Tree models, achieving 87.5% accuracy.	Dataset was limited to only 252 samples, and there was no clear mapping between land characteristics and cassava crop requirements.
Senagi et al. (2017)	Using Parallel RF for Predicting Land Suitability for Sorghum Crop Production	Evaluated multiple models (RF, LR, LDA, KNN, GNB, SVM); RF gave the highest accuracy (96%) using 10-fold cross-validation.	Crucial soil fertility features like pH, Nitrogen, and Phosphorus were not included.
Sarmadian et al. (2014)	SVM-Based Modeling of Land Suitability for Rainfed Agriculture	Used SVM on non-linear data; achieved RMSE of 3.72 and R ² of 0.84 for wheat suitability.	The model predicted only binary classes (moderately or marginally suitable), and the dataset size was not specified.
Bhimanpallewar & Narasinagrao (2017)	ML Approach to Assess Crop-Specific Suitability for Small/Marginal Scale Croplands	Used Decision Tree classifier with 98.9% accuracy.	Real-time data was used, but the method for collecting this data (e.g., sensors or manual entry) was not clarified.

In light of the gaps identified in previous studies and the persistent agricultural challenges facing Ethiopia, this research proposes an innovative approach that leverages machine learning to automate the prediction of land suitability for cereal crop cultivation. Traditional land evaluation methods, which rely heavily on manual processes and the FAO framework, tend to be labor-intensive and time-consuming. By developing and applying machine learning models, this study aims to enhance decision-making for agricultural professionals and support farmers in accurately determining land suitability levels. The proposed system not only increases precision but also significantly reduces the resources required such as time, labor, and cost thus contributing to improved productivity in cereal crop farming.

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1. Introduction

This chapter presents the research methodology used to develop a machine learning-based land suitability prediction system. It details the research design, study area, data source and data collection process, data annotation and labeling, feature selection, model training, and evaluation techniques. The methodology is designed to address the limitations of traditional land evaluation approaches and provide a scalable, automated framework for suitability classification.

3.2. Research Design

This study employs a quantitative experimental research design, focused on collecting and analyzing soil, climate, and topographic data to assess land suitability. The approach is used to explore the relationship between independent variables (land characteristics) and the dependent variable (suitability classes). Machine learning models are developed to validate the research objectives and automate land suitability classification, aligning with the quantitative method's goal of drawing data-driven conclusions (Kamiri & Mariga, 2021).

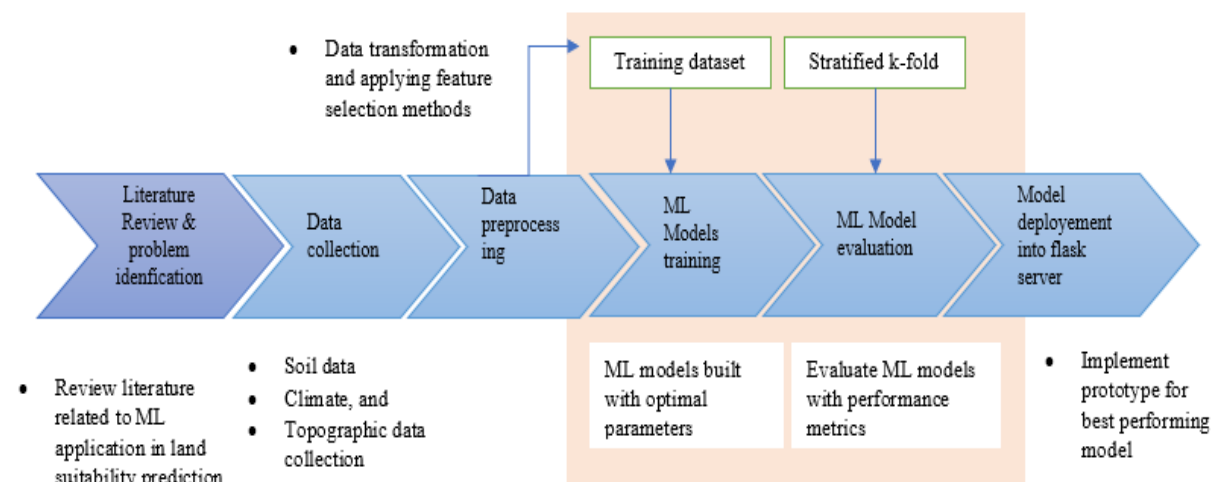


Figure 3.1 Workflow Diagram for Predicting Land Suitability

3.3. Study Area

The study was conducted in the Kellem Wollega Zone of the Oromia Region, Ethiopia, covering approximately 985,117 hectares. This area includes 12 districts, including Anfillo, Sadi Chanka, Dalle Sadi, Dalle Wabera, Gawo Kebe, Gidami, Hawa Gelan, Yemalogi Welel, Jimma Horro, Sayo, Dembi Dollo, and Lalo Kile and represents a diverse agro-ecological environment suitable for major cereal crops such as wheat, barley, teff, maize, and sorghum.

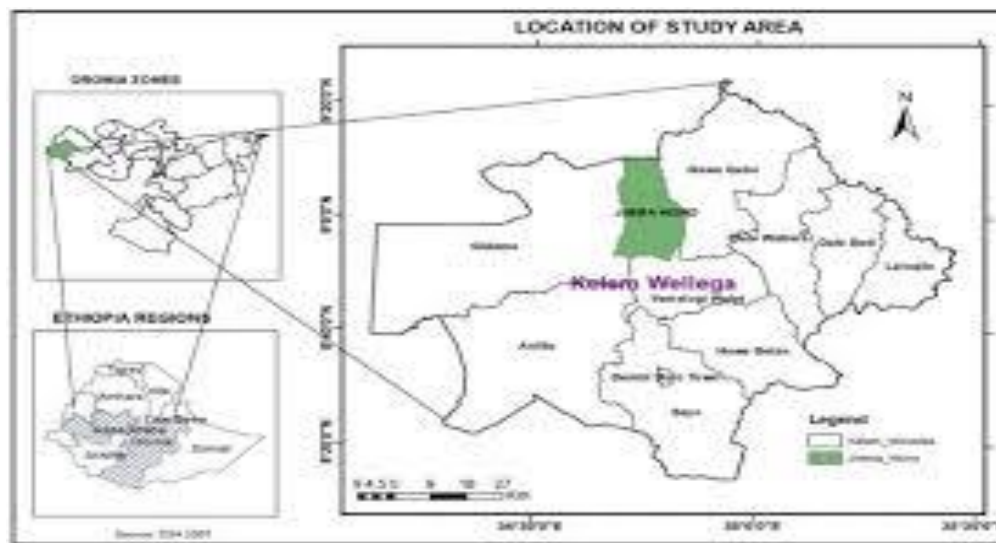


Figure 3.2 Location of study area.

3.4. Data Collection and Sources

The dataset used in this study was prepared based on soil, climate, and topographic information obtained from the Engineering Corporation of Oromia (ECO) in collaboration with the Oromia Agricultural Research Institute (OARI). ECO offers services including engineering consulting, community support, and land use planning. As part of its mandate, ECO conducts soil surveys and provides datasets to the Oromia Rural Land Administration and Use Bureau to support land evaluation, rural land use planning, and strategic environmental assessments. Data collection took place between June 2024 and December 2024, covering twelve woredas in the Kellem Wollega Zone, Oromia Region, Ethiopia. A total of 1,150 land units were sampled, and for each unit, 15 key attributes were recorded, including soil pH, texture, organic matter, slope, rainfall, and

temperature, among others. The original data files were in Excel and Word format and were transformed into a structured CSV file suitable for machine learning analysis.

3.5. Data Annotation and Labeling

Annotation of the data involved assigning a land suitability class to each land unit for specific cereal crops (wheat, barley, teff, maize, and sorghum). The labeling process was carried out using FAO land evaluation guidelines, which define specific thresholds for each parameter to classify suitability levels as S1 (Highly Suitable), S2 (Moderately Suitable), S3 (Marginally Suitable), N1 (Currently Not Suitable), and N2 (Permanently Not Suitable). For each crop, the collected land characteristics were compared against these FAO thresholds, and suitability classes were assigned accordingly.

3.5.1 Annotation Guidelines and Expert Validation

To ensure annotation quality, the labeling process followed a structured guideline developed in collaboration with experts from Oromia Agricultural Research Institute (OARI) and soil specialists at ECO. A manual annotation protocol was used to maintain consistency, and a random 10% sample of the labeled records was re-evaluated by an independent expert to validate correctness. Discrepancies were reviewed and resolved through expert consensus. This validation process helped improve the reliability of the labeled data, which was then used for model training and evaluation.

Using these guidelines, the dataset was labeled by mapping the crop requirements against the land attributes. The labeling was verified by agricultural experts to ensure data accuracy and consistency.

3.6. Data Preprocessing

Preprocessing data is a fundamental step in preparing it for machine learning, as it converts unprocessed information into an organized and refined format suitable for algorithmic analysis. The process includes data cleaning, managing missing values, and converting categorical variables to ensure the dataset is suitable for accurate analysis and prediction.

Noise Removal: This step involves identifying and correcting inaccurate, inconsistent, or incomplete data that may degrade the performance of the model. By eliminating such noise, the overall quality of the dataset is improved, which enhances model accuracy.

Managing Missing Values: Missing data can lower the quality and reliability of a model. Several techniques are used to address this issue, such as imputing missing values using statistical methods, predicting missing values based on other available data, or removing records with missing values if deemed non-essential.

Encoding Categorical Variables: The dataset includes non-numeric fields such as soil texture and crop name, which need to be converted into numerical form for processing by machine learning algorithms. In this study, label encoding is employed to transform categorical data into integer values, allowing the models to effectively interpret and process these attributes.

Handling imbalanced data: This study addresses class imbalance in multiclass classification by setting `class_weight` in the random forest algorithm's hyperparameter setting, using ensemble algorithms like gradient boosting, and using KNN instead of under-sampling or over-sampling techniques like SMOTE. These solutions prevent data loss, noise, and model overfitting, ensuring effective data handling.(Tahir et al., 2019).

3.6.1. Selection Methods Feature

Feature selection is essential for identifying the most important features in a dataset, which helps reduce dimensionality and enhances the performance and efficiency of machine learning models. These methods are broadly categorized into three types: filter, wrapper, and embedded approaches.

In this research, widely adopted techniques namely univariate selection (a filter method), Recursive Feature Elimination (RFE), and Sequential Forward Selection (both wrapper methods) are employed to select optimal features for effective model training (Shaikh, 2018).

Univariate Feature Selection (UFS)

UFS is a filter-based method that is fast and simple, ideal for getting a quick ranking of feature importance. It uses statistical tests (e.g., ANOVA F-value) to evaluate the relationship between each feature and the target variable independently. Used as a baseline method to quickly eliminate irrelevant features before applying more complex techniques(Khalid et al., 2014).

Recursive Feature Elimination with Cross-Validation (RFECV)

RFECV is a wrapper method that recursively removes the least important features and evaluates model performance using cross-validation. It helps prevent overfitting and improves generalization by evaluating subsets of features in combination. Provides a robust and data-driven approach to find the optimal number of features for the selected machine learning models(Guyon et al.,2002).

Sequential Forward Selection (SFS)

SFS starts with no features and adds one feature at a time, choosing the one that improves model performance the most at each step. It's useful when the dataset is not very large and you want to capture feature interactions. This method gave the best accuracy when combined with Gradient Boosting, making it the most effective in this specific use case(Molina et al., 2002).

3.6.2. Machine Learning Models

This research adopts supervised machine learning approaches to classify land suitability and enhance cereal crop productivity using a cleaned and labeled dataset. Three models Random Forest, Gradient Boosting, and KNN were selected for their proven performance in agricultural studies, ability to handle imbalanced data, and interpretability. They also show strong resilience when working with datasets that have class imbalance (Taghizadeh-Mehrjardi et al., 2020; Senagi et al., 2017; Sarmadian et al., 2014). Advanced models like SVM, XGBoost, and Deep Learning were not included due to:

- **SVM:** Computationally intensive for multi-class and large datasets.
- **XGBoost:** Requires extensive tuning; standard GB was simpler and effective.
- **Deep Learning:** Needs larger datasets and more resources than available.

These exclusions are justified by the study's focus on practical, interpretable models.

3.7. Predictive Model Evaluations

Model evaluation plays an essential role in the development of machine learning systems, as it assesses how accurately the model can make predictions on data it hasn't previously encountered. This step is vital for understanding the model's generalization capabilities and ensuring reliable performance in real-world applications.

This study uses cross-validation techniques to prevent overfitting and obtain reliable performance metrics. Specifically, Stratified K-Fold Cross-Validation is implemented, as it ensures that each fold preserves the original dataset's class distribution a necessary consideration due to the imbalanced nature of the data collected.

3.7.1. Performance Evaluation Metrics for Predictive Modeling

Assessing the effectiveness of a predictive model is a continuous process that aims to achieve optimal accuracy. Various commonly used metrics help evaluate classification models, such as accuracy, precision, recall, F1-score, and the confusion matrix. The confusion matrix specifically offers a detailed summary of a model's predictions by dividing them into four distinct groups, helping to better understand its overall performance.

True Positives (TP): These refer to positive tuples that have been correctly identified.

True Negatives (TN): The term "true negatives" (TN) describes the negative tuples that the classifier has accurately classified.

False Positives (FP): These refer to negative tuples that have been misclassified as positive.

False Negatives (FN): Positive tuples that have been mistakenly classified as negative.

Accuracy (ACC) Accuracy measures the overall proportion of correct predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{Equation 3.1}$$

Precision measures the proportion of tuples that are truly classified as positive. It can be calculated as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{Equation 3.2}$$

Recall quantifies the fraction of positive tuples that are classified correctly. It is also known as sensitivity or the true positive rate,

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{Equation 3.3}$$

F1-score merges precision and recall into one measure, making it especially valuable for evaluating models on imbalanced datasets.

$$F1 - score = 2 * \left(\frac{\text{precision} * \text{Recall}}{\text{precision} + \text{Recal}} \right) \quad \text{Equation 3.4}$$

Confusion Matrix In this study, it applied to a five-class classification system, where land suitability is categorized as highly suitable, moderately suitable, marginally suitable, currently not suitable, and permanently not suitable.

Table 3.3 Five class classifications Confusion matrix

Predicted Values						
Actual value		S1	S2	S3	N1	N2
	S1	True S1	False S2	False S3	False N1	False N2
	S2	False S1	True S2	False S3	False N1	False N2
	S3	False S1	False S2	True S3	False N1	False N2
	N1	False S1	False S2	False S3	True N1	False N2
	N2	False S1	False S2	False S3	False N1	True N2

CHAPTER FOUR

4. PROPOSED SOLUTION

4.1. Introduction

This chapter outlines the proposed machine learning-based framework developed to predict land suitability for cereal crops. It aims to resolve the issues highlighted earlier in the study and address the research questions by filling gaps identified in the literature. The overall framework of the system is described to explain the role and purpose of each individual component within the proposed solution.

4.2. System Architecture for Land Suitability Prediction

The system starts by collecting raw data and preparing it through several steps. This includes removing errors, filling in missing information, and converting any non-numeric data so that it can be used by machine learning models. After that, the most important features are selected using methods like univariate selection, recursive feature elimination, and forward selection. This helps improve model accuracy and speed.

The preprocessed dataset is used to train three machine learning models: Random Forest, Gradient Boosting, and K-Nearest Neighbors. Their performance is tested using a method called stratified 10-fold cross-validation. This ensures that all types of land classes are well represented in each test, leading to more reliable results (Novaes et al., 2021).

To find out which model works best, the system uses different performance measurements. After training and testing, the final model is deployed in a real-world setting where it can give quick and accurate recommendations. Farmers can enter data like soil type, rainfall, temperature, slope, and the type of cereal crop they want to grow. The system will then classify the land into one of five categories: highly suitable, moderately suitable, marginally suitable, currently not suitable, or permanently not suitable. This helps farmers choose the right crop such as maize, teff, wheat, barley, or sorghum for their land before planting.

This system also supports experts by giving them a fast and reliable way to assess land using key factors like soil quality, topography, and climate.

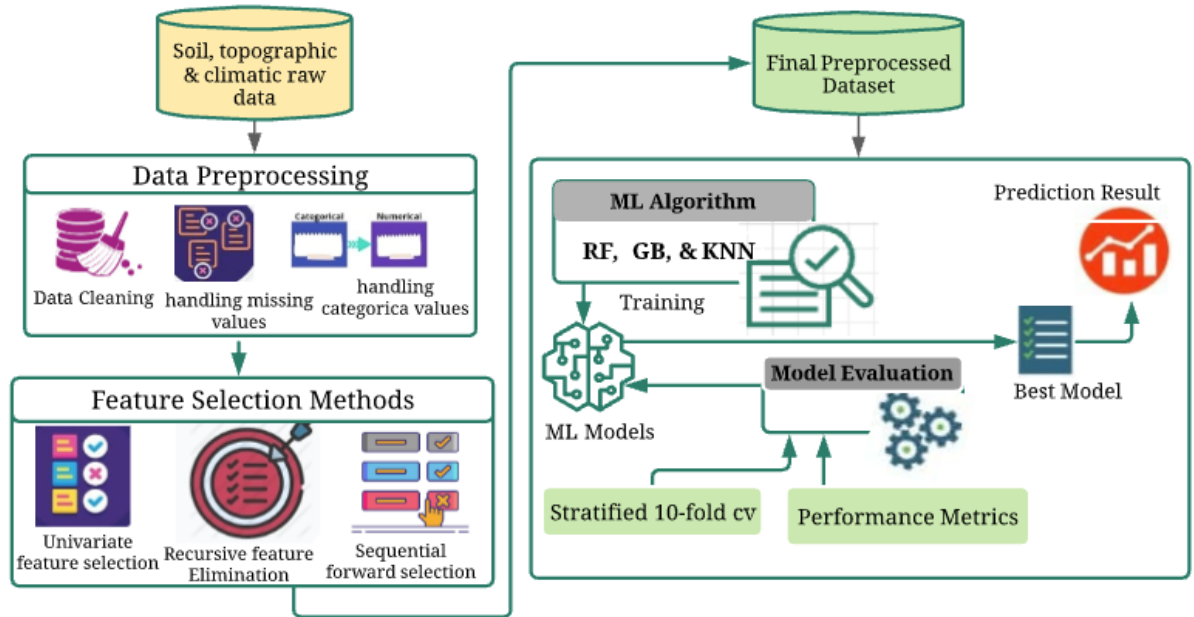


Figure 4.1 Architecture of proposed system

4.3. Data Preprocessing

Data preprocessing is a fundamental step that involves preparing raw data for machine learning by performing tasks such as cleaning, handling missing values, and managing categorical variables. In this study, the dataset after initial preparation is saved in a Comma-Separated Values (CSV) format with a .csv file extension. This format is particularly suitable because Python's built-in CSV module enables efficient parsing and manipulation of data during preprocessing.

The dataset collected for this research included inconsistencies and missing values that required correction and imputation. Additionally, categorical attributes such as soil texture and crop type needed to be encoded into numerical form to be compatible with machine learning algorithms. The preprocessing process ensures that the data is in a structured and consistent format, making it suitable for accurate modeling. A flow diagram illustrating the data preprocessing steps is presented below

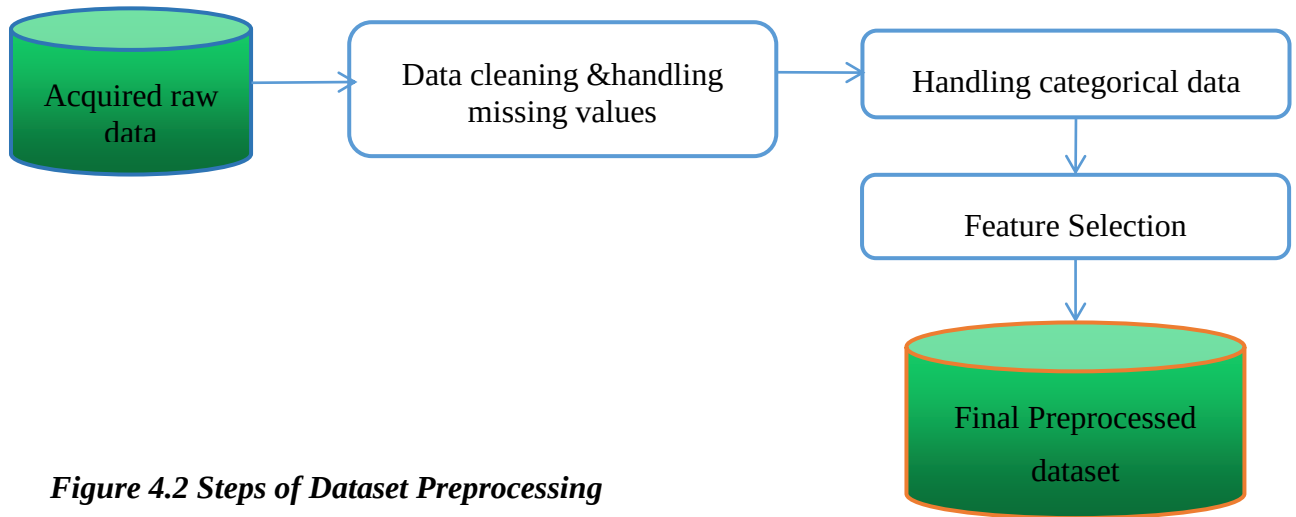


Figure 4.2 Steps of Dataset Preprocessing

4.3.1. Data Cleaning and Handling Missing Values

Data cleaning is an essential task aimed at correcting or eliminating inaccurate, improperly formatted, or incomplete entries within a dataset. The dataset used in this study contained a few instances of such errors, which were removed to maintain data quality. Furthermore, the dataset contained a minimal amount of missing data. These were addressed through data imputation techniques. Specifically, missing numerical values were filled using either the mean or median of the corresponding attribute based on the available non-missing data. No missing values were found in the categorical attributes of the dataset.

4.3.2. Handling Categorical Data

In this study, categorical variables like soil texture and crop name, which cannot be directly interpreted by machine learning models, were successfully transformed into numeric form using label encoding. Each unique category was assigned an integer value from 0 to $n-1$, enabling the algorithms to effectively utilize these features during training.

4.3.3. Feature Selection

This study employs three feature selection techniques to identify the most relevant predictors for land suitability classification. These techniques identified the most significant features, reduced redundancy and noise, and contributed to improved model accuracy while lowering the risk of overfitting and computational overhead.

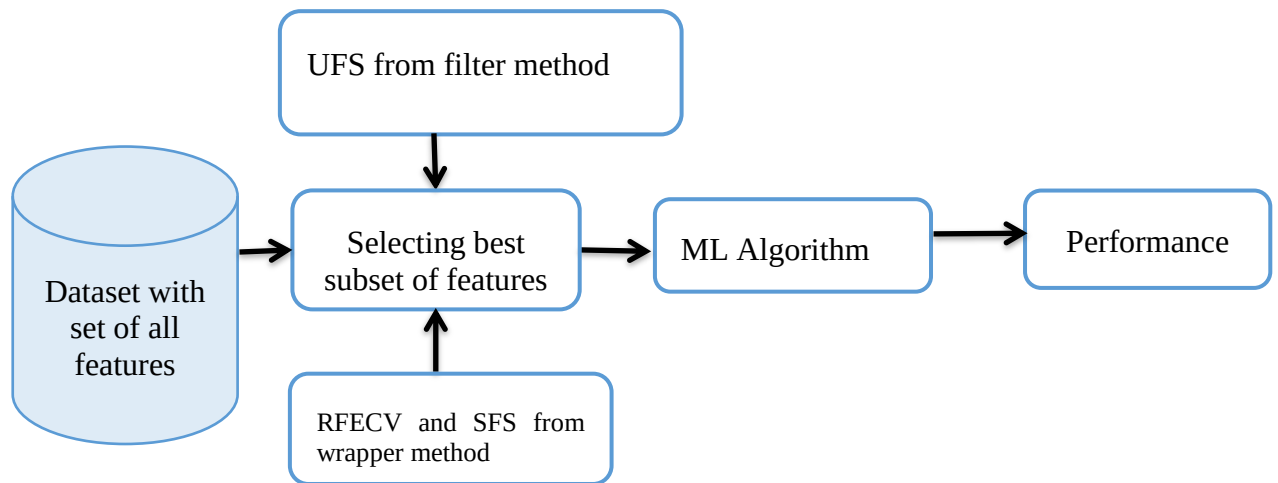


Figure 4.3: Feature Selection Workflow and Model Development

Univariate Feature Selection (UFS): Also referred to as ANOVA-based selection, this method uses statistical tests to independently rank and select the top-performing features using the SelectKBest function from Scikit-learn, without considering inter-feature relationships (Drotár et al., 2015).

Recursive Feature Elimination with Cross-Validation (RFECV) is a technique that improves model performance by selecting the most important features through iterative elimination and cross-validation. It helps reduce overfitting and enhance accuracy, particularly in high-dimensional datasets (Guyon et al., 2002).

Sequential Forward Selection (SFS): This method begins with no features and incrementally adds the feature that most reduces classification error at each step. The process continues until a pre-defined number of features is selected, optimizing model performance through repeated evaluation (Marcano-Cedeño et al., 2010).

4.4. Machine Learning Model Building

To automate land suitability classification a multi-class problem this study applies GB, RF and K_NN algorithms. Each model is trained and evaluated using stratified 10-fold cross-validation and various performance metrics to identify the most accurate model for final implementation.

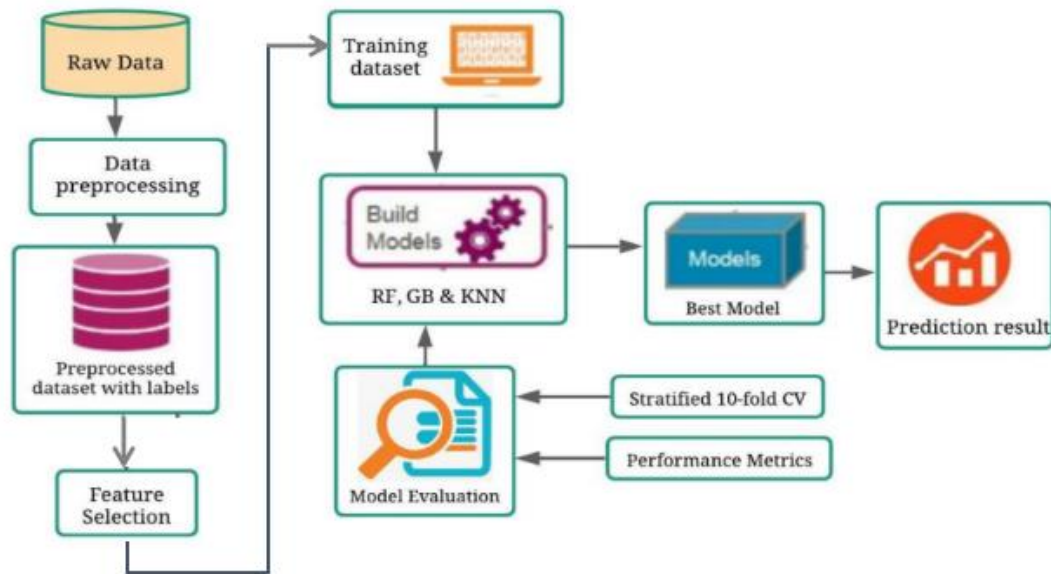


Figure 4.4 Model Building Flow Diagram.

Random Forest(RF): is operates by randomly selecting subsets of the training data to grow individual trees. Each tree independently provides a prediction, and the final decision is made based on the majority vote among them. Increasing the number of trees generally enhances model accuracy and robustness, making RF suitable for handling complex datasets.

Gradient Boosting Classifier (GB): is a sequential ensemble approach that integrates several weak learners commonly shallow decision trees into a single, highly predictive model. Each learner is trained to correct the mistakes of its predecessor, resulting in continuous performance improvement. This technique is known for its high predictive power, particularly in complex classification tasks, and is effective at managing both bias and variance, thereby reducing the risk of overfitting and underfitting (Bentéjac et al., 2021).

K-Nearest Neighbors (KNN):It operates on the principle that similar instances exist in close proximity within the feature space. The algorithm assigns a class to a new data point based on the majority label among its k closest neighbors, determined by calculating distances commonly using Euclidean or Manhattan metrics between the points. Given two feature vectors $A = (x_1, x_2, \dots, x_m)$ and $B = (y_1, y_2, \dots, y_m)$, these distance functions are used to assess similarity (Hu et al., 2016; Li & Zhao, 2020).

$$Euclidean\ dist(A, B) = \sqrt{\frac{\sum_{i=1}^m (x_i - y_i)^2}{m}} \quad \text{Equation 4.1}$$

Manhattan distance, also referred to as city block distance, is a widely utilized distance measure in the K_NN algorithm. It computes the separation between two points by summing the absolute differences of their Cartesian coordinates. The term "city block" originates from the way this metric simulates movement through a grid-patterned city, where traversal is limited to horizontal and vertical directions.

$$Manhattan\ dist(A, B) = \sum_{i=1}^m |x_i - y_i| \quad \text{Equation 4.2}$$

The number of neighboring data points taken into account while classifying a target instance is indicated by the K value in KNN. Choosing the right K, along with an appropriate distance metric and weighting approach, is vital for achieving high model accuracy. These parameters are refined through hyperparameter tuning methods, which aim to minimize classification errors and enhance overall model effectiveness. For a detailed overview of KNN's functioning, refer to Wang (2019).

4.4.1. Hyperparameter Tuning

Hyperparameter tuning is the process of identifying the optimal set of hyperparameters that enables a machine learning model to achieve its best possible performance. Proper configuration of these parameters is essential to achieving optimal model accuracy. Among the various methods available, Grid Search and Random Search are two of the most widely adopted approaches.

Grid Search is a conventional method that systematically explores a predefined subset of the hyperparameter space. It evaluates every possible combination of parameters in the grid, which can be computationally expensive, especially in high-dimensional spaces. However, its advantage lies in its simplicity and ability to be easily parallelized since the evaluations are independent of one another (Liashchynskiy, 2019).

In contrast, Random Search randomly selects combinations of hyperparameter values rather than exhaustively searching the entire space. This method is often more efficient than Grid Search, especially when only a few hyperparameters significantly influence model performance (Liashchynskiy, 2019). In this study, Random Search was employed for hyperparameter optimization.

4.5. Model Evaluation and Prototyping

To evaluate the performance of the developed models on unseen data, stratified 10-fold cross-validation was applied. This method maintains the original class distribution within each fold, allowing for a more balanced and dependable assessment. The models were trained on the complete dataset, and their predictive performance was evaluated using key metrics, including accuracy, precision, recall, F1-score, and confusion matrix. These indicators offered a detailed understanding of each model's accuracy and consistency in classifying land suitability.

4.5.1. Prototyping

Once the machine learning models are developed and evaluated, the next step is prototyping which involves testing the model in a real-world environment. After identifying the best-performing classifier, it is deployed using a Flask web server integrated with an html-based user interface. This prototype allows end-users such as agricultural experts, land evaluators, and farmers to interact with the model. By inputting soil test results, climate data, and topographic features, users can obtain automated and accurate land suitability predictions, enhancing decision-making in agricultural planning and management.

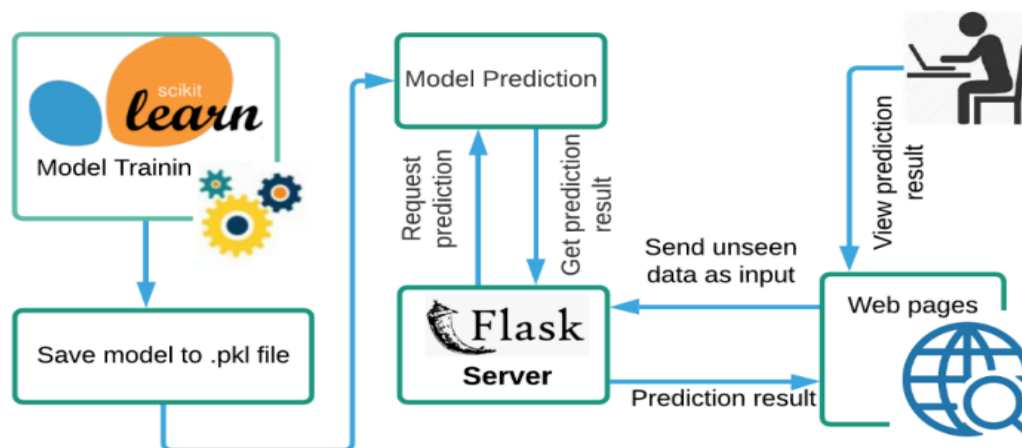


Figure 4.5 Land suitability prediction Deployment Architecture

CHAPTER FIVE

5. RESULT AND DISCUSSION

5.1. Overview

This chapter discusses the findings from the application of machine learning algorithms to evaluate land suitability for cereal crop cultivation. It starts by briefly describing the tools and technologies used during model development and deployment through a Flask-based web interface. The chapter then summarizes the dataset characteristics, including the distribution of suitability classes, and evaluates model performance with and without applying feature selection methods. Finally, it identifies the most effective model, which is then deployed, and compares the findings with existing studies to highlight the contributions and relevance of this research.

5.2. Working Environment

The implementation and experimentation of this study were carried out using the Python programming language, which was employed throughout all phases from data preprocessing to model development and evaluation. Python was chosen due to its widespread use and powerful libraries for machine learning applications. Additionally, an HTML-based web interface was developed to enable user interaction with the deployed model, allowing users to easily input data and view prediction results.

Table 5.1: Tools and Libraries Used in the Study

Tool/Library	Version	Description
Anaconda Navigator	4.10.1	A GUI-based tool for managing environments, packages, and launching applications without using the command line (Anaconda, 2021).
Jupyter Notebook	6.1.4	The Jupyter Notebook is a browser-based tool designed for interactive computing, enabling code writing, data visualization, and documentation within a single interface (Kluyver et al., 2016).
Python	3.8.5	Python is a versatile and user-friendly programming language widely used in scientific computing, especially for data analysis and artificial intelligence tasks (Van Rossum & Drake, 2009).
Microsoft Excel	2016	Spreadsheet software used for organizing and

		preparing datasets.
Microsoft Word	2016	Word processing software used to prepare thesis documentation.
Scikit-learn	0.24.1	Scikit-learn is an accessible and powerful Python package offering tools for implementing machine learning algorithms and performing data analysis (Pedregosa et al., 2011).
Pandas	1.1.3	A Python library for fast and flexible data manipulation and analysis (McKinney, 2010).
NumPy	1.19.2	A core package for numerical computing in Python, supporting large multi-dimensional arrays (Harris et al., 2020).
Matplotlib	3.3.2	Matplotlib is a widely-used Python plotting library that supports generating various types of data visualizations, including graphs and charts (Hunter, 2007).
Seaborn	0.11.0	Seaborn is an advanced Python library based on Matplotlib, that provides a simple, high-level interface for creating attractive and informative statistical graphics. (Waskom, 2021).
Flask	1.1.2	Flask is a lightweight and flexible Python web framework designed for building web applications quickly and with minimal code(Grinberg, 2018).

5.3. Deployment Environment

The implementation was carried out on a personal desktop equipped with an Intel(R) Celeron(R) CPU N2820 @ 2.13GHz, supported by 4 GB of RAM and a 931.51 GB storage drive. The computing environment was based on a 64-bit Windows 11 Pro operating system utilizing an x64 architecture. All Python libraries and tools mentioned earlier were installed and run on this setup.

5.4. Dataset Description

This study utilized data supplied by the Engineering Corporation of Oromia (ECO), a multidisciplinary organization based in Akaki Kality, Addis Ababa, Ethiopia. ECO is known for delivering services in areas such as agricultural development, professional training, land use evaluation, irrigation system design, and laboratory assessments of soil and water quality (ECO, 2021). The dataset specifically originates from a 2016 soil

testing project conducted on behalf of the Oromia Rural Land Administration and Use Bureau, concentrating on the Kellem Wollaga Zone. This dataset includes both raw soil test results and supporting survey reports that assisted in parameter understanding and accurate labeling.

The study aimed to classify land into five suitability levels for cereal crop cultivation. Accordingly, data spanning soil, climate, and topography was collected, consisting of 2024 records and 15 attributes. The main features are presented in table below.

Table 5.2 Description of dataset features

No.	Symbol	Feature Full Name	Type	Description
1	ST	Soil Texture	Categorical	Soil texture describes the mix of sand, silt, and clay in the soil, affecting its ability to retain water and support plant roots.
2	Tmin	Mean Minimum Temperature	Numerical	Mean minimum temperature indicates the average of the lowest temperatures recorded each day over a certain period.
3	Tmax	Mean Maximum Temperature	Numerical	Mean maximum temperature represents the typical daily peak temperature averaged over a set time.
4	RF	Annual Rainfall	Numerical	Total yearly precipitation (mm).
5	SD	Soil Depth	Numerical	Depth of soil available for root growth (cm).
6	SLP	Slope	Numerical	Land inclination percentage; influences erosion and drainage.
7	pH	Soil Acidity/Alkalinity	Numerical	Measure of soil pH level.
8	EC	Electrical Conductivity	Numerical	Indicates soil salinity level (dS/m).
9	K	Potassium Content	Numerical	Level of available potassium in soil (ppm).

10	CEC	Cation Exchange Capacity	Numerical	Soil's ability to hold and exchange cations (meq/100g).
11	TN	Total Nitrogen	Numerical	Total nitrogen content in soil (%).
12	OM	Organic Matter	Numerical	Organic content in the soil (%).
13	AvP	Available Phosphorus	Numerical	Level of phosphorus available for plant use (ppm).
14	CN	Crop Name	Categorical	Type of cereal crop (e.g., maize, teff, sorghum, wheat, barley).
15	SC	Suitability Class	Categorical	Target class representing land suitability (e.g., highly, moderately, etc.).

5.5. Dataset Class Distributions Result

The dataset utilized in this research comprises 2,024 entries, each classified according to the suitability levels for growing key cereal crops, including wheat, maize, sorghum, barley, and teff. These instances are categorized into five suitability classes based on the FAO land evaluation framework. Specifically, 271 entries fall under the Highly Suitable (S1) category, indicating optimal conditions for crop production. Another 362 instances are labeled as Moderately Suitable (S2), meaning the land can support crop growth but with some limitations. A total of 430 records are classified as Marginally Suitable (S3), signifying that the land is usable but has considerable constraints. Meanwhile, 476 instances are marked as Currently Not Suitable (N1), suggesting that the land is unsuitable for farming under present conditions without significant improvement. Lastly, 485 data entries are categorized under the Permanently Not Suitable (N2) class, signifying that these areas possess inherent limitations such as poor soil properties, extreme topography, or unfavorable climate conditions that render them unsuitable for crop production, even with substantial improvements or interventions. The detailed distribution of these classes is graphically presented in the figure below to provide a clearer understanding of the land suitability status within the study area.

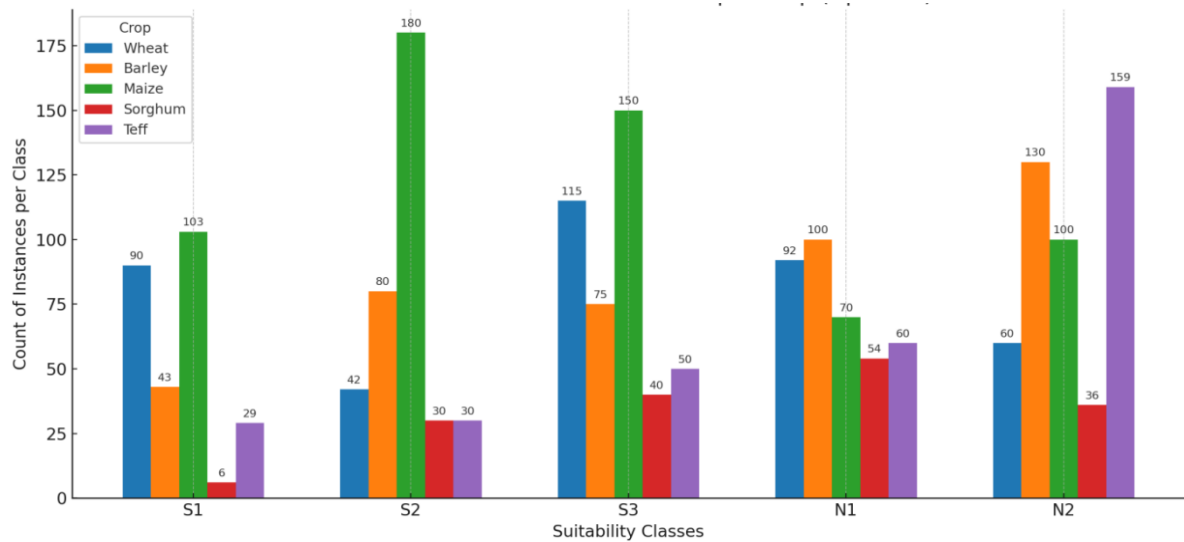


Figure 5.1 *distribution of dataset in each class for cereal crops*

Figure 5.1 illustrates a mild class imbalance in the dataset related to cereal crop suitability. This imbalance arises because a larger portion of land in the study region falls under unsuitable categories either temporarily or permanently compared to those classified as highly, moderately, or marginally suitable for cultivation.

5.6. Preprocessing Implementations

Preprocessing includes managing missing data, encoding categorical variables, and performing feature selection, all implemented using Python. A sample of the preprocessing code is provided in Appendix C:1.

5.7. Results of Feature Selection

This study applied three advanced feature selection methods UFS, RFECV, and SFS to identify the most relevant features and improve model accuracy. The selected features for each method and algorithm are summarized in the upcoming tables

Table 5.4 *Feature Selection Outcomes for Each Machine Learning Model*

Feature Selection Methods	ML Models	Selected features	Original features
Ufs	RF	13	15
	Gb	14	15
	Knn	14	15
Rfecv	RF	13	15

	Gb	6	15
Sfs	RF	9	15
	Gb	10	15
	Knn	6	15

5.8. Model Performance and Evaluation Results

The model evaluation employed a stratified 10-fold cross-validation approach, dividing the dataset into ten equal subsets. During each iteration, one subset was designated as the test set, while the remaining nine were utilized for training the model. The objective was to classify land into five different suitability categories, and the models' effectiveness was measured using multiple classification performance metrics.

5.8.1. Evaluation of Model Performance with Feature Selection

To improve the predictive capability of the model, three distinct feature selection approaches UFS, RFECV, and SFS were employed. The subsets of features obtained through each method were subsequently used to train the Gradient Boosting algorithm. The effectiveness of the model was then validated using a stratified 10-fold cross-validation strategy.

5.8.2. Performance of GB Model with F S Methods

The Gradient Boosting (GB) model was tested with the input features selected through UFS, RFECV, and SFS to examine how each technique influenced predictive performance. Using stratified 10-fold cross-validation, the performance of the model was quantified through confusion matrices and standard classification metrics, enabling a detailed comparison of each feature selection method's impact on model accuracy and reliability.

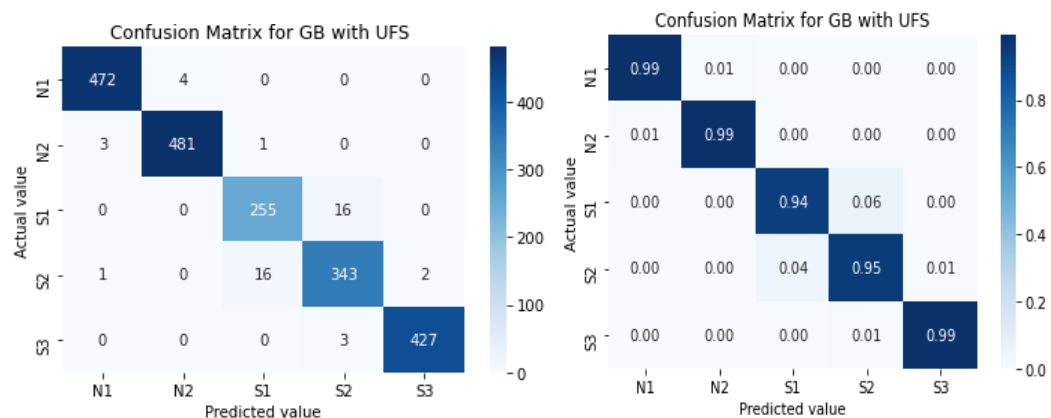


Figure 5.5 Performance of GB Model with UFS without and with Normalization

Using UFS, the GB model correctly classified most instances across all classes: 472 of 476 (N1), 481 of 485 (N2), 255 of 271 (S1), 343 of 362 (S2), and 427 of 430 (S3). However, it misclassified a few, including N1 as N2, and S1 as S2. Despite high accuracy, models using UFS showed slightly lower performance compared to those trained on the full feature set.

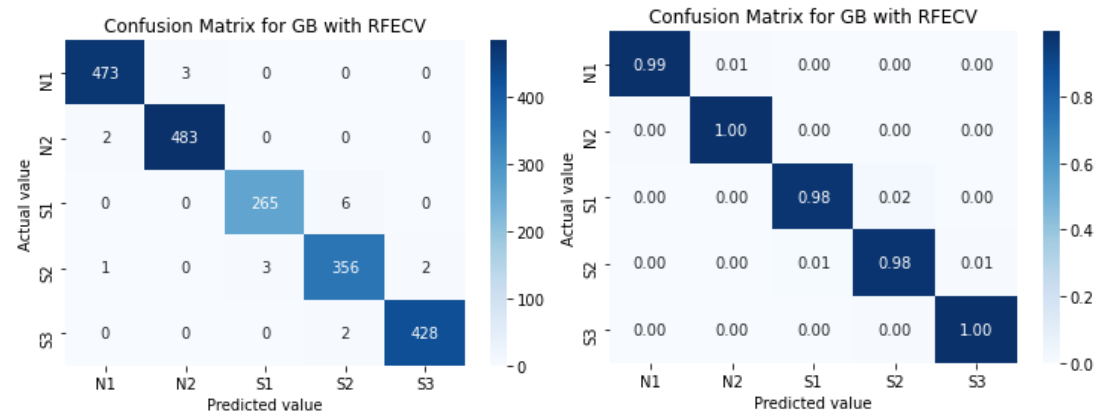


Figure 5.6 Performance of GB Model with RFECV without and with Normalization

With RFECV, GB demonstrated improved classification: 473 of 476 (N1), 483 of 485 (N2), 265 of 271 (S1), 356 of 362 (S2), and 428 of 430 (S3) were correctly predicted. Misclassifications were reduced, indicating better model performance than when using UFS or even the original features.

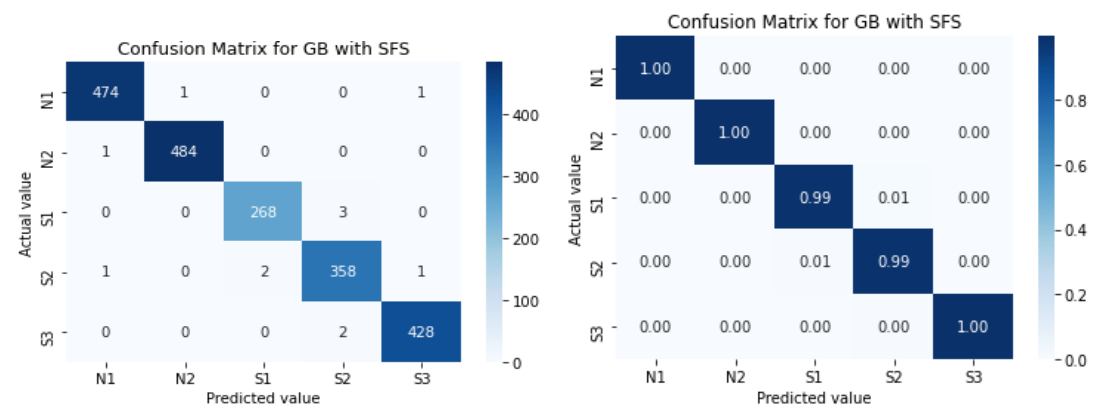


Figure 5.7 Performance of GB Model with SFS without and with Normalization

SFS yielded the best results for the GB model, with 474 of 476 (N1), 484 of 485 (N2), 268 of 271 (S1), 358 of 362 (S2), and 428 of 430 (S3) classified accurately. Minimal errors were observed, and the normalized confusion matrix confirmed consistently high accuracy across all classes.

Overall, among the feature selection methods tested, SFS proved to be the most effective, significantly enhancing the predictive power of the GB model. In comparison to UFS, RFECV, and even the full feature set, models trained with SFS-selected features particularly GB achieved the highest scores across all evaluation metrics. The table below summarizes model performance with and without feature selection using 10-fold cross-validation.

Table 5.5 summarized performance of all models with and without feature selection

Algorithms	Number of features	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RF	15	92.75	92.72	92.71	92.70
RF with UFS	13	93.45	93.43	93.42	93.41
RF with RFECV	13	93.59	93.57	93.56	93.55
RF with SFS	9	93.96	93.93	93.92	93.91
GB	15	94.15	94.13	94.11	94.1
GB with UFS	14	94.09	93.37	93.25	93.03
GB with RFECV	6	94.06	94.01	93.98	93.96
GB with SFS	10	94.41	94.37	94.34	94.35
KNN	15	92.74	92.3	92.4	92.3
KNN with UFS	14	92.64	92.14	92.25	92.14
KNN with SFS	6	93.33	92.97	92.98	92.92

5.9. Results of Hyperparameter Tuning

Although automatic hyperparameter tuning can be computationally intensive and memory-demanding, this study employed cross-validated randomized search rather than relying on manual tuning, which is often time-consuming and suboptimal. Manual tuning

typically lacks efficiency and may lead to inferior model performance due to limited parameter exploration.

To ensure robust model optimization, both Grid Search CV and Randomized Search CV methods were tested. The results demonstrated that Randomized Search CV outperformed Grid Search in terms of efficiency and effectiveness. Therefore, the hyperparameters recommended by the Randomized Search CV approach were adopted. The optimal hyperparameter values selected for each model through this method are summarized in Table 5.6 below.

Table 5.6: Selected Hyperparameters for RF,GB and KNN

Parameter RF	Value	Parameter GB	Value	Parameter KNN	Value
n_estimators	1200	n_estimators	500	n_neighbors	7
max_depth	15	max_depth	11	weights	distance
criterion	entropy	criterion	friedman_mse	algorithm	auto
bootstrap	True	min_samples_leaf	5	leaf_size	30
max_features	auto	min_samples_split	6		
min_samples_split	2				

5.10. Comparison

Numerous studies have implemented ML techniques in agriculture to improve crop productivity, enhance food security, alleviate poverty, and support economic development, particularly for farmers. Many of these works focus on evaluating land suitability based on soil, climate, and topographic data, following the FAO guidelines (Welteji, 2018). Traditional land evaluation methods such as GIS-based suitability mapping and Analytic Hierarchy Process (AHP) have long been used. However, they are time-consuming, expert-dependent, and not predictive. In contrast, this study's use of machine learning Random Forest, Gradient Boosting, and KNN demonstrates the

advantages of data-driven prediction, scalability, and automation. Although a direct numerical comparison was not implemented, this study proposes that machine learning can complement traditional GIS-AHP approaches by automating classification and improving repeatability. This research offers a contribution by introducing a machine learning-driven method. To assess land suitability for important cereal crops.

To highlight the study's strength, the findings are compared to various other research discussed in the literature. Komolafe et al. (2019) successfully assessed cassava land suitability using SVM and Decision Trees on a small dataset of 252 records, achieving 87.5% accuracy after 10-fold cross-validation. However, their analysis was unclear on how the dataset was categorized based on crop requirements. In contrast, our study utilized a bigger dataset of 2024 records with precise crop requirement mapping.

Similarly, Senagi et al. (2017) applied multiple classifiers including RF, LR, LDA, KNN, GNB, and SVM for sorghum suitability and reported 96% accuracy using RF with 10-fold CV. However, they did not apply feature selection, unlike this study which included three feature selection methods.. This work attained an accuracy of 94.41% using Gradient Boosting in combination with SFS for a more complex five-class classification task.

Sarmadian et al. (2014) employed SVM to assess rainfed wheat suitability using binary classification. While they achieved an RMSE of 3.72 and R^2 of 0.84, they did not use feature selection or multi-class classification, both of which are integrated in this work.

In another study, Bhimanpallewar and Narasinagrao (2017) utilized a significantly large dataset consisting of 74,737 records to assess Jowar (sorghum) suitability using a Decision Tree model. Using a basic holdout validation procedure, they obtained a high accuracy of 98.9%. Despite the size of their dataset, the current research surpasses it in both accuracy and methodological rigor by integrating advanced feature selection and hyperparameter tuning. The Gradient Boosting algorithm, combined with Sequential Forward Selection and validated using stratified 10-fold cross-validation, delivered an exceptional accuracy of 94.41%.

Additionally, Ogunde and Olanbo (2017) developed a web-based decision support system that employed a Decision Tree model for cassava land suitability classification. However, their model did not incorporate vital environmental factors such as climatic and

topographic variables, which are critical for accurate prediction. As a result, their model attained a relatively modest accuracy of 76.5%, highlighting the limitations of using an incomplete feature set.

A comprehensive comparison of these previous works and the proposed method in this study is summarized in Table 5.7 below

Table 5.7 Comparison Between the Proposed Approach and Prior Research Studies

Reference	Algorithm(s)	Feature Selection	Target Class	Performance	Year
Komolafe et al. (2019)	SVM, DT	Information Gain	5, classes	DT achieved 87.5% accuracy	2019
Senagi et al. (2017)	RF, LR, LDA, KNN, GNB, SVM	Not Specified	4, classes	RF achieved 96% accuracy	2017
Sarmadian et al. (2014)	SVM	Not Specified	Binary	RMSE = 3.72, R ² = 0.84	2014
Bhimanpallewar & Narasinagrao (2017)	DT	Not Specified	5, classes	DT achieved 98.9% accuracy	2017
Ogunde & Olanbo (2017)	DT	Not Specified	4, classes	DT achieved 76.5% accuracy	2017
Proposed Method (This Study)	RF, GB, KNN	UFS, RFECV, SFS	5, classes	GB with SFS achieved 94.41%	2025

This comparison highlights that the proposed approach using Gradient Boosting with Sequential Forward Selection (GB + SFS) yields superior performance compared to earlier works, demonstrating its effectiveness for multi-class land suitability prediction for cereal crops.

5.11. Result and Discussion

The research explored the effectiveness of three widely used ML algorithms: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbors (KNN). Each model was

tested in two configurations both with and without the use of feature selection strategies. The three feature selection techniques employed in this study included Univariate Feature Selection (UFS), which evaluates features individually based on statistical tests; Recursive Feature Elimination with Cross-Validation (RFECV), which iteratively removes less important features while validating performance; and Sequential Forward Selection (SFS), which incrementally builds the best feature subset by adding one feature at a time based on model improvement.

This comprehensive approach was designed not only to evaluate model accuracy but also to optimize feature relevance, reduce model complexity, and improve generalization performance in predicting land suitability for cereal crops.

Among the models, Gradient Boosting consistently delivered superior results, particularly when paired with SFS, achieving the highest performance (e.g., 94.41% accuracy and 94.35% F1-score) using only 10 selected features. UFS contributed little to performance gains, while KNN remained the least effective model, even after applying feature scaling.

Given its outstanding performance, the GB model integrated with SFS is recommended for predicting land suitability. To support real-world application, a web-based prototype was developed and deployed, offering a user-friendly interface for agricultural professionals and land evaluators. As illustrated in Figure 5.2, this GUI allows users to interact with the deployed model, supporting decision-making in practical agricultural planning.

The screenshot displays a web-based interface titled "Land Suitability Prediction". It features a grid of input fields for various parameters: Soil Texture (class) is set to "Clay Loam"; Mean maximum Temperature (°C) is "15"; Soil Depth (cm) is "200"; EC (dS/m) is "0.33"; Available Phosphorus (PPM) is "36.08"; Mean minimum Temperature (°C) is "12.5"; Annual Rainfall (mm) is "591.7"; Maximum Slope (%) is "5"; Potassium (%) is "1.75"; and Crop Name is set to "Wheat". Below the input fields are two buttons: "predict" and "Clear". At the bottom, a green message box with a close icon (x) displays the text: "Land is Highly Suitable[S1] for the selected crop".

Figure 5.2 GUI of the Land Suitability Prediction Prototype

CHAPTER SIX

6. CONCLUSION AND FUTURE WORK

6.1. Conclusion

This study aimed to predict land suitability for cereal crop production by leveraging machine learning (ML) techniques. Comprehensive datasets comprising soil physical and chemical properties, climatic data and topographic features were acquired from the Engineering Corporation of Oromia. These data were annotated with expert guidance and validated by agricultural professionals, resulting in a final dataset of 2,024 records and 15 attributes, including the target suitability class.

To tackle slight class imbalance in the dataset, the study employed several strategies: using the `class_weight='balanced'` parameter in the Random Forest model, selecting ensemble Gradient Boosting due to its robustness against imbalance, and incorporating K-Nearest Neighbors (KNN), which is less sensitive to class distribution. All models were evaluated using stratified 10-fold cross-validation to ensure equitable class representation across training and testing sets.

Data preprocessing included mean imputation for missing values and label encoding for categorical variables. Three feature selection methods were used, and models (Random Forest, Gradient Boosting, and KNN) were trained with and without feature selection. Hyperparameters were optimized using Randomized Search CV and some evaluation metrics. The findings indicated that Gradient Boosting consistently delivered superior performance compared to the other algorithms.

The best performance was achieved by Gradient Boosting combined with SFS, which yielded 94.41% accuracy, 94.37% precision, 94.34% recall, and a 94.35% F1-score using only 10 features. Therefore, the GB + SFS model is recommended for land suitability prediction, and a prototype application was developed to facilitate its use by agricultural experts, evaluators, and farmers.

6.2. Recommendation

Given the promising outcomes of this study, several recommendations and future directions are proposed:

- **Expand Model Coverage:** Future studies should include additional algorithms such as SVM, XGBoost, and deep learning to assess their performance in diverse agro-ecological zones.
- **Enhance Geographic Scope:** Include data from other Ethiopian regions to validate and generalize them.
- **Collaborate with Government Agencies:** Use the deployed prototype in land-use planning programs at the district or regional level.
- **Continuous Expert Involvement:** Engage agricultural experts for periodic review and updating of suitability thresholds and models.
- **Centralized Data Storage:** Agricultural institutions and research centers in Ethiopia often collect valuable data for specific studies, which are not systematically archived. Establishing a centralized data repository would ensure long-term preservation and accessibility of soil, climate, and topographic datasets, thereby supporting future research and decision-making.
- **Real-Time IoT Integration:** Future research should incorporate Internet of Things (IoT) systems for real-time data collection. Field-deployed sensors can directly measure soil nutrients, temperature, and rainfall conditions on-site. These real-time inputs could then feed into the predictive model, providing instant land suitability assessments to farmers via mobile alerts.
- **Crop Coverage Extension:** While this study focused on five staple cereals barley, sorghum, maize, wheat and teff the framework can be extended to other important crops such as rice, chickpeas, and lentils.

Future research should focus on collecting labeled data from other Ethiopian zones to perform external validation and assess the model's transferability. A possible extension is to create a nationwide land suitability prediction system using federated or transfer learning approaches. The integration of ML with spatial layers in GIS platforms could allow wider adoption in land planning offices and rural extension services.

6.3. Contribution of the Study

This study contributes significantly to the field of agricultural land evaluation by introducing a machine learning-based approach that enhances the speed, accuracy, and usability of land suitability classification. Key contributions include:

- Development of an annotated dataset with 2024 labeled records based on FAO standards and validated by agronomy experts.
- Evaluation and comparison of three machine learning algorithms (RF, GB, KNN) combined with three feature selection methods (UFS, RFECV, SFS).
- Identification of Gradient Boosting with Sequential Forward Selection as the best-performing model with an accuracy of 94.41%.
- Deployment of a Flask-based prototype application for real-time prediction of land suitability for five major cereal crops.
- Comparative analysis of machine learning against traditional GIS–AHP and FAO-based evaluation methods.

REFERENCES

- Abebe, S., & Sewnet, A. (2020). Rural land use problems and management options in Debre Tsyon Kebele, Ethiopia. *GeoJournal* 85(1), 145–157.
<https://doi.org/10.1007/s10708-018-9951-7>
- Agidew, A., & Singh, K. N. (2017). The implications of land use and land cover changes for rural household food insecurity in the Northeastern highlands of Ethiopia : the case of the Teleyayen sub - watershed. *Agriculture & Food Security*, 1–14.
<https://doi.org/10.1186/s40066-017-0134-4>
- Anaconda. (2021). Anaconda Navigator. <https://www.anaconda.com>
- Bagherzadeh, A., Ghadiri, E., Reza, A., Darban, S., & Gholizadeh, A. (2016). Land suitability modeling by parametric-based neural networks and fuzzy methods for soybean production in a semi-arid region. *Modeling Earth Systems and Environment*, 2(2), 1–11. <https://doi.org/10.1007/s40808-016-0152-4>
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. In *Artificial Intelligence Review* (Vol. 54, Issue 3). Springer Netherlands. <https://doi.org/10.1007/s10462-020-09896-5>
- Bhimanpallewar, R., & Narasinagrao, M. R. (2017). A machine learning approach to assess crop specific suitability for small/marginal scale croplands. *International Journal of Applied Engineering Research*, 12(23), 2–4.
- Bollegala, D. (2017). Dynamic feature scaling for online learning of binary classifiers. *Knowledge-Based Systems*, 129, 97–105. <https://doi.org/10.1016/j.knosys.2017.05.010>
- Bozdağ, A., Yavuz, F., & Günay, A. S. (2016). AHP and GIS based land suitability analysis for Cihanbeyli (Turkey) County. *Environmental Earth Sciences*, 75(9).
<https://doi.org/10.1007/s12665-016-5558-9>
- Brownlee, J. (2020). Data Preparation for Machine Learning. *Ambiguous Childhoods : Peer Socialisation, Schooling and Agency in a Zambian Village*, 1.1, 232–246.
- Drotár, P., Gazda, J., & Smékal, Z. (2015). An experimental comparison of feature selection methods on two-class biomedical datasets. *Computers in Biology and Medicine*, 66, 1–10. <https://doi.org/10.1016/j.compbiomed.2015.08.010>

ECO. (2021). *ECO RESEARCH TRAINING AND LABORATORY*.

<https://ecoet.org/research-training-and-laboratory/>

Chauhan, D. S. E. (2021). *Decision Tree Algorithm*.

<https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>

Fekadu, E., & Negese, A. (2020). GIS assisted suitability analysis for wheat and barley crops through AHP approach at Yikalo sub-watershed, Ethiopia. *Cogent Food & Agriculture*, 6(1), 1–22. <https://doi.org/10.1080/23311932.2020.1743623>

Gotor, E., Kozicka, M., Pagnani, T., Occelli, M., & Caracciolo, F. (2021). Understanding the link between the Productive Safety Net Program and agrobiodiversity cultivation in Ethiopia. *Policy Brief No. 51. Rome (Italy): Alliance of the Bioversity International and CIAT.*, 51, 1–10.

Gupta, C. (2019). Feature Selection and Analysis for Standard Machine Learning Classification of Audio Beehive Samples,. *Masters Thesis, Utah State University*, 68–70

Jennifer B. Morgan. (2013). *Plant-Soil Interactions: Nutrient Uptake*. (Science Department, York Technical College) & Erin L. Connolly (Department of Biological Sciences, University of South Carolina).

<https://www.nature.com/scitable/knowledge/library/plant-soil-interactions-nutrientuptake-105289112/>

Kahsay, A., Haile, M., Gebresamuel, G., & Mohammed, M. (2018). Land suitability analysis for sorghum crop production in northern semi-arid Ethiopia: Application of GIS-based fuzzy AHP approach. *Cogent Food & Agriculture*, 4(1), 1–4.

<https://doi.org/10.1080/23311932.2018.1507184>

Kamiri, J., & Mariga, G. (2021). Research Methods in Machine Learning: A Content Analysis. *International Journal of Computer and Information Technology*(2279-0764), 10(2). <https://doi.org/10.24203/ijcit.v10i2.79>

Komolafe, E. O., Awoyelu, I. O., & Ojetade, J. O. (2019). *Predictive Modeling for Land Suitability Assessment for Cassava Cultivation*. 9(1), 21–31.

<https://doi.org/10.5923/j.ac.20190901.02>

- Li, Z., & Zhao, C. (2020). KNN algorithm of early warning system for applied research course. *E3S Web of Conferences*, 218. <https://doi.org/10.1051/e3sconf/202021801030>
- Liashchynskiy, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS*. 1–11. <http://arxiv.org/abs/1912.06059>
- Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. *IECON Proceedings (Industrial Electronics Conference)*, 2–3. <https://doi.org/10.1109/IECON.2010.5675075>
- Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, 91, 2–3. <https://doi.org/10.1016/j.procs.2016.07.111>
- Mokarram, M., Hamzeh, S., Aminzadeh, F., & Zarei, A. R. (2015). Using machine learning for land suitability classification. *West African Journal of Applied Ecology*, 23(1), 63–73. <https://doi.org/10.4314/wajae.v23i1>
- Monster, F. (2017). *DK History: Early Farming*. Sandbox Networks, Inc., Publishing. <https://www.factmonster.com/history/us/dk-history-early-farming>
- Mujtaba, H. (2020). *Types of Cross Validation*. GreatLearning. <https://www.mygreatlearning.com/blog/cross-validation/>
- Mylavarapu, R., Bergeron, J., Wilkinson, N., & Hanlon, E. A. (2020). *Soil pH and Electrical Conductivity: A County Extension Soil Laboratory Manual*. 1, 1–10. <https://doi.org/10.32473/edis-ss118-2020>
- Neameh, J. B. (2003). Land evaluation for Land Use Planning with especial attention to sustainable fodder production in the Rouzeh Chai catchment of Orumiyeh area - Iran. *Masters Thesis, INTERNATIONAL INSTITUTE FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION ENSCHEDE, THE NETHERLANDS*, 1–10.
- Ng, S. C., Bongso, T. A., Liow, S. L., Edirisinghe, R., Tok, V., & Ratnam, S. S. (1992). Framework for Land Evaluation. *Journal of Assisted Reproduction and Genetics*, 9(3), 186–189. <https://doi.org/10.1007/BF01203810>
- Novaes, M. T., Ferreira de Carvalho, O. L., Guimarães Ferreira, P. H., Nunes Tiraboschi, T. L., Silva, C. S., Zambrano, J. C., Gomes, C. M., de Paula Miranda, E., Abílio de arvalho Júnior, O., & de Bessa Júnior, J. (2021). Prediction of secondary testosterone deficiency using machine learning: A comparative analysis of ensemble and base

classifiers, probability calibration, and sampling strategies in a slightly imbalanced dataset. *Informatics in Medicine Unlocked*, 23.

<https://doi.org/10.1016/j.imu.2021.100538>

Ogunde, A. O., & Olanbo, A. R. (2017). A web-based decision support system for evaluating soil suitability for cassava cultivation. *Advances in Science, Technology and Engineering Systems*, 2(1), 42–50.

<https://doi.org/10.25046/aj020105>

Onwonga, R., & Freyer, B. (2006). Impact of traditional farming practices on nutrient balances in smallholder farming systems of Nakuru District, Kenya. *Proceedings of the Tropentag Conference*, 11–13.

Perner, P. (2015). Advances in mass data analysis of images and signals in medicine, biotechnology, chemistry and food industry. In *10th International Conference, MDA , Proceedings* (pp. 12–13). Kenya Agricultural and Livestock Research Organization, Kenya.

Ritung S, Wahyunto, Agus F, H. H. (2007). *Land Suitability Evaluation with a case map of Aceh Barat District*. Indonesian Soil Research Institute and World Agroforestry Centre.

Sarmadian, F., Keshavarzi, A., Rooien, A., Zahedi, G., & Javadikia, H. (2014). Support Vector Machines Based-Modeling of Land Suitability Analysis for Rainfed Agriculture. *Journal of Geosciences and Geomatics*, 2(4), 2–3.

<https://doi.org/10.12691/jgg-2-4-4>

Senagi, K., Jouandeau, N., & Kamoni, P. (2017). Using parallel random forest classifier in predicting land suitability for crop production. *Journal of Agricultural Informatics*, 8(3). <https://doi.org/10.17700/jai.2017.8.3.390>

Seyoum Taffesse, A., Dorosh, P., & Gemessa, S. A. (2013). Crop production in Ethiopia: Regional patterns and trends. In *Food and Agriculture in Ethiopia: Progress and Policy Challenges* (Vol. 16, pp. 53–83). <https://doi.org/10.9783/9780812208610.53>

Shaikh, R. (2018). *Feature Selection Techniques in Machine Learning with Python*.

<https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>

- Staff Soil Survey. (2016). *Kellem Wollaga Zone Oromia Region Land Evaluation* (pp. 104–113).
- Tadesse, M., & Negese, A. (2020). Land suitability evaluation for sorghum crop by using GIS and AHP techniques in Agamsa sub-watershed, Ethiopia.
<https://doi.org/10.1080/23311932.2020.1743624>
- Taghizadeh-mehrjardi, R. (2007). Land Suitability Assessment. *Agricultural Sciences & Agronomy | Geoinformatic*, 1–4.
- Taghizadeh-Mehrjardi, R., Nabiollahi, K., Rasoli, L., Kerry, R., & Scholten, T. (2020). Land suitability assessment and agricultural production sustainability using machine learning models. *Agronomy*, 10(4). <https://doi.org/10.3390/agronomy10040573>
- Tahir, M. A. U. H., Asghar, S., Manzoor, A., & Noor, M. A. (2019). A Classification Model for Class Imbalance Dataset Using Genetic Programming. *IEEE Access*, 7, 5–6. <https://doi.org/10.1109/ACCESS.2019.2915611>
- Tashayo, B., Honarbakhsh, A., Akbari, M., & Eftekhari, M. (2020). Land suitability assessment for maize farming using a GIS-AHP method for a semi- arid region, Iran. *Journal of the Saudi Society of Agricultural Sciences*, 19(5), 332–338.
<https://doi.org/10.1016/j.jssas.2020.03.003>
- Tserkezou, I. S. A. and P. (2008). Cereal Waste Management: Treatment Methods and Potential Uses of Treated Waste. *Waste Management for the Food Industries*.
- Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science and Technology*, 26(1), 2–5.
- Wang, L. (2019). Research and Implementation of Machine Learning Classifier Based on KNN. *IOP Conf. Ser.: Mater. Sci. Eng.* 677 052038, 2–6.
<https://doi.org/10.1088/1757-899X/677/5/052038>
- Welteji, D. (2018). A critical review of rural development policy of Ethiopia: Access, utilization and coverage. *Agriculture and Food Security*, 7(1), 1–6.
<https://doi.org/10.1186/s40066-018-0208-y>
- Worldometer. (2021). *Ethiopian Population*. <https://www.worldometers.info/world->

population/ethiopia-population/

- Yohannes, H., & Soromessa, T. (2018). Land suitability assessment for major crops by using GIS-based multi-criteria approach in Andit Tid watershed, Ethiopia. *Cogent Food & Agriculture*, 4(1), 1–29. <https://doi.org/10.1080/23311932.2018.1470481>
- Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308–324. <https://doi.org/10.1016/j.trc.2015.02.019>
- Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, 6, 1–2. <https://doi.org/10.1109/ACCESS.2018.2789428>
- Zichen Wang. (2018). *Practical tips for class imbalance in binary classification*. <https://towardsdatascience.com/practical-tips-for-class-imbalance-in-binaryclassification-6ee29bcdb8a7>

APPENDICES

Appendix A: Selected features by each feature selection methods

Table A.1 List of Parameter selected by each feature Selection methods

Models	List of Selected feature by UFS	List of Selected features by RFECV	List of Selected Features by SFS
RF	'Annual Rainfall(mm)', 'Depth (cm)', 'K', 'Texture', 'mean min_temp', 'max_slope', 'min_slope', 'CEC', 'T_N', 'PH_water', 'mean_max temp', 'O_M', 'Av_p'	'Mean min_temp', 'Mean max_temp', Texture', 'Annual_Rainfall (mm)', 'Depth', 'PH_water', 'K', 'O_M', 'EC', 'T_N', 'Av_p', 'Sodicity (ESP)', 'Crop_name'	'Mean min_temp', 'Texture', 'Annual_Rainfall', 'Depth', 'EC', 'Mean max_temp', 'CEC', 'Sodicity(ESP)', 'Crop_name'
GB	'Annual Rainfall(mm)', 'Depth', 'K', 'Av_p', 'Texture', 'mean min_temp', 'max_slope', 'min_slope', 'CEC', 'T_N', 'PH_water', 'Mean max_temp', 'O_M', 'EC'	'Annual Rainfall(mm)', 'Depth', 'EC', 'K', 'O_M', 'Crop_name'	'Texture', 'mean min_temp', 'mean max_temp', 'Annual Rainfall(mm)', 'Depth', 'max_slope', 'K', 'EC', 'Av_p', 'Crop_name'
KNN	'Annual Rainfall(mm)', 'Depth', 'K', 'Av_p', 'Texture', 'mean min_temp', 'max_slope', 'min_slope', 'CEC', 'T_N', 'PH_water', 'Mean max_temp', 'O_M', 'EC'	-----	'Annual Rainfall(mm)', 'EC', 'K', 'T_N', 'O_M', 'Crop_name'

Appendix B: Result

Appendix B:1 Stratified 10-fold CV result

Accuracy Score of each Fold per iteration of GB

Accuracy of Fold 1 is 94.51

Accuracy of Fold 2 is 93.25

Accuracy of Fold 3 is 94.03

Accuracy of Fold 4 is 93.07

Accuracy of Fold 5 is 94.5

Accuracy of Fold 6 is 92.98

Accuracy of Fold 7 is 93.01

Accuracy of Fold 8 is 92.05

Accuracy of Fold 9 is 93.89

Accuracy of Fold 10 is 94.50

Mean Accuracy of all folds: 94.15

Accuracy Score of each Fold per iteration of GB with UFS

Accuracy of Fold 1 is 94.51

Accuracy of Fold 2 is 93.25

Accuracy of Fold 3 is 94.03

Accuracy of Fold 4 is 93.07

Accuracy of Fold 5 is 94.5

Accuracy of Fold 6 is 93.8

Accuracy of Fold 7 is 94.01

Accuracy of Fold 8 is 92.45

Accuracy of Fold 9 is 94.89

Accuracy of Fold 10 is 93.50

Mean Accuracy of all folds: 94.09

Accuracy Score of each Fold per iteration of GB with RFECV

Accuracy of Fold 1 is 94.01

Accuracy of Fold 2 is 93.2

Accuracy of Fold 3 is 94.0

Accuracy of Fold 4 is 92.0

Accuracy of Fold 5 is 94.5

Accuracy of Fold 6 is 93.5

Accuracy of Fold 7 is 95.0

Accuracy of Fold 8 is 92.45

Accuracy of Fold 9 is 93.8

Accuracy of Fold 10 is 94.0

Mean Accuracy of all folds: 94.06

Accuracy Score of each Fold per iteration of GB with SFS

Accuracy of Fold 1 is 94.51

Accuracy of Fold 2 is 93.25

Accuracy of Fold 3 is 94.03

Accuracy of Fold 4 is 94.07

Accuracy of Fold 5 is 93.5

Accuracy of Fold 6 is 94.5

Accuracy of Fold 7 is 95.01

Accuracy of Fold 8 is 93.45

Accuracy of Fold 9 is 93.89

Accuracy of Fold 10 is 94.01

Mean Accuracy of all folds: 94.41

Appendix B:2 Classification Reports

		precision	recall	f1-score	support
GB	0	0.99	0.99	0.99	476
	1	0.99	0.99	0.99	485
	2	0.99	0.98	0.99	271
	3	0.97	0.99	0.98	362
	4	1.00	0.99	0.99	430
	accuracy			0.99	2024
	macro avg	0.99	0.99	0.99	2024
	weighted avg	0.99	0.99	0.99	2024

GB with UFS

		precision	recall	f1-score	support
GB with RFECV	0	0.99	0.99	0.99	476
	1	0.99	0.99	0.99	485
	2	0.94	0.94	0.94	271
	3	0.95	0.95	0.95	362
		precision	recall	f1-score	support
	0	0.99	0.99	0.99	476
	1	0.99	1.00	0.99	485
	2	0.99	0.98	0.98	271
	3	0.98	0.98	0.98	362
	4	1.00	1.00	1.00	430
	accuracy			0.99	2024
	macro avg	0.99	0.99	0.99	2024
	weighted avg	0.99	0.99	0.99	2024

		precision	recall	f1-score	support
GB with SFS	0	1.00	1.00	1.00	476
	1	1.00	1.00	1.00	485
	2	0.99	0.99	0.99	271
	3	0.99	0.99	0.99	362
	4	1.00	1.00	1.00	430
	accuracy			0.99	2024
	macro avg	0.99	0.99	0.99	2024
	weighted avg	0.99	0.99	0.99	2024

Appendix C: Sample Code

Appendix C:1 Sample code for Preprocessing Implementations

#Feature scaling to build KNN model

```
scaler = MinMaxScaler(feature_range=(0, 1))
X1 = scaler.fit_transform(features)
```

```
#from sklearn.neighbors import KNeighborsClassifier
#instantiate the model
knn=KNeighborsClassifier(metric='manhattan', n_neighbors=7, weights='distance')
#Fit to train the model
knn.fit(X1, target)
```

#Handling categorical values

```
#Handling All categorical features at one
from sklearn.preprocessing import LabelEncoder
lb_encoder = LabelEncoder()
categorical_column = ls_dataset.columns.tolist()
for column in categorical_column:
    if ls_dataset[column].dtype == 'object':
        ls_dataset[column] = lb_encoder.fit_transform(ls_dataset[column])
```

Appendix C:2 Sample code for Hyperparameter tuning Implementations

#Hyperparameter tuning implementation for RF

```
: rf_classifier = ensemble.RandomForestClassifier(n_jobs=-1)
param_grid = {
    "n_estimators": [100,400,1000, 1200,1500],
    "max_depth": np.arange(1,20),
    "criterion": ["gini", "entropy"],
}
```

```
: rf_model = model_selection.RandomizedSearchCV(
    estimator = rf_classifier,
    param_distributions=param_grid,
    n_iter = 10,
    scoring="accuracy",
    n_jobs=1,
    cv=10)
```

```
: rf_model.fit(features, target)
print(rf_model.best_score_)
print(rf_model.best_estimator_.get_params())
```

#Hyperparameter tuning implementation for GB

```

from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import GradientBoostingClassifier
distributions = dict(n_estimators = [100, 500, 900, 1000],
                    max_depth = np.arange(1,20),
                    min_samples_leaf = np.arange(1,10),
                    min_samples_split = [2, 4, 6, 10],
                    criterion = ["friedman_mse", "mae"],
                    max_features = ['auto', 'sqrt', 'log2', None])
# Set up the random search with 10-fold cross validation
random_search_cv = RandomizedSearchCV(GradientBoostingClassifier(),
distributions, cv=10, n_iter=10, scoring="accuracy", n_jobs = -1, random_state=42)
random_search_cv.fit(features,target)
#To display score result with best parameter
random_search_cv.cv_results_
print(random_search_cv.best_score_)
print(random_search_cv.best_estimator_.get_params())

```

#Hyperparameter tuning implementation for KNN

```

from sklearn.model_selection import RandomizedSearchCV
weight_options = ['uniform', 'distance']
metric = ['euclidean', 'manhattan', 'hamming ']
k_range = list(range(3, 21))
# specify "parameter distributions"
param_dist = dict(n_neighbors=k_range, weights=weight_options, metric=metric)
rand = RandomizedSearchCV(KNeighborsClassifier(), param_dist, cv=10, scoring='accuracy', n_iter=10, random_state=42)
# fit with features and target
rand.fit(features, target)
# to display scores
rand.cv_results_
# examine the best model
print(rand.best_score_)
print(rand.best_params_)
print(rand.best_estimator_)

```