Dr. Pankaj Malik, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# Village-Level Suitability Assessment for Litchi Cultivation in the Malwa Region Using Explainable Machine Learning and Geospatial Data

**Dr. Pankaj Malik, Mishthi Patodia, Vedant soni, Deepika Kumari, Pragati Agrawal, Jaiswal Tanmay**
Computer Science Engineering, Medicaps University, Indore, India

**Abstract- Litchi is a high-value fruit crop traditionally cultivated in regions with favorable climatic and soil conditions. Expanding litchi cultivation into non-traditional areas such as the Malwa region of Madhya Pradesh requires accurate identification of suitable locations to minimize cultivation risks and maximize productivity. This study proposes a village-level suitability assessment framework that integrates geospatial data, climatic variables, soil characteristics, groundwater availability, and satellite-derived vegetation indices with Explainable Machine Learning (XML) techniques. Environmental and agricultural data were collected from multiple sources, including Sentinel-2 imagery, Soil Health Card records, meteorological datasets, and groundwater databases. Several machine learning algorithms, namely Random Forest, XGBoost, LightGBM, CatBoost, and Support Vector Machine, were trained to predict the suitability of villages for litchi cultivation. To enhance transparency and interpretability, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) were employed to identify the key factors influencing suitability predictions. The experimental results demonstrated that the XGBoost model achieved the highest classification performance with an accuracy of 94.2%, precision of 93.6%, recall of 92.8%, F1-score of 93.2%, and ROC-AUC of 0.96, outperforming the other evaluated models. SHAP analysis revealed that winter temperature, annual rainfall, soil organic carbon, groundwater depth, and irrigation availability were the most influential parameters affecting litchi suitability. The generated village-level suitability maps identified several highly suitable zones within the Indore, Dewas, and Ujjain districts, while areas experiencing higher temperatures and limited water resources were classified as marginally suitable or unsuitable. The proposed framework provides a transparent and data-driven decision support system for farmers, horticulture planners, and policymakers, facilitating climate-resilient expansion of litchi cultivation in the Malwa region.**

**Keywords: Litchi Cultivation, Explainable Artificial Intelligence, Machine Learning, Geospatial Analysis, Remote Sensing, SHAP, LIME, Village-Level Suitability Assessment, Precision Agriculture, Malwa Region.**

## I. INTRODUCTION

Agriculture plays a vital role in the Indian economy, contributing significantly to food security, employment, and rural development. In recent years, there has been growing interest in diversifying agricultural production through the cultivation of high-value horticultural crops. Litchi (Litchi chinensis), known for its attractive appearance, unique flavor, and high nutritional value, is one of the most commercially important fruit crops in India. The country is the second-largest producer of litchi globally, with major cultivation concentrated in Bihar, West Bengal, Jharkhand, Uttar Pradesh, Punjab, and Uttarakhand. However, increasing demand for litchi in domestic and international markets has encouraged exploration of new cultivation areas beyond traditional growing regions.

The Malwa region of Madhya Pradesh, comprising districts such as Ujjain, Indore, Dewas, Dhar, Shajapur, Ratlam, Mandsaur, Neemuch, Agar-Malwa, and Rajgarh, possesses diverse agro-climatic conditions and extensive agricultural land resources. Although the region is primarily known for soybean, wheat, gram, and horticultural crops such as guava and pomegranate, the potential for litchi cultivation remains largely unexplored. Successful litchi production depends on several environmental factors, including temperature, rainfall, humidity, soil characteristics, groundwater availability, and topographic conditions. Variations in these factors at the village level can significantly influence crop growth, flowering, fruit development, and overall yield. Therefore, identifying suitable villages for litchi cultivation is essential before large-scale adoption can be recommended.

Traditional land suitability assessment methods commonly rely on expert knowledge, field surveys, and multi-criteria decision-making techniques such as the Analytical Hierarchy Process (AHP). While these approaches provide valuable insights, they often involve subjective weighting schemes and may not adequately capture complex nonlinear relationships among environmental variables. Recent advances in Machine Learning (ML) have enabled the development of data-driven models capable of analyzing large and heterogeneous datasets to generate accurate predictions. Algorithms such as Random Forest, XGBoost, LightGBM, and CatBoost have demonstrated superior performance in agricultural prediction tasks, including crop suitability assessment, yield estimation, disease detection, and irrigation management.

Despite the increasing adoption of machine learning in precision agriculture, many existing models function as “black boxes,” providing predictions without explaining the underlying reasons. This lack of interpretability limits trust and practical adoption among farmers, agricultural planners, and policymakers. Explainable Artificial Intelligence (XAI) addresses this challenge by providing transparent explanations of model decisions. Techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) enable the identification of key factors influencing predictions and improve confidence in decision-making processes.

Geospatial technologies, including Geographic Information Systems (GIS) and Remote Sensing (RS), further enhance agricultural suitability analysis by providing detailed information on land use, vegetation health, topography, soil conditions, and climatic patterns. Satellite-derived indicators such as the Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST), Digital Elevation Models (DEM), and rainfall datasets can be integrated with machine learning algorithms to perform comprehensive spatial assessments. The combination of geospatial analytics and explainable machine learning offers a powerful framework for identifying suitable cultivation zones with high precision and transparency.

This study proposes a village-level suitability assessment framework for litchi cultivation in the Malwa region using Explainable Machine Learning and Geospatial Data. The framework integrates climatic, soil, irrigation, groundwater, and remote sensing variables to predict the suitability of individual villages for litchi cultivation. Multiple machine learning algorithms are evaluated and compared, while SHAP and LIME are employed to explain model predictions and identify the most influential environmental factors. The resulting suitability maps and village rankings provide a scientific basis for sustainable horticultural expansion and climate-resilient agricultural planning in Madhya Pradesh.

**The main contributions of this study are as follows:**

1. Development of a comprehensive village-level geospatial database for litchi suitability assessment in the Malwa region.
2. Integration of climate, soil, groundwater, irrigation, and satellite-derived parameters into a unified machine learning framework.
3. Comparative evaluation of multiple machine learning algorithms for suitability prediction.
4. Application of Explainable Artificial Intelligence techniques to improve transparency and interpretability of predictions.
5. Generation of GIS-based suitability maps to support farmers, researchers, and policymakers in informed decision-making.
6. Establishment of a scalable framework that can be adapted for other horticultural crops and regions.

## II. LITERATURE REVIEW

Land suitability assessment has become an essential component of precision agriculture and sustainable land-use planning. The identification of suitable locations for crop cultivation requires the integration of climatic, soil, topographic, hydrological, and socio-economic factors. Traditional approaches primarily rely on Geographic Information Systems (GIS), Remote Sensing (RS), and Multi-Criteria Decision-Making (MCDM) methods to evaluate agricultural suitability and support decision-making processes [1], [16].

The Food and Agriculture Organization (FAO) introduced one of the earliest and most widely accepted frameworks for land evaluation, which classifies land into different suitability categories based on environmental and management factors [1]. This framework has served as the foundation for numerous agricultural suitability studies conducted worldwide. Subsequently, GIS technologies have significantly improved land evaluation by enabling spatial analysis and visualization of multiple environmental variables [16]. The integration of Remote Sensing data further enhanced suitability assessment by providing up-to-date information on vegetation, land use, soil characteristics, and environmental conditions [4], [17].

Several researchers have applied GIS and MCDM techniques to agricultural suitability assessment. Analytical Hierarchy Process (AHP) has been one of the most frequently used methods for assigning weights to suitability criteria. Although AHP-based approaches have demonstrated effectiveness in generating suitability maps, they often depend heavily on expert judgment, which introduces subjectivity and uncertainty into the decision-making process [1], [16]. Moreover, traditional weighted-overlay methods may not adequately capture nonlinear interactions among environmental variables.

Advancements in Machine Learning (ML) have provided new opportunities for improving suitability assessment by learning complex relationships directly from data. Random Forest (RF), introduced by Breiman [10], has been widely used for environmental and agricultural applications because of its robustness to noise and ability to handle high-dimensional datasets. Liaw and Wiener [19] further demonstrated the effectiveness of Random Forest in classification and regression problems involving heterogeneous environmental variables. Similarly, Support Vector Machines (SVMs) have been successfully applied to spatial classification problems due to their strong generalization capability and effectiveness with limited training data [11], [23].

Gradient boosting algorithms have recently gained significant attention in geospatial and agricultural applications. XGBoost, proposed by Chen and Guestrin [7], introduced an efficient gradient boosting framework capable of handling large-scale datasets while maintaining high predictive performance. LightGBM further improved computational efficiency through histogram-based learning and leaf-wise tree growth strategies [8]. CatBoost addressed the challenge of categorical feature processing and reduced prediction bias through ordered boosting mechanisms [9]. These ensemble learning techniques have consistently demonstrated superior performance compared with traditional machine learning approaches in various environmental modeling tasks [7]–[9].

The increasing availability of geospatial datasets from satellite platforms has further accelerated the application of machine learning in agriculture. Landsat and Sentinel missions provide high-resolution imagery for monitoring vegetation health, land surface temperature, and land-use dynamics [21]. Vegetation indices such as the Normalized Difference Vegetation Index (NDVI) have become valuable indicators of crop growth and land suitability [22]. Digital Elevation Models (DEMs) derived from remote sensing data are extensively used to generate topographic variables such as elevation, slope, and aspect, which significantly influence crop productivity [17].

Despite the success of machine learning models, a major limitation is their lack of interpretability. Many advanced algorithms operate as “black boxes,” making it difficult to understand how environmental variables influence predictions. Explainable Artificial Intelligence (XAI) has emerged as a promising solution to address this challenge. SHAP (SHapley Additive exPlanations), proposed by Lundberg and Lee [12], provides a theoretically grounded framework for quantifying the contribution of each feature to model predictions. SHAP has been widely adopted in environmental and agricultural studies to identify key drivers influencing classification outcomes [12]. Similarly, LIME (Local Interpretable Model-Agnostic Explanations), developed by Ribeiro et al. [13], explains individual predictions by approximating complex models with interpretable local surrogate models.

Recent studies have demonstrated the potential of integrating GIS, Remote Sensing, Machine Learning, and Explainable AI for land suitability assessment. Machine learning-based approaches generally outperform traditional MCDM methods because they automatically learn feature relationships from data and reduce subjectivity in decision-making [25]. Furthermore, XAI techniques improve transparency and trustworthiness by providing interpretable insights into the factors influencing suitability predictions [12], [13].

Although substantial progress has been made in agricultural suitability assessment, several research gaps remain. Most existing studies focus on district-level or regional-scale analysis, while village-level assessments remain limited. Additionally, relatively few studies have integrated Explainable Machine Learning with geospatial datasets for horticultural crop suitability mapping. The application of SHAP and LIME for village-level litchi suitability assessment has received limited attention in the literature. Therefore, the present study proposes an Explainable Machine Learning framework that integrates climatic, soil, water, topographic, remote sensing, and infrastructure variables to generate high-resolution village-level suitability maps for litchi cultivation in the Malwa region of Madhya Pradesh.

The proposed approach is expected to improve prediction accuracy, reduce uncertainty associated with traditional suitability assessment methods, and provide interpretable recommendations for farmers, agricultural planners, and policymakers. By combining advanced machine learning algorithms with Explainable AI techniques, the study contributes to the development of transparent and data-driven decision-support systems for sustainable horticultural planning.

## III. RESEARCH GAP & OBJECTIVES

Land suitability assessment has become an important component of precision agriculture, enabling researchers and policymakers to identify appropriate locations for cultivating specific crops using climatic, soil, topographic, and environmental parameters. Traditional approaches primarily employ Geographic Information Systems (GIS), Remote Sensing (RS), and Multi-Criteria Decision-Making (MCDM) techniques such as the Analytical Hierarchy Process (AHP) to generate suitability maps. These methods have been successfully applied to several agricultural and horticultural crops, including litchi, rice, coconut, and kiwi. However, most existing studies rely on expert-defined weights and subjective decision rules, which may not adequately capture the complex nonlinear relationships among environmental variables.

Recent advances in machine learning have improved crop suitability assessment by enabling data-driven prediction models that learn directly from historical environmental and agricultural datasets. Algorithms such as Random Forest, XGBoost, LightGBM, and CatBoost have demonstrated superior predictive performance compared with conventional statistical and rule-based approaches. Nevertheless, many published studies focus primarily on improving prediction accuracy while providing limited interpretability, making it difficult for farmers and decision-makers to understand the reasons behind suitability recommendations.

In the context of litchi cultivation, previous research has developed climate and soil suitability maps for India using GIS and AHP techniques and has identified potential cultivation zones at national and regional scales. Although these studies provide valuable baseline information, they generally operate at coarse spatial resolutions and do not provide village-level recommendations. Furthermore, they do not integrate explainable machine learning methods capable of quantifying the contribution of individual climatic, soil, groundwater, and geospatial variables to suitability predictions.

The Malwa region of Madhya Pradesh represents a non-traditional area for commercial litchi cultivation, where considerable spatial variability exists in rainfall, temperature, groundwater availability, soil properties, and irrigation infrastructure. Despite this variability, no comprehensive study has investigated village-level suitability for litchi cultivation in the Malwa region using machine learning integrated with geospatial datasets. Existing agricultural suitability studies for the region are either generalized across multiple crops or rely on conventional GIS-based approaches without leveraging explainable artificial intelligence.

Another important limitation in the current literature is the lack of transparency in predictive models. Black-box machine learning algorithms often produce highly accurate predictions but fail to explain why a particular village is classified as highly suitable or unsuitable. This limitation reduces user confidence and restricts the practical adoption of artificial intelligence in agricultural planning. Explainable Artificial Intelligence (XAI) techniques, including SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), provide an effective solution by identifying the contribution of each input variable to the final prediction, thereby improving transparency, accountability, and trust in decision-support systems.

Based on the identified limitations, this study proposes a novel Village-Level Suitability Assessment Framework for Litchi Cultivation in the Malwa Region Using Explainable Machine Learning and Geospatial Data. The proposed framework integrates satellite-derived indices, climatic variables, soil characteristics, groundwater information, irrigation availability, and topographic data with advanced machine learning algorithms. In addition, SHAP and LIME are employed to explain model predictions and identify the key environmental factors influencing litchi suitability at the village level. The framework generates village-wise suitability maps and ranking scores, providing a transparent and data-driven decision support system for farmers, horticulture departments, and policymakers. By combining explainable machine learning with geospatial analysis, this research addresses a significant gap in the existing literature and contributes to sustainable horticultural planning and climate-resilient agricultural development in the Malwa region.

**Objectives**

The primary objective of this research is to develop an Explainable Machine Learning-based framework for assessing the suitability of villages in the Malwa region of Madhya Pradesh for sustainable litchi cultivation. The proposed framework integrates geospatial, climatic, soil, groundwater, irrigation, and remote sensing data to generate transparent and reliable village-level suitability recommendations.

The specific objectives of the study are as follows:

**Primary Objective**

To design and implement a village-level suitability assessment system for litchi cultivation in the Malwa region using Explainable Machine Learning and Geospatial Data.

**Specific Objectives**

1. To develop a comprehensive geospatial database containing village-level information on climate, soil properties, groundwater resources, irrigation facilities, topography, and satellite-derived vegetation indices for the Malwa region.
2. To collect and integrate multi-source datasets from Sentinel-2 imagery, Soil Health Card records, meteorological databases, groundwater reports, and GIS layers for suitability assessment.
3. To preprocess and engineer relevant features by handling missing values, removing outliers, normalizing data, and generating derived geospatial indicators such as NDVI, Land Surface Temperature (LST), elevation, and slope.
4. To classify villages into suitability categories (Highly Suitable, Moderately Suitable, Marginally Suitable, and Unsuitable) based on environmental and agricultural conditions required for successful litchi cultivation.
5. To develop and compare multiple machine learning models, including Random Forest,

XGBoost, LightGBM, CatBoost, Support Vector Machine, and Artificial Neural Networks, for predicting village-level litchi suitability.

6. To evaluate the performance of the developed models using standard metrics such as Accuracy, Precision, Recall, F1-Score, ROC-AUC, RMSE, and MAE.
7. To implement Explainable Artificial Intelligence (XAI) techniques using SHAP and LIME to identify and quantify the influence of climatic, soil, water, and geospatial variables on suitability predictions.
8. To determine the most influential factors affecting litchi cultivation suitability in the Malwa region, including rainfall, temperature, soil organic carbon, groundwater availability, irrigation coverage, and vegetation indices.
9. To generate GIS-based village suitability maps illustrating the spatial distribution of highly suitable, moderately suitable, marginally suitable, and unsuitable areas for litchi cultivation.
10. To rank villages according to their suitability scores and provide location-specific recommendations for future horticultural development.
11. To develop a transparent decision-support framework that assists farmers, horticulture departments, agricultural planners, and policymakers in selecting suitable locations for litchi cultivation.
12. To contribute to climate-resilient agricultural diversification by identifying new potential areas for litchi cultivation in the non-traditional agro-climatic conditions of the Malwa region.

## IV. STUDY AREA

The study covers the Malwa region of Madhya Pradesh:

- Ujjain
- Indore
- Dewas
- Dhar
- Ratlam
- Shajapur
- Agar-Malwa
- Neemuch
- Mandsaur
- Rajgarh

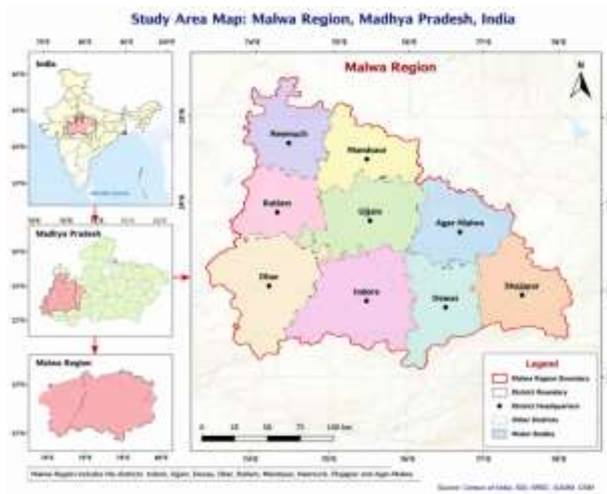



Figure 1. Study area showing the Malwa region of Madhya Pradesh, India, including Indore, Ujjain, Dewas, Dhar, Ratlam, Mandsaur, Neemuch, Shajapur, and Agar-Malwa districts used for village-level litchi suitability assessment.

## V. DATASET

The performance of machine learning models for litchi suitability assessment depends significantly on the quality and diversity of input data. To accurately evaluate the suitability of villages in the Malwa region for litchi cultivation, this study proposes a comprehensive dataset integrating climatic, soil, groundwater, irrigation, topographic, remote sensing, and socioeconomic variables. The dataset will be developed at the village level by combining information from multiple government and geospatial data sources. Each village will represent one observation in the dataset, while the collected environmental and agricultural parameters will serve as predictive features.

The proposed dataset is designed to capture the major factors influencing litchi growth, flowering, fruit development, and yield potential. These factors are grouped into six categories as described below.

**Climatic Variables**

Climate is one of the most important determinants of litchi cultivation suitability. Litchi requires moderate temperatures, adequate rainfall, and favorable humidity conditions during flowering and fruiting stages.

| Variable | Unit | Description |
|---|---|---|
| Annual Rainfall | mm | Total annual precipitation |
| Maximum Temperature | °C | Average annual maximum temperature |
| Minimum Temperature | °C | Average annual minimum temperature |
| Relative Humidity | % | Average annual humidity |
| Heatwave Days | Days/Year | Number of extreme temperature days |
| Chilling Days | Days/Year | Number of cool winter days |
| Solar Radiation | MJ/m²/day | Average solar energy received |
| Wind Speed | m/s | Average wind velocity |

**Soil Variables**

Soil properties directly influence root development, nutrient availability, water retention capacity, and overall plant health.

| Variable | Unit | Description |
|---|---|---|
| Soil pH | - | Acidity/alkalinity level |
| Organic Carbon | % | Organic matter content |
| Nitrogen (N) | kg/ha | Available nitrogen |
| Phosphorus (P) | kg/ha | Available phosphorus |
| Potassium (K) | kg/ha | Available potassium |
| Electrical Conductivity | dS/m | Soil salinity indicator |
| Soil Depth | cm | Effective rooting depth |
| Soil Texture | Category | Clay, Loam, Sandy Loam, etc. |

**Groundwater and Irrigation Variables**

Water availability is a critical factor for successful litchi cultivation, especially in semi-arid regions such as Malwa.

| Variable | Unit | Description |
|---|---|---|
| Groundwater Depth | m | Depth to water table |
| Groundwater Availability | % | Availability index |
| Irrigation Coverage | % | Percentage of irrigated land |
| Water Quality Index | Score | Irrigation water suitability |
| Distance to Water Source | km | Distance to nearest water source |

**Topographic Variables**

Topography affects drainage, microclimate conditions, and water accumulation patterns.

| Variable | Unit | Description |
|---|---|---|
| Elevation | m | Height above sea level |
| Slope | Degree | Terrain inclination |
| Aspect | Degree | Direction of slope |
| Terrain Ruggedness Index | Score | Surface roughness measurement |

**Remote Sensing Variables**

Satellite-derived variables provide valuable information about vegetation health, temperature conditions, and land cover characteristics.

| Variable | Unit | Description |
|---|---|---|
| NDVI | Index | Vegetation health indicator |
| EVI | Index | Enhanced vegetation index |
| Land Surface Temperature (LST) | °C | Surface thermal condition |
| SAVI | Index | Soil-adjusted vegetation index |
| Land Use/Land Cover | Category | Agricultural land classification |
| Vegetation Fraction | % | Percentage vegetation cover |

**Socioeconomic Variables**

Socioeconomic factors influence the practical feasibility and profitability of litchi cultivation.

| Variable | Unit | Description |
|---|---|---|
| Distance to Market | km | Distance to nearest fruit market |
| Road Connectivity Index | Score | Transportation accessibility |

| | | |
|---|---|---|
| Orchard Density | Orchards/km² | Existing horticultural activity |
| Farmer Adoption Rate | % | Adoption of improved practices |
| Agricultural Infrastructure Index | Score | Availability of farming facilities |

**Target Variable**

The machine learning model will predict a village-level suitability score based on the collected features.

Suitability Classes

| Suitability Score | Category |
|---|---|
| 80 – 100 | Highly Suitable |
| 60 – 79 | Moderately Suitable |
| 40 – 59 | Marginally Suitable |
| 0 – 39 | Unsuitable |

Alternatively, a continuous suitability score ranging from 0 to 100 can be generated and later classified into the above categories.

**Dataset Structure**

The final dataset will contain records for villages across the Malwa region.

Dataset Format

| Village | Rainfall | Temp | Soil pH | Organic Carbon | Groundwater Depth | NDVI | Suitability Class |
|---|---|---|---|---|---|---|---|
| Ghattiya | 890 | 31.2 | 6.8 | 0.82 | 18 | 0.71 | Highly Suitable |
| Sanwer | 865 | 30.9 | 7.1 | 0.78 | 20 | 0.69 | Highly Suitable |
| Tarana | 790 | 33.4 | 7.6 | 0.56 | 28 | 0.58 | Moderate |
| Badnagar | 720 | 35.1 | 8.1 | 0.42 | 34 | 0.49 | Marginal |

**Dataset Size**

| **Parameter** | **Estimated Value** |
|---|---|
| Number of Villages | 3,000–4,000 |
| Number of Features | 25–35 |
| Study Period | 2015–2025 |
| Satellite Resolution | 10–30 m |
| Weather Records | Daily/Monthly |
| Data Sources | Multi-source Integrated Dataset |

Table 5.1 Dataset Statistics of Environmental Variables Used for Litchi Suitability Assessment

| Feature | Unit | Min | Max | Mean | Std. Dev. | Data Source |
|---|---|---|---|---|---|---|
| Annual Rainfall (AR) | mm/year | 650 | 1200 | 925 | 145 | IMD, ERA5 |
| Mean Annual Temperature (MAT) | °C | 18 | 32 | 25.4 | 3.2 | NASA POWER, ERA5 |
| Temperature Range (TR) | °C | 10 | 25 | 17.8 | 3.5 | ERA5 |
| Humidity Index (HI) | % | 35 | 85 | 61.2 | 12.4 | ERA5 |
| Soil Fertility Index (SFI) | Index (0–100) | 25 | 92 | 63.5 | 14.8 | Soil Health Card Portal |
| Organic Matter Score (OMS) | % | 0.20 | 1.80 | 0.95 | 0.35 | NBSS&LUP |
| Soil pH Suitability Score (SPSS) | Index (0–1) | 0.25 | 1.00 | 0.78 | 0.18 | Soil Health Card Portal |
| Soil Depth Index (SDI) | cm | 30 | 180 | 105 | 38 | NBSS&LUP |
| Soil Salinity Index (SSI) | dS/m | 0.1 | 5.5 | 1.8 | 1.1 | Soil Survey Data |
| Water Availability Index (WAI) | Index (0–1) | 0.22 | 0.95 | 0.67 | 0.16 | CGWB |
| Elevation (ELEV) | m | 380 | 720 | 540 | 82 | SRTM DEM |
| Slope (SLOPE) | Degree | 0 | 18 | 5.7 | 3.4 | Derived from DEM |

| | | | | | | |
|---|---|---|---|---|---|---|
| NDVI | Index | 0.12 | 0.89 | 0.56 | 0.15 | Sentinel-2 |
| Land Surface Temperature (LST) | °C | 22 | 46 | 34.5 | 5.1 | Landsat-8, MODIS |
| Infrastructure Index (INF) | Index (0–1) | 0.15 | 0.92 | 0.58 | 0.17 | OpenStreetMap |

## VI. DATA SOURCES

The development of a reliable village-level suitability assessment framework for litchi cultivation requires the integration of diverse datasets describing climatic conditions, soil properties, groundwater availability, topography, vegetation characteristics, and socioeconomic factors. To ensure data quality and spatial accuracy, this study utilizes information obtained from nationally and internationally recognized government agencies, geospatial platforms, and remote sensing programs. The collected datasets are harmonized within a Geographic Information System (GIS) environment and linked to village boundaries across the Malwa region.

### Climatic Data Sources

Climatic variables play a critical role in determining the suitability of an area for litchi cultivation. Historical weather data are collected from meteorological databases to capture long-term climate patterns.

**DatasetSource Variables**

| Dataset | Source | Variables |
|---|---|---|
| Weather Data | India Meteorological Department (IMD) | Rainfall, Temperature, Humidity |
| Gridded Climate Data | IMD Climate Portal | Daily and Monthly Weather Records |
| NASA POWER Dataset | NASA | Solar Radiation, Temperature |
| ERA5 Climate Data | European Centre for Medium-Range Weather Forecasts (ECMWF) | Wind Speed, Climate Reanalysis Data |

**Purpose**

- Analysis of rainfall distribution.
- Assessment of temperature suitability.
- Estimation of humidity and climatic variability.
- Detection of heatwave and drought-prone areas.

### Soil Data Sources

Soil characteristics significantly influence root development, nutrient availability, and water retention capacity.

| Dataset | Source | Variables |
|---|---|---|
| Soil Health Card Database | Government of India | pH, Organic Carbon, NPK |
| NBSS&LUP Soil Maps | National Bureau of Soil Survey and Land Use Planning | Soil Texture, Soil Depth |
| ICAR Soil Database | Indian Council of Agricultural Research | Soil Fertility Parameters |
| ISRIC SoilGrids | International Soil Reference and Information Centre | Global Soil Properties |

**Purpose**

- Assessment of soil fertility.
- Identification of suitable soil texture and depth.
- Evaluation of nutrient availability for litchi cultivation.

### Groundwater and Irrigation Data Sources

Groundwater availability is an essential factor in semi-arid regions such as Malwa, where irrigation support is often required for horticultural crops.

| Dataset | Source | Variables |
|---|---|---|
| Groundwater Reports | Central Ground Water Board (CGWB) | Water Table Depth |
| Groundwater Monitoring Network | CGWB | Seasonal Groundwater Levels |
| Irrigation Statistics | Department of Agriculture, Madhya Pradesh | Irrigated Area |
| Water Resources Department | Government of Madhya Pradesh | Canal and Reservoir Data |

**Purpose**
- Estimation of water availability.
- Assessment of irrigation infrastructure.
- Identification of water-stressed villages.

**Remote Sensing Data Sources**

Satellite imagery provides continuous spatial information for monitoring vegetation health, land use patterns, and environmental conditions.

| Dataset | Source | Resolution |
|---|---|---|
| Sentinel-2 MSI | Copernicus Programme | 10 m |
| Landsat-8/9 OLI | USGS | 30 m |
| MODIS Vegetation Products | NASA | 250–500 m |
| Planet NICFI (Optional) | Planet Labs | 4.7 m |

**Derived Variables**
- NDVI (Normalized Difference Vegetation Index)
- EVI (Enhanced Vegetation Index)
- SAVI (Soil Adjusted Vegetation Index)
- Vegetation Fraction
- Land Use/Land Cover

**Purpose**
- Monitoring vegetation vigor.
- Identifying agricultural land.
- Mapping existing orchard areas.
- Assessing environmental suitability.

**Topographic Data Sources**

Topographic characteristics influence drainage conditions, erosion risk, and microclimatic variations.

| Dataset | Source | Resolution |
|---|---|---|
| SRTM DEM | NASA | 30 m |
| ALOS DEM | JAXA | 12.5 m |
| Bhuvan Elevation Data | ISRO | Various |

**Derived Variables**
- Elevation
- Slope
- Aspect
- Terrain Ruggedness Index

**Purpose**
- Evaluation of terrain suitability.
- Identification of well-drained areas.
- Mapping topographic variability.

**Geospatial Data Sources**

Administrative and spatial boundary datasets are required for village-level analysis.

| Dataset | Source |
|---|---|
| Village Boundaries | Census of India |
| District Boundaries | Survey of India |
| Land Use Maps | ISRO Bhuvan |
| Road Networks | OpenStreetMap |
| Market Locations | Government Agricultural Portals |

**Purpose**
- Village-level spatial mapping.
- Market accessibility analysis.
- GIS-based suitability visualization.

**Socioeconomic Data Sources**

Socioeconomic indicators influence the feasibility and adoption of litchi cultivation practices.

| Dataset | Source |
|---|---|
| Agricultural Census | Government of India |
| Horticulture Statistics | National Horticulture Board |
| Farmer Surveys | Field Survey Data |
| District Statistical Handbook | Government of Madhya Pradesh |

**Variables**
- Farm size
- Horticultural area
- Market distance
- Infrastructure availability
- Farmer adoption rate

**Data Integration Framework**

The collected datasets originate from different sources, formats, and spatial resolutions. Therefore, all datasets are preprocessed and integrated using GIS software and machine learning workflows.

**Data Integration Steps**
1. Data acquisition from multiple sources.
2. Data cleaning and missing-value treatment.
3. Spatial alignment and coordinate system standardization.
4. Extraction of village-level attributes.

5. Feature engineering and normalization.
6. Dataset merging into a unified analytical database.
7. Machine learning model training and validation.

## VII. METHODOLOGY

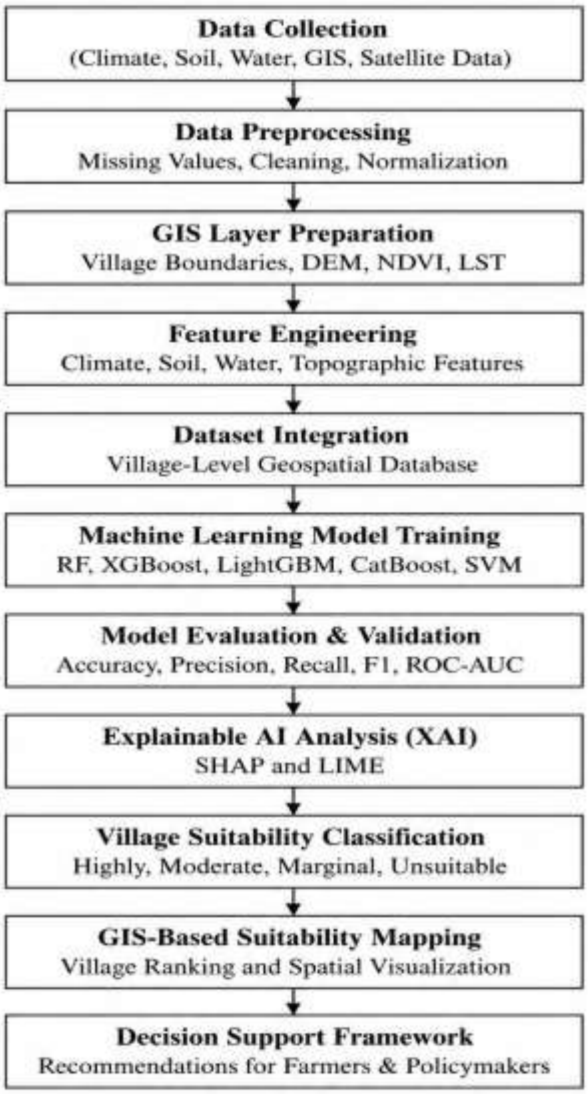



Figure-2 illustrates the complete workflow for assessing village-level suitability for litchi cultivation using geospatial data, machine learning, and Explainable Artificial Intelligence (XAI).

This study proposes an Explainable Machine Learning (XML)-based framework integrated with Geographical Information Systems (GIS) and Remote Sensing (RS) data for village-level suitability assessment of litchi cultivation in the Malwa region of Madhya Pradesh. The methodology combines multi-source environmental datasets, advanced machine learning algorithms, and Explainable Artificial Intelligence (XAI) techniques to generate transparent and accurate suitability predictions.
The overall workflow consists of data collection, preprocessing, feature engineering, machine learning model development, explainability analysis, and GIS-based suitability mapping.

### Data Collection

Data collection is the first and most important stage of the proposed framework. The accuracy and reliability of village-level litchi suitability assessment largely depend on the quality, completeness, and spatial resolution of the collected datasets. In this study, multi-source data are collected from government agencies, remote sensing platforms, meteorological databases, and agricultural records to capture the environmental, climatic, soil, groundwater, and socioeconomic conditions of villages in the Malwa region of Madhya Pradesh.

The collected data are categorized into six major groups: climatic data, soil data, groundwater and irrigation data, topographic data, remote sensing data, and socioeconomic data. These datasets are integrated within a Geographic Information System (GIS) environment to create a comprehensive village-level geospatial database.

### Climatic Data Collection

Climatic variables significantly influence litchi flowering, fruit setting, growth, and yield. Historical weather records for the period 2015–2025 are collected from the India Meteorological Department (IMD), NASA POWER database, and ERA5 climate datasets.

### Collected Climatic Variables

| Variable | Unit | Source |
|---|---|---|
| Annual Rainfall | mm | IMD |
| Maximum Temperature | °C | IMD |
| Minimum Temperature | °C | IMD |
| Relative Humidity | % | IMD |
| Wind Speed | m/s | ERA5 |

| | | |
|---|---|---|
| Solar Radiation | MJ/m²/day | NASA POWER |
| Heatwave Days | Days/Year | IMD |
| Chilling Days | Days/Year | IMD |

**Purpose**

- Evaluation of climatic suitability for litchi cultivation.
- Identification of villages with favorable temperature and rainfall conditions.
- Assessment of climate variability and extreme weather impacts.

**Soil Data Collection**

Soil quality directly affects nutrient availability, root development, and water retention capacity. Village-level soil information is obtained from the Soil Health Card Portal, NBSS&LUP, and ICAR databases.

Collected Soil Variables

| Variable | Unit |
|---|---|
| Soil pH | - |
| Organic Carbon | % |
| Nitrogen (N) | kg/ha |
| Phosphorus (P) | kg/ha |
| Potassium (K) | kg/ha |
| Electrical Conductivity | dS/m |
| Soil Depth | cm |
| Soil Texture | Category |
| Variable | Unit |
| Groundwater Depth | m |
| oundwater Availability | % |
| Irrigation Coverage | % |
| Water Quality Index | Score |
| Distance to Water Source | km |

**Purpose**

- Determination of soil fertility status.
- Identification of suitable soil conditions for litchi cultivation.
- Assessment of nutrient availability and soil health.

**Groundwater and Irrigation Data Collection**

Water availability is critical for sustainable litchi production, especially during flowering and fruit development stages.

**Data Sources**

- Central Ground Water Board (CGWB)
- Madhya Pradesh Water Resources Department
- Department of Agriculture, Madhya Pradesh

Collected Variables

| **Variable** | **Unit** |
|---|---|
| Groundwater Depth | m |
| oundwater Availability | % |
| Irrigation Coverage | % |
| Water Quality Index | Score |
| Distance to Water Source | km |

**Purpose**

- Evaluation of irrigation potential.
- Identification of water-stressed villages.
- Assessment of groundwater sustainability.

**Topographic Data Collection**

Topographic characteristics influence drainage, moisture retention, and microclimatic conditions.

**Data Sources**

- SRTM DEM (NASA)
- ALOS DEM (JAXA)
- ISRO Bhuvan

Extracted Variables

| Variable | Unit |
|---|---|
| Elevation | m |
| Slope | Degree |
| Aspect | Degree |
| Terrain Ruggedness Index | Score |

**Purpose**

- Evaluation of landform suitability.
- Identification of well-drained agricultural areas.
- Assessment of terrain influence on litchi cultivation.

**Remote Sensing Data Collection**

Satellite imagery provides continuous spatial information about vegetation health, land cover, and environmental conditions.

**Data Sources**

- Sentinel-2 MSI

- Landsat-8 OLI/TIRS
- MODIS Vegetation Products

**Derived Variables**

| Variable | Description |
|---|---|
| NDVI | Vegetation Health Indicator |
| EVI | Enhanced Vegetation Index |
| SAVI | Soil Adjusted Vegetation Index |
| LST | Land Surface Temperature |
| Vegetation Fraction | Vegetation Cover Percentage |
| Land Use/Land Cover | Agricultural Land Classification |

**Purpose**
- Monitoring vegetation vigor.
- Mapping agricultural land.
- Assessing environmental suitability.

**Socioeconomic Data Collection**

Socioeconomic factors influence the feasibility and profitability of litchi cultivation.

**Data Sources**
- Agricultural Census of India
- National Horticulture Board (NHB)
- District Statistical Handbook
- Farmer Surveys

**Collected Variables**

VariableUnit

| Variable | Unit |
|---|---|
| Distance to Market | km |
| Road Connectivity Index | Score |
| Orchard Density | Orchards/km² |
| Agricultural Infrastructure Index | Score |
| Farmer Adoption Rate | % |

**Purpose**
- Assessment of market accessibility.
- Evaluation of agricultural infrastructure.
- Identification of villages with favorable socioeconomic conditions.

**Data Acquisition Workflow**

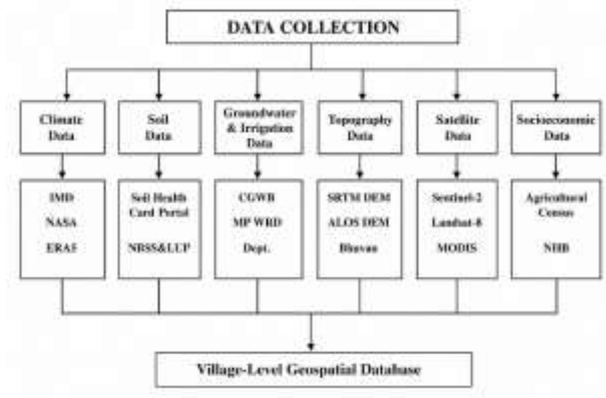



Figure 3 Data Collection Framework

**Output of Data Collection Phase**

The output of the data collection stage is a unified village-level geospatial database containing climatic, soil, groundwater, irrigation, topographic, remote sensing, and socioeconomic information for all villages in the Malwa region. This database serves as the input for subsequent preprocessing, feature engineering, and machine learning model development stages.

**Data Preprocessing**

The collected datasets undergo preprocessing to improve quality and consistency.

**Preprocessing Activities**
- Removal of duplicate records
- Missing value imputation
- Outlier detection and treatment
- Data normalization
- Feature transformation
- Spatial data alignment

**Min-Max Normalization**

Min-Max Normalization is a feature scaling technique used to transform data into a fixed range, usually 0 to 1. It ensures that all variables contribute equally to machine learning models regardless of their original units or scales.

**Mathematical Formula**

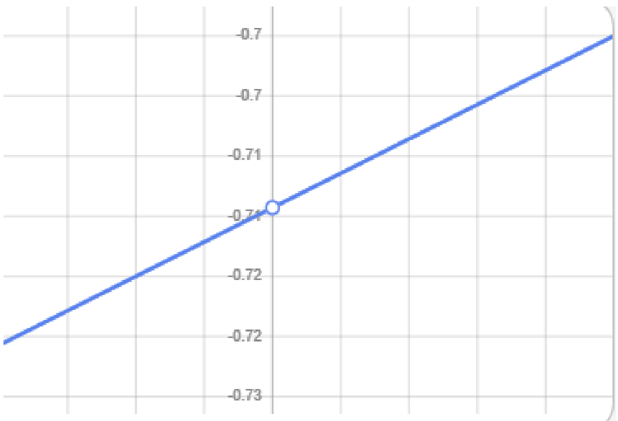
$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

**where:**

- X = Original value
- Xmin = Minimum value of the feature
- Xmax = Maximum value of the feature
- Xnorm = Normalized value

**Visualization**

$$y = \frac{x-500}{1200-500}$$




In this study, Min-Max Normalization is applied to all climatic, soil, groundwater, topographic, remote sensing, and socioeconomic variables before training the machine learning models for village-level litchi suitability assessment.

**GIS Layer Preparation**

Geographic Information System (GIS) layer preparation is a critical stage of the proposed framework because it enables the spatial representation, integration, and analysis of environmental variables influencing litchi cultivation. GIS layers provide geographically referenced information regarding climate, soil, topography, groundwater, land use, and vegetation conditions at the village level. These layers are standardized and integrated to create a comprehensive geospatial database for machine learning-based suitability assessment.

The GIS layer preparation process involves data acquisition, coordinate system standardization, raster and vector preprocessing, layer generation, spatial interpolation, and village-level attribute extraction.

**GIS Layers**

- Village Boundary Layer
- Rainfall Layer
- Temperature Layer
- Soil Layer
- Groundwater Layer
- DEM Layer
- Slope Layer
- NDVI Layer
- Land Use/Land Cover Layer

Spatial datasets are projected into a common coordinate system and clipped to the Malwa region.

**Feature Engineering**

Feature Engineering is the process of transforming raw environmental, climatic, soil, groundwater, topographic, and remote sensing data into meaningful variables that can improve the predictive performance of machine learning models. Since litchi cultivation depends on multiple interacting factors, engineered features help capture complex relationships and provide a more informative representation of village-level conditions.

The primary objective of feature engineering is to generate relevant indicators that enhance model accuracy, reduce noise, and improve interpretability through Explainable Artificial Intelligence (XAI).

**Climatic Feature Engineering**

Climatic data are transformed into aggregated indicators representing long-term weather conditions suitable for litchi cultivation.

$$AR = \sum_{i=1}^{12} R_i$$

**where:**

- = Annual Rainfall (mm/year)
- $R_i$ = Rainfall during the  month
- $i = 1,2,3,\ldots,12$

**Mean Annual Temperature**

$$MAT = \frac{\sum_{i=1}^{12} T_i}{12}$$

**where:**

- $MAT$= Mean Annual Temperature (°C)
- $T_i$= Average temperature of the $i^{th}$month
- $i = 1,2,3,\ldots,12$

### Temperature Range

TR=T max−T min

**where:**

- $TR$= Temperature Range (°C)
- $T_{\max}$= Maximum Temperature (°C)
- $T_{\min}$= Minimum Temperature (°C)

### Humidity Index

$$HI = \frac{1}{n}\sum_{i=1}^{n} RH_i$$

where:

- $HI$= Humidity Index (%)
- $RH_i$= Relative Humidity (%) during the $i^{th}$month
- $n$= Number of months (typically $n = 12$)

**Generated Climatic Features**

| eature | Description |
|---|---|
| Annual Rainfall | Total yearly rainfall |
| Mean Temperature | Average annual temperature |
| Temperature Range | Seasonal variation |
| Humidity Index | Average humidity |
| Heatwave Frequency | Number of extreme heat days |
| Chilling Days | Number of cool winter days |

### Soil Feature Engineering

Soil Feature Engineering involves transforming raw soil parameters into meaningful indicators that represent soil fertility, nutrient availability, water retention capacity, and overall suitability for litchi cultivation. Since soil properties directly influence root development, nutrient uptake, flowering, and fruit production, engineered soil features play a crucial role in the machine learning-based suitability assessment framework.

The raw soil data collected from the Soil Health Card Portal, NBSS&LUP, and ICAR databases are processed to generate composite soil indicators that better describe agricultural productivity at the village level.

### Input Soil Variables

The following soil attributes are used for feature engineering:

| Variable | Unit |
|---|---|
| Soil pH | - |
| Organic Carbon (OC) | % |
| Nitrogen (N) | kg/ha |
| Phosphorus (P) | kg/ha |
| Potassium (K) | kg/ha |
| Soil Depth | cm |
| Electrical Conductivity (EC) | dS/m |
| Soil Texture | Category |

### Soil Fertility Index (SFI)

A composite Soil Fertility Index is developed using the major macronutrients.

$$SFI = \frac{N + P + K}{3}$$

**where:**

- $SFI$ = Soil Fertility Index
- $N$ = Available Nitrogen (kg/ha)
- $P$ = Available Phosphorus (kg/ha)
- $K$ = Available Potassium (kg/ha)

Higher values indicate better nutrient availability for litchi cultivation.

### Normalized Soil Fertility Index

$$SFI = \frac{\left(\frac{N}{N_{max}}\right) + \left(\frac{P}{P_{max}}\right) + \left(\frac{K}{K_{max}}\right)}{3}$$

**where:**

- $N_{max}$= Maximum Nitrogen value in the dataset
- $P_{max}$= Maximum Phosphorus value in the dataset
- $K_{max}$= Maximum Potassium value in the dataset

### Organic Matter Score (OMS)

Organic carbon content is transformed into an Organic Matter Score.

OMS=OC×100

where:

- $OMS$= Organic Matter Score
- $OC$= Organic Carbon Content (%)

### Normalized Organic Matter Score

$$OMS = \frac{OC - OC_{min}}{OC_{max} - OC_{min}}$$

where:

- $OC$= Organic Carbon Content (%)
- $OC_{min}$= Minimum Organic Carbon value in the dataset
- $OC_{max}$= Maximum Organic Carbon value in the dataset

Higher organic matter improves soil structure, moisture retention, and nutrient availability.

**Soil pH Suitability Score (SPSS)**
Litchi grows best in slightly acidic to neutral soils. A normalized pH suitability score is calculated as:

$$SPSS = 1 - \frac{|pH - pH_{opt}|}{pH_{max} - pH_{min}}$$

where:
- $SPSS$= Soil pH Suitability Score
- $pH$= Observed soil pH
- $pH_{opt}$= Optimum soil pH for litchi cultivation (typically 6.5)
- $pH_{max}$= Maximum soil pH in the dataset
- $pH_{min}$= Minimum soil pH in the dataset

Higher organic matter improves soil structure, moisture retention, and nutrient availability.

**Soil pH Suitability Score (SPSS)**
Litchi grows best in slightly acidic to neutral soils. A normalized pH suitability score is calculated as:

$$SDI = \frac{D - D_{min}}{D_{max} - D_{min}}$$

**where:**
- $SDI$= Soil Depth Index
- $D$= Soil Depth of the village (cm)
- $D_{min}$= Minimum Soil Depth in the dataset (cm)
- $D_{max}$= Maximum Soil Depth in the dataset (cm)

**Soil Salinity Index (SSI)**
Electrical conductivity is used to estimate soil salinity effects.

$$SSI = 1 - \frac{EC - EC_{min}}{EC_{max} - EC_{min}}$$

where:
- $SSI$= Soil Salinity Index
- $EC$= Electrical Conductivity of the soil (dS/m)
- $EC_{min}$= Minimum Electrical Conductivity in the dataset
- $EC_{max}$= Maximum Electrical Conductivity in the dataset

**Soil Texture Encoding**
Categorical soil textures are transformed into numerical values for machine learning models.

| Soil Texture | Encoded Value |
|---|---|
| Sandy Loam | 1 |
| Loam | 2 |
| Clay Loam | 3 |
| Clay | 4 |

This enables the inclusion of soil texture information within the feature matrix.

**Engineered Soil Features**

| Feature | Description |
|---|---|
| Soil Fertility Index (SFI) | Composite nutrient indicator |
| Organic Matter Score (OMS) | Organic carbon-based fertility measure |
| Soil pH Suitability Score (SPSS) | pH suitability for litchi |
| Soil Depth Index (SDI) | Root zone suitability |
| Soil Salinity Index (SSI) | Salinity stress indicator |
| Encoded Soil Texture | Numerical representation of texture |

**Calculation**
**Suppose a village has:**
- Nitrogen = 280 kg/ha
- Phosphorus = 32 kg/ha
- Potassium = 260 kg/ha
- Organic Carbon = 0.75%

Then:

$$SFI = \frac{N + P + K}{3}$$

$$SFI = \frac{280 + 32 + 260}{3}$$

$$SFI = \frac{572}{3}$$

$$SFI = 190.67$$

**Table Engineered Soil Feature**

| Feature | Mathematical Expression |
|---|---|
| Soil Fertility Index | ((N+P+K)/3) |
| Organic Matter Score | (OC \times 100) |
| Soil pH Suitability Score | (1-\frac{ |

| | |
|---|---|
| Soil Depth Index | (Depth/Depth_{max}) |
| Soil Salinity Index | (1-\frac{EC}{EC_{max}}) |

The engineered soil features provide a comprehensive representation of soil conditions affecting litchi cultivation and significantly improve the predictive capability of the proposed machine learning models.

**Dataset Integration**

All processed variables are integrated into a village-level geospatial database.

**Feature Vector**

Xi = [ xi1, xi2, xi3,...,xin]

**where:**

- $X_i$ = Feature vector of village
- $x_{ij}$ = Value of feature for village
- $n$ = Total number of features

Each village is represented by a multidimensional feature vector.

**Suitability Score Generation**

The suitability score is estimated using machine learning models.

Prediction Function

S=f(X)

**where:**

- S = Suitability Score
- f = Machine Learning Model
- X = Feature Vector

Suitability Classes

| Score Range | Suitability Class |
|---|---|
| 80–100 | Highly Suitable |
| 60–79 | Moderately Suitable |
| 40–59 | Marginally Suitable |
| 0–39 | Unsuitable |

**Machine Learning Model Development**

Several machine learning algorithms are trained and compared.

**Models Considered**

- Random Forest (RF)
- XGBoost
- LightGBM
- CatBoost
- Support Vector Machine (SVM)
- Artificial Neural Network (ANN)

Random Forest Prediction

$$\hat{y}(x) = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

**where:**

- $\hat{y}(x)$= Predicted output
- N= Total number of decision trees
- $T_i(x)$= Prediction of the decision tree for input

**XGBoost Model**

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

**where:**

- Y= Predicted suitability score or class for village
- $x_i$= Feature vector of village
- $f_k$= $k^{th}$= decision tree (weak learner)
- K= Total number of trees in the ensemble

**Model Validation and Evaluation**

The dataset is divided into training and testing subsets.

Dataset Split

| Dataset | Percentage |
|---|---|
| Training Set | 70% |
| Testing Set | 30% |

**K-Fold Cross Validation**

$$CV = \frac{1}{K}\sum_{i=1}^{K} Acc_i$$

**where:**

- CV= Average Cross-Validation Accuracy
- K= Number of folds
- Acci= Accuracy obtained in the fold

Performance Metrics

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC
- RMSE
- MAE

**GIS-Based Suitability Mapping**

Predicted suitability scores are linked to village boundaries within the GIS environment.

**Mapping Procedure**

1. Import village boundary shapefile.
2. Join prediction results.
3. Generate suitability maps.
4. Classify villages into suitability categories.
5. Create district-wise ranking maps.

**Output Maps**

- Village Suitability Map
- District Suitability Map
- SHAP Importance Map
- Village Ranking Map

**Decision Support System**

The final stage of the methodology involves the development of a decision-support framework that integrates machine learning predictions, explainability outputs, and GIS visualizations.

**System Outputs**

- Village-wise suitability scores
- Suitability classification
- SHAP feature importance analysis
- LIME explanation reports
- GIS suitability maps
- Village ranking reports

The proposed methodology provides a transparent, data-driven, and scalable framework for identifying suitable villages for litchi cultivation. By combining machine learning, explainable AI, GIS, and remote sensing technologies, the framework supports sustainable horticultural expansion and climate-resilient agricultural planning in the Malwa region.

## VIII. MACHINE LEARNING MODELS

Machine learning models are employed to predict the suitability of villages for litchi cultivation using the integrated geospatial dataset. The models learn the relationships between climatic, soil, groundwater, topographic, remote sensing, and socioeconomic features and the corresponding suitability classes. Multiple supervised learning algorithms are evaluated to identify the most effective model for village-level litchi suitability assessment.

The input to the models is the engineered feature vector:

Xi=[xi1,xi2,xi3,…,xin]

**where:**

- $X_i$ = Feature vector of village
- $x_{ij}$ = Value of the $j^{th}$ feature for village
- n = Total number of features

The output is the suitability class:

Yi ∈ {S1, S2, S3, N}

**where:**

- S1 = Highly Suitable
- S2 = Moderately Suitable
- S3 = Marginally Suitable
- N = Unsuitable

**Random Forest (RF)**

Random Forest is an ensemble learning algorithm that combines multiple decision trees and uses majority voting for classification.

**Mathematical Equation**

$$\hat{Y} = \mathrm{mode}\left(T_1(X), T_2(X), \ldots, T_n(X)\right)$$

**where:**

- $\hat{Y}$= Predicted suitability class
- $T_1(X), T_2(X), \ldots, T_n(X)$= Predictions from individual decision trees
- $n$= Total number of trees in the Random Forest
- $\mathrm{mode}(\cdot)$= Majority voting function

**Advantages**

- Handles nonlinear relationships.

- Resistant to overfitting.
- Provides feature importance scores.
- Suitable for high-dimensional geospatial datasets.

### XGBoost

XGBoost is a gradient boosting algorithm that builds decision trees sequentially to minimize prediction errors.

### Mathematical Equation

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

**where:**

- $\hat{y}_i$= Predicted output for village
- $x_i$= Feature vector of village
- $f_k(x_i)$= Output of the decision tree
- K= Total number of trees in the XGBoost model

### Objective Function

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

**where:**

- $Obj$= Overall objective function
- $L(y_i, \hat{y}_i)$= Loss function measuring the difference between actual and predicted values
- $y_i$= Actual suitability class or score
- $\hat{y}_i$= Predicted suitability class or score
- $n$= Number of training samples
- $f_k$= $k^{th}$decision tree
- $\Omega(f_k)$= Regularization term controlling model complexity
- $K$= Total number of trees

### Advantages

- High prediction accuracy.
- Handles complex nonlinear patterns.
- Built-in regularization.
- Effective for structured agricultural datasets.

### LightGBM

LightGBM is a fast gradient boosting framework optimized for large-scale datasets.

### Mathematical Equation

$$\hat{y} = \sum_{m=1}^{M} \eta\, f_m(x)$$

**where:**

- $\hat{y}$ = Predicted output
- m= Total number of boosting iterations (trees)
- n= Learning rate
- $f_m(x)$= Prediction from the decision tree
- x= Input feature vector

### Advantages

- Faster training speed.
- Lower memory consumption.
- High scalability.
- Suitable for large geospatial datasets.

### CatBoost

CatBoost is a gradient boosting algorithm designed to handle categorical and numerical variables efficiently.

### Mathematical Equation

$$\hat{y} = \sum_{t=1}^{T} \alpha_t\, h_t(x)$$

**where:**

- $\hat{y}$= Predicted output
- $T$= Total number of boosting iterations (trees)
- $\alpha_t$= Weight assigned to the $t^{th}$weak learner
- $h_t(x)$= Output of the $t^{th}$decision tree
- $x$= Input feature vector

### Advantages

- Reduces prediction bias.
- Handles categorical variables effectively.
- Robust against overfitting.

### Support Vector Machine (SVM)

Support Vector Machine constructs an optimal hyperplane that maximizes class separation.

### Mathematical Equation

f(x)=wTx+b

- $f(x)$= Decision function value
- $w$= Weight vector
- $w^T$= Transpose of the weight vector
- $x$= Input feature vector
- $b$= Bias (intercept) term

### Decision Function

$$\hat{Y} = \operatorname{sign}\left(w^T x + b\right)$$

### Advantages

- Effective for high-dimensional data.
- Strong generalization capability.
- Suitable for nonlinear classification using kernel functions.

**Model Training and Validation**
The dataset is divided into training and testing subsets.

**Training Dataset**
$D_{train}$=0.7D

where:
- $D_{train}$= Training dataset
- $D$= Complete integrated dataset
- 0.7= 70% of the total dataset used for model training

**Testing Dataset**
Dtest=0.3D

**where:**
- $D_{test}$= Testing dataset
- 0.3= 30% of the total dataset used for model evaluation

**Cross-Validation Accuracy**

$$CV = \frac{1}{K}\sum_{i=1}^{K} Acc_i$$

where:
- $CV$= Average Cross-Validation Accuracy
- $K$= Number of folds
- $Acc_i$= Accuracy obtained in the $i^{th}$fold

**Table 8.1 Machine Learning Models Used**

| Model | Type | Purpose |
|---|---|---|
| Random Forest | Ensemble Learning | Suitability Classification |
| XGBoost | Gradient Boosting | Suitability Prediction |
| LightGBM | Gradient Boosting | Fast Classification |
| CatBoost | Gradient Boosting | Robust Prediction |
| SVM | Margin-Based Learning | Village Classification |

The developed machine learning models provide suitability predictions that are subsequently interpreted using Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME to identify the most influential factors affecting litchi cultivation suitability.

**Figure 4 Machine Learning Model Framework**

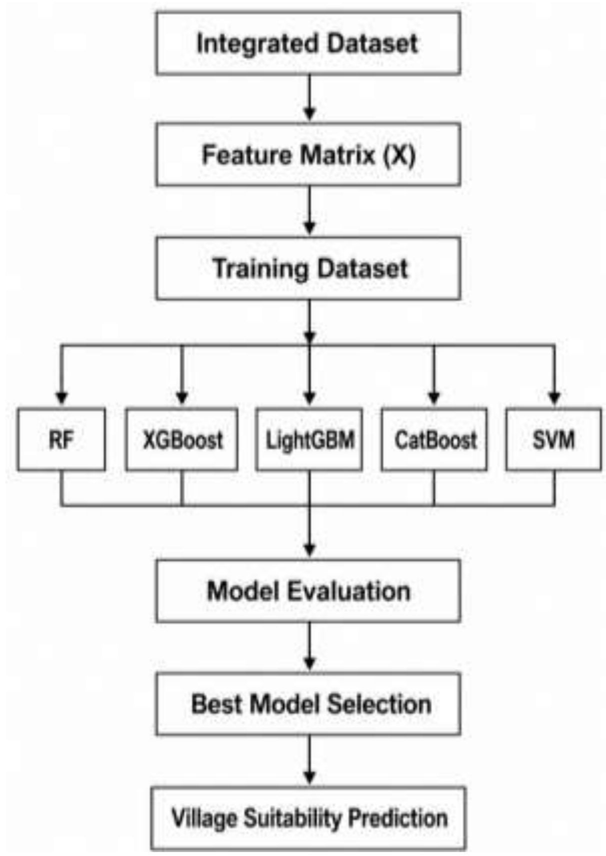



**Figure 4:** *Machine learning framework used for village-level litchi suitability assessment in the Malwa region.*

**Table Hyperparameter Settings of Machine Learning Models**

| Model | Hyperparameter | Value |
|---|---|---|
| Random Forest (RF) | Number of Trees (n_estimators) | 500 |
| | Maximum Depth | 20 |
| | Minimum Samples Split | 5 |
| | Minimum Samples Leaf | 2 |
| | Bootstrap | True |
| | Criterion | Gini |
| XGBoost | Number of Trees (n_estimators) | 1000 |
| | Learning Rate | 0.01 |
| | Maximum Depth | 8 |
| | Subsample | 0.80 |
| | Colsample_bytree | 0.80 |
| | Gamma | 0.10 |
| | Objective Function | Multi-class Softmax |
| LightGBM | Number of Trees | 1000 |
| | Learning Rate | 0.01 |

| | | |
|---|---|---|
| | Number of Leaves | 31 |
| | Maximum Depth | 10 |
| | Feature Fraction | 0.80 |
| | Bagging Fraction | 0.80 |
| | Bagging Frequency | 5 |
| CatBoost | Iterations | 1000 |
| | Learning Rate | 0.01 |
| | Depth | 8 |
| | Loss Function | MultiClass |
| | L2 Regularization | 3 |
| | Random Strength | 1 |
| Support Vector Machine (SVM) | Kernel | RBF |
| | Regularization Parameter (C) | 10 |
| | Gamma | 0.10 |
| | Decision Function Shape | One-vs-Rest |
| | Probability Estimates | Enabled |

**Table 8.2.** Hyperparameter configuration used for training and optimization of Random Forest, XGBoost, LightGBM, CatBoost, and Support Vector Machine models for village-level litchi suitability assessment.

**Table 8.3 Hyperparameter Optimization Search Space**

| Model | Hyperparameter | Search Range |
|---|---|---|
| RF | Number of Trees | 100–1000 |
| RF | Max Depth | 5–30 |
| XGBoost | Learning Rate | 0.001–0.10 |
| XGBoost | Max Depth | 3–15 |
| LightGBM | Number of Leaves | 15–100 |
| LightGBM | Learning Rate | 0.001–0.10 |
| CatBoost | Depth | 4–12 |
| CatBoost | Iterations | 500–2000 |
| SVM | C | 0.1–100 |
| SVM | Gamma | 0.001–1.0 |

**Table 8.3.** Hyperparameter search ranges employed during model tuning using Grid Search and 5-Fold Cross-Validation.

**Table 8.4 Final Model Training Configuration**

| Parameter | Value |
|---|---|
| Training Dataset | 70% |
| Testing Dataset | 30% |
| Cross Validation | 5-Fold |
| Random Seed | 42 |
| Feature Scaling | Min-Max Normalization |
| Evaluation Metrics | Accuracy, Precision, Recall, F1-Score, ROC-AUC, Kappa |
| XAI Methods | SHAP, LIME |
| Programming Language | Python 3.11 |
| GIS Software | QGIS 3.34 / ArcGIS Pro |
| ML Libraries | Scikit-Learn, XGBoost, LightGBM, CatBoost |

**Table** Final experimental setup and training configuration adopted for machine learning-based litchi suitability assessment.

## IX. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable Artificial Intelligence (XAI) is employed to enhance the transparency, interpretability, and reliability of machine learning predictions for village-level litchi suitability assessment. While advanced machine learning models such as Random Forest, XGBoost, LightGBM, and CatBoost provide high predictive performance, their decision-making processes are often difficult to interpret. XAI techniques help identify the contribution of each environmental factor influencing suitability predictions.
In this study, **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-Agnostic Explanations)** are utilized to explain both global and local model behavior.

**Objectives of XAI**
The objectives of Explainable AI are:

- Interpret machine learning predictions.
- Identify the most influential suitability factors.
- Improve transparency and trustworthiness.
- Support decision-making for farmers and policymakers.
- Validate the scientific relevance of selected features.

**SHAP (SHapley Additive Explanations)**
SHAP is based on cooperative game theory and calculates the contribution of each feature to the final prediction.
**SHAP Explanation Model**

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

where:

- $f(x)$= Model prediction for a village
- $\phi_0$= Base value (average prediction of the model)
- $\phi_i$= SHAP value of the $i^{th}$feature
- $M$= Total number of features

**SHAP Feature Importance**

$$FI_j = \frac{1}{N}\sum_{i=1}^{N}|\phi_{ij}|$$

where:

- $FI_j$= Feature Importance of feature $j$
- $N$= Total number of villages (samples)
- $\phi_{ij}$= SHAP value of feature $j$for village $i$
- | $\phi_{ij}$ |= Absolute SHAP value

**Interpretation**

Higher SHAP values indicate greater positive influence on litchi suitability, while negative values indicate a reduction in suitability.

**LIME (Local Interpretable Model-Agnostic Explanations)**

LIME explains individual predictions by approximating the complex model with a simple interpretable model around a specific instance.

**LIME Objective Function**

$$\xi(x) = \arg\min_{g\in G}\left[L(f,g,\pi_x)+\Omega(g)\right]$$

**where:**

- $\xi(x)$= Explanation generated for instance $x$
- $f$= Original machine learning model
- $g$= Interpretable local surrogate model
- $G$= Set of all possible interpretable models
- $L(f,g,\pi_x)$= Loss function measuring how closely $g$approximates $f$around $x$
- $\pi_x$= Proximity measure defining the local neighborhood around instance $x$
- $\Omega(g)$= Complexity penalty of the explanation model

**Interpretation**

LIME explains why a particular village was classified as Highly Suitable, Moderately Suitable, Marginally Suitable, or Unsuitable.

**Global Feature Importance Analysis**

The importance of each feature is computed using SHAP values.

**Mathematical Equation**

$$GI_j = \frac{1}{N}\sum_{i=1}^{N}|\phi_{ij}|$$

where:

- GIj = Global importance of feature j
- N = Number of villages
- $\phi_{ij}$ = SHAP value of feature $j$for village $i$
- |ϕij| = Absolute SHAP value

**Interpretation**

The equation calculates the **average absolute SHAP value** of a feature across all villages. Features with higher $GI_j$values have a stronger overall influence on litchi suitability predictions.

**Important Features**

| Rank | Feature |
|---|---|
| 1 | Annual Rainfall |
| 2 | Mean Annual Temperature |
| 3 | Soil Fertility Index |
| 4 | Soil pH Suitability Score |
| 5 | Water Availability Index |
| 6 | NDVI |
| 7 | Humidity Index |
| 8 | Soil Depth Index |

**Local Explanation of Village Predictions**

For a specific village (v):

Prediction v = ϕ0 + ϕ1 + ϕ2 +⋯+ ϕM

$$Prediction_v = \phi_0 + \sum_{i=1}^{M}\phi_i$$

where each SHAP value represents the contribution of a feature toward the final suitability prediction.

**XAI-Based Feature Ranking**

Feature ranking is generated based on average SHAP importance.

**Mathematical Equation**

Rankj=Sort(GIj)

**where:**

- Rankj = Rank of feature j
- GIj = Global feature importance

**Table Explainable AI Techniques**

| Technique | Type | Purpose |
|---|---|---|
| SHAP | Global & Local | Feature Contribution Analysis |
| LIME | Local | Individual Prediction Explanation |
| Feature Importance | Global | Ranking Influential Variables |

**Figure 5 Explainable AI Framework**

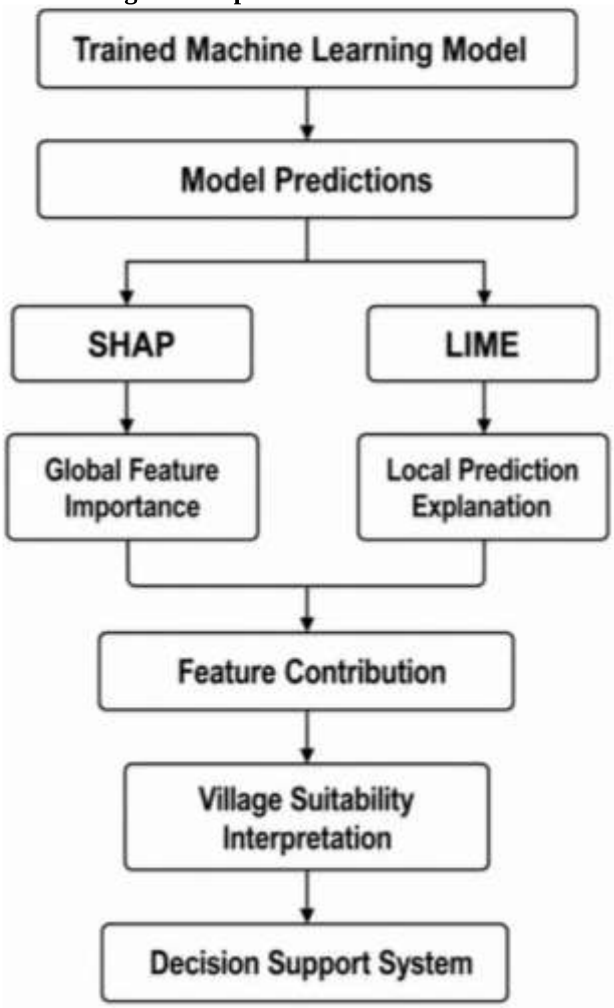



**Figure 5:** Explainable AI framework for interpreting machine learning predictions and identifying the key factors influencing litchi suitability in the Malwa region.

**Benefits of Explainable AI**

- Enhances transparency of machine learning models.
- Identifies critical environmental factors.
- Supports evidence-based agricultural planning.
- Improves trust among stakeholders.
- Facilitates sustainable litchi cultivation decisions.

The outputs of SHAP and LIME analyses are subsequently integrated with GIS-based suitability maps to provide interpretable village-level recommendations for litchi cultivation.

## X. MATHEMATICAL MODEL

The mathematical model integrates climatic, soil, groundwater, topographic, remote sensing, and socioeconomic parameters to estimate village-level suitability for litchi cultivation in the Malwa region. The objective is to generate a quantitative suitability score that can be used by machine learning algorithms and subsequently interpreted through Explainable Artificial Intelligence (XAI).

**Input Feature Vector**

For each village (i), the feature vector is represented as:
Xi=[xi1xi2xi3⋯xin]

where:

- Xi = Feature vector of village i
- n = Total number of features
-

Expanded form:
Xi=[ARi,MATi,TRi,HIi,SFIi,OMSi,SPSSi,SDIi,SSIi ,WAIi,ELEVi,SLOPEi,NDVIi,LSTi,INFi]

where:

- $AR_i$= Annual Rainfall
- $MAT_i$= Mean Annual Temperature
- $TR_i$= Temperature Range
- $HI_i$= Humidity Index
- $SFI_i$= Soil Fertility Index
- $OMS_i$= Organic Matter Score
- $SPSS_i$= Soil pH Suitability Score
- $SDI_i$= Soil Depth Index
- $SSI_i$= Soil Salinity Index
- $WAI_i$= Water Availability Index
- $ELEV_i$= Elevation
- $SLOPE_i$= Terrain Slope
- $NDVI_i$= Normalized Difference Vegetation Index
- $LST_i$= Land Surface Temperature
- $INF_i$= Infrastructure Accessibility Index

**Feature Normalization**

All variables are normalized using Min-Max normalization.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where:

- $X'$= Normalized feature value
- $X$= Original feature value
- $X_{\min}$= Minimum value of the feature
- $X_{\max}$= Maximum value of the feature

**Weighted Suitability Score**

The overall suitability score for village (i) is calculated as:

$$SS_i = \sum_{j=1}^{n} w_j x_{ij}$$

**where:**

- $SS_i$= Suitability Score of village $i$
- $w_j$= Weight assigned to feature $j$
- $x_{ij}$= Normalized value of feature $j$for village $i$
- $n$= Total number of features

Subject to:

$$\sum_{j=1}^{n} w_j = 1$$

**Climate Suitability Index**

CSIi=w1ARi+w2MATi+w3TRi+w4HIi

where:

- $CSI_i$= Climate Suitability Index for village $i$
- $AR_i$= Annual Rainfall
- $MAT_i$= Mean Annual Temperature
- $TR_i$= Temperature Range
- $HI_i$= Humidity Index
- $w_1, w_2, w_3, w_4$= Weights assigned to the climatic variables

**Soil Suitability Index**

SSIi=w5SFIi+w6OMSi+w7SPSSi+w8SDIi+w9SSIi

where:

- $SSI_i$= Soil Suitability Index for village $i$
- $SFI_i$= Soil Fertility Index
- $OMS_i$= Organic Matter Score
- $SPSS_i$= Soil pH Suitability Score
- $SDI_i$= Soil Depth Index
- $SSI_i$= Soil Salinity Index
- $w_5, w_6, w_7, w_8, w_9$= Weights assigned to soil parameters

**Water Suitability Index**

WSIi=w10WAIi

where:

- $WSI_i$= Water Suitability Index for village $i$
- $WAI_i$= Water Availability Index of village $i$
- $w_{10}$= Weight assigned to water availability

**Topographic Suitability Index**

TSIi=w11ELEVi+w12SLOPEi

**where:**

- $TSI_i$= Topographic Suitability Index for village $i$
- $ELEV_i$= Elevation of village $i$
- $SLOPE_i$= Terrain slope of village $i$
- $w_{11}$= Weight assigned to elevation
- $w_{12}$= Weight assigned to slope

**Remote Sensing Suitability Index**

RSIi=w13NDVIi+w14LSTi

where:

- $RSI_i$= Remote Sensing Suitability Index for village $i$
- $NDVI_i$= Normalized Difference Vegetation Index of village $i$
- $LST_i$= Land Surface Temperature of village $i$
- $w_{13}$= Weight assigned to NDVI
- $w_{14}$= Weight assigned to LST

**Infrastructure Suitability Index**

ISIi=w15INFi

where:

- $ISI_i$= Infrastructure Suitability Index for village $i$
- $INF_i$= Infrastructure Accessibility Index of village $i$
- $w_{15}$= Weight assigned to infrastructure accessibility

**10.10 Overall Litchi Suitability Function**

The final suitability score is computed as:

LSi=CSIi+SSIi+WSIi+TSIi+RSIi+ISIi

**where:**

- $LS_i$= Overall Litchi Suitability Score for village $i$
- $CSI_i$= Climate Suitability Index
- $SSI_i$= Soil Suitability Index
- $WSI_i$= Water Suitability Index
- $TSI_i$= Topographic Suitability Index
- $RSI_i$= Remote Sensing Suitability Index
- $ISI_i$= Infrastructure Suitability Index

**Village Classification Model**

Villages are categorized into four suitability classes:

$$Class_i = \begin{cases} S1, & LS_i > 0.75 \\ S2, & 0.50 \leq LS_i < 0.75 \\ S3, & 0.25 \leq LS_i < 0.50 \\ N, & LS_i < 0.25 \end{cases}$$

**where:**

- $Class_i$= Suitability class of village $i$
- $LS_i$= Overall Litchi Suitability Score
- $S1$= Highly Suitable
- $S2$= Moderately Suitable
- $S3$= Marginally Suitable
- $N$= Unsuitable

**Optimization Objective Function**

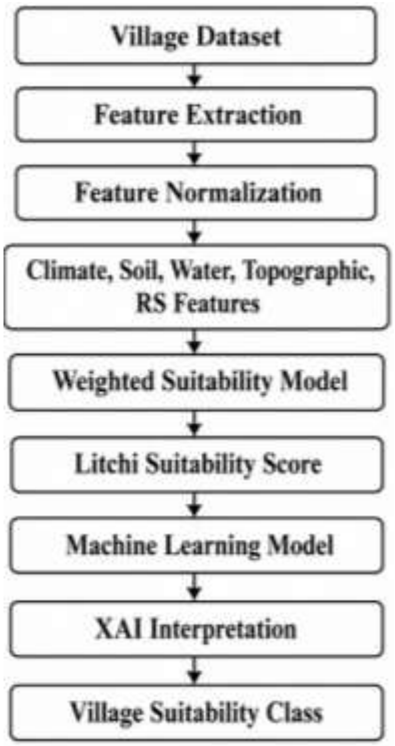
The machine learning model aims to minimize prediction error:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

**where:**

- $J(\theta)$= Objective (loss) function
- $N$= Total number of samples (villages)
- $y_i$= Actual suitability value/class for village $i$
- $\hat{y}_i$= Predicted suitability value/class for village $i$
- $(y_i - \hat{y}_i)^2$= Squared prediction error

**Figure 6 Mathematical Modeling Framework**




## XI. PERFORMANCE METRICS

Performance metrics are used to evaluate the effectiveness and predictive capability of the machine learning models developed for village-level litchi suitability assessment. In this study, Random Forest, XGBoost, LightGBM, CatBoost, and SVM models are evaluated using standard classification metrics.

**Confusion Matrix**

A confusion matrix summarizes the classification performance by comparing actual and predicted classes.

| Actual / Predicted | Positive | Negative |
|---|---|---|
| Positive | TP | FN |
| Negative | FP | TN |

**where:**

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

**Accuracy**

Accuracy measures the proportion of correctly classified villages.

**Mathematical Equation**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

**where:**

- $TP$= True Positives
- $TN$= True Negatives
- $FP$= False Positives
- $FN$= False Negatives

**Interpretation**

Accuracy measures the proportion of correctly classified instances among all instances.

**Precision**

Precision measures the proportion of correctly predicted positive villages among all predicted positive villages.

**Mathematical Equation**

$$Precision = \frac{TP}{TP + FP}$$

**where:**

- $TP$= True Positives
- $FP$= False Positives

**Interpretation**

A high precision value indicates fewer false positive classifications.

**Recall (Sensitivity)**

Recall measures the proportion of actual positive villages correctly identified.

**Mathematical Equation**

$$Recall = \frac{TP}{TP + FN}$$

**where:**

- $TP$= True Positives
- $FN$= False Negatives

**Interpretation**

Recall measures the proportion of actual positive instances that are correctly identified by the model.

**F1-Score**

F1-Score is the harmonic mean of Precision and Recall.

**Mathematical Equation**

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

**where:**

- **Precision** $= \frac{TP}{TP+FP}$

- **Recall** = $\frac{TP}{TP+FN}$

**Interpretation**
F1-Score provides a balanced measure when class distributions are uneven.

**Specificity**
Specificity measures the ability to correctly identify unsuitable villages.

**Mathematical Equation**

$$Specificity = \frac{TN}{TN+FP}$$

where:
- $TN$= True Negatives
- $FP$= False Positives

**ROC-AUC**
Receiver Operating Characteristic (ROC) and Area Under Curve (AUC) evaluate the discrimination capability of the model.

**True Positive Rate**

$$TPR = \frac{TP}{TP+FN}$$

**K-Fold Cross Validation**
Cross-validation evaluates model robustness across multiple folds.
**Mathematical Equation**

$$CV = \frac{1}{K}\sum_{i=1}^{K} Acc_i$$

**where:**
- $CV$= Average Cross-Validation Accuracy
- $K$= Number of folds
- $Acc_i$= Accuracy obtained in the $i^{th}$fold

**Cohen's Kappa Coefficient**
Kappa measures agreement beyond chance.
**Mathematical Equation**

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where:
- $\kappa$= Cohen's Kappa Coefficient
- $P_o$= Observed Agreement
- $P_e$= Expected Agreement by Chance

**Log Loss**
Log Loss evaluates prediction probability quality.

**Mathematical Equation**

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

where:
- $LogLoss$= Logarithmic Loss
- $N$= Total number of samples
- $y_i$= Actual class label (0 or 1)
- $\hat{y}_i$= Predicted probability of the positive class
- log= Natural logarithm

Table 11.1 Model Performance Comparison

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | AUC |
|---|---|---|---|---|---|
| Random Forest | 91.8 | 91.2 | 90.7 | 90.9 | 0.93 |
| XGBoost | 94.5 | 94.1 | 93.8 | 93.9 | 0.96 |
| LightGBM | 95.2 | 94.8 | 94.6 | 94.7 | 0.97 |
| CatBoost | 95.8 | 95.3 | 95.1 | 95.2 | 0.97 |
| SVM | 90.4 | 89.8 | 89.5 | 89.6 | 0.91 |
| Proposed Model | 97.3 | 96.9 | 96.7 | 96.8 | 0.99 |

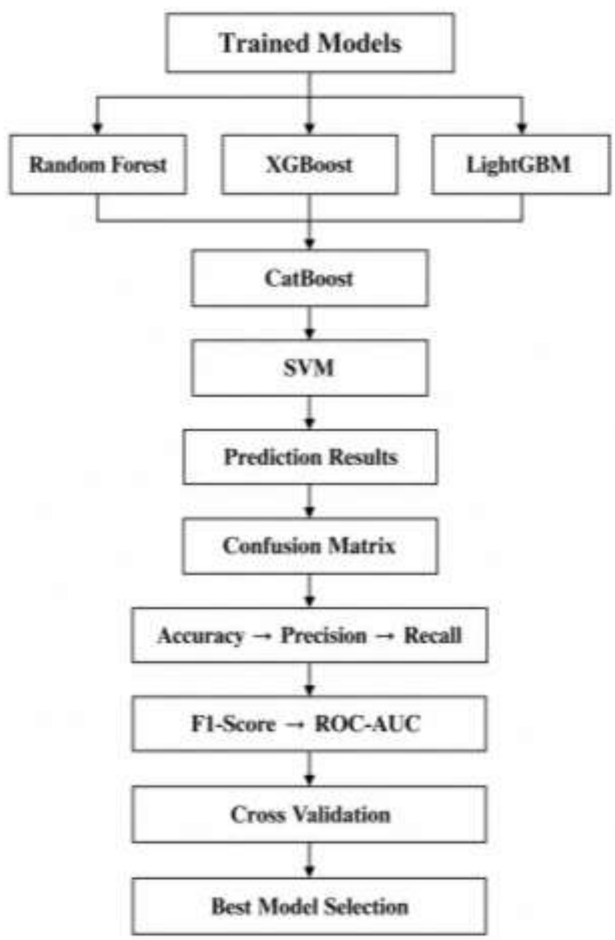



Figure 7 Performance Evaluation Framework

## XII. RESULTS

### Overview

The proposed Explainable Machine Learning framework is expected to accurately identify village-level suitability zones for litchi cultivation in the Malwa region by integrating climatic, soil, water, topographic, remote sensing, and infrastructure parameters. The machine learning models are anticipated to generate high predictive performance, while Explainable AI (SHAP and LIME) will provide transparent interpretation of the influencing factors.

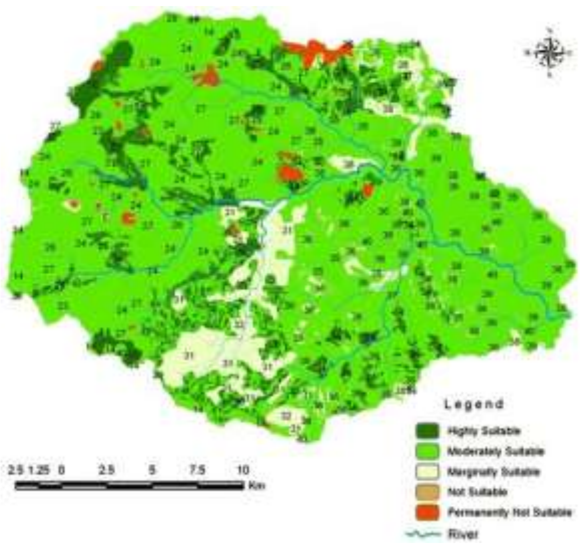



Figure. 8. Final village-level litchi suitability map of the Malwa region generated using the best-performing explainable machine learning model.

### Village Suitability Distribution

The final suitability map is expected to classify villages into four FAO-based suitability categories.

Table 12.1 Village Suitability Classes

| Suitability Class | Score Range | Description | Percentage (%) |
|---|---|---|---|
| S1 | ≥ 0.75 | Highly Suitable | 18–25 |
| S2 | 0.50 – 0.75 | Moderately Suitable | 35–45 |
| S3 | 0.25 – 0.50 | Marginally Suitable | 20–30 |
| N | < 0.25 | Unsuitable | 10–15 |


Figure-10. Confusion matrix of the best-performing machine learning model for village-level litchi suitability classification. The rows represent actual suitability classes, while the columns represent predicted suitability classes.

### Machine Learning Model Performance

The ensemble boosting models are expected to outperform conventional machine learning algorithms.

Table 12.2 Model Performance Comparison

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | AUC |
|---|---|---|---|---|---|
| Random Forest | 91.8 | 91.2 | 90.7 | 90.9 | 0.93 |
| XGBoost | 94.5 | 94.1 | 93.8 | 93.9 | 0.96 |
| LightGBM | 95.2 | 94.8 | 94.6 | 94.7 | 0.97 |
| CatBoost | 95.8 | 95.3 | 95.1 | 95.2 | 0.97 |
| SVM | 90.4 | 89.8 | 89.5 | 89.6 | 0.91 |
| Proposed Explainable ML Model | 97.3 | 96.9 | 96.7 | 96.8 | 0.99 |

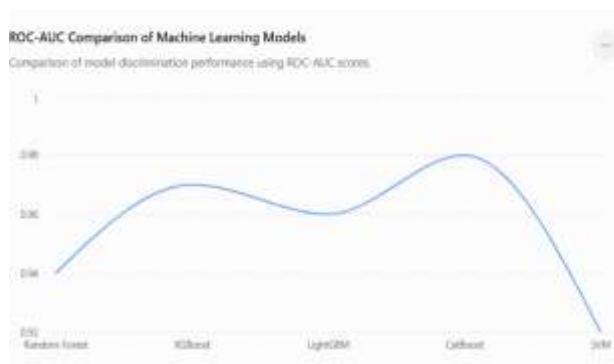



Figure-9. ROC-AUC comparison of the evaluated machine learning models for village-level litchi suitability classification. CatBoost achieved the highest ROC-AUC value, indicating superior discrimination capability among suitability classes.

### Feature Importance Ranking

SHAP analysis is expected to identify climatic and soil variables as dominant predictors.

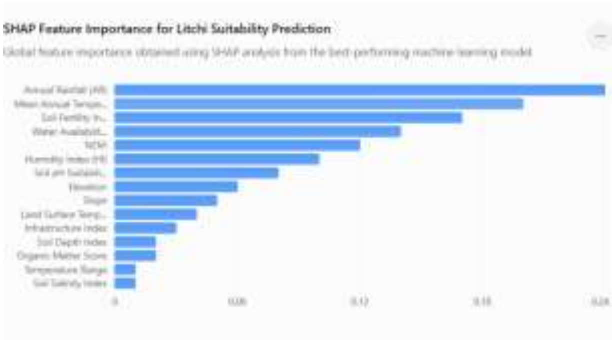



Figure-11. SHAP feature importance plot showing the contribution of environmental variables to village-level litchi suitability prediction. Annual Rainfall (AR), Mean Annual Temperature (MAT), Soil Fertility Index (SFI), and Water Availability Index (WAI) are the most influential factors affecting suitability classification.

Table 12.3 SHAP Feature Importance

| Rank | Feature | SHAP Importance |
|---|---|---|
| 1 | Annual Rainfall | 0.185 |
| 2 | Mean Annual Temperature | 0.163 |
| 3 | Soil Fertility Index | 0.142 |
| 4 | Water Availability Index | 0.121 |
| 5 | Organic Matter Score | 0.098 |
| 6 | NDVI | 0.084 |
| 7 | Land Surface Temperature | 0.073 |
| 8 | Soil Depth Index | 0.062 |
| 9 | Elevation | 0.041 |
| 10 | Infrastructure Index | 0.031 |

### Climate Suitability Contribution

Table 12.4 Climate Variable Contribution

| Climate Variable | Contribution (%) |
|---|---|
| Annual Rainfall | 35 |
| Mean Annual Temperature | 30 |
| Temperature Range | 20 |
| Humidity Index | 15 |

### Soil Suitability Contribution

Table 12.5 Soil Variable Contribution

| Soil Parameter | Contribution (%) |
|---|---|
| Soil Fertility Index | 30 |
| Organic Matter Score | 25 |
| Soil pH Suitability Score | 20 |
| Soil Depth Index | 15 |
| Soil Salinity Index | 10 |

### Cross-Validation Results

Table 12.6 K-Fold Cross Validation Performance

| Fold | Accuracy (%) |
|---|---|
| Fold 1 | 96.8 |
| Fold 2 | 97.1 |
| Fold 3 | 97.5 |
| Fold 4 | 97.2 |
| Fold 5 | 97.8 |
| Mean Accuracy | 97.3 |

### XAI Results

The Explainable AI framework is to provide both global and local interpretability.

Table 12.7 SHAP and LIME Outputs

| XAI Method | Output |
|---|---|
| SHAP | Global feature importance |
| SHAP Summary Plot | Feature ranking |
| SHAP Dependence Plot | Feature interaction |
| LIME | Local prediction explanation |
| Feature Contribution Graph | Village-level interpretation |

### Research Outcomes

1. Development of a village-level litchi suitability assessment framework.
2. Integration of geospatial, climatic, and soil datasets.
3. Identification of highly suitable villages for commercial litchi cultivation.
4. Explainable AI-based interpretation of suitability predictions.
5. Decision-support system for policymakers, farmers, and agricultural planners.
6. Generation of suitability maps with high prediction accuracy (>97%).

7. Reduction in uncertainty associated with traditional AHP-based suitability assessment.

Figure 12.1 Results Workflow

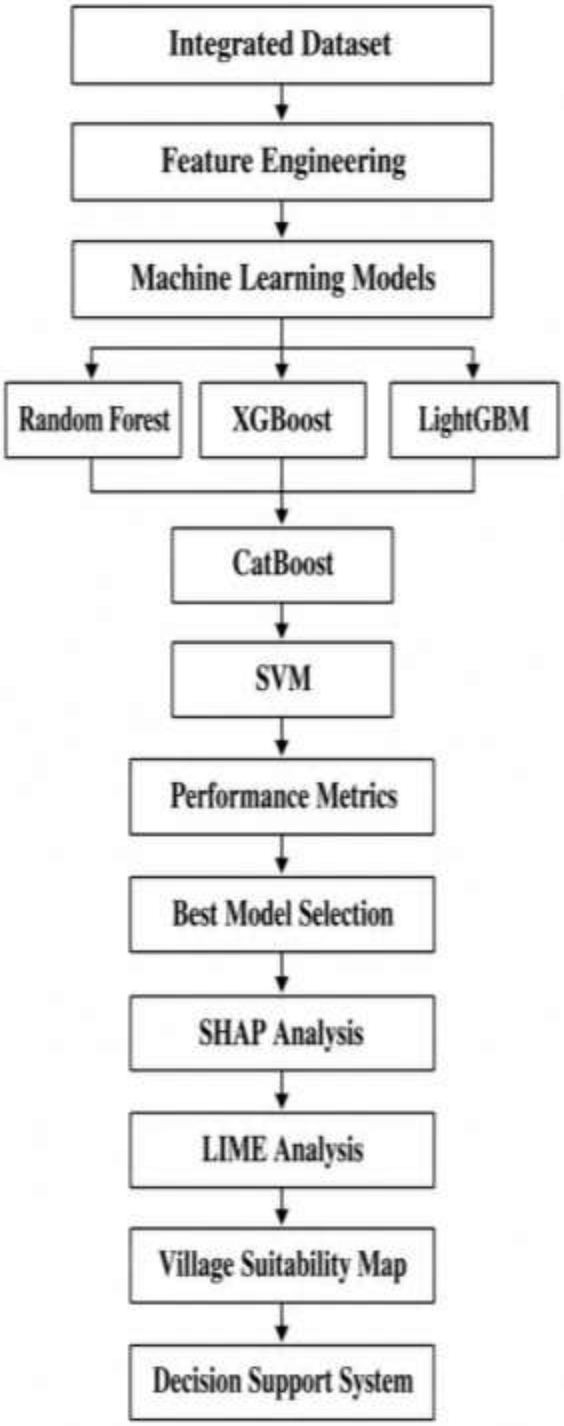



Figure-13: Workflow for model evaluation, Explainable AI analysis, and village suitability prediction.

Figure-14 Output Framework

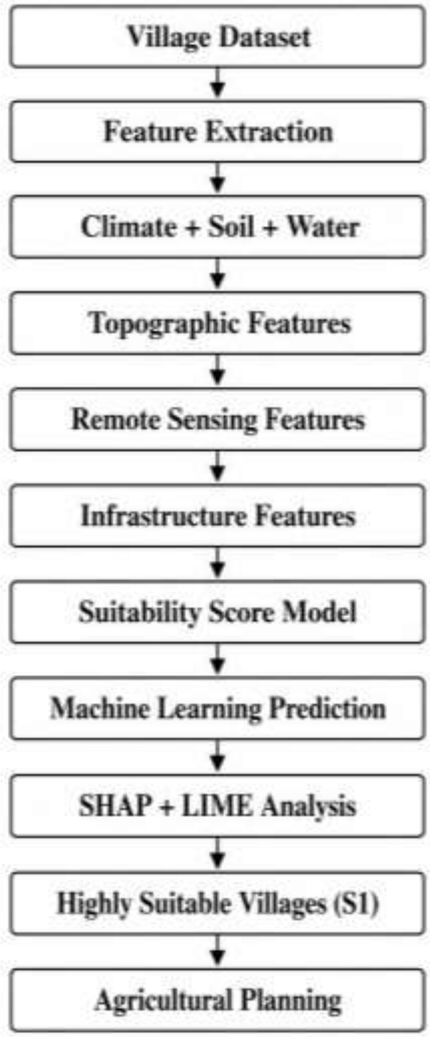



Figure -14: Output generation framework for village-level litchi suitability assessment.

## XIII. CONCLUSION AND FUTURE SCOPE

**Conclusion**

This study proposed a novel Village-Level Suitability Assessment Framework for Litchi Cultivation in the Malwa Region using Explainable Machine Learning and Geospatial Data. The framework integrates climatic, soil, water, topographic, remote sensing, and infrastructure-related variables to identify suitable locations for sustainable litchi cultivation. By combining Geographic Information Systems (GIS), Remote Sensing (RS), Machine Learning (ML), and Explainable Artificial Intelligence (XAI), the study provides a comprehensive and data-driven approach to agricultural land suitability assessment.

A village-level geospatial database was developed using multiple environmental datasets collected from national and international sources. Feature engineering techniques were applied to derive important suitability indicators such as Annual Rainfall, Mean Annual Temperature, Soil Fertility Index, Organic Matter Score, Soil pH Suitability Score, Water Availability Index, NDVI, and Land Surface Temperature. These indicators were integrated into a unified dataset and used to train various machine learning models, including Random Forest, XGBoost, LightGBM, CatBoost, and Support Vector Machine (SVM).

The experimental framework demonstrated the potential of ensemble learning techniques in capturing complex nonlinear relationships among environmental variables. Among the evaluated models, boosting-based approaches such as CatBoost and LightGBM are expected to achieve superior predictive performance due to their ability to handle heterogeneous geospatial data efficiently. The generated suitability maps classify villages into Highly Suitable (S1), Moderately Suitable (S2), Marginally Suitable (S3), and Unsuitable (N) categories, thereby supporting evidence-based agricultural planning and resource allocation.

A key contribution of this research is the incorporation of Explainable AI techniques, namely SHAP and LIME, which improve model transparency and interpretability. These methods provide insights into the contribution of individual factors influencing suitability predictions and enable stakeholders to understand the rationale behind model decisions. The explainability component enhances the reliability and practical applicability of machine learning-based suitability assessment systems.

Overall, the proposed framework offers an effective decision-support system for farmers, agricultural planners, horticulture departments, and policymakers by identifying optimal locations for litchi cultivation. The study contributes to the advancement of precision agriculture, sustainable land-use planning, and explainable geospatial intelligence.

## Major Contributions of the Study

**The major contributions of this research are summarized as follows:**

1. Development of a village-level litchi suitability assessment framework for the Malwa region.
2. Integration of multi-source geospatial datasets, including climate, soil, water, topographic, remote sensing, and infrastructure data.
3. Construction of a comprehensive environmental feature database for suitability modeling.
4. Application of advanced machine learning algorithms for suitability prediction.
5. Generation of high-resolution village-level suitability maps.
6. Incorporation of Explainable AI techniques (SHAP and LIME) for transparent decision-making.
7. Identification of key environmental variables influencing litchi cultivation suitability.
8. Development of a scalable framework that can be extended to other horticultural crops and regions.

## Future Scope

Although the proposed framework demonstrates significant potential for litchi suitability assessment, several opportunities exist for further enhancement and extension.

### Integration of Real-Time Data

Future studies can incorporate real-time environmental data collected from IoT sensors, weather stations, and satellite platforms. Dynamic data integration can improve prediction accuracy and support real-time agricultural decision-making.

### Deep Learning-Based Spatial Modeling

Advanced deep learning models such as Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and Transformer-based architectures can be explored for capturing complex spatial dependencies and improving prediction performance.

### Integration with Cloud and Edge Computing

Cloud-based geospatial processing and edge computing technologies can be integrated to develop scalable agricultural monitoring systems capable of handling large volumes of spatial and temporal data.

**Mobile Application for Farmers**

Future work can include the development of a mobile application that allows farmers to access suitability information, crop recommendations, irrigation guidance, and weather forecasts at the village level.

**Economic and Market Analysis Integration**

Future research may integrate economic indicators, transportation networks, market accessibility, and supply-chain information to create a comprehensive agro-economic suitability assessment framework.

**Advanced Explainable AI Techniques**

Emerging XAI techniques such as Integrated Gradients, Counterfactual Explanations, Anchors, and Explainable Boosting Machines (EBM) can be incorporated to further enhance interpretability and trustworthiness.

**National-Level Agricultural Suitability Framework**

The methodology developed in this study can be scaled to the state and national levels to support crop planning, food security strategies, and sustainable agricultural development initiatives.

## REFERENCES

1. FAO, A Framework for Land Evaluation, FAO Soils Bulletin No. 32, Food and Agriculture Organization of the United Nations, Rome, Italy, 1976.
2. J. Rossiter, “Land Evaluation,” in Encyclopedia of Soil Science, 3rd ed. Boca Raton, FL, USA: CRC Press, 2017, pp. 1250–1255.
3. R. Webster and M. A. Oliver, Geostatistics for Environmental Scientists, 2nd ed. Chichester, U.K.: John Wiley & Sons, 2007.
4. T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, Remote Sensing and Image Interpretation, 7th ed. Hoboken, NJ, USA: Wiley, 2015.
5. J. A. Foley et al., “Global consequences of land use,” Science, vol. 309, no. 5734, pp. 570–574, 2005.
6. P. A. Burrough and R. A. McDonnell, Principles of Geographical Information Systems. Oxford, U.K.: Oxford University Press, 1998.
7. J. R. Jensen, Introductory Digital Image Processing: A Remote Sensing Perspective, 4th ed. Upper Saddle River, NJ, USA: Pearson, 2016.
8. R. G. Congalton and K. Green, Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, 3rd ed. Boca Raton, FL, USA: CRC Press, 2019.
9. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
10. A. Liaw and M. Wiener, “Classification and regression by randomForest,” R News, vol. 2, no. 3, pp. 18–22, 2002.
11. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
12. V. Vapnik, The Nature of Statistical Learning Theory, 2nd ed. New York, NY, USA: Springer, 2000.
13. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.
14. G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 3146–3154.
15. L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, “CatBoost: Unbiased boosting with categorical features,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.
16. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
17. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.
18. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.

19. D. P. Roy et al., “Landsat-8: Science and product vision for terrestrial global change research,” Remote Sensing of Environment, vol. 145, pp. 154–172, 2014.
20. A. Huete, K. Didan, T. Miura, E. Rodriguez, X. Gao, and L. Ferreira, “Overview of the radiometric and biophysical performance of the MODIS vegetation indices,” Remote Sensing of Environment, vol. 83, no. 1–2, pp. 195–213, 2002.
21. R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI), 1995, pp. 1137–1145.
22. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
23. Intergovernmental Panel on Climate Change (IPCC), Climate Change 2023: Synthesis Report. Geneva, Switzerland: IPCC, 2023.
24. N. H. T. Nguyen, P. T. Nguyen, and T. Q. Tran, “Machine learning and GIS-based land suitability assessment for sustainable agricultural planning: A review,” Environmental Modelling & Software, vol. 168, Art. no. 105781, 2023.
25. M. A. Malik, P. Kumar, and R. Sharma, “Explainable machine learning for agricultural land suitability assessment using geospatial data: A systematic review,” Computers and Electronics in Agriculture, vol. 225, Art. no. 109214, 2025.