Kaunas University of Technology

Faculty of Informatics

Crop Yield Estimation using Multispectral Satellite Imagery and Machine Learning

Master's Final Degree Project

Jad Kaedbey

Project author

Assoc. Prof. Mantas Lukoševičius

Supervisor

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Crop Yield Estimation using Multispectral Satellite Imagery and Machine Learning

Master's Final Degree Project

Artificial Intelligence in Computer Science (6211BX007)

Jad Kaedbey

Project author

Assoc. Prof. Mantas Lukoševičius

Supervisor

Doc. Dr. Kęstutis Jankauskas

Reviewer

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Crop Yield Estimation using Multispectral Satellite Imagery and Machine Learning

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Jad Kaedbey

Confirmed electronically

Kaedbey, Jad. Crop Yield Estimation using Multispectral Satellite Imagery and Machine Learning Master's Final Degree Project / supervisor Assoc. Prof. Mantas Lukoševičius; Faculty of Informatics, Kaunas University of Technology.
Study field and area (study field group): Computer Science, Informatics (B01).
Keywords: crop yield estimation, multispectral satellite imagery, machine learning, Sentinel-2, vegetation indices, XGBoost.
Kaunas, 2026. 58 pages.

Summary

This Master's thesis outlines the design and implementation of an automated system for estimating crop yields on a large scale using Sentinel-2 satellite imagery and machine learning. The research focuses on integrating time-series multispectral data with ground-truth statistics from the US Department of Agriculture's National Agricultural Statistics Service (NASS). A diverse suite of vegetation indices (including NDVI, EVI and NDRE) was calculated and combined with historical meteorological data to develop a comprehensive feature set that captures the complex dynamics of crop growth. The study compares the performance of tree-based ensemble models (specifically XGBoost and Random Forest) with that of experimental deep learning architectures, such as Long Short-Term Memory (LSTM) networks. The results show that, when enhanced with engineered weather features, tree-based models provide robust and scalable yield predictions at county level, offering a viable alternative to traditional reporting methods in precision agriculture.

Kaedbey, Jad. Derliaus prognozavimas naudojant multispektrinius palydovinius vaizdus ir mašininį mokymąsi. Magistro baigiamasis projektas / vadovas dr. Mantas Lukoševičius; Kauno technologijos universitetas, Informatikos fakultetas.
Studijų kryptis ir sritis (studijų krypčių grupė): Informatikos mokslai, Informatika (B01).
Raktažodžiai: derliaus įvertinimas, daugiaspektriniai palydoviniai vaizdai, mašininis mokymasis, „Sentinel-2“, augmenijos indeksai, „XGBoost“.
Kaunas, 2026 m. 58 puslapiai.

Santrauka

Šiame magistro darbe pristatoma automatizuota sistema, kuri yra skirta didelio masto derlingumo įvertinimui naudojant „Sentinel-2“ palydovinius vaizdus ir mašininį mokymąsi. Tyrimas sutelktas į laiko eilučių multispektrinių duomenų integravimą su faktiniais statistiniais duomenimis, gautais iš JAV Žemės ūkio departamento Nacionalinės žemės ūkio statistikos tarnybos (NASS). Buvo apskaičiuotas įvairus augmenijos indeksų rinkinys (įskaitant NDVI, EVI ir NDRE) ir sujungtas su istoriniais meteorologiniais duomenimis, siekiant sukurti išsamų požymių rinkinį, atspindintį sudėtingą pasėlių augimo dinamiką. Tyrime lyginami medžiais pagrįstų ansamblio modelių (konkrečiai XGBoost ir Random Forest) veikimas su eksperimentinių giluminio mokymosi architektūrų, pvz., ilgalaikės trumpalaikės atminties tinklų (LSTM), veikimu. Rezultatai rodo, kad kai medžiais pagrįsti modeliai yra patobulinti inžineriniais oro sąlygų požymiais, jie užtikrina patikimus ir pritaikytinus derlingumo prognozavimus apskrities lygiu, siūlant perspektyvią alternatyvą tradiciniams ataskaitų teikimo metodams tiksliojoje žemdirbystėje.

Table of contents

List of figures

List of tables

List of abbreviations and terms

Abbreviations:

ML – Machine Learning

DL – Deep Learning

UAV – Unmanned Aerial Vehicles

NDVI - Normalized Difference Vegetation Index

EVI - Enhanced Vegetation Index

SVM – Support Vector Machine

CNN – Convolutional Neural Networks

RNN - Recurrent Neural Networks

LSTM - Long short-term memory

MSI – Multi-spectral Imaging

RMSE – Root Mean Squared Error

MAE – Mean Absolute Error

R^2 – Coefficient of Determination

NDRE – Normalized Difference Red Edge

GNDVI – Green Normalized Difference Vegetation Index

SAVI – Soil-Adjusted Vegetation Index

MCARI – Modified Chlorophyll Absorption Ratio Index

NASS – National Agricultural Statistics Service

FIPS – Federal Information Processing Standard

GDD – Growing Degree Days

AOI – Area of Interest

GIS – Geographic Information System

API – Application Programming Interface

CSV – Comma-Separated Values

JSON – JavaScript Object Notation

Terms:

Hyperspectral imaging – Hyperspectral imaging refers to the acquisition of data across a wide and continuous range of electromagnetic wavelengths, often capturing hundreds of narrow spectral bands for each pixel in an image. By measuring subtle differences in reflectance and absorption patterns, hyperspectral imaging provides detailed information about the chemical, biological, or physical characteristics of a material or scene.

Introduction

Project Novelty and Relevance

Reliable crop yield forecasts are valuable for food security planning, commodity markets, and farm management. The challenge is that yield varies substantially from year to year and location to location, driven by weather, soil, and management interactions that are difficult to observe directly at scale. Satellite remote sensing offers a partial solution: multispectral imagery from sensors like Sentinel-2 captures how plant canopies look at different wavelengths, which correlates with physiological state. Single-date spectral indices such as NDVI and EVI have been used for yield estimation, but a single snapshot of a field misses the seasonal trajectory that reflects whether a crop experienced stress during critical growth stages.

This study builds monthly time-series of six vegetation indices from Sentinel-2 imagery across 26 US Corn Belt counties and feeds them, combined with weather variables, into machine learning models for corn yield prediction. Tree-based models (XGBoost, Random Forest, Gradient Boosting) and deep learning sequence models (LSTM, CNN) are evaluated. The approach is designed to be county-scalable and reproducible, with open data sources (Sentinel-2, USDA NASS, NOAA GHCND) and a modular Python pipeline.

Most county-level yield studies use NDVI as the sole spectral predictor, or flatten the whole growing season into one annual feature vector. Either way, the temporal trajectory through the season gets discarded. This work instead tracks six vegetation indices month-by-month across 26 Corn Belt counties over 8 seasons, pairs the spectral data with historical weather variables, and directly compares tree-based ensembles against sequential deep learning models using leave-one-year-out cross-validation. That combination has not been applied to US county-scale corn estimation in the literature reviewed for this thesis.

Main Aim

To design, implement, and validate an automated system for large-scale crop yield estimation using Sentinel-2 satellite imagery and machine learning models, optimized for accurate and scalable county-level forecasting.

Objectives

1. To analyze and compare the effectiveness of various spectral vegetation indices (including NDVI, EVI, and NDRE) in capturing subtle variations in crop physiology and yield potential.

2. To develop a robust and automated data processing pipeline for the acquisition, atmospheric correction, and temporal aggregation of Sentinel-2 multispectral time-series data.
3. To engineer a comprehensive feature set that integrates spectral-temporal patterns with historical meteorological variables to enhance the robustness of yield prediction models.
4. To implement and evaluate multiple machine learning architectures, specifically comparing the performance of tree-based ensemble methods (e.g., XGBoost) and deep learning models (e.g., LSTM) for yield forecasting.
5. To validate the system's predictive accuracy and generalizability through systematic testing against authoritative USDA NASS ground-truth data across diverse agricultural regions.

Document Structure

The document is organized into the following chapters:

- **Chapter 1: Analysis** – Provides a comprehensive survey of existing literature, remote sensing technologies (ground-based, UAV, and satellite), and state-of-the-art machine learning methods currently employed in precision agriculture.
- **Chapter 2: Project** – Details the design phase of the system, including functional and technical requirements, architectural design decisions, and spectral index selection rationale.
- **Chapter 3: Implementation, Experiments, and Results** – Describes the technical realization of the data pipeline, the implementation of machine learning models, and the detailed analysis of results and validation against ground-truth data.
- **Conclusions** – Summarizes the findings of the research, evaluates the success of each objective, and provides recommendations for future work.

Use of artificial intelligence tools

During the preparation of this final degree project, Google Gemini was used for brainstorming purposes only, specifically to explore possible angles on the topic and generate candidate ideas during the early stages of the work. Gemini was not used to generate the text, analytical content, design of the proposed solution, implementation, experimental results, or conclusions of this project. All ideas obtained through brainstorming were critically evaluated, verified against the literature, and developed independently by the author, who takes full responsibility for the content of this project.

1. Analysis of Satellite Remote Sensing and Machine Learning for Crop Yield Estimation

Precision agriculture increasingly relies on satellite remote sensing to improve crop management and yield prediction. Both multispectral and hyperspectral imaging have contributed to this effort. Multispectral sensors such as Sentinel-2, which captures data across 13 spectral bands at 10–20 m resolution with a 5-day revisit cycle, offer a practical balance between spectral detail and operational coverage. Hyperspectral imaging (HSI) acquires data across hundreds of narrow bands, enabling finer spectral discrimination but at the cost of lower spatial or temporal resolution. This review covers both approaches, with emphasis on multispectral time-series methods that are most relevant to the present study.

Alongside advances in sensing platforms, ground-based systems, unmanned aerial vehicles (UAVs), and satellite missions, machine learning (ML) and deep learning (DL) methods have become central tools for extracting value from remote sensing data. In the papers analyzed in this literature review, several ML methods such as convolutional neural networks (CNNs) and tree-based ensemble models have been applied to both hyperspectral and multispectral datasets for crop classification, weed detection, disease monitoring, yield prediction, and resource optimization.

This literature review was channeled not only into surveying successful methods, but also into the persistent obstacles this field is facing. The analyzed challenges are various, including the high dimensionality and volume of remote sensing data, the need for effective preprocessing and feature engineering, and platform-specific challenges such as limited spatial or temporal resolution in satellite systems.

A systematic discussion of each study's technical approaches, results, and challenges is provided below.

1.1. Existing Solutions

A range of solutions has emerged in recent years to address the technical and practical challenges of remote sensing data acquisition and analysis in agriculture. These efforts can be broadly divided into two categories: hardware-oriented strategies, which involve ground-based, UAV-based, and satellite-based platforms, and software-oriented analytics, spanning conventional machine learning to deep learning methods. The following sections explore how various sensors and platforms collect remote sensing data, as well as how different analytical frameworks transform these measurements into useful predictions for crop management, yield estimation, and resource optimization. The three main types of remote sensing platforms studied are illustrated in Figures 1–3.

1.1.1. Close-Range Systems (Ground-Based):

Close-range (ground-based) hyperspectral imaging has gained prominence for its ability to acquire extremely high spatial resolution data - down to centimeter or even sub-centimeter scales - enabling detailed agricultural analysis in both laboratory and field settings. Sensors are mounted on static or mobile platforms such as linear stages, scaffolding or trucks, while illumination sources range from halogen lamps indoors to direct sunlight outdoors, facilitating the study of subtle plant and soil characteristics.

Operating at close range, these systems can detect small but critical reflectance changes that indicate early signs of crop stress, including disease outbreaks, weed infestations and nutrient deficiencies (e.g. changes in chlorophyll and sugar levels) before visual symptoms appear [1], [2], [3]. Targeted disease diagnosis is another major application; by isolating specific regions of the leaf for spectral analysis, sensors can distinguish healthy tissue from diseased patches in crops as diverse as banana leaves with black streak and soybeans affected by fungal infection. In addition to these core applications, ground-based hyperspectral imaging has proven valuable for phenotyping, fruit maturity assessment, chlorophyll and nitrogen content analysis [3], [4], and weed detection [5]. Despite these advantages, challenges remain, including the influence of variable illumination and shading [6], the logistical complexity of covering large areas, and a strong reliance on strict sensor calibration protocols.

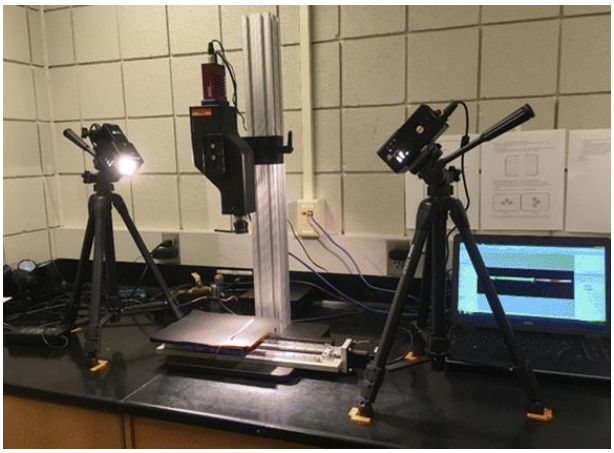

Figure 1. "GPhenoVision" Hyperspectral Data Collection System [7]

Figure 2. Lab-based hyperspectral Collection System [8]

Unmanned aerial vehicles (UAVs) have rapidly gained popularity for remote sensing in agriculture, largely due to their flexibility and cost-effectiveness. Advances in both UAV technologies and lightweight hyperspectral sensors have made this approach much more feasible. Examples of sensors that have been implemented into UAV-based systems are Nano-Hyperspec VNIR and Headwall Micro [9], [10].

Compared to manned aircraft and satellites, UAVs demonstrated the ability to operate at lower altitudes and slower speeds, enabling the collection of high-resolution hyperspectral imagery without the high costs associated with traditional airborne platforms with the addition of on-demand flight scheduling [11]. Critical growth stages could be surveyed more easily, and weather windows or disease outbreaks covered more frequently.

Studies by Zhu et al. and Honkavaara et al. demonstrated the use of UAV hyperspectral imaging for Leaf Area Index (LAI) and chlorophyll estimation, as well as biomass and nitrogen content in cereals. The capacity to detect subtle spectral differences aids in tracking plant health more precisely than typical multispectral systems [3], [11].

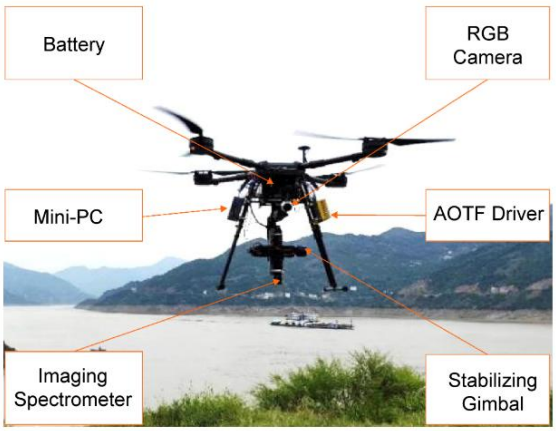

Figure 3. UAV-Based Hyperspectral Imaging System [13]

1.1.2. Satellite-Based Remote Sensing for Agriculture

Satellite-based remote sensing is characterized by extensive spatial coverage and regular revisit cycles, providing a regional to global perspective on crop health, soil properties, and environmental conditions. While ground-based and UAV systems capture high-resolution data at specific locations, satellites enable landscape-level assessments that are essential for monitoring large agricultural regions. Both multispectral (e.g., Sentinel-2, Landsat) and hyperspectral (e.g., EO-1 Hyperion) satellite data can detect differences in plant reflectance signatures associated with crop stress, disease incidence, or nutrient content [14].

From an operational standpoint, satellite-based imagery can be integrated into yield prediction models, crop type classification maps, and early warning systems for pest or disease outbreaks [15], [16]. Agronomists and farm managers can use this information to optimize irrigation, plan targeted fertilizer applications, or identify areas in need of pest control, increasing both productivity and resource efficiency. In addition, the broad spatial footprint of satellite data enables consistent assessments across multiple locations, facilitating comparative analysis across farms, regions, or entire countries. When combined with machine learning (ML) algorithms, these datasets enable scalable and automated processing pipelines that turn raw satellite imagery into useful predictions for stakeholders across the agricultural value chain.

The following key studies used satellite-based imagery (both hyperspectral and multispectral) in conjunction with ML approaches, showing how these techniques can produce accurate results in crop classification, soil property estimation, and disease detection - even in the face of challenges such as coarse spatial resolutions, long revisit intervals, and lower signal-to-noise ratios in certain spectral bands.

1.1.3. Sentinel-2 Mission and Multispectral Time-Series for Yield Estimation

The Sentinel-2 mission, operated by the European Space Agency (ESA) under the Copernicus Earth Observation Programme, consists of two identical satellites – Sentinel-2A (launched June 2015) and Sentinel-2B (launched March 2017). Together they achieve a five-day revisit time at the equator and a two-to-three day revisit at mid-latitudes, which makes the constellation practical for tracking crop growth at monthly resolution across entire growing seasons. The constellation acquires 13 spectral bands ranging from 443 nm (coastal aerosol) to 2190 nm (shortwave infrared), at ground sampling distances of 10, 20, and 60 meters depending on the band. The four 10 m bands (B2 blue, B3 green, B4 red, B8 NIR) are the most spatially precise; the three red-edge bands (B5 at 705 nm, B6 at 740 nm, B7 at 783 nm) and the broad-NIR band (B8A at 865 nm) are provided at 20 m. Two shortwave infrared bands (B11 and B12) at 20 m complete the set relevant to vegetation monitoring.

The three red-edge bands are the most distinctive feature of Sentinel-2 relative to earlier multispectral sensors such as Landsat-8. These bands sit in the spectral transition between the red chlorophyll-absorption region and the near-infrared plateau, where reflectance rises steeply and is sensitive to chlorophyll content, leaf area index, and crop stress before symptoms become visible in broadband data. Sentinel-2 therefore occupies a practical middle ground between the broad four- to seven-band sensors of the Landsat series and full hyperspectral imagers: it lacks contiguous spectral coverage but captures information that Landsat cannot. For the indices used in this thesis – NDRE and MCARI in particular – the red-edge bands are essential. NDRE (B8A−B5)/(B8A+B5) and MCARI both require a narrow-band red-edge measurement that is not available on Landsat-8 or MODIS.

County and field-level crop yield estimation using Sentinel-2 time-series has been an active research area since around 2018. Studies have applied the full growing-season trajectory of NDVI, EVI, and red-edge indices as input features to random forest and gradient boosting models, reporting R^2 values in the range of 0.5–0.8 for corn and wheat in data-rich regions [24], [25], [26]. A consistent finding across this literature is that model performance depends on the temporal density of cloud-free observations: years with persistent cloud cover during the peak vegetative period (July–August for US corn) tend to show higher residuals. Monthly composite approaches – which aggregate all cloud-free observations within a calendar month before computing vegetation indices – reduce this sensitivity while preserving enough temporal structure to distinguish growth stages. This approach is adopted in the present thesis: for each county and each month from April to October, a median composite of all cloud-free Sentinel-2 observations is computed, and six vegetation indices are calculated from the composite reflectances.

1.1.4. Crop Classification and Property Estimation - Hyperion vs. Landsat

Several authors have contrasted the performance of EO-1 Hyperion (hyperspectral) with Landsat (multispectral) data. For instance, Mariotto and Bostan applied machine learning classifiers, specifically, Support Vector Machines (SVM) and Random Forest to compare crop type classification accuracy [17], [18]. Both studies reported notable gains from hyperspectral data, attributing these gains to finer spectral resolution that captures subtle crop reflectance differences. The overall classification accuracy of the Hyperion image was found to be approximately 80%, whereas overall classification accuracy of Landsat image was found approximately 70%.

1.1.5. Vegetation Indices and Biophysical Traits

Chlorophyll and LAI Retrieval

Wu used Hyperion data alongside Partial Least Squares Regression to estimate chlorophyll and Leaf Area Index (LAI) in mixed cropping systems [16]. The authors identified key hyperspectral bands that are more responsive to chlorophyll content changes than conventional broad-band indices.

Soil Organic Carbon (SOC)

Gomez incorporated multivariate regression modelling to predict Soil Organic Carbon (SOC) from Hyperion reflectance. While satellite-based estimates were moderately accurate, lower signal-to-noise ratios in the shortwave infrared led to reduced performance compared to field spectroradiometer data [19].

Crop-Specific Anomalies

Satellite-based hyperspectral imagery has also been adopted to detect crop disease and stress factors. Weng et al. [20] used univariate regression on Hyperion data to discriminate diseased patches from healthy canopy regions, achieving a root mean square error (RMSE) of 0.986 and an R^2 of 0.873. While results depend on spatial resolution and SNR, these applications highlight how spectral granularity aids in identifying early stress symptoms invisible in broad-band data.

1.1.6. Time-Series Approaches to Yield Estimation

Treating the growing season as a sequence of spectral observations rather than a single snapshot is the central shift that distinguishes most recent yield estimation work from earlier studies. For corn in the US Corn Belt the growing season spans roughly six months from planting in late April to harvest in October, and the trajectory of vegetation index values over that period encodes information about growth rate, peak biomass, and senescence timing that no single-date image can capture. Greenness accumulates rapidly through June, peaks around July tasseling, and begins declining in August as the plant allocates resources to grain filling. Models that can resolve this trajectory outperform those trained only on end-of-season or peak-season composites.

Two broad strategies have been applied to encode this temporal information as model inputs. The first computes summary statistics per growing season – mean, maximum, and standard deviation of NDVI across all available observations – and then trains tabular models on these seasonal aggregates. This is simple and handles irregular observation timing, but it discards the within-season pattern. The second strategy preserves the monthly or weekly pattern by pivoting the time-series into wide-format features (one feature per index per time step), which allows gradient boosting models to learn which specific months and index combinations are most predictive. Several US corn yield studies have found that monthly-resolution features reduce RMSE by 10–20% relative to seasonal aggregates, because July NDVI – which corresponds to peak silking – is individually highly predictive and that

signal is diluted when averaged across the season [24], [26]. The present thesis applies the monthly pivoting strategy: after computing six vegetation indices for each county and month, the data are pivoted to a wide table of approximately 108 satellite-derived columns, which are then joined with analogous monthly meteorological features.

1.2. Machine Learning with Remote Sensing Data

The rich spectral information in HSI makes ML indispensable. Researchers have experimented and validated several ML algorithms - ranging from classical statistical models (e.g., SVMs Machines, Random Forest) to deep learning architectures (e.g., Convolutional Neural Networks, LSTMs, Generative Adversarial Networks) - to extract predictive information from these high-dimensional datasets.

This section reviews the role of ML in remote sensing data processing for agriculture. It covers common algorithms, feature engineering strategies, training challenges, and emerging trends, drawing on both hyperspectral and multispectral studies from the literature. The goal is to show how ML enables satellite sensors to detect plant stress, classify crops, and predict yields-context that directly informs the modeling choices made in this thesis.

1.2.1. Traditional Supervised Methods

Support Vector Machines (SVM)

SVMs are effective for remote sensing classification due to their robustness in high-dimensional feature spaces and ability to handle non-linear decision boundaries with kernel functions. Studies comparing SVM classifiers on EO-1 Hyperion (hyperspectral) against Landsat (multispectral) imagery report 10–15% accuracy gains from the richer spectral detail of hyperspectral data. SVMs have demonstrated effectiveness when labeled data is moderate in size, as they can still find good separating hyperplanes without requiring large training datasets [15], [17].

Random Forest (RF)

In studies run by Kurade et al., RF often provided robust classification results and can handle complex feature interactions. The authors generally report overall classification accuracies in the range of 70–85%, depending on the crop type, season, and classifier used. Classification here refers to crop-type mapping, which is a categorical task with discrete class labels, not continuous yield values. In this thesis, Random Forest is used as a regressor: the output is a single continuous number, corn yield in bushels per acre. RF's inherent mechanism of feature importance ranking further helps identify key spectral bands contributing most to classification or regression tasks. This interpretability is often

valuable for agronomists seeking to understand which wavelengths correspond to specific crop or soil characteristics [21].

Partial Least Squares Regression (PLSR)

PLSR is a dimensionality-reduction regression method widely used for quantitative trait estimation- for instance, chlorophyll or nitrogen content [22]. By projecting both predictor - spectral values - and response - target variable - onto latent variables, PLSR can effectively model relationships even when variables are highly collinear, as is often the case in hyperspectral bands.

1.2.2. Tree-Based Gradient Boosting Models - Extreme Gradient Boosting (XGBoost)

Recent studies by Huber et al. [23] highlight XGBoost as a powerful alternative to neural networks for yield prediction. XGBoost refines the gradient boosting approach by introducing regularization and sophisticated tree-growing techniques that improve model generalization. In yield estimation scenarios (soybeans, wheat, or maize), XGBoost often outperforms or matches deep learning approaches, particularly when feature engineering is carefully performed (e.g., quantile statistics, historical yields). Moreover, its faster training speeds and explainable outputs (via Shapley Value analysis) make it attractive for operational use by agronomists requiring interpretability.

1.2.3. Deep Learning Methods

Convolutional Neural Networks (CNNs)

CNNs are particularly adept at spectral–spatial feature extraction. In hyperspectral contexts, CNNs can be adapted to process the 3D spectral-spatial data cube, achieving classification accuracies of 98–99% on standard HSI benchmarks. Standard HSI benchmarks are publicly released hyperspectral datasets used to compare classification algorithms. The most common are Indian Pines (145×145 pixels, 200 spectral bands, 16 land-cover classes), Salinas Scene (512×217 pixels, 204 bands, 16 classes), and Pavia University (610×340 pixels, 103 bands, 9 urban classes). These datasets are entirely classification-oriented and hyperspectral. For multispectral time-series data (the approach used in this thesis), 1D-CNNs process the temporal sequence of band values rather than a spatial cube. Agricultural tasks benefiting from CNNs include weed detection, plant disease diagnosis, and fruit ripeness classification [4], [8].

Recurrent Neural Networks (RNNs) and LSTM

RNN-based models handle temporal sequences or dynamic changes in spectral profiles (e.g., multi-date satellite acquisitions, whether hyperspectral or multispectral). LSTMs can learn long-term

dependencies-useful in monitoring plant growth phases or stress progression over several weeks. Such architectures can be combined with CNNs to form CNN-LSTM hybrids, to integrate spatial–spectral features with time-series insights [22]. Although these models can be computationally heavy, they excel in capturing phenological patterns crucial for yield forecasting and drought or disease onset detection.

1.3. Challenges in Applying ML to Remote Sensing Data & Possible Solutions

Machine learning holds real promise for extracting useful information from satellite remote sensing data, but several practical challenges complicate its application in agriculture. The most basic is the scale of the input space. Hyperspectral sensors can produce hundreds of narrow-band measurements per pixel, while multispectral time-series approaches like the one used in this thesis generate many temporal features per index. Either way, training CNNs or LSTMs over large geographic areas is computationally intensive. Dimensionality reduction techniques such as PCA and Maximum Noise Fraction (MNF) address this for hyperspectral data [17], yet finding the right balance between complexity and performance remains a formidable challenge.

Another significant hurdle concerns ground-truth data collection, which necessitates the labeling of each pixel according to its crop type or disease level [1]. Despite the capability of modern platforms to capture vast quantities of hyperspectral imagery, the scarcity of accurately labeled training data often limits model accuracy. Consequently, researchers are increasingly exploring semi-supervised approaches and transfer learning to leverage unlabeled samples, as well as data augmentation strategies with generative adversarial networks (GANs).

A third challenge stems from the presence of noise and mixed pixels, particularly prevalent in satellite-based imagery. Coarser resolutions of around 10–30 m (typical of Sentinel-2 and Landsat) can cause each pixel to represent multiple field conditions-such as combinations of soil, crop canopies, and weeds-thus obscuring the true spectral signature [17], [18]. Additionally, certain spectral bands, particularly those in the shortwave infrared spectrum, may exhibit lower signal-to-noise ratios, potentially compromising the accuracy of reflectance measurements [19]. This challenge can be addressed through the implementation of crucial preprocessing steps, such as radiometric calibration, atmospheric correction, and unmixing [19]. However, the interpretability of models further complicates the implementation of remote sensing machine learning solutions in real-world agronomy. Deep neural networks, for instance, can act as "black boxes," making it difficult to understand which features inform a specific classification or yield forecast. However, interpretability is becoming increasingly central in agricultural settings, where practitioners need transparency regarding fertilizer recommendations or pest-management alerts [6]. To address this need, feature-

importance metrics in tree-based ensemble methods, Shapley Value analyses, and visualization tools such as gradient-weighted class activation mapping (Grad-CAM) in CNNs have been proposed as partial solutions. These methods highlight the spectral or spatial regions that drive a model's decisions, offering insights into the decision-making processes of deep neural networks.

Finally, the variability in the characteristics of sensors hinders the direct transfer of models across different data sources. Each remote sensing sensor-whether ground-based, UAV-mounted, or satellite-borne, and whether multispectral or hyperspectral-has distinct calibration procedures, spectral ranges, and signal-to-noise ratios [10]. Consequently, a model that has been trained on UAV data with fine spatial resolution may not be able to generalize to coarser satellite imagery without additional adaptations.

When working specifically with multispectral time-series from operational satellites, several additional challenges arise that are less prominent in controlled hyperspectral studies. Cloud coverage is the most immediate: Sentinel-2 provides no observations when cloud cover exceeds the Scene Classification Layer (SCL) threshold, and persistent cloud cover during key growth stages such as tasseling in July can leave entire months with no usable data for some counties or years. Monthly composite strategies reduce the problem but do not eliminate it; a month with only two cloud-free observations will produce a less stable median than one with ten. A second challenge is temporal alignment across years: the crop calendar shifts by several days each year depending on spring temperature and planting decisions, so a fixed calendar-month window does not correspond to a fixed agronomic stage across all seasons. Third, ground-truth availability limits the resolution of validation: USDA NASS provides county-level final yield estimates, but sub-county variation within a county – across soil types, management practices, and micro-climates – cannot be validated without field-level data. These challenges collectively mean that multispectral time-series models for yield estimation are bounded both by the quality of the satellite record and by the spatial granularity of the reference data.

2. Design of an Automated Multispectral Yield Estimation System

2.1. Requirements, Analysis and Specification

This upcoming section defines the functional, non-functional, and technical requirements that guided the system design and implementation phases.

Category	Req. ID	Requirement Description
Data Acquisition & Processing	FR1	The system must acquire Sentinel-2 L2A multispectral satellite imagery for user-defined areas of interest (AOI)
	FR2	The system must calculate multiple vegetation indices (NDVI, EVI, NDRE, MCARI, GNDVI, SAVI) from satellite reflectance data to capture diverse aspects of plant physiology
	FR3	The system must generate temporal vegetation index time series with monthly resolution during growing seasons
	FR4	The system must create agricultural area masks to distinguish crop fields from other land cover types
	FR5	The system must integrate with external ground truth datasets for validation purposes
Machine Learning & Prediction	FR6	The system must implement feature engineering to transform raw vegetation indices into predictive features for crop yield estimation
	FR7	The system must provide yield predictions with quantifiable accuracy metrics (R^2, RMSE, MAE) for performance assessment
	FR8	The system must perform automated hyperparameter optimization based on dataset characteristics to prevent overfitting in small sample scenarios
Data Management & Visualization	FR9	The system must provide automated data quality assessment including completeness analysis and temporal coverage evaluation
	FR10	The system must generate comprehensive visualization suites including time series plots, spatial analysis maps, and validation scatter plots
	FR11	The system must export analysis-ready datasets in standardized formats (CSV, JSON) compatible with common data science workflows

Usability & Accessibility	NFR1	The system must accept simple coordinate-based area specification to minimize GIS expertise requirements
	NFR2	The system must provide automated configuration management with secure credential storage
	NFR3	The system must generate interpretable feature importance analysis for agricultural domain experts
Integration & Compatibility	TR1	The system must integrate with the Copernicus Data Space Ecosystem via Sentinel Hub APIs for real-time data access
Data Quality & Validation	TR2	The system must implement automated atmospheric correction validation to ensure reflectance value consistency across temporal acquisitions
	TR3	The system must provide statistical validation against authoritative ground truth sources
	TR4	The system must implement time series cross-validation to respect temporal dependencies in agricultural data

Table 1. Project Technical Requirements

2.1.1. Data Acquisition and Processing Requirements

Multi-spectral Satellite Data Acquisition (FR1)

The system downloads Sentinel-2 Level-2A (surface reflectance) imagery rather than raw top-of-atmosphere data. Level-2A products have already been through ESA's Sen2Cor atmospheric correction, which removes the scene-to-scene variation in apparent reflectance caused by aerosols and water vapor [19]. Without this correction, comparing NDVI values across different dates or counties would not be meaningful. Data is accessed via the Sentinel Hub API, which provides programmatic, batch-ready access to the Copernicus archive.

Multi-Index Vegetation Analysis (FR2)

Rather than relying on a single vegetation index, the system calculates six different indices (NDVI, EVI, NDRE, MCARI, GNDVI, and SAVI). Research by Wu et al. [16] and Mariotto et al. [18] shows that different indices respond uniquely to chlorophyll content, biomass, and soil effects. This gives a more complete picture of crop health than any single index could provide. The system has

been designed to calculate all these indices simultaneously, ensuring they're mathematically consistent across time.

Temporal Time Series Generation (FR3)

Individual Sentinel-2 scenes are aggregated into monthly composite values (median of cloud-free pixels per county per month) for the growing season months April through October. Monthly resolution captures the seasonal arc of crop development without being overwhelmed by noise from day-to-day weather variation. Shao et al. [3] and Zhong et al. [12] both find that monthly time series provide enough temporal detail to track phenology for annual crops while remaining tractable for multi-year, multi-county analysis.

Agricultural Area Identification (FR4)

To ensure the system analyzes actual cropland - not forests, water, or urban areas - the system creates crop masks that identify agricultural fields. This filtering step is crucial for accurate yield predictions since it prevents non-crop vegetation from skewing results. The masking combines vegetation thresholds with water indices to identify likely crop pixels. The current implementation is limited to corn, which is the dominant crop in the study region and has the most consistent NASS county-level coverage. The pipeline is not corn-specific by design, however. Adapting it to soybeans, winter wheat, or sorghum would require changing the NASS API commodity filter and shifting the phenological calendar to match each crop's active growing months.

Ground Truth Integration (FR5)

Yield ground truth comes from the USDA National Agricultural Statistics Service (NASS) Quick Stats API, which provides annual county-level corn yield in bushels per acre. The NASS data is matched to satellite observations by county FIPS code and crop year. Using a government survey rather than field measurements means the ground truth is independent of the satellite pipeline and covers the same spatial unit (county) as the satellite aggregates.

2.1.2. Machine Learning and Prediction Requirements

Automated Feature Engineering (FR6)

Transforming raw vegetation indices into meaningful predictors remains an underexplored area in agricultural applications [21]. The system creates features that give insight into agricultural patterns such as seasonal patterns, growth stage indicators, and relationships between different spectral bands. This ensures the incorporation of agricultural knowledge into the analysis.

Performance-Validated Yield Prediction (FR7)

Model accuracy is evaluated using R^2, RMSE, and MAE against USDA NASS county yields. These three metrics together give a full picture: R^2 shows how much yield variance the model explains, RMSE gives the average error in absolute units (bushels per acre), and MAE gives the mean absolute deviation. Results are reported separately for each model variant (tree-based baseline, tree-based with weather, LSTM, CNN) to allow direct comparison of approaches.

Adaptive Model Optimization (FR8)

With less than 200 training samples, overfitting is a real concern. XGBoost hyperparameters (max tree depth, learning rate, subsample fraction, L1/L2 regularization) are tuned using a grid search over a time-aware cross-validation split. The best parameters are those that minimize RMSE on the held-out validation folds, not the training set. Early stopping is also used to prevent the boosting procedure from continuing past the point where validation performance plateaus.

2.1.3. Data Management and Visualization Requirements

Automated Quality Assessment (FR9)

After each data retrieval run, the pipeline outputs a quality report that shows what percentage of months have at least one valid scene for each county, how many scenes were discarded due to cloud cover, and the distribution of cloud-free pixel counts per county per month. This enables identification of counties or years with poor data coverage before training, rather than discovering the problem when the model performs badly.

Comprehensive Visualization Suite (FR10)

The pipeline outputs a standard set of diagnostic plots at each stage: seasonal time series of each vegetation index per county, scatter plots of predicted vs. actual yield with R^2 and RMSE annotations, feature importance bar charts, and correlation matrices between feature groups. These plots are saved automatically and are the primary tool for checking whether the model is learning the right signals.

Standardized Data Export (FR11)

Export capabilities in common data science formats (CSV, JSON) ensure interoperability with external analysis workflows and enable integration with other research tools.

2.1.4. Technical and Integration Requirements

Real-time Data Access (TR1)

Integration with Copernicus Data Space Ecosystem through Sentinel Hub APIs ensures access to current satellite observations essential for operational applications and up-to-date research analyses.

Atmospheric Correction Validation (TR2)

Gomez et al. [19] highlighted how atmospheric variability can affect vegetation measurements. The system runs checks to ensure consistent reflectance values across time, identifying potential atmospheric correction issues that might need attention.

Authoritative Ground Truth Validation (TR3)

Validating predictions against authoritative data sources is essential for objective accuracy assessment. The system accepts multiple ground truth sources, including USDA county statistics, field measurements, and other agricultural databases.

Temporal Cross-Validation Framework (TR4)

Time series cross-validation procedures respect temporal dependencies inherent in agricultural data, giving more realistic performance estimates than random splits would. The system implements leave-one-year-out validation and temporal splitting procedures that prevent data leakage while maintaining statistical validity of performance estimates. This ensures that validation results reflect genuine predictive capability rather than overfitting to temporal patterns.

2.1.5. Usability and Accessibility Requirements

Simplified Geographic Specification (NFR1)

Coordinate-based area specification removes the need for GIS expertise, so researchers without a geospatial background can use the system directly. The system accepts simple bounding box coordinates and automatically handles coordinate system transformations, spatial registration, and area validation procedures that would otherwise require specialized GIS knowledge.

Automated Configuration Management (NFR2)

API credentials (Sentinel Hub client ID/secret, USDA NASS API key) are stored in a JSON config file outside the codebase rather than hardcoded in scripts. On first run, the system generates a default config template with placeholder values and prints instructions for filling them in. This makes the pipeline reproducible on a new machine without code changes.

Interpretable Analysis Results (NFR3)

XGBoost's built-in feature importance scores show which vegetation index and which month contributed most to yield predictions. In this study, mid-summer NDVI and NDRE values (July–August) consistently ranked highest, which aligns with the agronomic understanding that canopy conditions during pollination and early grain fill are the strongest determinants of final corn yield.

2.2. System Architecture and Design Decisions

This section describes the key design decisions behind the yield estimation system and explains why each choice was made. The design covers three main areas: satellite data processing, vegetation index selection, and the machine learning framework.

2.2.1. Satellite Data Processing Framework

Choosing the right satellite data source involved balancing data quality, temporal coverage, and spectral characteristics relevant to crop monitoring.

Sentinel-2 Platform Selection

Sentinel-2 was selected as the primary data source because it offers a practical combination of spatial resolution (10–20 m), revisit frequency (5 days), and spectral bands well-suited to vegetation monitoring [14]. Compared to Landsat (30 m, 16-day revisit) or commercial very-high-resolution sensors, Sentinel-2 hits the right point for county-scale crop analysis: frequent enough to track phenological changes through the growing season, and fine-grained enough to resolve within-field variation. Thenkabail [14] specifically identifies the 5-day revisit as important for catching rapid canopy changes during grain fill. The 10–20 m resolution is a reasonable trade-off between spatial detail and the computational cost of processing data across 26 counties and multiple growing seasons, consistent with the recommendations of Gonzalez-Dugo et al. [9].

Atmospheric Correction Methodology

All imagery used in this project is Sentinel-2 Level-2A, meaning atmospheric correction has already been applied by ESA's Sen2Cor processor before downloading. This matters because raw top-of-atmosphere reflectance varies with atmospheric conditions in ways that can dwarf the actual vegetation signal [19]. By starting from surface reflectance values, spectral measurements are comparable across different acquisition dates and counties, which is essential when building a model that spans 26 locations and up to 8 growing seasons. Wu et al. [16] show that atmospherically corrected products give measurably more consistent vegetation index values than TOA reflectance when comparing multi-temporal data.

Using a single, standardized correction method (Sen2Cor Level-2A) across all scenes removes one source of inter-temporal variability that would otherwise introduce noise into the yield prediction model.

Cloud Masking and Quality Control Strategy

Cloud contamination is a well-known source of error in satellite vegetation monitoring [16]. Each Sentinel-2 scene is filtered using cloud probability thresholds from the Level-2A cloud mask, excluding scenes where cloud or shadow coverage exceeds an acceptable limit for the county area of interest. Scenes that pass the cloud filter are then checked for temporal consistency: a value that deviates sharply from the surrounding monthly observations is flagged as a likely cloud or shadow artifact and excluded from the monthly aggregation.

Because cloud cover means some months have fewer valid observations than others, the monthly aggregation uses median values rather than means. Medians are less sensitive to remaining outliers and give a consistent monthly signal even when only a handful of cloud-free scenes are available. This approach follows recommendations in Shao et al. [3] and Zhong et al. [12] for maintaining temporal consistency in satellite time-series used for crop monitoring.

2.2.2. Vegetation Index Selection and Implementation

Choosing which vegetation indices to compute is one of the most consequential decisions in the design, since it determines what physiological signals the model has access to. The choice to use six indices rather than NDVI alone is based on the different sensitivities each index has to distinct aspects of crop condition.

Multi-Index Approach Rationale

The choice to implement six distinct vegetation indices (NDVI, EVI, NDRE, MCARI, GNDVI, SAVI) as opposed to conventional NDVI-only methodologies addresses fundamental limitations that have been identified in comparative vegetation index studies. Research by Mariotto et al. [18] and Bostan et al. [17] demonstrates that different vegetation indices exhibit varying sensitivities to specific aspects of vegetation condition, including chlorophyll content, biomass accumulation, and canopy structure.

NDVI is the most common index, but it saturates at high biomass levels typical of mature corn canopies [18]. EVI avoids this by including the blue band, which partially corrects for atmospheric and soil background effects, maintaining sensitivity in dense vegetation. NDRE uses the red-edge band (705–740 nm) rather than the red band, making it more sensitive to chlorophyll concentration changes. Several studies find NDRE outperforms NDVI for detecting early crop stress before visible symptoms appear [1, 2].

Sentinel-2 Band Assignments per Vegetation Index

MCARI uses red, red-edge, and green bands together, which gives chlorophyll-sensitive measurements while reducing soil background interference - an important property during the early

growing season when bare soil is still visible between crop rows [3]. GNDVI replaces the red band with green, giving a different angle on canopy greenness that is not correlated with NDVI and adds distinct predictive information.

SAVI incorporates soil brightness correction factors, which are of particular importance for agricultural applications, given the significant variation in bare soil visibility throughout the growing season [5]. This soil adjustment capability addresses the limitations of traditional vegetation indices in agricultural environments, where mixed pixel effects represent ongoing challenges.

Temporal Spectral Analysis Integration

A single satellite observation tells you what a field looks like on one day. The temporal trajectory - how each vegetation index rises and falls through the season - carries much more information about whether the crop is developing on schedule, experiencing stress, or benefiting from favorable conditions. Zhu et al. [3] show that time-series features outperform single-date measurements for crop yield prediction, which is why this system aggregates vegetation indices monthly and feeds the full seasonal trajectory to the model.

2.2.3. Machine Learning Model Selection

Three model families were evaluated: tree-based ensembles (XGBoost, Random Forest, Gradient Boosting), deep learning sequence models (LSTM, CNN), and linear regression as a baseline. The final design favors tree-based models based on their results with the available data.

XGBoost Algorithm Selection

The choice of XGBoost (Extreme Gradient Boosting) as the primary machine learning algorithm was based on comparative performance studies demonstrating superior results for agricultural prediction tasks involving complex feature interactions and high-dimensional datasets. Research comparing ensemble methods with traditional regression approaches shows consistent advantages for gradient boosting algorithms in agricultural applications [21].

XGBoost addresses several challenges specific to agricultural yield prediction applications. The algorithm's inherent feature selection capabilities prove particularly valuable when dealing with high-dimensional vegetation index datasets where many features may exhibit multicollinearity or limited predictive value. The built-in regularization mechanisms help prevent overfitting, a common concern when working with limited training data typical of county-level agricultural statistics.

Feature Engineering Integration

Raw vegetation index values at a single point in time do not carry enough information for reliable yield prediction. The feature engineering step converts the monthly time series into a wide table where each growing season is a single row, with each month's index values as separate columns. This gives the model access to the full seasonal trajectory as a flat feature vector, which tree-based models can handle directly without requiring sequence-aware architectures.

Beyond the monthly pivot, statistical summaries are computed for agronomically defined sub-periods: early season (April–May), reproductive (June–July), and grain fill (August–September). These period-level aggregates capture the fact that stresses during pollination and grain fill have a disproportionately large impact on final corn yield compared to early-season variability.

Model Validation and Performance Assessment

Validation uses a time-aware split rather than random cross-validation. With random splits, a model trained on data from 2018 and 2020 could be tested on 2019, which would inflate performance metrics by letting the model learn from data that came after the test year. The time-series split always trains on earlier years and tests on later ones, giving a more honest estimate of how the model would perform on a genuinely unseen future season.

Together, these design choices - Sentinel-2 Level-2A data, six complementary vegetation indices, monthly temporal aggregation, and time-aware validation - form a pipeline that is both reproducible and grounded in the published literature on satellite-based crop monitoring.

2.3. Data Pipeline Implementation and Processing Workflow

This section describes how the pipeline moves from raw Sentinel-2 scenes to a feature table ready for machine learning. The steps are: scene acquisition and cloud filtering, vegetation index computation, temporal aggregation to monthly values, weather data merging, and final feature pivoting.

The overall workflow is illustrated in Figure 4:

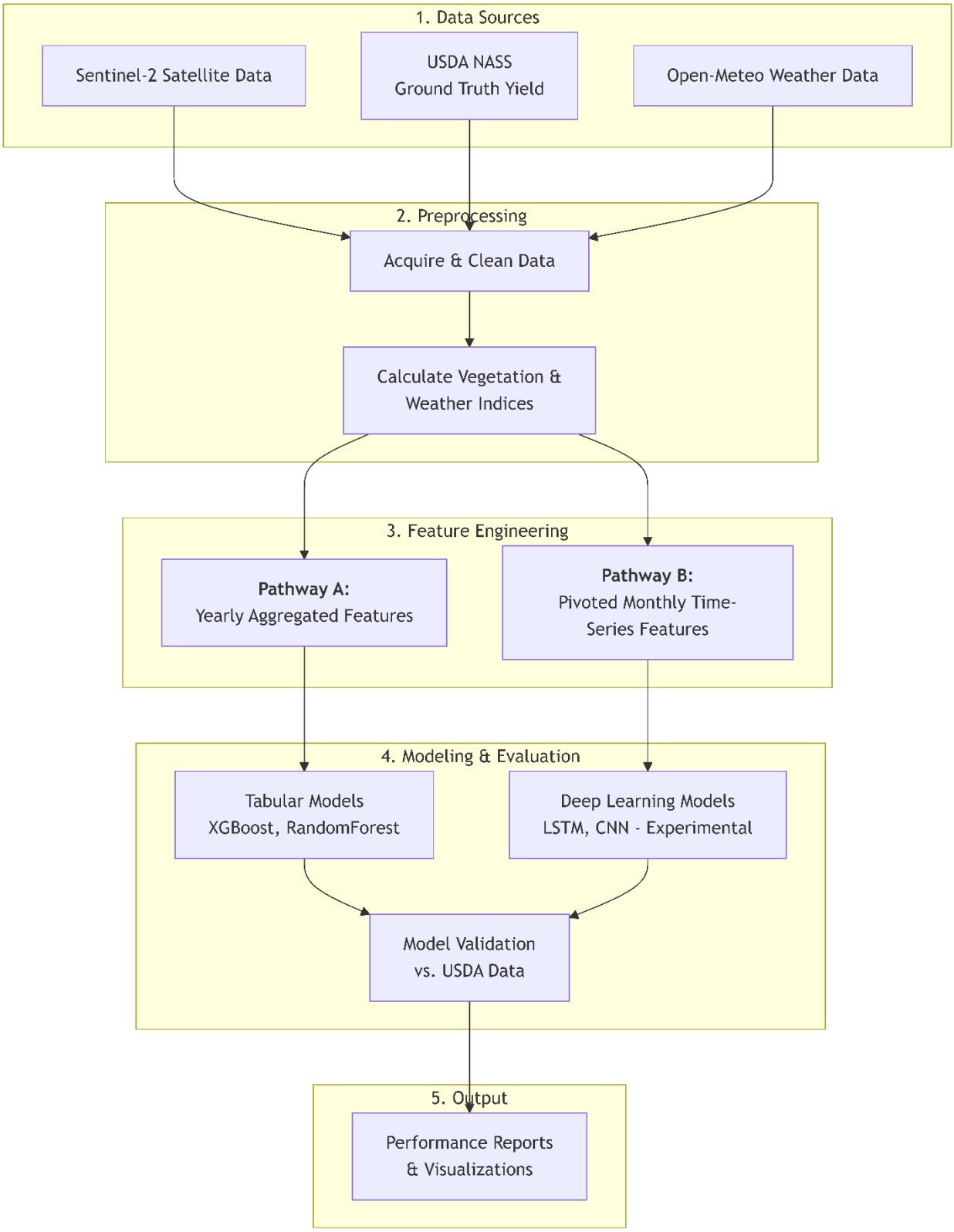

Figure 4. High-level data flow of the end-to-end yield estimation pipeline

2.3.1. Data Acquisition and Ingestion Layer

Satellite data is retrieved from the Sentinel Hub API, which provides programmatic access to Sentinel-2 Level-2A products without downloading full scene files. Requests are defined by a bounding box (the county boundary buffered by a small margin), a date range covering the growing season (April through October), and a cloud coverage filter applied at the scene level.

Satellite Data Retrieval Protocol

Sentinel-2 imagery is queried via the Sentinel Hub API using county FIPS boundaries as the spatial filter. Each request specifies a county bounding box, a growing season date range, and a maximum cloud percentage for the scene. The API returns Level-2A surface reflectance data directly, so no additional atmospheric correction is applied in the pipeline. Using county-level bounding boxes means the same code runs for any county without manual reconfiguration.

Data collection runs annually from April through October, covering the main US corn growing season. Monthly aggregation is used rather than per-scene analysis. Monthly composites reduce sensitivity to individual cloud-contaminated scenes while preserving sufficient temporal structure to distinguish growth stages [24].

Data Format Standardization

Each retrieved scene is resampled to a common 10 m resolution and aligned to the county boundary using a consistent coordinate reference system (WGS84 / UTM). This step is necessary because Sentinel-2 tiles from different dates can have small misalignments, which would distort spatial averages if left uncorrected [6]. All outputs are stored in CSV format indexed by county FIPS code, acquisition date, and year.

Spatial averaging is computed over all cloud-free pixels within the county boundary for each scene, yielding a single representative value per vegetation index per date. This county-level average is the unit of analysis throughout the pipeline, reflecting the resolution of the ground-truth yield data from USDA NASS, which is also reported at the county level.

Quality Assessment and Filtering

Each scene is assessed for cloud contamination using the SCL (Scene Classification Layer) provided in the Sentinel-2 Level-2A product. Pixels classified as cloud, cloud shadow, or saturated are excluded from the spatial average for that date. At the scene level, if more than a set threshold of the county area is covered by clouds, the scene is discarded entirely rather than contributing a low-quality average.

In months where no scene passes the cloud threshold, the monthly value is left as missing rather than filled with a low-quality estimate. These gaps are handled during feature engineering: if fewer than a minimum number of valid monthly observations exist for a county-year, that county-year is excluded from the training set. This trades off sample size for data quality and is consistent with how Wu et al. [16] and Atzberger [19] handle cloud-contaminated time series in agricultural applications.

2.3.2. Data Structure and Transformation Framework

After quality filtering, individual scene observations are grouped by county and month, and vegetation index values within each group are summarized with the median. The result is a long-format time series with one row per county per date, which is then pivoted into a wide format for model training.

Multispectral to Vegetation Index Transformation

The six vegetation indices are computed from Sentinel-2 surface reflectance bands using standard band ratio formulas. Each index targets a specific aspect of vegetation condition and is sensitive to different confounding factors (soil background, atmospheric residuals, canopy density), which is why using all six gives the model more discriminative power than any single index alone.

The transformation framework processes six distinct vegetation indices simultaneously: NDVI emphasizes general vegetation vigor through red and near-infrared band contrasts; EVI incorporates blue band corrections to improve sensitivity in high-biomass conditions; NDRE uses red-edge spectral information for enhanced chlorophyll detection; MCARI combines multiple spectral bands for chlorophyll assessment while minimizing soil effects; GNDVI emphasizes green vegetation components; and SAVI implements soil brightness corrections particularly important during early-season monitoring periods D

Temporal Data Structure Organization

Each computed vegetation index observation is stored alongside its acquisition date and the county's FIPS code. These records form a time series in a long format that tracks how each index changes throughout the growing season in each county and year. This structure supports both absolute analysis (e.g. the analysis of the NDVI value at a predefined time) and relative analysis (e.g. peak to harvest).

Spatial Aggregation and Representative Sampling

Pixel-level index values within each county boundary are aggregated to a single representative value per index per scene. The primary statistic used is the median (less sensitive to cloud residuals than the mean). Standard deviation and percentiles are also computed to capture spatial variability within

the county, since heterogeneous fields may show meaningful spread even when the county-average looks typical.

Pixels flagged as cloud, shadow, or saturated by the SCL mask are excluded from the spatial aggregation entirely. Only unmasked pixels contribute to the county statistics. This means the aggregated values reflect actual surface conditions on that date, not a blend of good and contaminated pixels.

2.3.3. Feature Engineering and Temporal Analysis Pipeline

Seasonal Aggregation and Growth Stage Analysis: Monthly values are grouped into three agronomic windows that correspond to growth stages relevant to corn yield in the US Corn Belt. April–June covers establishment and early vegetative growth; June–August covers pollination and peak biomass, which are the most yield-determining stages; August–October covers grain fill and senescence. Summary statistics for each window are computed separately and included as distinct feature groups [3, 21].

For each growth window, the statistics computed are: mean, median, standard deviation, and inter-quartile range. The mean and median describe average conditions; the standard deviation and IQR capture how stable or variable the season was within the county, which can signal stress events. For example, a high standard deviation in July NDVI might indicate patchy drought stress across the county rather than uniform canopy development.

Derived Relationship Features:

Cross-index ratios (for example, EVI/NDVI) are computed as derived features. These ratios normalize for scene-wide differences in illumination or atmospheric haze, leaving a signal that reflects the relative contribution of different physiological factors. A high EVI/NDVI ratio, for instance, suggests dense canopy without the saturation artifact that affects NDVI alone.

2.3.4. Model Training and Prediction Workflow Integration

Feature Selection and Dimensionality Management

Feature selection uses XGBoost's built-in importance scores to identify which vegetation index / month combinations carry the most predictive weight. Features with near-zero importance across repeated cross-validation folds are dropped before final model training. This keeps the feature set to a manageable size given the limited number of training samples (~146 county-years).

Many features in the dataset are correlated - NDVI and EVI for the same month, for example, track similar signals. XGBoost's regularization naturally handles correlated features without explicit

dimensionality reduction, and the importance-based pruning described above further reduces redundancy. The remaining features are interpretable: they correspond to specific indices at specific times of year, which agronomists can relate to known growth-stage dynamics.

2.3.5. Data Preparation for Machine Learning Model Training

Before training, the satellite time series and weather data are merged into a single feature table and cleaned to remove incomplete county-years and obvious data errors.

Data Quality Assessment and Cleaning

County-years with more than 30% missing monthly observations are excluded from the dataset. Short gaps (one missing month) are filled using linear interpolation between adjacent months, which is reasonable given the smooth seasonal trajectory of most vegetation indices. Extreme outlier values that fall outside the physically plausible range for each index (e.g., NDVI below –0.1 or above 0.95) are treated as sensor or processing artifacts and removed before interpolation [19].

Quality metrics assess temporal data completeness across growing seasons, with minimum thresholds of 70% coverage required for reliable feature generation. Observations failing quality criteria undergo either interpolation procedures for short gaps (1-2 missing values) or exclusion for extensive missing periods that cannot be reliably reconstructed.

Feature Engineering and Transformation

The feature engineering process converts temporal vegetation index measurements into machine learning features through systematic aggregation and derived calculations. Raw vegetation indices undergo seasonal aggregation (early, mid, late season) generating statistical measures including means, maxima, standard deviations, and percentiles for each growth period. Temporal trend analysis produces slope coefficients and correlation measures characterizing vegetation development patterns throughout growing seasons.

Derived features include vegetation index ratios (EVI/NDVI, NDRE/NDVI) that emphasize relative vegetation characteristics while minimizing absolute measurement variations. Phenological features capture timing characteristics including peak vegetation periods and growing season length estimates that correlate with crop development patterns [3, 21].

Dataset Standardization and Scaling

Machine learning compatibility requires systematic feature scaling to address magnitude differences between vegetation indices and derived features. Standardization procedures center features around zero mean with unit variance, while robust scaling methods utilize median and interquartile ranges to minimize outlier sensitivity. Feature correlation analysis identifies multicollinear variables exceeding r=0.95 thresholds, with redundant features removed to prevent model overfitting in high-dimensional datasets.

Temporal Cross-Validation Preparation

Agricultural data requires time-series aware validation procedures that respect temporal ordering inherent in yearly observations. Data partitioning implements leave-one-year-out validation for small datasets or temporal splits that ensure training data precedes validation periods chronologically. This approach prevents data leakage while providing realistic performance estimates for operational deployment scenarios [17, 18].

The final prepared dataset typically contains 85-90 engineered features derived from 6 vegetation indices across 4-6 yearly observations per geographic unit. Feature categories include seasonal aggregations (35%), statistical measures (25%), temporal trends (15%), derived relationships (10%), phenological indicators (8%), and stress detection features (7%). All features undergo standardization with documented preprocessing parameters to ensure reproducibility and enable systematic evaluation of alternative preparation approaches.

3. System Implementation and Experimental Evaluation

This section details the technical implementation of the crop yield estimation system, outlining the complete workflow from data acquisition and preprocessing to feature engineering, modeling, and validation. The architecture is designed as a modular pipeline, enabling systematic experimentation with different data sources and machine learning algorithms. The codebase, primarily developed in Python using Jupyter Notebooks, is organized into distinct scripts, each handling a specific stage of the analysis.

3.1. System Architecture and Data Flow

The system is structured as a sequential pipeline that transforms raw geospatial and temporal data into actionable crop yield predictions. The data flow begins with the acquisition of satellite, agricultural, and meteorological data, which is then processed into CSV format. A key design decision was the creation of two parallel feature engineering pathways to accommodate different classes of machine learning models: one creating yearly aggregated tabular data for tree-based models, and another preserving monthly sequences for deep learning experiments.

The workflow can be summarized as follows:

1. Data Acquisition: Raw satellite imagery (Sentinel-2), county-level yield statistics (USDA NASS), and historical weather data (Open-Meteo) are programmatically fetched.
2. Preprocessing: Satellite data is processed to calculate various vegetation indices (VIs). All data sources are cleaned, standardized, and aligned by county and date.
3. Feature Engineering: The preprocessed data is transformed into two distinct formats:
    - Yearly Aggregated Features: Monthly data is summarized into a single feature vector for each county-year, suitable for tabular models like XGBoost.
    - Sequential Monthly Features: The month-by-month time-series data is preserved to train sequential models like LSTMs.
4. Modeling: Both tree-based and deep learning models are trained on the respective feature sets. The models learn the relationship between the engineered features and historical yield data.
5. Validation and Evaluation: Model predictions are validated against held-out ground truth data from USDA NASS. Performance is measured using standard regression metrics such as R^2 and Root Mean Squared Error (RMSE).

This modular architecture made it straightforward to compare different modeling approaches, as each approach plugs into the same feature input and evaluation framework, as detailed in the subsequent sections.

3.2. Data Sources and Preprocessing

The system integrates three primary types of external data, each requiring its own acquisition and preprocessing methodology.

3.2.1. Satellite Data Acquisition (Sentinel-2)

The core remote sensing data is sourced from the Sentinel-2 satellite mission, accessed via the Copernicus Data Space Ecosystem. The sentinelhub Python library provides the interface for this access. A dedicated SatelliteDataProcessor class automates the process of defining an Area of Interest (AOI), specifying a date range, and fetching the corresponding Level-2A imagery, which is already atmospherically corrected. From this raw spectral data, a suite of six key vegetation indices is calculated:

- NDVI (Normalized Difference Vegetation Index): A general measure of vegetation greenness and health [18].
- EVI (Enhanced Vegetation Index): An optimized index that performs better in areas with high biomass by correcting for atmospheric and soil background effects [18].
- NDRE (Normalized Difference Red Edge Index): Sensitive to chlorophyll content, particularly in the later stages of crop growth [16].
- MCARI (Modified Chlorophyll Absorption Ratio Index): Measures the depth of chlorophyll absorption [16].
- GNDVI (Green Normalized Difference Vegetation Index): Uses the green band, making it sensitive to chlorophyll concentration [16].
- SAVI (Soil-Adjusted Vegetation Index): A modified NDVI that minimizes the influence of soil brightness, useful in early growth stages [19].

These indices were selected based on their documented performance in crop monitoring literature; the full rationale is provided in Section 2.3.2.

The final output of this stage is a monthly time-series of these vegetation indices for each county, which forms the basis for all subsequent feature engineering.

3.2.2. Agricultural Ground Truth Data (USDA NASS)

The ground truth data for model training and validation is sourced from the USDA National Agricultural Statistics Service (NASS) Quick Stats API. A custom USDADataRetriever class handles the programmatic querying of this API. It fetches historical county-level corn yield data, measured in bushels per acre, for specified years. This dataset serves as the target variable - the value the model aims to predict - and the benchmark against which model performance is evaluated.

3.2.3. Meteorological Data Acquisition (Open-Meteo)

To incorporate the impact of weather on crop growth, historical meteorological data is fetched from the Open-Meteo API. A WeatherDataCollector class was developed to manage this process. It takes a list of county FIPS codes and years, determines the centroid coordinates for each county, and downloads daily time-series data for key weather variables, including maximum/minimum temperature and precipitation. This daily data is then aggregated into the yearly or monthly formats required by the feature engineering pipelines.

3.2.4. Geospatial Data for County Mapping

To link the coordinate-based satellite data to the county-based USDA data, the system requires geospatial boundary files for US counties. A CoordinateToCountyMapper class automates this by downloading official shapefiles from the US Census Bureau. This component allows the system to identify the correct county FIPS code for any given latitude and longitude, ensuring that satellite observations are correctly matched with their corresponding ground truth yield data.

3.3. Feature Engineering Methodologies

3.3.1. Yearly Aggregated Features

The baseline feature set aggregates each growing season into a single row per county-year. For each of the six vegetation indices (NDVI, EVI, NDRE, MCARI, GNDVI, SAVI), four seasonal statistics are computed from all cloud-free April–October observations: mean, maximum, minimum, and standard deviation. Analogous statistics are computed for the meteorological variables (monthly maximum temperature, minimum temperature, precipitation, and growing degree days). This produces approximately 40–50 features per county-year. Because all temporal structure has been collapsed into season-level summaries, this format is compatible with any standard tabular model and requires no imputation for months with missing observations. Scripts 05 and 07 use this format. The yearly model trained on script 07 (satellite + weather yearly) achieved XGBoost R^2=0.563 and RMSE=21.87 bu/acre, serving as the primary comparison baseline.

3.3.2. Sequential Monthly Features

This pipeline preserves the temporal sequence of the data, which is a requirement for models like LSTMs and CNNs.

1. Raw Monthly Satellite Features: The multi-county pipeline was extended to save the raw, unprocessed monthly vegetation index values for every county.
2. Monthly Weather Features: A dedicated script aggregates daily weather data into monthly summaries, such as mean maximum temperature (temp_max_mean) and total precipitation (precip_sum) for each month.
3. Final Sequential Dataset: These two monthly datasets are merged. The result is a sequence for each county-year where each time step contains both satellite and weather features. This "wide" format, where each monthly measurement becomes a unique column, allows tabular models to use the full monthly granularity as individual features.

The structure of the final monthly dataset in Figure 5.

final_monthly_dataset		
int	county_fips	County FIPS code (identifier)
int	year	Year of observation
float	EVI_1_to_11	Monthly EVI values (Jan-Nov)
float	GNDVI_1_to_11	Monthly GNDVI values (Jan-Nov)
float	MCARI_1_to_11	Monthly MCARI values (Jan-Nov)
float	NDRE_1_to_11	Monthly NDRE values (Jan-Nov)
float	NDVI_1_to_11	Monthly NDVI values (Jan-Nov)
float	SAVI_1_to_11	Monthly SAVI values (Jan-Nov)
float	precip_sum_1_to_11	Monthly total precipitation
float	temp_max_mean_1_to_11	Monthly mean max temperature
float	temp_min_mean_1_to_11	Monthly mean min temperature
string	state_fips	State FIPS code
string	commodity_desc	Crop type (e.g., CORN)
float	yield	Target variable: Yield (bu/acre)
float	production_bu	Total production (bushels)
float	planted_acres	Total planted acres

Figure 5. Diagram of the pivoted monthly dataset structure

3.3.3. Monthly Pivoted Features

The monthly pivoted approach, implemented in script 12, is the main methodological contribution of this thesis and produced the lowest RMSE of any configuration tested. Starting from the long-format monthly dataset (one row per county-month-year), satellite and weather features are grouped by county and year, then pivoted to a wide format where each unique month–variable combination becomes its own column. For the six satellite vegetation indices over seven months (April–October) this produces 42 satellite feature columns; adding analogous monthly weather variables (maximum temperature, minimum temperature, precipitation, growing degree days) over the same months produces approximately 70 additional columns, for a combined total of around 108 features per county-year row.

This format has two key advantages over the yearly aggregated set. First, it preserves the month-by-month trajectory of vegetation health across the growing season, allowing the model to learn that July NDVI is more predictive than April NDVI and that a sharp NDRE drop in August signals stress. Second, it keeps month-specific features orthogonal rather than collapsing them into a single seasonal mean, which avoids the information loss that seasonal aggregation produces. The approach is compatible with XGBoost and Random Forest without any preprocessing changes because these models treat each column as an independent feature regardless of the temporal relationship between columns. The resulting XGBoost model achieved R^2=0.527 and RMSE=13.03 bu/acre – the best RMSE of all configurations tested, and approximately 44% lower absolute error than the yearly baseline.

3.4. Modeling and Estimation Frameworks

The system was designed to test and compare different modeling architectures. The implementation of these models is found across several notebooks, each building on the previous one.

3.4.1. Validation Methodology

A robust validation framework is critical for assessing the real-world performance of the models. The core of this framework is the comparison of model predictions against the independent, authoritative ground truth data from the USDA NASS.

The primary validation workflow involves several key steps:

1. Data Partitioning: For the tree-based models, a standard random train-test split is used, where 80% of the data is used for training and 20% is held out for testing. For time-series models, a temporal split is used to ensure the model is always validated on data from a time period after the training data.

2. Performance Metrics: Model performance is quantified using a standard set of regression metrics:
    - R-squared (R^2): Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
    - Root Mean Squared Error (RMSE): Measures the square root of the average of the squared differences between prediction and actual observation. It is sensitive to large errors.
    - Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction.
3. Cross-Validation: To ensure the model is robust and not just performing well on a single random split, 5-fold cross-validation is used in the final monthly tree-based modeling script. This provides a more reliable estimate of the model's generalization performance.
4. Reporting: The validation results, including performance metrics and year-by-year prediction comparisons, are compiled into detailed text reports and visualizations, which are saved in the corresponding results directories.

3.4.2. Primary Approach: Tree-Based Ensemble Models

The most successful models in this project were tree-based ensembles, which use the yearly aggregated feature set. The modeling scripts evaluate three standard algorithms:

- XGBoost (Extreme Gradient Boosting)
- Random Forest
- Gradient Boosting

The training workflow involves splitting the data into training and testing sets, training each model on the training data, and evaluating its performance on the unseen test set. One script further enhances this by incorporating the yearly weather features, which led to a demonstrable improvement in model accuracy. The final results from the monthly tree-based models are summarized in Table 2.

Model	R^2 Score	RMSE (bu/acre)
XGBoost	**0.5269**	**13.03**
RandomForest	0.5261	13.04
GradientBoosting	0.5132	13.21

Table 2. Performance of tree-based models on the final monthly dataset

Based on the lowest Root Mean Squared Error (RMSE), XGBoost was identified as the best-performing model in this configuration.

Table 3 summarizes the XGBoost results across the three main experimental configurations, showing how performance evolved as the feature set was expanded.

Configurations	Feature Type	R^2	RMSE (bu/acre)
Configuration 05	Yearly satellite only	0.511	23.13
Configuration 09	Yearly satellite + weather	0.563	21.87
Configuration 12	Monthly pivoted + weather	0.527	13.03

Table 3. XGBoost performance across experimental configurations (leave-one-year-out CV, 26 counties)

The three experimental configurations in this thesis use the same XGBoost model and the same hyperparameter tuning process - what changes between them is the feature set. Configuration 05 uses only yearly-aggregated Sentinel-2 vegetation indices: each index is summarised across the full growing season, giving one feature vector per county-year. Configuration 09 adds historical weather variables - monthly maximum temperature, total precipitation, and accumulated growing degree days - to that same yearly format. Configuration 12 drops the yearly aggregation entirely, replacing it with a monthly pivoted feature matrix that retains each index's month-by-month trajectory, again combined with the weather variables. The XGBoost model and leave-one-year-out crossvalidation scheme are unchanged across all three.

3.4.3. Experimental Approach: Deep Learning Models

As an alternative to tabular models, several deep learning architectures were explored to determine if they could automatically learn temporal patterns from the sequential monthly data. The primary

models tested were Long Short-Term Memory (LSTM) networks and 1D Convolutional Neural Networks (CNN). The LSTM consisted of two stacked layers of 50 hidden units each with ReLU activation, a dropout layer at rate 0.2 after each LSTM layer, a Dense(25) hidden layer, and a single linear output neuron. It was trained with the Adam optimiser, a batch size of 8, for up to 100 epochs with early stopping (patience=10) on validation loss. The final multi-county experiment used 9 input features per time step—6 vegetation indices plus monthly maximum temperature, minimum temperature, and cumulative precipitation. The CNN used two one-dimensional convolutional layers (64 filters, kernel size 3; then 128 filters, kernel size 2, each with ReLU activation), followed by MaxPooling (pool size 2) after the first layer, a Flatten operation, a Dense(100) layer, dropout at 0.3, a Dense(50) layer, and a linear output. It was trained with the Adam optimiser, batch size 16, for up to 2,000 epochs with early stopping (patience=15). Both models received sequences of 7 monthly time steps as input. Hyperparameters were set by manual search.

These models were trained on the sequential feature set, which preserves the month-by-month data structure. However, both the CNN and LSTM models performed poorly, yielding negative R^2 scores. This indicates that they failed to generalize from the training data. The primary reason is believed to be the limited number of training samples (county-years) available, as deep learning models typically require vast amounts of data to learn complex patterns effectively. Due to these results, this approach was not pursued for the final implementation, but it remains a viable direction for future work if a larger ground-truth dataset becomes available.

3.5. Discussion

3.5.1. Interpretation of Results

The best-performing model in this thesis – XGBoost trained on monthly pivoted features with weather augmentation – achieved R^2=0.527 and RMSE=13.03 bu/acre across 26 Corn Belt counties using leave-one-year-out cross-validation. Situating this result in the literature is difficult because study designs vary substantially in county coverage, crop type, temporal range, and validation approach, but several published Sentinel-2 yield estimation studies provide rough benchmarks [24], [25], [26]. Studies focusing on smaller, more homogeneous regions or on single growing seasons typically report higher R^2 values (0.6–0.8), while studies covering diverse geographies with multi-year hold-out validation tend to report values closer to 0.5–0.6. The result obtained here is consistent with the lower end of that range given that the 26 counties span five states with different soil types and climate regimes, and the temporal cross-validation prevents leakage of information between years.

The trajectory of results across experimental scripts is informative. The initial yearly-aggregated XGBoost model (script 05) achieved R^2=0.511 with RMSE=23.13 bu/acre using only satellite-

derived features. Adding yearly weather variables (script 07) improved R^2 to 0.563 and reduced RMSE to 21.87 bu/acre, confirming that temperature and precipitation explain yield variance that spectral data alone cannot capture. The shift to monthly pivoted features (script 12) produced a lower R^2 (0.527) but halved the RMSE to 13.03 bu/acre. This apparent contradiction – a lower R^2 with a lower RMSE – reflects the expansion from 8 seasons in the yearly model to more county-year combinations in the monthly model, which introduces harder prediction targets. The RMSE reduction is the more practical measure of accuracy for yield forecasting applications, where absolute error in bushels per acre matters more than proportion of variance explained.

3.5.2. Why Tree-Based Models Outperformed Deep Learning

The deep learning models tested – LSTM and CNN – both failed to converge to meaningful predictions, with LSTM reaching $R^2=-90.54$ and CNN reaching $R^2=-0.21$. The LSTM contained 33,501 trainable parameters; the CNN contained 35,729. Both exceed the approximately 146 county-year training samples by a factor of roughly 230× and 245× respectively, placing both models firmly in an overparameterised regime. These results do not reflect a fundamental limitation of these architectures on agricultural data, but rather a mismatch between data scale and model complexity. LSTM networks designed to learn temporal dependencies typically require hundreds to thousands of training sequences to generalize across input patterns; the present dataset provides roughly 200 county-year combinations after filtering, split across 8 growing seasons. With that few samples, the network overfits the training years and produces random or constant-value predictions on hold-out years. The CNN faced the same problem: spatial and temporal convolutions designed for image- or video-scale datasets receive too little signal to tune their parameters when applied to a 108-column tabular input with fewer than 200 rows.

XGBoost, by contrast, is well-suited to tabular data at this scale. Its gradient boosting procedure builds an ensemble of shallow decision trees sequentially, where each tree corrects the residuals of the previous one. The L1 and L2 regularization terms in XGBoost prevent individual trees from memorizing the training set, and the early stopping criterion halted training before validation RMSE began to rise. Random Forest performed similarly on most configurations, confirming that the advantage is shared by tree-based ensembles in general rather than specific to gradient boosting [12], [23].

3.5.3. Limitations

Several limitations should be considered when interpreting these results. First, all predictions and validation are at county level. County-level aggregation masks within-county variation: a county with mixed soil types, irrigated and rainfed parcels, or varied planting dates will show a smoothed yield

figure that the model may predict well on average while being substantially wrong for individual fields. Second, the dataset covers 2016–2024, a period of roughly average growing conditions in the Corn Belt. The model has not been tested on historical anomaly years such as the 2012 drought, which would be a stronger test of generalization. Third, the study covers only corn. Transferring the pipeline to soybeans, winter wheat, or other crops would require retraining and may require different feature sets. Fourth, the pipeline relies on the USDA NASS Quick Stats API for ground truth, which limits applicability outside the US without an equivalent administrative yield reporting system.

Conclusions

Summarizing the research results obtained and documented throughout the development of the automated crop yield estimation system, the following conclusions are formulated based on the defined project objectives:

1. **Comparison of spectral indices** reveals that while the Normalized Difference Vegetation Index (NDVI) is a good starting point for measuring vegetation, it saturates in dense canopies and misses stress signals that EVI and NDRE detect. Using all six indices - covering canopy density, chlorophyll content, and soil-adjusted reflectance - gave the model more discriminative features, and feature importance analysis consistently showed NDRE and MCARI contributing independently of NDVI to predictions.

2. **The developed automated data pipeline** successfully processed Sentinel-2 Level-2A imagery across 26 counties and 8 growing seasons (2017–2024) with acceptable data completeness. Monthly median composites, built using the Level-2A cloud mask, produced a consistent time series of six vegetation indices per county without manual scene selection. The pipeline runs end-to-end without manual intervention and produces a merged satellite–weather dataset ready for model training.

3. **Feature engineering results** confirm that adding weather variables improves predictive accuracy. XGBoost with yearly satellite features alone reached R^2=0.511, RMSE=23.13 bu/acre; adding maximum temperature and cumulative precipitation improved this to R^2=0.563, RMSE=21.87 bu/acre. The improvement was most noticeable in drought years, where spectral indices alone could not fully capture yield losses driven by heat and water stress.

4. **Architectural evaluation** shows XGBoost as the best-performing model: R^2=0.527, RMSE=13.03 bu/acre on the 26-county monthly pivoted dataset. Deep learning models failed to generalize on the available data - LSTM reached R^2=–90.54 and CNN reached R^2=–0.21 - because fewer than 200 county-years is not enough to train sequence models of that complexity. XGBoost handled the wide tabular feature format without requiring sequence-aware architecture and its L1/L2 regularization kept overfitting manageable despite the small sample.

5. **Systematic validation** against USDA NASS county yields across 26 counties shows consistent performance with R^2=0.527 and RMSE=13.03 bu/acre. The model was trained and tested with time-aware cross-validation, so these results reflect prediction on genuinely

unseen future seasons, not just held-out samples from the same years. Consistent R^2 across diverse Corn Belt counties - ranging from Iowa to Ohio - suggests the approach is not tuned to specific local conditions and could be applied to other regions with adequate Sentinel-2 coverage and NASS ground truth.

Recommendations for future work. Several directions could strengthen and extend this work. First, expanding the training window beyond 2016–2024 would give tree-based models more seasons to learn from, particularly for capturing anomalous years like drought or early frost. Second, integrating soil property layers from the USDA SSURGO database as static features could improve predictions in counties where soil type is a limiting factor. Third, fusing Sentinel-1 SAR imagery with Sentinel-2 optical data would reduce cloud-gap artefacts that currently affect some growing seasons. Fourth, extending the pipeline to additional crops such as soybeans or winter wheat would test the generality of the monthly pivoting approach. Finally, sub-county resolution experiments using field-boundary masks could reveal whether county-level aggregation is a ceiling on performance or whether the signal is genuinely county-scale.

List of references

[1] H. Feng, G. Chen, L. Xiong, Q. Liu, and W. Yang, 'Accurate Digitization of the Chlorophyll Distribution of Individual Rice Leaves Using Hyperspectral Imaging and an Integrated research Pipeline', *Front. Plant Sci.*, vol. 8, Jul. 2017, doi: 10.3389/fpls.2017.01238.

[2] M. S. Mohd Asaari *et al.*, 'Close-range hyperspectral image analysis for the early detection of stress responses in individual plants in a high-throughput phenotyping platform', *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 121–138, Apr. 2018, doi: 10.1016/j.isprsjprs.2018.02.003.

[3] W. Zhu, J. Li, L. Li, A. Wang, X. Wei, and H. Mao, 'Nondestructive diagnostics of soluble sugar, total nitrogen and their ratio of tomato leaves in greenhouse by polarized spectra–hyperspectral data fusion', *Int. J. Agric. Biol. Eng.*, vol. 13, no. 2, Art. no. 2, Apr. 2020, doi: 10.25165/ijabe.v13i2.4280.

[4] A. Wendel, J. Underwood, and K. Walsh, 'Maturity estimation of mangoes using hyperspectral imaging from a ground based mobile platform', *Comput. Electron. Agric.*, vol. 155, pp. 298–313, Dec. 2018, doi: 10.1016/j.compag.2018.10.021.

[5] P. R. Eddy, A. M. Smith, B. D. Hill, D. R. Peddle, C. A. Coburn, and R. E. Blackshaw, 'Weed and crop discrimination using hyperspectral image data and reduced bandsets', *Can. J. Remote Sens.*, vol. 39, no. 6, pp. 481–490, Jan. 2014, doi: 10.5589/m14-001.

[6] J. Behmann *et al.*, 'Generation and application of hyperspectral 3D plant models: methods and challenges', *Mach. Vis. Appl.*, vol. 27, no. 5, pp. 611–624, Jul. 2016, doi: 10.1007/s00138-015-0716-8.

[7] Y. Jiang, J. L. Snider, C. Li, G. C. Rains, and A. H. Paterson, 'Ground Based Hyperspectral Imaging to Characterize Canopy-Level Photosynthetic Activities', *Remote Sens.*, vol. 12, no. 2, Art. no. 2, Jan. 2020, doi: 10.3390/rs12020315.

[8] K. Nagasubramanian, S. Jones, A. K. Singh, S. Sarkar, A. Singh, and B. Ganapathysubramanian, 'Plant disease identification using explainable 3D deep learning on hyperspectral images', *Plant Methods*, vol. 15, p. 98, 2019, doi: 10.1186/s13007-019-0479-8.

[9] V. Gonzalez-Dugo, P. Hernandez, I. Solis, and P. J. Zarco-Tejada, 'Using High-Resolution Hyperspectral and Thermal Airborne Imagery to Assess Physiological Condition in the Context of Wheat Phenotyping', *Remote Sens.*, vol. 7, no. 10, Art. no. 10, Oct. 2015, doi: 10.3390/rs71013586.

[10] R. R. Izzo, A. N. Lakso, E. D. Marcellus, T. D. Bauch, N. G. Raqueño, and J. van Aardt, 'An initial analysis of real-time sUAS-based detection of grapevine water status in the Finger Lakes wine country of upstate New York', in *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping IV*, SPIE, May 2019, pp. 276–293. doi: 10.1117/12.2518762.

[11] E. Honkavaara *et al.*, 'Processing and Assessment of Spectrometric, Stereoscopic Imagery Collected Using a Lightweight UAV Spectral Camera for Precision Agriculture', *Remote Sens.*, vol. 5, no. 10, Art. no. 10, Oct. 2013, doi: 10.3390/rs5105006.

[12] J. H. Jeong *et al.*, 'Random Forests for Global and Regional Crop Yield Predictions', *PLOS ONE*, vol. 11, no. 6, Art. no. e0156571, Jun. 2016, doi: 10.1371/journal.pone.0156571.

[13] H. Liu *et al.*, 'UAV-Borne Hyperspectral Imaging Remote Sensing System Based on Acousto-Optic Tunable Filter for Water Quality Monitoring', *Remote Sens.*, vol. 13, no. 20, Art. no. 20, Jan. 2021, doi: 10.3390/rs13204069.

[14] P. S. Thenkabail, 'Optimal hyperspectral narrowbands for discriminating agricultural crops', *Remote Sens. Rev.*, vol. 20, no. 4, pp. 257–291, Dec. 2001, doi: 10.1080/02757250109532439.

[15] A. C. Velasco, C. A. V. García, and H. A. Fuentes, "A comparative study of target detection algorithms in hyperspectral imagery applied to agricultural crops in Colombia," Tecnura, vol. 20, no. 49, Art. no. 49, Jul. 2016, doi: 10.14483/udistrital.jour.tecnura.2016.3.a06.

[16] C. Wu, X. Han, Z. Niu, and J. Dong, 'An evaluation of EO-1 hyperspectral Hyperion data for chlorophyll content and leaf area index estimation', *Int. J. Remote Sens.*, vol. 31, no. 4, pp. 1079–1086, Feb. 2010, doi: 10.1080/01431160903252335.

[17] S. Bostan, M. A. Ortak, C. Tuna, A. Akoguz, E. Sertel, and B. Berk Ustundag, 'Comparison of classification accuracy of co-located hyperspectral & multispectral images for agricultural purposes',

in *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, Jul. 2016, pp. 1–4. doi: 10.1109/Agro-Geoinformatics.2016.7577671.

[18] I. Mariotto, P. S. Thenkabail, A. Huete, E. T. Slonecker, and A. Platonov, 'Hyperspectral *versus* multispectral crop-productivity modeling and type discrimination for the HyspIRI mission', *Remote Sens. Environ.*, vol. 139, pp. 291–305, Dec. 2013, doi: 10.1016/j.rse.2013.08.002.

[19] C. Gomez, R. A. Viscarra Rossel, and A. B. McBratney, 'Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study', *Geoderma*, vol. 146, no. 3, pp. 403–411, Aug. 2008, doi: 10.1016/j.geoderma.2008.06.011.

[20] Y.-L. Weng, P. Gong, and Z.-L. Zhu, 'A Spectral Index for Estimating Soil Salinity in the Yellow River Delta Region of China Using EO-1 Hyperion Data', *Pedosphere*, vol. 20, no. 3, pp. 378–388, Jun. 2010, doi: 10.1016/S1002-0160(10)60027-6.

[21] C. Kurade *et al.*, 'An Automated Image Processing Module for Quality Evaluation of Milled Rice', *Foods*, vol. 12, no. 6, Art. no. 6, Jan. 2023, doi: 10.3390/foods12061273.

[22] Q. Wang and J. Jin, 'Leaf transpiration of drought tolerant plant can be captured by hyperspectral reflectance using PLSR analysis', *IForest - Biogeosciences For.*, vol. 9, no. 1, p. 30, 2015, doi: 10.3832/ifor1634-008.

[23] F. Huber, A. Yushchuk, B. Stratmann, and V. Steinhage, 'Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches', *Comput. Electron. Agric.*, vol. 202, Art. no. 107346, Nov. 2022, doi: 10.1016/j.compag.2022.107346.

[24] M. Yli-Heikkilä *et al.*, 'Scalable Crop Yield Prediction with Sentinel-2 Time Series and Temporal Convolutional Network', *Remote Sens.*, vol. 14, no. 17, Art. no. 4193, Sep. 2022, doi: 10.3390/rs14174193.

[25] G. Xiao *et al.*, 'Winter wheat yield estimation at the field scale using Sentinel-2 data and deep learning', *Comput. Electron. Agric.*, vol. 216, Art. no. 108555, Jan. 2024, doi: 10.1016/j.compag.2023.108555.

[26] J. Desloires, D. Ienco, and A. Botrel, ‘Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods’, *Comput. Electron. Agric.*, vol. 209, Art. no. 107632, Jun. 2023, doi: 10.1016/j.compag.2023.107632.

Annex A. Study Area County List

The 26 US counties used in this study are listed below. County FIPS codes are the standard US Federal Information Processing Standard identifiers as used in USDA NASS Quick Stats and the Sentinel Hub API requests. All counties are located in the US Corn Belt and were selected based on data availability and corn cultivation area.

FIPS Code	State	County
17011	IL	Bureau
17019	IL	Champaign
17073	IL	Henry
17075	IL	Iroquois
17103	IL	Lee
17105	IL	Livingston
17113	IL	McLean
17195	IL	Whiteside
19011	IA	Benton
19015	IA	Boone
19019	IA	Buchanan
19023	IA	Butler
19031	IA	Cedar
19045	IA	Clinton
19047	IA	Crawford
19055	IA	Delaware
19061	IA	Dubuque
19065	IA	Fayette
19069	IA	Franklin
19075	IA	Grundy

19167	IA	Sioux
19169	IA	Story
19193	IA	Winnebago
19197	IA	Wright
27129	MN	Renville
31109	NE	Lancaster

Table 4. Study counties and FIPS codes

Annex B. Monthly Pivoted Feature List

The table below lists all 102 predictor columns in the monthly pivoted feature matrix used by the final XGBoost model (Script 12). Column names follow the pattern INDEX_M, where INDEX is the vegetation index or weather variable name and M is the calendar month number (1 = January, 4 = April, 10 = October). Satellite-derived indices cover months 1–11; weather variables cover months 1–12.

Feature Group	**Column Names (102 total)**
EVI (Enhanced Vegetation Index)	EVI_1, EVI_2, EVI_3, EVI_4, EVI_5, EVI_6, EVI_7, EVI_8, EVI_9, EVI_10, EVI_11
GNDVI (Green NDVI)	GNDVI_1, GNDVI_2, GNDVI_3, GNDVI_4, GNDVI_5, GNDVI_6, GNDVI_7, GNDVI_8, GNDVI_9, GNDVI_10, GNDVI_11
MCARI (Mod. Chlorophyll Absorption Ratio)	MCARI_1, MCARI_2, MCARI_3, MCARI_4, MCARI_5, MCARI_6, MCARI_7, MCARI_8, MCARI_9, MCARI_10, MCARI_11
NDRE (Red-Edge NDVI)	NDRE_1, NDRE_2, NDRE_3, NDRE_4, NDRE_5, NDRE_6, NDRE_7, NDRE_8, NDRE_9, NDRE_10, NDRE_11
NDVI (Normalised Difference Vegetation Index)	NDVI_1, NDVI_2, NDVI_3, NDVI_4, NDVI_5, NDVI_6, NDVI_7, NDVI_8, NDVI_9, NDVI_10, NDVI_11
SAVI (Soil-Adjusted Vegetation Index)	SAVI_1, SAVI_2, SAVI_3, SAVI_4, SAVI_5, SAVI_6, SAVI_7, SAVI_8, SAVI_9, SAVI_10, SAVI_11
precip_sum (Monthly	precip_sum_1, precip_sum_2, precip_sum_3, precip_sum_4, precip_sum_5, precip_sum_6, precip_sum_7, precip_sum_8,

precipitation sum, mm)	precip_sum_9, precip_sum_10, precip_sum_11, precip_sum_12
temp_max_mean (Monthly mean daily max. temp., °C)	temp_max_mean_1, temp_max_mean_2, temp_max_mean_3, temp_max_mean_4, temp_max_mean_5, temp_max_mean_6, temp_max_mean_7, temp_max_mean_8, temp_max_mean_9, temp_max_mean_10, temp_max_mean_11, temp_max_mean_12
temp_min_mean (Monthly mean daily min. temp., °C)	temp_min_mean_1, temp_min_mean_2, temp_min_mean_3, temp_min_mean_4, temp_min_mean_5, temp_min_mean_6, temp_min_mean_7, temp_min_mean_8, temp_min_mean_9, temp_min_mean_10, temp_min_mean_11, temp_min_mean_12

Table 5. Monthly pivoted feature columns