Evaluation of a genetic programming - based framework for NIR estimation from RGB bands

Diego Saqui
Institute of Science, Technology and Innovation (ICTIN)
Federal University of Lavras
São Sebastião do Paraíso - MG, Brazil
diego.saqui@ufla.br

Henrique L. M. Monteiro
Institute of Science, Technology and Innovation (ICTIN)
Federal University of Lavras
São Sebastião do Paraíso - MG, Brazil
henrique.monteiro@ufla.br

Rafael R. M. Ribeiro
Institute of Science, Technology and Innovation (ICTIN)
Federal University of Lavras
São Sebastião do Paraíso - MG, Brazil
rafaelribeiro@ufla.br

Lucas E. de O. Aparecido
Department of Agricultural Engineering
Federal Institute of Southern Minas Gerais
Muzambinho - MG, Brazil
lucas.aparecido@ifsuldeminas.edu.br

Steve T. M. Ataky
University of Quebec in Montreal
Montreal - QC, Canada
steve.ataky@nca.ufma.br

Danton D. Ferreira
Department of Automation (DAT)
Federal University of Lavras
Lavras - MG, Brazil
danton@ufla.br

***Abstract*—Remote sensing using the Near Infrared (NIR) band is a well-established technique for assessing plant health in precision agriculture. However, traditional approaches require costly multi- and hyperspectral sensors, limiting their accessibility for small and medium-sized farms. This study proposes a low-cost alternative by estimating NIR reflectance from standard RGB imagery using Genetic Programming (GP). GP automatically evolves mathematical models to predict NIR values, which are then used to compute the Normalized Difference Vegetation Index (NDVI)—a widely used indicator of plant vigor and health. The method was evaluated on three benchmark hyperspectral datasets (Indian Pines, Salinas, and Pavia University) across two scenarios: (1) models trained per dataset and (2) a generalized model trained on combined datasets. GP demonstrated strong performance, achieving Pearson correlation coefficients up to 0.9059 and spectral angle errors below 0.001 radians. NDVI estimations were particularly accurate, with correlations exceeding 0.97 in some cases. Due to its simplicity, interpretability, and low computational requirements, the proposed approach offers a practical and scalable solution for vegetation monitoring in resource-constrained agricultural environments, enabling small and medium-sized producers to leverage spectral analysis without expensive sensors.**

***Index Terms*—Genetic programming, NIR estimation, vegetation indices, precision agriculture, remote sensing, low-cost solutions**

I. Introduction

Unmanned aerial vehicles (UAVs) have recently become increasingly popular in applications like aerial photography, environmental monitoring, oil spill detection, and precision agriculture [1]. In particular, data acquisition without physical contact, using multispectral and hyperspectral sensors mounted on drones, enables the extraction of spectral information about crop health [2], [3].

Multispectral images allow for the acquisition of information across dozens of bands, while hyperspectral images capture data in hundreds of bands [4], [5]. Some of these bands, although invisible to the human eye, carry essential information for constructing vegetation indices (VIs). VIs are typically computed using a visible spectrum band and a Near Infrared (NIR) band. NIR, in particular, covers spectral regions from 780 to 2500 nm with information that is not visible to the human eye but allows for the acquisition of spectral data where vegetation exhibits high reflectance levels [6], [7]. The spectral band around 800 nm in the NIR region is one of the most frequently used in agricultural applications, especially in the calculation of VIs. This band exhibits high reflectance in healthy vegetation, enabling efficient differentiation between vegetated and non-vegetated areas [8], [9].

VIs are equations combining different spectral bands to indicate vegetation activity [10]. An example is the Normalized Difference Vegetation Index (NDVI), an index that allows the assessment of plant vigor and health, identifying areas affected by water stress, pests, and diseases. It is one of the most well-known and widely accepted indices among farmers [11]. The NDVI is calculated using NIR (approximately 800 nm) and Red (approximately 670 nm) band (visible spectrum) these bands being historically adopted for offering strong spectral contrast [7].

However, the acquisition of NIR data generally requires specialized sensors, which can be expensive and complex to integrate. These sensors, when embedded in drones, enable the capture of data such as NIR but come with high costs, making their adoption unfeasible for small and medium-sized agricultural producers. To overcome this limitation, recent studies have explored the estimation of NDVI values directly from conventional RGB images, eliminating the need for NIR data or adopting approaches that estimate NIR to subsequently obtain NDVI [7], [12]. This is especially useful in scenarios where budget constraints or limited access to equipment are key concerns.

Some works consider machine learning algorithms, such as k-nearest neighbors (KNN) for spectral classification [13]; and

generative adversarial networks (GANs), such as the Pix2Pix network for converting RGB images into spectral indices [12], [14]. While these approaches have yielded promising outcomes, interpreting the outputs of such models remains a challenge, and the dependence on large datasets for training persists.

VIs are generated through arithmetic combinations of spectral bands that exhibit distinct reflectance patterns of vegetation activity [10]. Genetic Algorithms (GAs) are a metaheuristic strategy that explores combinatorial problems and tends to present good results where they are applied. In precision agriculture, particularly with multi- and hyperspectral images, GAs and derived algorithms such as Non-dominated Sorting Genetic Algorithm (NSGA-II) and multiobjective evolutionary algorithm based on decomposition (MOEA/D) are utilized to select bands while enhancing the classification and categorization process of different agricultural crops [15]–[17]. Genetic programming (GP) is a technique that enables computers to solve problems without explicit programming, using GAs to automatically generate computer programs and mathematical equations [18]. GP is a type of evolutionary learning that mimics natural selection, evolving candidate solutions through mutation and recombination.

GP has been used to develop customized spectral indices tailored to specific agricultural and environmental needs. It has been used to generate new VIs that are better correlated with the C factor of the RUSLE equation, allowing for a more accurate estimation of soil erosion by considering senescent and dead vegetation, which traditional indices such as NDVI and EVI do not adequately capture [19]. It has also been employed in the development of a spectral index that improves bare soil identification, standing out for its ability to automatically generate a classification threshold, resulting in high accuracy in differentiating various land covers [20]. Additionally, it has been applied in the agroindustry to create new hyperspectral indices aimed at identifying dry matter in agricultural crops. By using correlation coefficients as fitness functions, the methodology enabled the generation of simple and interpretable indices, based on the combination of a few hyperspectral bands, facilitating non-invasive analysis of the internal composition of agricultural products [21].

This study explores a GP - based framework to estimate NIR from visible bands and, subsequently, calculate NDVI. This framework enables the automatic discovery of mathematical formulas and transformation functions to estimate spectral information from RGB images, eliminating the need for real NIR data while maintaining high accuracy in vegetation health assessment. We expect GP to uncover complex relationships between visible spectral bands and NIR, offering an accessible and low-cost solution for precision agriculture applications. GP's flexibility supports the development of interpretable models that require less data, addressing challenges associated with conventional machine learning and neural network-based techniques. This is particularly beneficial for small and medium-sized farmers, who will be able to perform advanced vegetation analysis without the need for expensive NIR sensors.

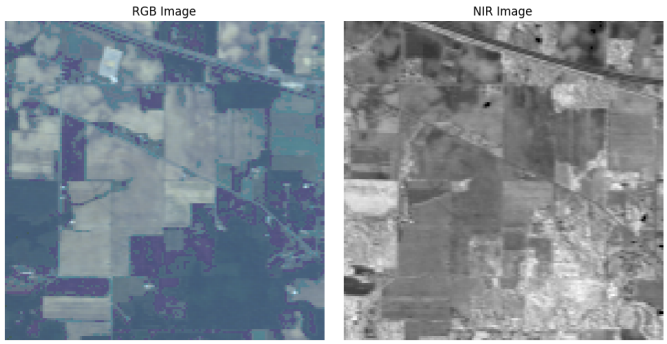

Fig. 1. Indian Pines image with color (RGB) and NIR bands.

The remainder of this paper is organized as follows: Section II discusses the spectral datasets and preprocessing steps used in this study. Section III provides a detailed description of the proposed methodology, including the genetic programming framework and evaluation metrics. Section IV presents the results and discussions, comparing the estimated NIR and NDVI with ground truth data. Finally, Section V discusses the implications of the findings and outlines future research directions.

II. Spectral Datasets and preprocessing

Because there is no large, standardized hyperspectral dataset publicly available, this study used well-established datasets in formats such as .hdr and .lan for extracting spectral information. These images include ground truth data, which associate each pixel with specific land cover types—such as vegetation, bare soil, minerals, and water. However, these labels were not directly relevant to the goals of this work. All pixels were considered, regardless of their assigned classes, since the goal was to estimate the NIR band through equations combining various mathematical operations on the visible bands. Therefore, the NIR band was used directly as the ground truth for evaluating the models.

For spectral analysis, three publicly available hyperspectral scenes widely referenced in the literature were chosen: Indian Pines, Salinas, and Pavia University.

A. Indian Pines

The Indian Pines dataset (Fig. 1) was collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over an agricultural area in the Indian Pines region, Indiana, USA. It contains 220 spectral bands in the 400–2500 nm range, with a spatial resolution of 20 meters per pixel. After removing bands with strong water absorption, 200 bands are typically used for analysis. The dataset includes 16 land cover classes, primarily agricultural crops such as corn, soybeans, and alfalfa, as well as forested areas and bare soil.

B. Salinas

The Salinas dataset was also acquired by the AVIRIS sensor over the agricultural region of Salinas Valley, California, USA. It contains 224 spectral bands in the 400–2500 nm range,

with a higher spatial resolution of 3.7 meters per pixel. After removing water absorption bands, 204 bands are used for analysis. This dataset is characterized by 16 land cover classes, representing different types of vegetation, agricultural crops, and bare soil. It is widely used for hyperspectral image classification due to its complexity and rich spectral information.

C. Pavia University

The Pavia University (Pavia U.) dataset was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the city of Pavia, Italy. It contains 102 spectral bands in the 430–860 nm range, with a spatial resolution of 1.3 meters per pixel. The dataset covers an urban area with 9 land cover classes, including buildings, asphalt roads, vegetation, and bare soil. Due to its complexity and diversity of urban materials, the Pavia U. dataset is widely used in hyperspectral classification studies in urban environments [22].

These datasets include both visible and near-infrared (NIR) spectral information, which are key for this study. Among them, the visible bands and one NIR band (centered near 800 nm) were used. Data acquisition was performed using the Spectral Python (SPy) library, enabling the loading and manipulation of spectral data [23].

The spectral data were normalized to make the input features comparable in scale for the evolutionary model. Normalization was performed independently for each spectral band, scaling the values to the [0, 1] range. This preprocessing step is essential to prevent bands with larger numeric ranges from dominating the optimization process and to improve the convergence and stability of the model.

III. Genetic Programming - based Framework

The optimization process was implemented using the Distributed Evolutionary Algorithms in Python (DEAP) library [24], focusing on the Genetic Programming (GP) - based framework for spectral modeling and reconstruction. The goal is to evolve a mathematical function, represented as an expression tree with operators, to estimate the NIR band from the visible bands.

Each individual (candidate solution) was structured to represent an equation composed of visible bands (in the RGB ranges) and algebraic expressions.

A. Genetic Operations

The genetic operations considered were the following:

- Mathematical expressions composed of basic algebraic operations (addition, subtraction, multiplication, and protected division)
- Trigonometric functions (sine and cosine)
- Statistical functions (arithmetic and weighted mean)

B. Evaluation Function

The quality of the generated individuals was evaluated using a weighted fitness function that balances accuracy and model simplicity. Three components were considered:

- Pearson Correlation Coefficient: Measures the linear relationship between the actual and estimated NIR values.
- Root Mean Squared Error (RMSE): Indicates the average magnitude of the estimation error — the lower, the better.
- Size Penalty: Penalizes overly complex individuals to discourage overfitting.

The fitness function was defined as a weighted combination of these elements, as shown in Eq. (1):

$$\text{fit} = 0.6 \cdot (1 + \text{Correlation}) + 0.4 \cdot (1 - \text{RMSE}) - \text{SizePenalty} \quad (1)$$

This formulation prioritizes models that closely follow the true NIR values, minimize prediction errors, and maintain structural simplicity.

C. Algorithm parameters and Evolutionary Process

The algorithm was configured with the following parameters, which were chosen empirically:

- Population size: 200 individuals where each individual represents a mathematical expression encoded as a tree structure.
- Number of generations: 200
- Crossover rate: 70%
- Mutation rate: 20%
- Selection: Tournament with size 3
- Elitism: 15 best individuals with the highest fitness values in each generation, ensuring that the best solutions were carried over to the next population without modification.

The evolution was conducted over 200 generations, with each generation recording the average, minimum, and maximum values of the fitness function.

The parameters were selected based on preliminary tests and prior knowledge. Trial-and-error experiments were conducted to assess how population size, number of generations, and genetic operators affected solution quality and convergence speed. The final values aimed to balance exploration, exploitation, and computational cost. Tournament selection (size 3) and elitism were used to retain the best individuals and maintain selective pressure.

D. Extraction of the Best Model

At the end of the evolution, the best model found was extracted and converted into a computational function using the DEAP library's expression compiler. This function was then used to estimate the NIR values of the test pixels, and its performance was assessed by comparing the estimated values with the actual NIR values (ground truth in this context).

E. Distribution of Pixels into Training and Testing Sets

To ensure a robust evaluation, the hyperspectral image pixels were randomly divided into training (80%) and testing (20%) sets while preserving the original spectral distribution. The random selection was performed at the pixel level (not spatial regions) to avoid bias related to the spatial distribution of pixels. Each pixel was treated as an independent sample, with its visible spectral bands as features and the NIR band

value as the label. This approach ensures that the model learns general spectral relationships without overfitting to local patterns, while the testing set provides an unbiased evaluation of generalization capability. The randomization was controlled by a fixed seed for reproducibility, and normalization (min-max) was applied separately to each set to prevent data leakage.

F. Experimental Setup and Evaluation Metrics

Two experiments were conducted to evaluate the performance and generalization capacity of the GP - based framework in estimating the NIR band from visible spectral data:

- **Experiment 1**: Each dataset (Indian Pines, Salinas, and Pavia U.) was independently split into training and testing subsets. A separate GP - based framework was trained for each dataset, resulting in three distinct equations for NIR estimation—one per image.
- **Experiment 2**: A single training set was created by combining the training portions of all three datasets. The GP - based framework was trained once using this unified dataset, producing a single generalized equation. This equation was then evaluated separately on the test set of each individual dataset (Indian Pines, Salinas, and Pavia U.), enabling us to evaluate how well it generalizes to different scenes.

To assess the quality of the NIR predictions in both experiments, four metrics were computed on the test data:

- **Mean Squared Error (MSE)**: Represents the average of the squared differences between the actual and estimated NIR values. Lower values indicate smaller prediction errors.
- **Root Mean Squared Error (RMSE)**: The square root of the MSE, providing an error estimate in the same scale as the reflectance data. It facilitates direct interpretation of the magnitude of the errors.
- **Pearson Correlation Coefficient** (r): Measures the strength of the linear relationship between the estimated and actual NIR values. Values closer to 1 indicate a strong positive correlation.
- **Spectral Angle Mapper (SAM)**: Computes the angle (in radians) between the predicted and actual spectral vectors. Smaller angles imply higher spectral fidelity, which is important in preserving the reflectance signature.

IV. Results and Discussions

A. Experiment 1

1) Fitness Evolution Across Datasets and Equation Analysis: The graph in Figure 2 illustrates the evolution of fitness over generations of the GP - based framework applied to the training data of three different datasets: (a) Indian Pines, (b) Salinas, and (c) Pavia U. In each subfigure, the green line represents the maximum fitness, the blue line indicates the average fitness, and the red line shows the minimum fitness observed in the population at each generation.

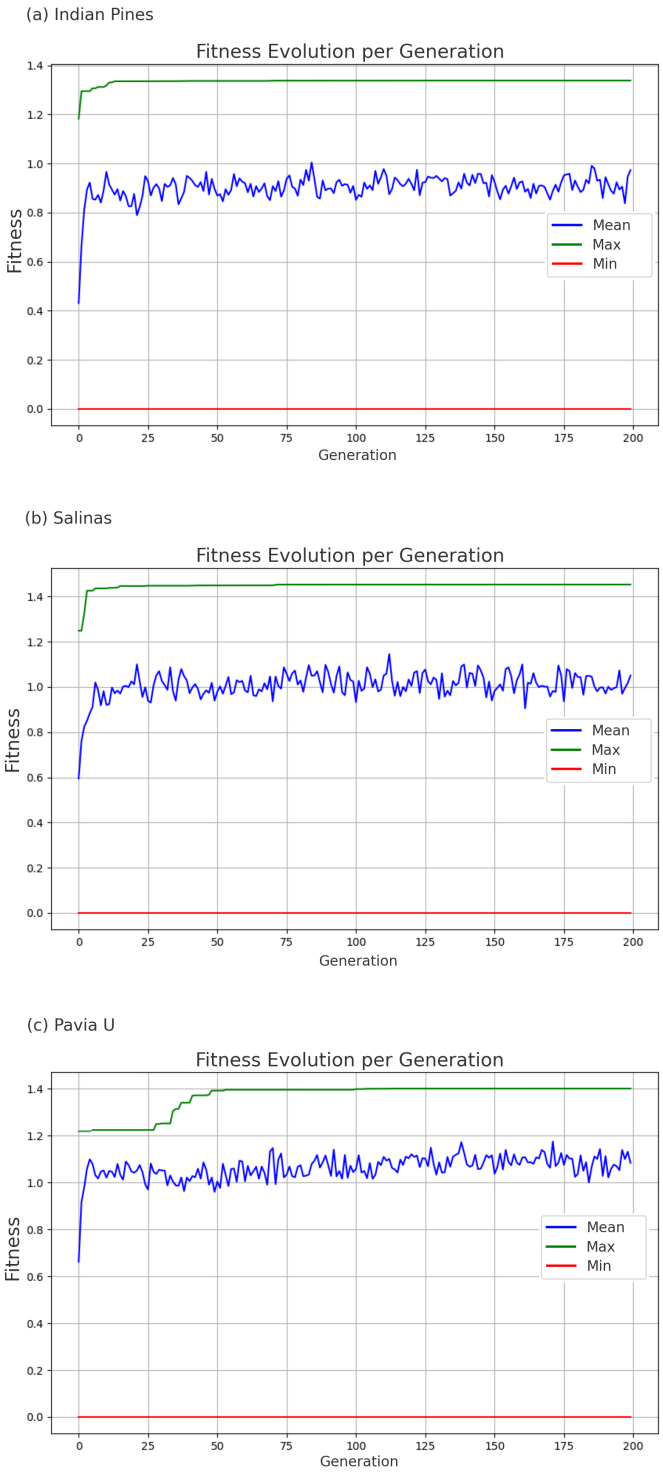

Fig. 2. Fitness progression across generations for all datasets: Indian Pines, Salinas and Pavia U. Colors represent average (blue), maximum (green), and minimum (red) fitness.

A consistent increase in both average and maximum fitness values is observed across all datasets, indicating that the population of solutions progressively improved during the training process. The minimum fitness values (red lines), however,

remained at zero throughout all generations. This behavior results from the algorithm's strategy of discarding invalid individuals—those containing inconsistencies in the fitness equation—by assigning them the worst possible fitness value. Although these problematic solutions continued to emerge occasionally, they did not hinder the overall progress of the evolutionary process.

Among the datasets, Indian Pines (Figure 2a) demonstrated the fastest convergence, with fitness values stabilizing by generation 25. Salinas (Figure 2b) reached convergence slightly later, around generation 75. In contrast, Pavia U. (Figure 2c) required a substantially longer period, converging only after nearly 110 generations. This indicates that each dataset presents a different level of difficulty for the evolutionary search.

The similarity between the average and maximum fitness values in the final generations indicates that the evolution stabilized and no significant improvements were found.

In addition to fitness convergence, the derived equations for each dataset provide insight into the nature of the solutions generated by the genetic programming approach. The final expressions are listed below:

$$\text{Indian Pines: } \hat{y} = \sin\left(\max\left(B_{16}, \ln\left(\left|\ln\left(|B_8| + \varepsilon\right)\right| + \varepsilon\right)\right)\right) \tag{2}$$

This equation is the most complex among the three, featuring:

- Nested logarithmic functions, with absolute values and a small constant $\varepsilon = 10^{-10}$ to prevent undefined operations;
- A maximum operator $\max(B_{16}, \cdot)$, choosing the higher value between a spectral band and the transformed expression;
- A sine function applied to the entire structure, introducing strong nonlinearity and oscillatory behavior.

$$\text{Salinas: } \hat{y} = \sqrt{B_{31}} - \frac{B_{30}}{B_{19}} \tag{3}$$

This expression is comparatively simpler and includes:

- A square root of band B_{31}, adding moderate nonlinearity;
- A division of two spectral bands B_{30}/B_{19}, a common structure in vegetation indices;
- A subtraction combining both terms to form the final index.

$$\text{Pavia U.: } \hat{y} = \sqrt{(B_{19} - 0.1 \cdot B_{27}) - \tan(B_{31})} \tag{4}$$

This equation consists of:

- A weighted difference between bands: $B_{19} - 0.1 \cdot B_{27}$;
- A tangent transformation on band B_{31}, introducing nonlinearity and possible sensitivity to small spectral variations;
- A square root over the whole expression, affecting the final scale and possibly regularizing extreme values.

This shows that the GP-based method generated equations specific to the spectral characteristics of each dataset. The Indian Pines equation involved deeply nested and highly nonlinear elements; the Salinas equation emphasized spectral ratios and basic arithmetic, while the Pavia U. equation combined moderate nonlinearity with weighted band differences.

This functional diversity demonstrates that no shared or consistent equation pattern emerged in Experiment 1, where each dataset was modeled independently. Such variation suggests that the algorithm adapted specifically to the unique spectral distributions of each dataset, which motivated Experiment 2 to explore the feasibility of a generalizable expression.

2) Performance Metrics Comparison: The quantitative results for NIR and NDVI predictions across the three datasets are presented in Tables I and II, respectively. All evaluations were performed on test data, ensuring that the metrics reflect the model's generalization capability and are not the result of overfitting.

Regarding the NIR prediction results in Table I, the lowest MSE and RMSE were achieved on the Pavia U. dataset, with values of 0.0035 and 0.0590, respectively. These results indicate that the model was highly accurate in approximating the NIR band in this dataset. The corresponding Pearson's correlation coefficient of 0.8359 further supports this observation, confirming a strong linear relationship between the predicted and actual values. Similarly, the SAM value of 0.0001 radians indicates that the spectral signatures were preserved with minimal angular deviation.

The Salinas dataset also showed excellent NIR prediction performance, with intermediate MSE (0.0104), RMSE (0.1021), and a high correlation coefficient (0.9059), suggesting that the model effectively captured the spectral characteristics despite the increased spectral complexity of this dataset. For Indian Pines, while the overall performance was still acceptable—with a Pearson correlation of 0.7575 and a SAM of 0.0007 radians—the MSE (0.0426) and RMSE (0.2065) were notably higher, indicating greater difficulty in modeling the NIR band in this case. This may be attributed to the higher heterogeneity and noise typically present in the Indian Pines scene.

In contrast, Table II shows that NDVI prediction yielded overall stronger correlations, particularly in the Indian Pines dataset, where Pearson's r=0.9802 and RMSE was only 0.0865. This result is expected given that NDVI, being a normalized index, tends to be more stable and less sensitive to absolute reflectance errors. The Salinas dataset also showed strong performance (RMSE = 0.0927; r=0.9384), while PaviaU exhibited slightly higher error (RMSE = 0.1342) and a lower correlation (0.8908), yet still within acceptable limits for vegetation monitoring applications.

It is important to note that SAM values for NDVI prediction were substantially higher than those for NIR, particularly in Salinas (0.7500 radians) and PaviaU (0.6295 radians), suggesting that while NDVI was predicted with high numerical accuracy, the underlying spectral shape may have experienced moderate angular distortions. These discrepancies are likely

TABLE I
NIR PREDICTION PERFORMANCE ACROSS DATASETS

Metric	Indian Pines	Salinas	PaviaU
MSE	0.0426	0.0104	0.0035
RMSE	0.2065	0.1021	0.0590
Pearson's r	0.7575	0.9059	0.8359
SAM (rad)	0.0007	0.0001	0.0001

TABLE II
NDVI PREDICTION PERFORMANCE ACROSS DATASETS

Metric	Indian Pines	Salinas	PaviaU
MSE	0.0075	0.0086	0.0180
RMSE	0.0865	0.0927	0.1342
Pearson's r	0.9802	0.9384	0.8908
SAM (rad)	0.2007	0.7500	0.6295

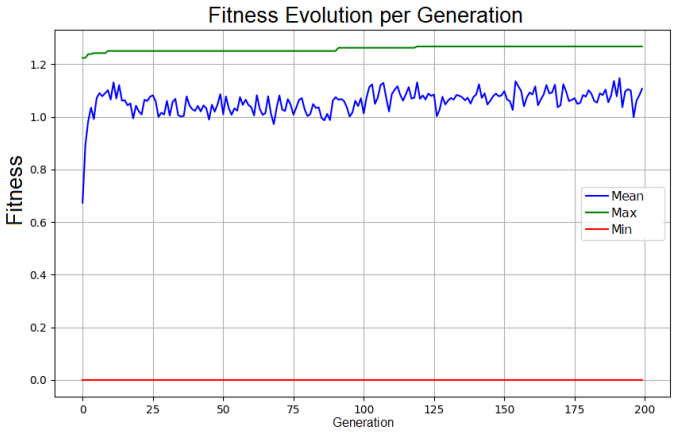

Fig. 3. Fitness progression across generations for Experiment 2 using all datasets combined. Lines indicate the average (blue), maximum (green), and minimum (red) fitness values per generation.

due to local spectral variations near vegetation transition zones (e.g., soil–vegetation boundaries), where NDVI values change more abruptly.

Overall, the results confirm the model's effectiveness in predicting both NIR reflectance and NDVI values across diverse scenes. The consistent performance on unseen test data demonstrates its potential applicability in remote sensing tasks such as vegetation analysis and land cover classification.

B. Experiment 2

1) Fitness Evolution Across Datasets and Equation Analysis: Figure 3 presents the evolution of fitness across 200 generations in Experiment 2, where a single equation was evolved to generalize over all datasets simultaneously (Indian Pines, Salinas, and Pavia U.).

A rapid increase in both average and maximum fitness values is observed within the first 20 generations. From that point onward, the maximum fitness remains nearly constant, indicating that the population quickly reached a region of optimal or near-optimal solutions. The average fitness continues to fluctuate slightly but remains relatively high throughout the remaining generations, suggesting consistent performance across the population. The minimum fitness (red line), as in Experiment 1, remains at zero due to the penalization of invalid individuals during training.

This convergence behavior demonstrates that the evolutionary process was effective even in the presence of data from multiple datasets, which increased the diversity and complexity of the search space. Despite these challenges, the algorithm succeeded in producing a simple and interpretable equation that generalized well.

The final expression obtained in Experiment 2 is:

$$\hat{y} = B_3 + (B_1 - B_{31}) \tag{5}$$

This equation is notably simpler than the ones derived individually in Experiment 1. It is a purely linear combination of three spectral bands, without any nonlinear transformations or nested operations. The structure:

- Adds the value of band B_3 to the difference $B_1 - B_{31}$;
- Reflects a direct linear weighting of bands from different spectral regions;
- May be interpreted as a form of normalized band ratio, commonly found in spectral indices.

The simplicity of this expression suggests that the GP - based framework, when trained on all datasets simultaneously, favored robust and generalizable structures over complex, dataset-specific formulations. Although this equation may not capture all nuances present in each dataset individually, its general nature could be advantageous for deployment in scenarios with limited knowledge about the underlying spectral distribution.

Overall, the results of Experiment 2 highlight the potential of evolving shared models that balance simplicity and performance, especially when aiming for broader applicability across heterogeneous remote sensing data.

2) Performance Metrics Comparison: The quantitative results for NIR and NDVI predictions obtained in Experiment 2 are summarized below. As in the previous experiment, all evaluations were conducted on test data from each dataset separately, ensuring that the reported metrics reflect the generalization capability of the model trained on the combined data.

TABLE III
NIR PREDICTION PERFORMANCE ACROSS DATASETS (EXPERIMENT 2)

Metric	Indian Pines	Salinas	PaviaU
MSE	0.0419	0.0461	0.0117
RMSE	0.2047	0.2147	0.1081
Pearson's r	0.2507	0.4654	-0.2098
SAM (rad)	0.0007	0.0001	0.0001

TABLE IV
NDVI PREDICTION PERFORMANCE ACROSS DATASETS (EXPERIMENT 2)

Metric	Indian Pines	Salinas	PaviaU
MSE	0.0116	0.0214	0.0424
RMSE	0.1077	0.1464	0.2058
Pearson's r	0.9720	0.8449	0.7103
SAM (rad)	0.2557	0.7520	0.7857

Regarding the NIR band estimation (Table III), the best result in terms of RMSE and MSE was obtained for the PaviaU dataset, with an MSE of 0.0117 and RMSE of 0.1081. Although this represents the lowest absolute error, the Pearson correlation was weak ($r = -0.2098$), indicating that the prediction values may not follow the expected linear trend of the ground truth. Nevertheless, the SAM value remained extremely low (0.0001 radians), suggesting that the predicted spectral signatures were still angularly aligned with the actual ones.

For Indian Pines, the model achieved an MSE of 0.0419 and RMSE of 0.2047, comparable to the values from Experiment 1. The Pearson correlation was modest ($r = 0.2507$), and the SAM remained consistent at 0.0007 radians. The Salinas dataset showed the highest error among the three, with MSE = 0.0461 and RMSE = 0.2147, while the correlation was slightly better ($r = 0.4654$) and the SAM was again extremely low (0.0001 radians). These results suggest that, although the spectral angles remained well preserved, the use of a single general equation led to decreased correlation and increased error compared to dataset-specific models.

In NDVI prediction (Table IV), the model trained on the combined dataset showed better correlation values. The Indian Pines dataset stood out with excellent performance: MSE = 0.0116, RMSE = 0.1077, and a very high correlation coefficient ($r = 0.9720$). The SAM value of 0.2557 radians indicates good spectral agreement, though slightly less precise than the angular consistency observed in the NIR predictions.

The Salinas dataset also showed robust NDVI prediction: MSE = 0.0214, RMSE = 0.1464, and correlation $r = 0.8449$, with a SAM of 0.7520 radians. For PaviaU, the NDVI results were somewhat weaker: MSE = 0.0424, RMSE = 0.2058, and correlation $r = 0.7103$, with a SAM of 0.7857 radians. Although the correlation values remained acceptable for remote sensing tasks, the angular deviations (SAM) were relatively high, particularly in heterogeneous areas like those in Salinas and PaviaU scenes.

Overall, the model derived in Experiment 2 using a single general equation, demonstrated strong angular fidelity (low SAM values) in NIR prediction across all datasets, confirming its ability to preserve spectral structure. However, the numerical error and correlation were generally lower than those obtained in Experiment 1, especially for PaviaU, where the NIR correlation was even negative.

For NDVI prediction, the unified model performed well, particularly for Indian Pines, and remained competitive for Salinas and PaviaU, indicating that NDVI, as a normalized index, is more resilient to the use of a generalized equation.

These findings highlight a trade-off: while a general equation can preserve spectral angles and simplify deployment, it may lead to loss of dataset-specific accuracy in absolute terms. Nevertheless, the unified approach remains viable for broader applications requiring scalability and interpretability.

V. Conclusion

This study introduced a Genetic Programming (GP)-based framework for estimating Near Infrared (NIR) reflectance from visible RGB bands, offering a cost-effective alternative to traditional multispectral sensors in precision agriculture. By evolving mathematical models automatically, the method accurately predicted NIR values using only RGB data, enabling the computation of critical vegetation indices such as NDVI. Experiments on multiple benchmark hyperspectral datasets demonstrated strong performance, with high correlation coefficients and low spectral distortion—particularly in NDVI estimation. While dataset-specific models achieved the highest accuracy, the generalized model also performed competitively, indicating its potential for broader applicability. The interpretability of the GP-derived equations distinguishes this method from black-box machine learning approaches, enhancing transparency for end-users.

Future work may focus on integrating the approach with UAV-captured RGB imagery to assess real-world applicability under field conditions. Expanding the training data to encompass a wider variety of agricultural environments could further improve robustness. Hybrid models that combine GP with lightweight neural networks may also enhance predictive performance while maintaining low computational demands. The method's affordability and minimal hardware requirements make it well-suited for small and medium-sized farms, helping bridge the gap between advanced spectral analysis and accessible agricultural technologies. Ultimately, this work contributes to the democratization of precision agriculture by empowering farmers with interpretable, low-cost tools for crop monitoring and decision-making. Future studies could also explore multi-objective optimization to better balance trade-offs between accuracy, complexity, and resource efficiency.

Acknowledgment

The authors would like to thank the Federal University of Lavras (UFLA), Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), National Council for Scientific and Technological Development (CNPQ) and Minas Gerais State Research Support Foundation (FAPEMIG) for supporting this work.

References

[1] B. Alzahrani, O. S. Oubbati, A. Barnawi, M. Atiquzzaman, and D. Alghazzawi, "Uav assistance paradigm: State-of-the-art in applications and challenges," *Journal of Network and Computer Applications*, vol. 166, p. 102706, 2020.

[2] D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," *Biosystems engineering*, vol. 114, no. 4, pp. 358–371, 2013.

[3] D. C. Tsouros, S. Bibi, and P. G. Sarigiannidis, "A review on uav-based applications for precision agriculture," *Information*, vol. 10, no. 11, p. 349, 2019.

[4] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[5] J. Liang, X. Li, P. Zhu, N. Xu, and Y. He, "Hyperspectral reflectance imaging combined with multivariate analysis for diagnosis of sclerotinia stem rot on arabidopsis thaliana leaves," *Applied Sciences*, vol. 9, no. 10, p. 2092, 2019.

[6] N. Stasenko, I. Shukhratov, M. Savinov, D. Shadrin, and A. Somov, "Deep learning in precision agriculture: Artificially generated vnir images segmentation for early postharvest decay prediction in apples," *Entropy*, vol. 25, no. 7, p. 987, 2023.

[7] D. C. de Lima, D. Saqui, S. Ataky, L. A. d. C. Jorge, E. J. Ferreira, and J. H. Saito, "Estimating agriculture nir images from aerial rgb data," in *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part I 19*. Springer, 2019, pp. 562–574.

[8] A. Bannari, D. Morin, F. Bonn, and A. Huete, "A review of vegetation indices," *Remote sensing reviews*, vol. 13, no. 1-2, pp. 95–120, 1995.

[9] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote sensing of Environment*, vol. 8, no. 2, pp. 127–150, 1979.

[10] A. Viña, A. A. Gitelson, A. L. Nguy-Robertson, and Y. Peng, "Comparison of different vegetation indices for the remote assessment of green leaf area index of crops," *Remote sensing of environment*, vol. 115, no. 12, pp. 3468–3478, 2011.

[11] Y. Xu, Y. Yang, X. Chen, and Y. Liu, "Bibliometric analysis of global ndvi research trends from 1985 to 2021," *Remote Sensing*, vol. 14, no. 16, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/16/3967

[12] G. Ortiz-Torres, M. A. Zurita-Gil, J. Y. Rumbo-Morales, F. D. J. Sorcia-Vázquez, J. J. Gascon Avalos, A. F. Pérez-Vidal, M. B. Ramos-Martinez, E. Martínez Pascual, and M. A. Juárez, "Integrating actuator fault-tolerant control and deep-learning-based ndvi estimation for precision agriculture with a hexacopter uav," *AgriEngineering*, vol. 6, no. 3, pp. 2768–2794, 2024. [Online]. Available: https://www.mdpi.com/2624-7402/6/3/161

[13] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[14] D. C. de Lima, D. Saqui, S. A. T. Mpinda, and J. H. Saito, "Pix2pix network to estimate agricultural near infrared images from rgb data," *Canadian Journal of Remote Sensing*, vol. 48, no. 2, pp. 299–315, 2022.

[15] D. Saqui, J. H. Saito, A. D. C. Lucio, E. J. Ferreira, D. C. Lima, and J. P. Herrera, "Methodology for band selection of hyperspectral images using genetic algorithms and gaussian maximum likelihood classifier," in *2016 international conference on computational science and computational intelligence (CSCI)*. IEEE, 2016, pp. 733–738.

[16] D. Saqui, J. H. Saito, D. C. de Lima, L. A. d. C. Jorge, E. J. Ferreira, S. T. Ataky, and F. Fambrini, "Nsga2-based method for band selection for supervised segmentation in hyperspectral imaging," in *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE, 2019, pp. 3580–3585.

[17] D. Saqui, J. H. Saito, D. C. De Lima, L. M. D. V. Cura, and S. T. M. Ataky, "Incorporated decision-maker-based multiobjective band selection for pixel classification of hyperspectral images," *Advances in Electrical and Computer Engineering*, vol. 19, no. 4, pp. 21–28, 2019.

[18] J. R. Koza, "Survey of genetic algorithms and genetic programming," in *Wescon conference record*. Western Periodicals Company, 1995, pp. 589–594.

[19] C. Puente, G. Olague, M. Trabucchi, P. D. Arjona-Villicaña, and C. Soubervielle-Montalvo, "Synthesis of vegetation indices using genetic programming for soil erosion estimation," *Remote Sensing*, vol. 11, no. 2, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/2/156

[20] J. Muñoz, C. Cobos, and M. Mendoza, "Vegetation index based on genetic programming for bare ground detection in the amazon," in *Advances in Computational Intelligence*, F. Castro, S. Miranda-Jiménez, and M. González-Mendoza, Eds. Cham: Springer International Publishing, 2018, pp. 259–271.

[21] A. Illanes, L. Rodríguez-Rolón, S. Esquivel, L. F. Rivera, I. Peña, and C. N. Sánchez, "Generation of hyperspectral indices for non-invasive crop property analysis through genetic programming," in *2024 IEEE 6th International Conference on BioInspired Processing (BIP)*, 2024, pp. 1–6.

[22] H. R. S. Scenes, "Hyperspectral remote sensing scenes," *Hyperspectral_Remote_Sensing_Scenes# Salinas-A_scene*, 2021.

[23] T. Boggs, "Spectral python," *Software. Available from*, 2016.

[24] F.-A. Fortin, F.-M. De Rainville, M.-A. G. Gardner, M. Parizeau, and C. Gagné, "Deap: Evolutionary algorithms made easy," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2171–2175, 2012.