Avaliação de Modelos Clássicos de Aprendizado de Máquina para Classificação de Estágios da Soja

1st Vinícius Tessele
dept. name of organization (of Aff.)
1Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas
São José do Rio Preto-SP. Brazil
https://orcid.org/0000-0002-0662-9200

***Abstract*—Phenological monitoring of soybean is essential for efficient crop management, allowing optimization of input use and increased productivity. This study presents a comparative evaluation of classical machine learning models applied to the automatic classification of soybean phenological stages, based on manual texture and color descriptors. Features were extracted using HOG, LBP, and HSV methods, followed by dimensionality reduction through PCA. The evaluated classifiers included Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Logistic Regression. Experiments using 5-fold cross-validation showed that the Random Forest model achieved the best overall performance, with an average accuracy of 0.875, followed by SVM with 0.900 in validation. The results demonstrate the feasibility of applying classical algorithms to automatic soybean stage identification, providing a foundation for future research involving deep neural networks and computer vision techniques in precision agriculture.**

***Index Terms*—Soybean, Machine Learning, Classification, Random Forest, SVM, Phenology**

I. Introdução

O monitoramento fenológico da soja permite compreender o desenvolvimento da cultura ao longo do seu ciclo de vida, com isso otimizar o manejo agrícola de modo a impactar no custo de produção e na qualidade do produto final. Sendo a soja uma das principais commodities agrícolas do Brasil, com relevância econômica nacional e internacional. O monitoramento da cultura é essencial para garantir a produtividade. A fenologia descreve as fases de crescimento da planta, desde a emergência até a maturação. A identificação precisa dos estágios fenológicos permite o produtor planejar com eficiência o uso de insumos, momento de irrigação e colheita. A identificação automática dos estágios da soja e de outras plantas é um desafio à variabilidade visual das plantas e às condições ambientais em campo.

A agricultura de precisão usa a IA para processar e interpretar dados obtidos de sensores, imagens e sistemas de monitoramento, viabilizando o uso de algoritmos de aprendizado de máquina para minimizar perdas e melhorar a produtividade. Este estudo contribui ao comparar abordagens clássicas de aprendizado supervisionado aplicadas à fenologia da soja, fornecendo uma base para aplicações futuras com redes neurais profundas e visão computacional. O objetivo deste trabalho é avaliar e comparar diferentes modelos clássicos de aprendizado de máquina na classificação automática dos estágios fenológicos da soja, a partir de descritores de textura e cor extraídos de imagens. Assim, este estudo contribui para o desenvolvimento de métodos de classificação fenológica automatizada, oferecendo subsídios para futuras aplicações comerciais. Os Resultados obtidos demonstram que o aprendizado de máquina e visão computacional utilizando modelos clássicos, foram bastante satisfatórios, demonstrando capacidade de distinção entre os estágios fenológicos com abordagem mais simples e de menor custo computacional.

II. Fundamentação Teórica

O monitoramento automatizado na agricultura já se tornou uma realidade, e o uso de técnicas de IA e visão computacional vem sendo amplamente empregado em diversas culturas. Nesse contexto, os métodos de aprendizado de máquina têm se destacado por sua capacidade de processar grandes volumes de dados e extrair informações relevantes para apoio à decisão no campo. A IA, de forma geral, busca desenvolver sistemas capazes de aprender com dados e tomar decisões de forma autônoma. Dentro desse domínio, o Aprendizado de Máquina (Machine Learning) permite que algoritmos identifiquem padrões em dados e realizem previsões com base em exemplos anteriores. Técnicas clássicas, como Support Vector Machines (SVM), Random Forests (RF) e Multilayer Perceptrons (MLP), têm sido utilizadas em tarefas de classificação supervisionada. Mais recentemente, o Aprendizado Profundo (Deep Learning), especialmente com o uso de Redes Neurais Convolucionais (CNNs), tem ampliado significativamente a precisão em aplicações de visão computacional. No entanto, é importante destacar que os métodos clássicos ainda apresentam vantagens relevantes, sobretudo em cenários com conjuntos de dados limitados, devido ao menor custo computacional e à menor necessidade de parametrização. No caso específico da cultura da soja, diversos fatores podem influenciar seu desenvolvimento e produtividade. Entre eles, destacam-se as plantas daninhas, que acarretam perdas significativas devido à competição por luz, nutrientes e água, além de dificultarem a colheita, atuarem como hospedeiras de pragas e doenças e exercerem efeitos alelopáticos sobre a cultura [2].

Para representação numérica das imagens, optou-se pela utilização de descritores manuais de textura e cor, visando

capturar as caracteristicas relevantes para diferenciação dos estágios fenológicos da soja. Histograma de Gradientes Orientados (HOG) foi empregado por sua capacidade de descrever variações locais de bordas e formas, sendo particularmente eficiente na distinção de estruturas geométricas presentes em folhas e vagens. O Local Binary Pattern (LBP) descreve padrões de textura a partir da vizinhança de pixels, sendo robusto a variações de iluminação. Complementarmente, o espaço de cor HSV foi empregado para representar a coloração média das imagens, uma característica importante na diferenciação entre fases vegetativas e reprodutivas da soja.

Como as técnicas de extração de características produzem vetores de alta dimensionalidade, aplicou-se a Análise de Componentes Principais (PCA) para reduzir o número de variáveis e eliminar redundâncias. O PCA projeta os dados em um novo espaço de componentes ortogonais que explicam a maior variância possível, preservando a informação essencial e reduzindo o custo computacional do treinamento.

Entre os modelos de aprendizado supervisionado utilizados, destacam-se:

SVM: busca encontrar o hiperplano que melhor separa as classes no espaço de características. Random Forest: combina múltiplas árvores de decisão para aumentar a robustez e reduzir o sobreajuste. De acordo com [5], a aplicação de algoritmos de Machine Learning como SVM e RF tem se mostrado eficaz na detecção de doenças em plantas de soja, permitindo a distinção entre exemplares saudáveis e infectados por meio de características visuais não invasivas. O estudo demonstrou que o desempenho dos modelos é influenciado pela seleção de atributos, e que a combinação dessas técnicas proporciona maior precisão e robustez no diagnóstico automatizado.

KNN: classifica novas amostras com base na proximidade em relação aos vizinhos mais próximos.

Para [6], a combinação de técnicas de visão computacional e algoritmos de aprendizado supervisionado, como o SVM e o KNN, permite detectar e quantificar com eficiência doenças foliares em plantas de soja. No estudo, imagens foram segmentadas e analisadas para extrair características de cor e textura, e posteriormente classificadas em folhas sadias e infectadas. Os resultados mostraram que o SVM apresentou melhor desempenho (87,3%) em comparação ao KNN (83,6%), evidenciando maior capacidade de separação entre classes e menor confusão espectral.

MLP: rede neural com múltiplas camadas densas, capaz de aprender relações não lineares entre as variáveis. No contexto agrícola, Conforme apresentado em [3] o MLP foi aplicado em conjunto com outros algoritmos, como Random Forest (RF) e Redes Neurais Convolucionais Temporais (1D-TempCNN), para o mapeamento do uso e cobertura da terra (LULC) a partir de séries temporais de imagens de satélite. O estudo demonstrou que o MLP apresentou desempenho competitivo, com exatidão global superior a 95%, embora a arquitetura 1D-TempCNN tenha apresentado melhor capacidade de generalização entre diferentes áreas agrícolas.

Segundo [4], algoritmos como SVM, RF e Redes Neurais apresentam desempenhos distintos conforme as características dos dados. Enquanto SVM e RF mostram maior robustez ao ruído e boa capacidade de generalização, as Redes Neurais exigem maior tempo de treinamento e ajuste de parâmetros, o que pode limitar seu uso prático. O estudo conclui que não há um classificador universalmente superior, pois a eficácia de cada modelo depende do tipo e da complexidade dos dados analisados.

Estudos recentes demonstram que modelos baseados em Redes Neurais Convolucionais são capazes de identificar com elevada precisão os estágios fenológicos da soja, como o estágio R7, auxiliando na determinação do momento mais adequado para a dessecação. Essa etapa é essencial para otimizar a colheita, melhorar a eficiência e a qualidade das sementes, bem como reduzir perdas e custos operacionais [1]. Segundo Silva et al. [8], o uso de índices espectrais provenientes de diferentes resoluções espaciais pode auxiliar na identificação precisa dos estágios fenológicos da soja, permitindo um monitoramento mais eficiente do desenvolvimento da cultura. Conforme demonstrado por Miranda et al. [7], a aplicação de plataformas de imagem aérea acopladas a índices de vegetação e técnicas de aprendizado de máquina representa uma abordagem promissora para fenotipagem de soja em alto rendimento. McCormick et al. [9] demonstraram que a combinação de modelos baseados em conhecimento e modelos de aprendizado de máquina, aplicada a um conjunto de dados intercontinental, permitiu reduzir o erro médio absoluto na predição da fenologia da soja para cerca de 4,41 e 5,27 dias nos estágios R1 e R7, respectivamente. Por sua vez, Rigalli et al. [10] mostraram que a detecção quase em tempo real dos estágios fenológicos da soja pode ser alcançada por meio da reflectância hiperespectral proximal associada a algoritmos de aprendizado de máquina.

III. Materiais e Métodos

Pesquisa caracterizada como pesquisa aplicada de natureza experimental. Esta seção descreve as etapas metodológicas para o desenvolvimento do sistema de classificação automática de estágios fenológicos da soja. O conjunto de dados (dataset) foi construido a partir de imagens capturadas com câmera de smartphone em lavouras de soja localizadas no município de Toledo, Paraná (Brasil), durante a safra 2024/2025. As imagens foram obtidas em condições de campo, sob diferentes intesidades de iluminação, ângulo e tipos de solo. O processo do desenvolvimento foi dividido em cinco fases.

1) construção da base de imagens.
2) pré-processamento das imagens.
3) extração de caracteristicas.
4) treinamento dos modelos de aprendizado de máquina.
5) avaliação de desempenho.

Cada uma dessas etapas é detalhada a seguir.

- **Base de dados**
  - Total: 2068 imagens
  - 6 classes correspondendo aos estágios: R1_R2 (437 imagens), R5_R6 (256), R7_R8 (421), V1_V2 (437), V3_V4 (256) e VE_VC (261).

- **Pré-processamento**
  - O conjunto de imagens passou um pré-processamento, para padronizar e reduzir ruídos e uniformizar as entradas.
  - Redimensionamento: 128×128 px
  - Conversão para escalas HSV e tons de cinza
  - Extração de HOG, LBP e média HSV
- **Extração de Características**
  - Dimensão final do vetor: 8121 features
  - Normalização: `StandardScaler`
  - Redução de dimensionalidade com PCA (100 componentes)
- **Modelos Avaliados**
  - SVM (kernel RBF)
  - Random Forest (200 árvores)
  - KNN (k=5)
  - Regressão Logística
  - MLP (camadas 128–64, 500 iterações)
- **Avaliação**
  - Divisão 80/20 treino-teste
  - Validação cruzada 5-fold
  - Métricas: *accuracy*, *precision*, *recall*, *f1-score* e *matriz de confusão*.

O experimento foi realizado na plataforma Google Colab, utilizando o ambiente Python 3 com suporte a GPU.

IV. Resultados e Discussão

Os experimentos foram conduzidos utilizando os cinco modelos de aprendizado do máquina descrito nos materiais e métodos.

TABLE I
Desempenho dos Modelos de Aprendizado de Máquina

Modelo	Acurácia (Teste)	Acurácia (Validação 5-Fold)
Random Forest	**0.874**	**0.875**
SVM (RBF)	0.843	0.900
MLP Neural Network	0.833	0.840
Regressão Logística	0.800	0.768
KNN (k = 5)	0.662	0.709

A Tabela 1 apresenta as médias das acurácias obtidas na fase de teste e na validação cruzada (5-fold). Observa-se que a RF apresentou o melhor despenho geral, com acurácia de 0,874 bi teste e 0,875 na validação cruzada, seguido pelo SVM com kernel RBF, que alcançou 0,843 e 0,900, respectivamente.

Conforme mostra a Tabela 2, o modelo RF alcançou valores elevados de precisão e revocação para as classes R1_R2 e R7_R8, o que demonstra excelente capacidade de identificar os estágios reprodutivos da soja. No entanto, observou-se desempenho inferior nas classes V3_V4 e VE_VC, cujos valores de F1-score foram 0,75 e 0,67, respectivamente. Essa diferença pode estar associada à semelhança visual entre os estágios vegetativos e às variações de iluminação nas imagens capturadas. De modo geral, o RF demonstrou robustez frente

TABLE II
Relatório de Classificação - Random Forest

Classe	Precisão	Revocação	F1-Score	Amostras
R1_R2	0.89	0.98	0.93	88
R5_R6	1.00	0.92	0.96	51
R7_R8	0.92	1.00	0.96	84
V1_V2	0.78	0.94	0.85	88
V3_V4	0.97	0.61	0.75	51
VE_VC	0.78	0.60	0.67	52
Acurácia Global	–	–	**0.87**	414

à variabilidade dos dados, sendo o modelo mais indicado para a classificação fenológica automática da soja neste estudo.

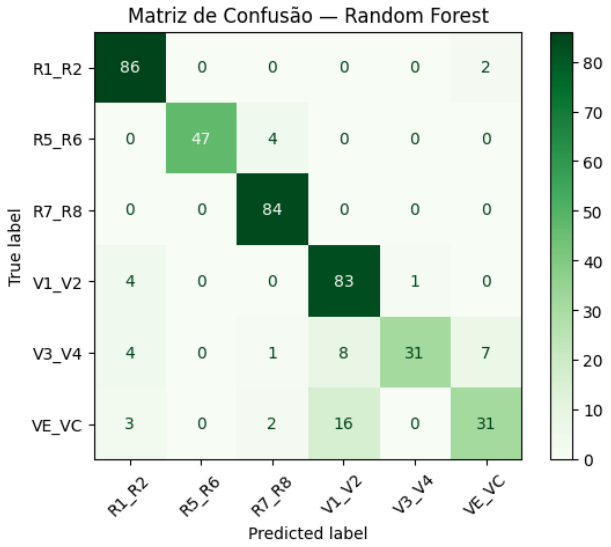

Fig. 1. Matriz de confusão do modelo RF para a classificação dos estágios fenológicos da soja.

A Figura 1 apresenta a matriz de confusão referente ao modelo RF, evidenciando seu desempenho na classificação. Observa-se que as classes R1_R2, R5_R6, R7_R8 e V1_V2 obtiveram altas taxas de acerto, com a maioria das amostras corretamente classificadas em suas respectivas categorias. Em contrapartida, verificam-se confusões significativas entre as classes V3_V4 e VE_VC, nas quais algumas amostras foram incorretamente atribuídas a estágios vizinhos. Esse comportamento pode ser explicado pela semelhança morfológica entre essas fases vegetativas e pelas variações de textura e iluminação nas imagens originais. Assim, a matriz de confusão confirma a consistência do modelo RF, reforçando os resultados apresentados na Tabela 2.

A Figura 2 apresenta amostras representativas de cada classe fenológica utilizada no conjunto de dados da soja. As imagens evidenciam a variação morfológica e cromática entre os estágios de desenvolvimento: desde a emergência das plântulas (VE_VC), passando pelas fases vegetativas (V1_V2, V3_V4), até os estágios reprodutivos iniciais (R1_R2) e avançados (R5_R6, R7_R8). Observa-se a evolução progressiva da densidade foliar, do fechamento do dossel e da coloração das

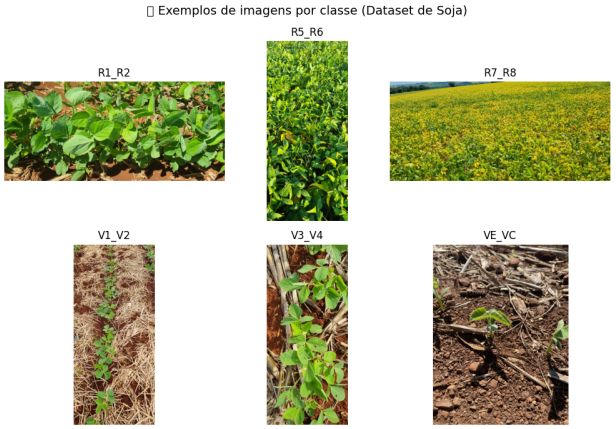

Fig. 2. Exemplos de imagens por classe.

folhas, aspectos que justificam o uso de descritores de textura e cor para caracterização fenológica. Esse conjunto de imagens constitui a base visual para o treinamento e a validação dos modelos de aprendizado de máquina aplicados neste estudo.

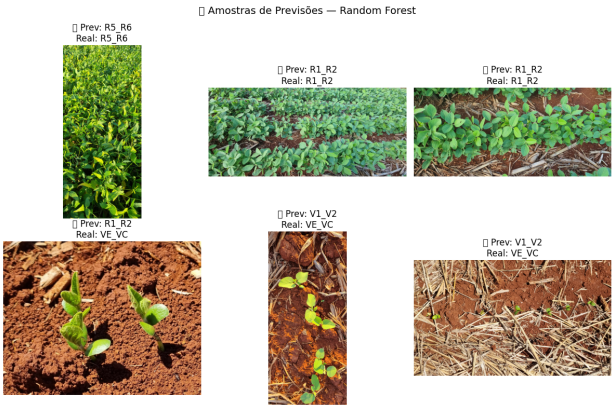

Fig. 3. Amostras de Previsões - Random Forest.

A Figura 3 apresenta exemplos de previsões realizadas pelo modelo RF em imagens reais de soja coletadas em campo. É possível observar que o classificador obteve acertos nos estágios intermediários, como R1_R2 e R5_R6, demonstrando boa capacidade de generalização para fases vegetativas e reprodutivas bem definidas. Entretanto, ocorreram erros nas fases iniciais (VE_VC), onde a confusão entre os estágios V1_V2 e VE_VC indica que as diferenças morfológicas sutis das plântulas recém-emergidas dificultam a distinção automática. Esses resultados ilustram tanto o potencial quanto os desafios do uso de algoritmos clássicos de aprendizado supervisionado em condições de campo, onde variações de iluminação e textura do solo afetam a classificação.

REFERENCES

[1] V. Tessele, M. B. Oliveira Júnior, and J. E. Schulz, "Treinamento de redes neurais artificiais para a identificação do momento de dessecação da soja," *Revista Caribeña de Ciencias Sociales*, vol. 13, no. 1, 2024. Available: https://doi.org/10.55905/rcssv13n9-005.

[2] A. C. S. Lima, J. S. Teixeira, F. A. Araújo, G. A. B. Alves, M. S. Melo, D. L. Souza, G. N. Souza Júnior, and M. B. Braga, "Detecção e mapeamento de vegetação daninha em áreas de cultivo de soja por meio de Inteligência Artificial e Sensoriamento Remoto no Sudeste Paraense," *Observatorio de la Economía Latinoamericana*, 2024. Available: https://doi.org/10.55905/oelv22n3-009.

[3] J. A. F. Vieira Netto, *Uso de Inteligência Computacional na Fenotipagem de Soja*, Universidade Federal de Viçosa (UFV), Viçosa, MG, 2024. Available: https://locus.ufv.br/server/api/core/bitstreams/357afd98-c720-4784-ba4f-519c9d6098d0/content.

[4] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, pp. 341–367, 2020. Available: https://doi.org/10.4236/jdaip.2020.84020.

[5] Z. Zainab, I. A. J. Al Ali, B. M. H. Mustafa, G. K. Mustafa, and R. Jaleel, "An intelligent model to combat soybean plant disease based on Random Forest and Support Vector Machine algorithms," *Fusion: Practice and Applications*, 2025. Available: https://doi.org/10.54216/fpa.2025.05020.

[6] S. Jadhav, V. R. Udup, and S. B. Patil, "Soybean leaf disease detection and severity measurement using multiclass SVM and KNN classifier," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 4077–4091, 2019. Available: https://doi.org/10.11591/ijece.v9i5.pp4077-4091.

[7] M. C. de C. Miranda, A. H. Aono, T. G. Fagundes, et al., "High-throughput phenotyping and machine learning techniques in soybean breeding: Exploring the potential of aerial imaging and vegetation indices," Agronomy Journal, vol. 117, e70012, 2025. [Online]. Available: https://doi.org/10.1002/agj2.70012

[8] A. A. da Silva, F. C. dos S. Silva, C. M. Guimarães, I. A. Saleh, J. F. da Cruz Neto, M. A. El-Tayeb, M. A. Abdel-Maksoud, J. G. Aguilera, H. AbdElgawad, and A. M. Zuffo, "Spectral indices with different spatial resolutions in recognizing soybean phenology," PLOS ONE, vol. 19, no. 9, Sep. 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0305610

[9] R. F. McCormick, S. K. Truong, J. Rotundo, A. P. Gaspar, D. Kyle, F. van Eeuwijk, and C. D. Messina, "Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models," in silico Plants, vol. 3, no. 1, p. diab004, Mar. 2021. [Online]. Available: https://doi.org/10.1093/insilicoplants/diab004

[10] N. F. Rigalli, E. Montero Bulacio, M. Romagnoli, J. M. Enrico, ... et al. , "Near real-time soybean phenology detection using proximally sensed hyperspectral canopy reflectance and machine learning methods," International Journal of Remote Sensing, vol. 46, no. 3, pp. 1-24, Apr. 2025. [Online]. Available: https://doi.org/10.1080/01431161.2025.2487228