

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE ENGENHARIA
CAMPUS DE ILHA SOLTEIRA

REGIMAR GARCIA DOS SANTOS

**CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO OS SEUS ATRIBUTOS
FISIOLÓGICOS USANDO APRENDIZAGEM DE MÁQUINA E DIFERENTES
SENSORES ESPECTRAIS**

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

REGIMAR GARCIA DOS SANTOS

CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO OS SEUS ATRIBUTOS FISIOLÓGICOS USANDO APRENDIZAGEM DE MÁQUINA E DIFERENTES SENSORES ESPECTRAIS

Tese apresentada à Faculdade de Engenharia de Ilha Solteira – Unesp como parte dos requisitos para obtenção do título de Doutor. Especialidade: Sistemas de Produção.

Paulo Eduardo Teodoro
Orientador

Larissa Pereira Ribeiro Teodoro
Coorientadora

Ilha Solteira
2025

FICHA CATALOGRÁFICA

Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

S237c Santos, Regimar Garcia dos.
Classificação de genótipos de soja quanto os seus atributos fisiológicos usando aprendizagem de máquina e diferentes sensores espectrais / Regimar Garcia dos Santos. -- Ilha Solteira: [s.n.], 2025
71 f. : il.

Tese (Doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Engenharia, Ilha Solteira. Área de conhecimento: Sistemas de Produção, 2025

Orientador: Paulo Eduardo Teodoro

Coorientador: Larissa Pereira Ribeiro Teodoro

Inclui bibliografia

1. Fenotipagem fisiológica. 2. Aprendizagem de máquina. 3. Sensores hiperspectrais. 4. Classificação de genótipos.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: Classificação de genótipos de soja quanto os seus atributos fisiológicos usando aprendizagem de máquina e diferentes sensores espectrais

AUTOR: REGIMAR GARCIA DOS SANTOS

ORIENTADOR: PAULO EDUARDO TEODORO

COORDINADORA: LARISSA PEREIRA RIBEIRO TEODORO

Aprovado como parte das exigências para obtenção do Título de Doutor em Agronomia, área: Sistemas de Produção pela Comissão Examinadora:

Documento assinado digitalmente
 **PAULO EDUARDO TEODORO**
Data: 03/11/2025 17:07:57-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. PAULO EDUARDO TEODORO (Participação Virtual)
Campus de Chapadão do Sul / Universidade Federal de Mato Grosso do Sul - UFMS

Prof. Dr. FABIO HENRIQUE ROJO BAIO (Participação Virtual)
Departamento de Agronomia / Universidade Federal de Mato Grosso do Sul - UFMS

Documento assinado digitalmente
 **ALAN MARIO ZUFFO**
Data: 03/11/2025 17:47:38-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. ALAN MARIO ZUFFO (Participação Virtual)
Centro de Estudos Superiores de Balsas / Universidade Estadual do Maranhão - UEMA

Documento assinado digitalmente
 **JORGE GONZALEZ AGUILERA**
Data: 03/11/2025 17:58:00-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. JORGE GONZÁLEZ AGUILERA (Participação Virtual)
Departamento de Agronomia / Universidade Estadual do Mato Grosso do Sul - UEMS

Documento assinado digitalmente
 **CARLOS ANTONIO DA SILVA JUNIOR**
Data: 03/11/2025 17:14:23-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. CARLOS ANTONIO DA SILVA JUNIOR (Participação Virtual)
Departamento de Geografia / Universidade do Estado de Mato Grosso - UNEMAT

Ilha Solteira, 03 de novembro de 2025.

DEDICATÓRIA

Dedico este trabalho aos meus pais, Marli Batista dos Santos e Lindomar Garcia dos Santos, pilares da minha formação e exemplos de dedicação e perseverança. À minha mãe, Marli, deixo uma gratidão especial, pois sua força, amor incondicional e incansável apoio sempre me guiaram nos momentos de maior desafio. Estendo ainda esta dedicação ao meu orientador, Prof. Dr. Paulo Eduardo Teodoro, cuja orientação firme, paciência e ensinamentos foram essenciais para a realização deste trabalho. Que esta conquista também represente um tributo àqueles que veem na educação a mais poderosa ferramenta de mudança.

"A educação é o passaporte para o futuro, pois o
amanhã pertence àqueles que se preparam hoje."

Malcolm X

AGRADECIMENTOS

Agradeço, em primeiro lugar, a Deus, fonte de inspiração, força e esperança em todos os momentos desta caminhada.

Estendo minha gratidão aos meus pais, Marli Batista dos Santos e Lindomar Garcia dos Santos, que sempre foram alicerce e exemplo de dedicação. Seria falha qualquer tentativa de expressar de forma justa o quanto sou grato à minha mãe. Em nenhuma situação, problema ou dificuldade que enfrentei deixou de sorrir e, carinhosamente, me dizer: “Você nasceu para vencer, filho meu. Eu acredito e tenho muito orgulho de você.” Palavras que renovaram minhas forças e me sustentaram até este momento.

Às minhas irmãs, Raquel Benta dos Santos e Regina Garcia dos Santos, sou igualmente grato pelo apoio constante, por me ouvirem falar de temas que, por vezes, fugiam ao seu entendimento, mas que ainda assim escutavam com paciência, demonstrando interesse e entusiasmo pelo meu trabalho. Agradeço também pela oportunidade de me deixarem provar o sentimento de ser quase como um pai para meus sobrinhos.

À minha esposa, Isabela Camila Porfiro de Almeida, expresso meu reconhecimento profundo. Obrigado por todos os dias que enfrentamos a distância e pela infinita paciência nos momentos em que estive ausente, mesmo estando fisicamente ao seu lado, imerso em leituras, trabalhos e estudos. Sinto-me abençoado por dividir contigo o peso do mundo e por multiplicar as alegrias e conquistas. Sobretudo, sou grato por ter a oportunidade de todos os dias, te escolher e ser também a tua escolha.

Desejo ainda expressar minha gratidão à Dra. Dthenifer Cordeiro Santana, que, além de possuir um carisma ímpar, humor inabalável e ser uma companhia sempre agradável, é exemplo de pesquisadora séria e dedicada. Sua amizade e parceria me ajudaram a crescer como pessoa e como pesquisador. Trabalhar ao seu lado me proporcionou lições valiosas que levo diariamente em minha jornada acadêmica e profissional.

Sou profundamente grato também aos colegas e amigos do grupo GEMP, que contribuíram de maneira significativa para o desenvolvimento deste trabalho. Registro especial apreço a Izabela Cristina de Oliveira, Elber Vinicius Martins Silva e Paulo Henrique Menezes das Chagas, pela amizade e pela parceria ao longo desta trajetória.

Vai ao meu orientador, Professor Dr. Paulo Eduardo Teodoro, minha mais sincera gratidão. Sua dedicação, confiança e ensinamentos foram decisivos para minha formação acadêmica e pessoal. Ter a oportunidade de contar com sua orientação não apenas acelerou minha graduação e fortaleceu minha trajetória no mestrado, como também foi determinante para a decisão de seguir o doutorado. Sua postura, que alia amizade fora do ambiente acadêmico à clareza em separar os momentos de descontração das responsabilidades profissionais, ensinou-me que a verdadeira grandeza de um educador está em reconhecer e tratar seus alunos com respeito, afeto e humanidade.

Estendo igualmente minha gratidão à Professora Dra. Larissa Pereira Teodoro, cuja dedicação e exemplo serviram de inspiração desde o início de minha caminhada científica. Sua atuação como orientadora no mestrado e coorientadora no doutorado foi fundamental não apenas para o desenvolvimento do trabalho, mas também para o meu amadurecimento acadêmico, influenciando de forma marcante a decisão de prosseguir no doutorado. Sua seriedade, ética e generosidade constituem referências que levarei por toda a vida.

Por fim, este trabalho só foi possível graças ao apoio concedido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), por meio do Código de Financiamento 001. Agradeço à CAPES pelo suporte financeiro que tornou viável a realização deste projeto acadêmico.

RESUMO

A fenotipagem de alta precisão associada a algoritmos de aprendizagem de máquina torna possível explorar com maior eficiência a variabilidade genética da soja. Ao reduzir o tempo e a subjetividade nas avaliações, essa abordagem acelera a seleção de genótipos superiores e potencializa os programas de melhoramento. No capítulo 1, foram classificados 32 genótipos F₃ de soja com base em variáveis fisiológicas utilizando sensor entre os comprimentos de onda VIS-NIR (400~825 nm). Os dados de reflectância foram agrupados em 20 bandas representativas e foram medidas fotossíntese líquida, concentração interna de CO₂, condutância estomática e transpiração. Os genótipos foram agrupados por k-means em dois grupos chamados de cluster C1 e C2, e esses grupos foram usados como variável de saída nos modelos de aprendizagem de máquina. Os algoritmos foram avaliados utilizando como entrada os comprimentos de onda contínuo e médias de bandas, e avaliados por porcentagem de classificação correta e F-score. Os resultados indicam que o conjunto de 32 genótipos dividiram-se em dois clusters com 20 genótipos e outro com 12; C2 apresentou maiores médias de condutância estomática, concentração de CO₂ e transpiração. Para fotossíntese líquida não houve diferença estatística significativa entre os grupos, ainda que a média para essa variável tenha sido superior no C1. Na classificação, os algoritmos máquina de vetor suporte (SVM) (63%), e regressão logística (RL) (66%), obtiveram maior percentual de classificações corretas (CC), quando se utilizou o espectro completo; já o algoritmo J48 (63%), destacou-se com as médias das bandas, apresentando melhor performance com dados agregados, indicando que a escolha do tipo de entrada influencia a eficiência dos algoritmos. O capítulo 2 avaliou 32 populações F₃ de soja com espectrorradiômetro em 2100 bandas espectrais (350~2500 nm). As medidas dos caracteres fisiológicos foram feitas aos 60 dias após a emergência e as mesmas folhas foram avaliadas utilizando o sensor espectral. Os dados foram analisados tanto na forma de espectro contínuo quanto como bandas agregadas. O cluster C1 apresentou maior fotossíntese e eficiência no uso da água, enquanto C2 exibiu maior condutância estomática e transpiração. Na classificação, o espectro contínuo obteve desempenho superior às bandas agregadas. Os algoritmos de aprendizado de máquinas J48 e REPTree alcançaram as maiores CC e F-scores, seguidos por SVM e redes neurais; Random Forest e regressão logística apresentaram menor desempenho.

Palavras-Chave: Fenotipagem fisiológica; Aprendizagem de máquina; Sensores hiperespectrais; Classificação de genótipos.

ABSTRACT

High-precision phenotyping combined with machine learning algorithms enables a more efficient exploration of soybean genetic variability. By reducing the time and subjectivity of evaluations, this approach accelerates the selection of superior genotypes and enhances breeding programs. Chapter 1 presents a critical review of the literature, situating the research topic within the current state of the art and establishing the theoretical foundations that underpin this dissertation. In chapter 2, 32 soybean genotypes were classified based on physiological variables using a VIS–NIR sensor (400~825 nm). The reflectance data were grouped into 20 representative bands, and measurements included net photosynthesis, internal CO₂ concentration, stomatal conductance, and transpiration. Genotypes were clustered by k-means into two groups, which were then used as output variables in the machine-learning models. The algorithms were tested with continuous wavelength inputs and band averages and evaluated using the percentage of correct classifications and F-score. Results indicated that the set of 32 genotypes split into one cluster with 20 genotypes and another with 12; cluster 2 exhibited higher mean stomatal conductance, CO₂ concentration and transpiration, whereas cluster 1 showed slightly higher photosynthesis. In classification, support vector machine (SVM) and logistic regression achieved higher accuracy when the full spectrum was used; the J48 algorithm performed best with the band averages. The superior performance of J48 with aggregated data indicates that the choice of input type influences algorithm efficiency. Chapter 3 evaluated 32 F₃ soybean populations with a hyperspectral spectroradiometer (350~2500 nm). Physiological traits were measured 60 days after emergence, and the same leaves were evaluated spectrally. The data were analyzed both as continuous spectra and as aggregated bands. Cluster 1 showed higher photosynthesis and water-use efficiency, whereas cluster 2 displayed higher stomatal conductance and transpiration. In classification, the continuous spectrum outperformed the aggregated bands. J48 and REPTree achieved the highest accuracies and F-scores, followed by SVM and neural networks; Random Forest and logistic regression exhibited lower performance.

Keywords: physiological phenotyping; machine learning; hyperspectral sensors; genotype classification.

SUMÁRIO

1. INTRODUÇÃO GERAL	8
2. REVISÃO BIBLIOGRÁFICA	10
2.1. Índices Espectrais Aplicados À Fenotipagem	10
2.2. Aplicação do Espectrorradiômetro na Fenotipagem Vegetal	12
2.3. Agrupamento De Genótipos Com Base Em Atributos Fisiológicos E Espectrais	13
2.4. Características Fisiológicas E Seus Papéis Na Fenotipagem Espectral	14
2.5. Relação Entre Reflectância Foliar E Respostas Fisiológicas	16
2.6. Modelos De Classificação	17
2.7. Modelos De Regressão	18
2.8. Redução De Dimensionalidade E Validação	19
Referências	19
CAPÍTULO 1 - CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO AO DESEMPENHO FISIOLÓGICO USANDO SENSOR VIS-NIR	24
RESUMO	24
ABSTRACT	25
1. INTRODUÇÃO	26
2. MATERIAL E MÉTODOS	28
3. RESULTADOS	32
4. DISCUSSÃO	39
5. CONCLUSÃO	41
REFERÊNCIAS	42
CAPÍTULO 2 - USO DE SENSOR VIS-NIR-SWIR PARA CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO AO SEU DESEMPENHO FISIOLÓGICO	48
Resumo	48
ABSTRACT	49
1. INTRODUÇÃO	50
2. MATERIAL E MÉTODOS	51
3. RESULTADOS	55
4. DISCUSSÃO	61
5. CONCLUSÃO	62
REFERÊNCIAS	63

1. INTRODUÇÃO GERAL

A espectroscopia tem se destacado como uma ferramenta promissora na agricultura de precisão, além de se destacar em outras tarefas que apoiam a agricultura, especialmente na fenotipagem de plantas. Essa tecnologia permite a captura de centenas de bandas estreitas e contínuas no espectro eletromagnético, situadas sobretudo no óptico, entre 350 e 2500 nm, possibilitando a detecção de variações sutis em pigmentos, estrutura foliar e conteúdo hídrico das plantas (SARIĆ et al., 2022; FALCIONI et al., 2020). Diferentemente dos sensores multiespectrais, que operam com bandas fixas e limitadas, os sensores hiperespectrais oferecem flexibilidade na seleção de comprimentos de onda, aumentando a precisão na análise de atributos fisiológicos (TEODORO et al., 2024; ZHANG et al., 2023).

O amplo espectro obtido por esses sensores permite identificar padrões bioquímicos e estruturais com elevado grau de detalhamento já que cada banda espectral responde a componentes específicos, como clorofila, carotenoides, água ou lignina, enquanto a imagem fornece informações sobre textura e morfologia foliar (ZHANG et al., 2023; WANG et al., 2023). Essa capacidade de discriminação espectral tem sido aplicada com sucesso na classificação de sementes, detecção de doenças e avaliação de estresses abióticos, com acurácias superiores a 90 % em culturas como soja (*Glycine max* (L.) Merrill), milho (*Zea mays* L.), e amendoim (*Arachis hypogaea* L.), (JEONG et al., 2024; GUI et al., 2021).

A integração entre espectroscopia hiperespectral, fisiologia vegetal e algoritmos de aprendizado de máquina tem ampliado as possibilidades de caracterização não destrutiva de genótipos em diferentes ambientes. O uso de espectrorradiômetros portáteis, permite a coleta de espectros foliares em campo sem danificar as amostras e com alta sensibilidade a variações em fotossíntese líquida (A), condutância estomática (gs), concentração interna de CO₂ (Ci) e eficiência no uso da água (WUE) (ADEGOYE et al., 2023; FALCIONI et al., 2020; PENG et al., 2019).

Esses dados espectrais, quando processados por algoritmos de classificação ou regressão, têm demonstrado capacidade de identificar genótipos superiores, contribuindo para programas de melhoramento genético e estratégias de manejo mais eficientes (LIAKOS et al., 2018; YOOSEFZADEH-NAJAFABADI et al., 2021; CORRALES et al., 2022). A combinação entre sensores de alta resolução espectral e modelos computacionais avançados representa um avanço significativo na fenotipagem digital, permitindo a seleção de cultivares mais produtivos e resilientes com maior rapidez e precisão (SANTANA et al., 2022; TEODORO et al., 2024).

Esses avanços reforçam o papel da espectroscopia como ferramenta essencial na fenotipagem de plantas em larga escala. Nesse contexto, esta tese teve como objetivo geral avaliar o potencial de diferentes sensores espectrais, associados a algoritmos de aprendizagem de máquina, para a classificação de genótipos de soja com base em atributos fisiológicos. Como objetivos específicos, buscou-se: (I) investigar a capacidade de sensores VIS-NIR e hiperespectrais em discriminar genótipos com diferentes padrões fisiológicos; (II) comparar o desempenho de algoritmos supervisionados de classificação e regressão na identificação de genótipos de acordo com sua resposta fisiológica e (III) avaliar o impacto do tipo de entrada dos dados comprimentos de onda contínuos e bandas espectrais médias sobre a acurácia dos modelos.

2. REVISÃO BIBLIOGRÁFICA

2.1. Índices Espectrais Aplicados À Fenotipagem

Os índices espectrais são construídos a partir de combinações matemáticas da reflectância em bandas específicas do espectro eletromagnético. Eles permitem inferir, de forma indireta e não destrutiva, características fisiológicas das plantas. Esses índices têm sido amplamente empregados para estimar variáveis como biomassa, teor de clorofila, conteúdo hídrico, área foliar e vigor, tornando-se ferramentas essenciais na fenotipagem de alto rendimento (LIU et al., 2020; FU et al., 2022).

Entre eles, destaca-se o Índice de Vegetação por Diferença Normalizada (NDVI), calculado como $(\text{NIR}-R)/(\text{NIR}+R)$, amplamente utilizado para estimar biomassa e cobertura vegetal. Embora o artigo clássico de Rouse et al. (1974) tenha estabelecido sua formulação, estudos recentes confirmam que o NDVI continua sendo um indicador robusto de área foliar e produtividade em culturas como milho e soja (FLORES et al., 2020; TEODORO et al., 2024; MOTA et al., 2024). Outra variação é o Green NDVI (GNDVI), que substitui a banda vermelha pela verde, aumentando a sensibilidade ao teor de clorofila e reduzindo a influência do solo. Índices baseados na banda verde têm se mostrado mais eficazes em culturas densas, como soja e canola, por evitarem a saturação em condições de elevada biomassa (SANTANA et al., 2022).

Já o Índice de Diferença Normalizada da Borda do Vermelho (RedEdge, NDRE), definido por $(R790 - R720)/(R790 + R720)$, utiliza a região da borda do vermelho para capturar variações em clorofila e nitrogênio foliar. Este índice tem mostrado elevada correlação com conteúdo de clorofila em estágios avançados de desenvolvimento e vem substituindo citações anteriores de Tucker et al. (2005), (LI et al., 2021; YANG et al., 2021). A reflectância fotossintética pode ser estimada pelo Photochemical Reflectance Index (PRI), calculado como $(R531 - R570)/(R531 + R570)$. Tradicionalmente descrito por Gamon et al. (1992), esse índice foi revisado com sucesso por Sukhova et al. (2022), que demonstraram modificações do PRI mais sensíveis a estresses por seca e calor. Outros estudos relatam a utilidade do PRI para monitorar a dissipação de energia e a eficiência do uso da luz em diversas culturas (SUKHOVA et al., 2022; LIU et al., 2020b).

Quando se busca estimar o conteúdo hídrico, é indicada a utilização de métodos como o Índice de Diferença Normalizada da Água clássico (NDWI), o Índice de Diferença Normalizada

da Água hiperespectral (NDWI-Hyp) e o Índice da Banda da Água (WBI), que exploram absorções na região do infravermelho próximo (NIR) e do infravermelho de ondas curtas (SWIR). Estudos recentes substituem referências antigas (GAO, 1996; SERRANO et al., 2000) e mostram que a combinação de bandas do NIR e do SWIR permite detectar déficit hídrico de forma precoce em soja, colza e videira (AHMAD; ALVINO, 2021; CRUSIOL et al., 2021a; PANG et al., 2020; WEN et al., 2023). Estudos comparativos demonstram que a escolha do índice influencia diretamente a sensibilidade às variações fisiológicas. Em soja e amendoim, índices baseados no red-edge, como o Índice de Diferença Normalizada do Red-Edge (NDRE), apresentaram correlações mais fortes com clorofila e nitrogênio foliar do que o Índice de Vegetação por Diferença Normalizada (NDVI) convencional.

O Índice de Vegetação por Diferença Normalizada no Verde (GNDVI) mostrou maior sensibilidade ao teor de clorofila e menor saturação em áreas com alta densidade de cobertura vegetal. No monitoramento de estresse hídrico, índices que combinam bandas do NIR e do SWIR, como NDWI, WBI e índice de estresse por umidade (MSI), têm se mostrado eficazes na detecção precoce de déficit hídrico, antes mesmo da manifestação de sintomas visuais (QUEMADA et al., 2021; PANG et al., 2020). Em colza (*Brassica napus L.*), por exemplo, o NDWI apresentou forte correlação com o teor de água foliar, permitindo intervenções agronômicas mais precisas (VILLACRÉS et al., 2021; CHANDRASEKAR et al., 2024).

A acurácia dos índices também depende da configuração espectral dos sensores utilizados. Estudos comparativos demonstram que a presença de bandas adicionais no rededge (705 e 740 nm) é essencial para estimativas precisas de clorofila; quando essas bandas são removidas, a performance dos modelos de predição diminui significativamente (DA SILVA JUNIOR et al., 2018; SINGH et al., 2025). Além disso, a combinação de múltiplos índices em modelos multivariados tem potencial para melhorar a correlação com variáveis fisiológicas complexas, como fotossíntese líquida (A), condutância estomática (gs) e eficiência no uso da água (WUE), ampliando a capacidade de discriminação entre genótipos (FENG et al., 2024; CRUSIOL et al., 2021). Portanto, os índices representam ferramentas versáteis e eficazes para inferência de atributos fisiológicos em oleaginosas, contribuindo para a seleção de genótipos mais eficientes e adaptados a diferentes condições ambientais.

2.2. Aplicação do Espectrorradiômetro na Fenotipagem Vegetal

O espectrorradiômetro é amplamente utilizado em estudos agrônômicos por sua capacidade de captar reflectância espectral em alta resolução, cobrindo a faixa de 350 a 2500 nm com mais de 2000 bandas espectrais (ASD Inc., 2010). O equipamento é composto por uma sonda de contato acoplada a uma fonte de luz interna, o que elimina interferências atmosféricas e permite medições rápidas e precisas tanto em laboratório quanto em campo (CHESHKOVA., 2022). A coleta de dados com esses sensores segue protocolos padronizados, incluindo calibração com painel branco de referência, repetição de leituras em diferentes posições da folha e controle do ângulo de observação e da distância entre a sonda e a amostra (FAN et al., 2019). Essas práticas garantem a reprodutibilidade dos espectros e a comparabilidade entre amostras, mesmo sob condições ambientais variáveis.

Diversos estudos demonstram a aplicabilidade dos dados coletados com esses sensores na fenotipagem de diversas culturas. Em soja e amendoim, por exemplo, espectros foliares obtidos com o equipamento foram utilizados para estimar parâmetros taxa máxima de carboxilação da Rubisco ($V_{c,max}$) e a taxa máxima de transporte de elétrons (J_{max}), por meio de modelos de regressão multivariada, com coeficientes de determinação entre 0,50 e 0,70 (KAKATI et al., 2022; CRUSIOL et al., 2021). Esses resultados indicam que a variabilidade genética entre genótipos contribui para a precisão dos modelos espectrais. Em estudos de qualidade de sementes de soja armazenadas, espectros obtidos com espectrorradiômetro foram submetidos a pré-processamentos como derivadas de primeira ordem (FD), correção multiplicativa de espalhamento (MSC) e normalização por variância (SNV), além de técnicas de seleção de variáveis como Abordagem Iterativa de Redução de Espaço de Variáveis (VISSA) e Amostragem Competitiva Adaptativa Reponderada (CARS).

Esses dados alimentaram modelos de aprendizado de máquina como Máquina de Vetores de Suporte (SVM) e outros algoritmos com diferentes abordagens, alcançando resultados promissores na predição de ganho genético da cultura da soja (SANTANA et al., 2023). Esses sensores também tem sido utilizado para monitorar estresse hídrico em culturas como colza e canola. Experimentos apresentados por Luo et al. (2024) mostraram que espectros foliares coletados com o equipamento permitiram detectar variações no conteúdo de água antes da manifestação de sintomas visuais, utilizando índices como Índice de diferença normalizada da

água e Water Band Index. Essa abordagem possibilitou intervenções agronômicas mais precisas e antecipadas (ZAHARAN et al., 2022; BADZMIEROWSKI et al., 2019).

Por ser uma técnica com uma ampla possibilidade de aplicações, os dados espectrais tem sido utilizado em videiras e culturas arbóreas sob déficit hídrico, o espectrorradiômetro foi utilizado para correlacionar padrões espectrais com variáveis fisiológicas como fotossíntese líquida (A), condutância estomática (gs) e eficiência no uso da água (WUE). Genótipos com maior resiliência hídrica apresentaram assinaturas espectrais distintas (GONZÁLEZ-FLOR et al., 2019). Esses exemplos evidenciam que os sensores são uma ferramenta indispensável para estudos fisiológicos em oleaginosas, permitindo a coleta de dados espectrais de alta qualidade com potencial para aplicação em programas de melhoramento genético, manejo de estresses abióticos e controle de qualidade pós-colheita.

2.3. Agrupamento De Genótipos Com Base Em Atributos Fisiológicos E Espectrais

O agrupamento de genótipos vegetais é uma abordagem amplamente utilizada em estudos de fenotipagem com o objetivo de identificar padrões fisiológicos e espectrais que permitam classificar materiais genéticos com base em sua resposta ao ambiente. Essa técnica parte da premissa de que genótipos com características fisiológicas semelhantes tendem a apresentar assinaturas espectrais correlacionadas, o que possibilita a formação de grupos homogêneos (CRUSIOL et al., 2021; BEEGUM et al., 2024; MÉRIDA-GARCÍA et al., 2024).

Variáveis fisiológicas como fotossíntese líquida (A), condutância estomática (gs), concentração interna de CO₂ (Ci), transpiração (E) e eficiência no uso da água (WUE) apresentam ampla variação intra e intergenótipos em oleaginosas, sendo sensíveis a fatores genéticos e ambientais (GU et al., 2019; TORTOSA et al., 2022). A análise multivariada dessas variáveis, por meio de técnicas como Análise de Componentes Principais (PCA) e agrupamento não supervisionado (K-means), tem sido eficaz na separação de genótipos com diferentes estratégias fisiológicas (DHANYA et al., 2024; SANTANA et al., 2022). Em estudo conduzido por Buchailot et al. (2022), genótipos de soja foram agrupados com base em espectros foliares e variáveis fisiológicas, revelando dois grupos distintos: um com maior eficiência hídrica e outro com maior condutância estomática. A análise de componente principais explicou mais de 85 % da variância total, e os clusters formados apresentaram diferenças significativas em fotossíntese líquida, condutância estomática e eficiência do uso da água.

Outro exemplo é o trabalho de Yang et al. (2021), que aplicaram PCA e K-means para classificar genótipos de colza sob estresse hídrico. Os grupos formados apresentaram assinaturas espectrais distintas nas regiões do rededge e do SWIR, associadas ao conteúdo de clorofila e água foliar. Essa abordagem permitiu identificar genótipos com maior tolerância ao déficit hídrico antes da manifestação de sintomas visuais. Em videiras submetidas a déficit hídrico, genótipos com maior WUE foram agrupados separadamente por PCA, explicando cerca de 78 % da variância total. Esses genótipos apresentaram menor condutância estomática e maior fotossíntese líquida, indicando uma estratégia de conservação hídrica com manutenção da assimilação de carbono (TORTOSA et al., 2022).

Além das técnicas clássicas de agrupamento, abordagens mais recentes têm integrado aprendizado de máquina para melhorar a classificação de genótipos. O algoritmo SUnSeT, por exemplo, separou o espectro complexo em múltiplos espectros básicos em 420 acessos de soja, extraindo 199 variáveis espectrais e classificando os genótipos com alta acurácia (JEONG et al., 2024). Em amendoim, modelos empilhados (*stacked ensemble learning*) combinaram seleção de comprimentos de onda e regressões para classificar variedades e detectar doenças, alcançando 98,03 % de acurácia com tempo de processamento inferior a 0,4 s (WU et al., 2022). Redes neurais convolucionais profundas, como ResNet e NASNetLarge, também têm sido utilizadas para classificar genótipos de soja com base em imagens hiperespectrais, atingindo acurácias entre 90,6 % e 97,2 % respectivamente (DHANYA et al., 2024; SANTANA et al., 2023). Esses modelos conseguem integrar informações espectrais e texturais, ampliando a capacidade de discriminação entre genótipos.

Esses exemplos demonstram que o agrupamento de genótipos com base em atributos fisiológicos e espectrais, aliado a técnicas de aprendizado de máquina, representa uma abordagem robusta para a seleção de materiais genéticos adaptados a diferentes condições ambientais. Essa estratégia contribui para o avanço da fenotipagem digital e para o desenvolvimento de cultivares mais eficientes e resilientes.

2.4. Características Fisiológicas E Seus Papéis Na Fenotipagem Espectral

A fenotipagem fisiológica de plantas envolve a avaliação de variáveis relacionadas à troca gasosa, eficiência fotossintética e uso de recursos hídricos. Essas características são fundamentais para compreender o desempenho dos genótipos em diferentes condições ambientais e têm sido

amplamente utilizadas em estudos com espectroscopia hiperespectral e aprendizado de máquina (CRUSIOL et al., 2021; BUCHAILLOT et al., 2022; LI et al., 2021). A fotossíntese líquida (A) representa a taxa de assimilação de CO₂ pela planta e está diretamente relacionada à produtividade. Genótipos com maior A tendem a apresentar maior capacidade de conversão de energia luminosa em biomassa, sendo considerados mais eficientes em ambientes com disponibilidade hídrica e luminosa adequada (TORTOSA et al., 2022).

Estudos com soja e amendoim demonstraram que espectros foliares obtidos com o sensor multiespectral podem ser utilizados para estimar A por meio de modelos de regressão multivariada, como PLSR, com R² entre 0,50 e 0,70 (BUCHAILLOT et al., 2022; CRUSIOL et al., 2021). Em videiras sob estresse hídrico, genótipos com maior eficiência do uso da água (WUE), mantiveram fotossíntese líquida (A), elevada mesmo com redução da condutância estomática (gs), (LIU et al., 2020). A região do rededge (700–750 nm) e o NIR (750–1300 nm) são sensíveis à estrutura do mesófilo e ao conteúdo de clorofila, sendo úteis para inferir a capacidade fotossintética. Em *Populus alba*, aumentos em V_c, max e nitrogênio foliar reduziram a reflectância no rededge, evidenciando a relação entre espectro e A (DHANYA et al., 2024; PENG et al., 2019).

A condutância estomática (gs) regula o fluxo de CO₂ para o interior da folha e a saída de vapor d'água para a atmosfera. Estômatos mais abertos favorecem a fotossíntese, mas aumentam a transpiração, enquanto estômatos mais fechados conservam água, mas podem limitar a assimilação de carbono (BREWER et al., 2012). Em genótipos de soja e amendoim, variações em gs foram correlacionadas com alterações espectrais no rededge e NIR, permitindo estimativas indiretas por modelos espectrais (CRUSIOL et al., 2021; BUCHAILLOT et al., 2022). Em videiras sob déficit hídrico, genótipos com maior WUE apresentaram gs moderado, mantendo A sem comprometer a conservação hídrica (TORTOSA et al., 2022). Índices como NDRE, PRI e REIP têm sido utilizados para inferir a abertura estomática, sendo sensíveis à clorofila e à atividade fotossintética (SUKHOVA et al., 2022; LI et al., 2021; DHANYA et al., 2024).

A concentração interna de CO₂ (C_i) reflete o equilíbrio entre a entrada de CO₂ pelos estômatos e sua fixação pelos cloroplastos. Valores elevados de C_i podem indicar limitação na atividade carboxilativa, enquanto valores baixos sugerem alta eficiência fotossintética (ARAU S et al., 2018). Metanálises indicam que o aumento da concentração atmosférica de CO₂ reduz gs em até 40 %, elevando a WUE e alterando o C_i (SAURER et al., 2023). Genótipos com maior eficiência carboxilativa mantêm C_i moderado mesmo sob elevação de CO₂, o que pode ser inferido

por espectros no rededge e SWIR (DHANYA et al., 2024). A eficiência no uso da água (WUE) é definida como a quantidade de carbono assimilado por unidade de água transpirada. Essa variável integra A e g_s , sendo um indicador chave para seleção de genótipos adaptados a ambientes com restrição hídrica (CRUSIOL et al., 2021).

Genótipos mais eficientes no uso da água mantêm a fotossíntese líquida elevada mesmo com a condutância estomática reduzida, demonstrando capacidade de conservar água sem comprometer a produtividade (TORTOSA et al., 2022; TAVARES et al., 2022). Elevações de CO_2 atmosférico tendem a aumentar a eficiência do uso da água, mas podem gerar desequilíbrios nutricionais (ARAU S et al., 2018). Índices espectrais como o Normalized Difference Water Index (NDWI), o Water Band Index (WBI) e o Moisture Stress Index (MSI), que combinam bandas do NIR e do SWIR, têm sido utilizados para estimar WUE de forma não destrutiva, permitindo a triagem precoce de genótipos com maior resiliência hídrica (LU et al., 2020; PENG et al., 2019).

2.5. Relação Entre Reflectância Foliar E Respostas Fisiológicas

A reflectância foliar é resultado de interações biofísicas e bioquímicas entre a luz incidente e os constituintes internos da folha, como pigmentos, estrutura celular e conteúdo hídrico. A espectroscopia hiperespectral permite capturar essas interações com alta resolução, possibilitando inferir variáveis fisiológicas de forma não destrutiva e com elevada precisão (SARIĆ et al., 2022).

As bandas do espectro visível (VIS: 400–700 nm) são fortemente influenciadas pela presença de pigmentos fotossintéticos como clorofila-a, clorofila-b e carotenoides. A absorção nessas faixas está diretamente relacionada à atividade fotossintética, sendo útil para estimar a fotossíntese líquida (A) e o teor de clorofila (DHANYA et al., 2024; SUKHOVA et al., 2022). A região do rededge (700–750 nm), por sua vez, é sensível a variações no conteúdo de clorofila e nitrogênio foliar, sendo amplamente utilizada em índices como NDRE e REIP (DHANYA et al., 2024; SILVA JUNIOR et al., 2020). No infravermelho próximo (NIR: 750–1300 nm), a reflectância é modulada pela estrutura do mesófilo e pela densidade celular, permitindo inferir atributos como condutância estomática (g_s) e eficiência fotossintética. Já no infravermelho de ondas curtas (SWIR: 1300–2500 nm), a absorção é dominada por água, lignina e celulose, sendo útil para estimar o conteúdo hídrico foliar e a eficiência no uso da água (PENG et al., 2019; ADEGOYE et al., 2023).

Estudos com soja demonstraram que deslocamentos na posição do rededge estão associados a aumentos nos teores de clorofila e nitrogênio; quando sensores sem essas bandas foram utilizados, a acurácia na estimativa de pigmentos diminuiu significativamente (CRUSIOL et al., 2021). Em amendoim e colza, índices que combinam bandas do NIR e do SWIR, como NDWI, WBI e MSI, foram eficazes na detecção precoce de estresse hídrico, permitindo intervenções antes da murcha visível (WU et al., 2022; LU et al., 2020; ADEGOYE et al., 2023). A relação entre reflectância e fotossíntese líquida (A) foi corroborada por experimentos com oleaginosas. Genótipos com maior WUE apresentaram padrões espectrais distintos, com menor reflectância no rededge e maior absorção no SWIR, agrupando-se separadamente em análises multivariadas como PCA e regressão por mínimos quadrados parciais (PLSR) (BUCHAILLOT et al., 2022).

Em *Populus alba*, aumentos em V_c , max e nitrogênio foliar reduziram a reflectância nas bandas do verde e rededge, reforçando a utilidade dessas regiões espectrais para estimar a capacidade fotossintética (DHANYA et al., 2024; PENG et al., 2019). Esses achados indicam que os padrões espectrais são específicos para cada genótipo e podem ser explorados para selecionar materiais com maior eficiência fisiológica e adaptabilidade. Além disso, a combinação de bandas do VIS, NIR e SWIR em modelos multivariados melhora a correlação com variáveis agrônomicas e fisiológicas, que apresentam alta complexidade, ampliando a capacidade de discriminação entre genótipos (LIU et al., 2020; YOOSEFZADEH-NAJAFABADI et al., 2021; DE OLIVEIRA et al., 2023).

2.6. Modelos De Classificação

O avanço da espectroscopia na agricultura tem sido acompanhado pela crescente adoção de algoritmos de aprendizado de máquina (ML), que oferecem soluções robustas para lidar com a alta dimensionalidade e complexidade dos dados espectrais. Esses modelos têm sido aplicados com sucesso em tarefas de classificação de genótipos, detecção de doenças, predição de atributos fisiológicos e avaliação da qualidade de sementes em oleaginosas como soja, amendoim e colza (LIAKOS et al., 2018; YOOSEFZADEH-NAJAFABADI et al., 2021; BELGIU & DRĂGUȚ, 2016).

Algoritmos de classificação supervisionada são amplamente utilizados para identificar genótipos com base em suas assinaturas espectrais. Redes neurais convolucionais profundas

(CNNs), como ResNet e NASNetLarge, têm sido aplicadas na classificação de sementes de soja com imagens hiperespectrais, alcançando acurácias entre 90,6 % e 97,2 % (DHANYA et al., 2024; SANTANA et al., 2023). Esses modelos conseguem integrar informações espectrais e texturais, ampliando a capacidade de discriminação entre genótipos.

O algoritmo SUnSeT, baseado em unmixing espectral, foi utilizado para extrair 199 variáveis de imagens hiperespectrais de 420 acessos de soja, permitindo a classificação com alta acurácia (JEONG et al., 2024). Em amendoim, modelos empilhados (stacked ensemble learning) combinaram seleção de comprimentos de onda e regressões para classificar variedades e detectar mofo, atingindo 98,03 % de acurácia com tempo de processamento inferior a 0,4 s (WU et al., 2022).

Além das redes profundas, algoritmos clássicos como Support Vector Machines (SVM), Random Forest (RF), regressão logística (LR), árvores de decisão (J48, M5P) e redes neurais artificiais (ANN) continuam sendo amplamente utilizados. Em estudos com soja, esses modelos foram aplicados para classificar genótipos com base em variáveis fisiológicas e espectrais, com desempenho variável conforme o tipo de entrada e o pré-processamento adotado (SANTANA et al., 2022; CORRALES et al., 2022; BELGIU & DRĂGUȚ, 2016).

2.7. Modelos De Regressão

Algoritmos de regressão têm sido empregados para estimar características fisiológicas a partir de espectros foliares. Modelos como Partial Least Squares Regression (PLSR), regressão bayesiana, LASSO e Elastic Net foram utilizados para prever parâmetros como V_c , $\max$, $J_{\max}$, fotossíntese líquida (A), condutância estomática (g_s) e eficiência do uso da água (WUE), com coeficientes de determinação entre moderado e alto em soja e amendoim (BUCHAILLOT et al., 2022; CRUSIOL et al., 2021; ALVES et al., 2019).

Na predição da qualidade de sementes de soja, modelos de máquina de vetor suporte (SVM), combinados com pré-processamentos como derivada de primeira ordem, correção multiplicativa de espalhamento (MSC), normalização por variância (SNV) e seleção de variáveis (VISSA, CARS, SPA) atingiram resultados de moderado a alto, evidenciando a precisão dos modelos espectrais (SANTANA et al., 2022). A sensibilidade às bandas do rededge melhora as correlações entre espectros e teores de clorofila e nitrogênio, enquanto índices que combinam NIR e SWIR auxiliam na estimativa de água foliar e WUE (ADEGOYE et al., 2023).

2.8. Redução De Dimensionalidade E Validação

Devido à alta dimensionalidade dos dados hiperespectrais, técnicas de redução e seleção de variáveis são essenciais para evitar sobreajuste e melhorar a interpretabilidade dos modelos. O unmixing espectral utilizado no algoritmo SUnSeT reduz a redundância e extrai bandas informativas (JEONG et al., 2024). Métodos como a Análise de Componentes Principais (PCA), a Seleção Mínima Absoluta e Operador de Contração (LASSO) e o Seletor Regressor Competitivo Adaptativo (CARS) têm sido empregados para compressão de dados e eliminação de colinearidade, (DHANYA et al., 2024; LIU et al., 2020).

Comparações entre algoritmos mostram que Random Forest (RF) e máquina vetor de suporte (SVM) são menos sensíveis à dimensionalidade e geram modelos robustos, enquanto redes profundas exigem maiores volumes de dados para evitar sobreajuste (WU et al., 2022). A validação cruzada, especialmente o k-fold estratificado com múltiplas repetições, tem sido amplamente utilizada para garantir a generalização dos modelos em estudos com soja e amendoim (YOOSEFZADEH-NAJAFABADI et al., 2021; LIAKOS et al., 2018).

Esses resultados demonstram que o uso de aprendizado de máquina na análise de dados espectrais representa uma abordagem eficaz para a classificação e predição de atributos fisiológicos em oleaginosas. A escolha do algoritmo, o tipo de entrada espectral e o pré-processamento adotado são fatores determinantes para o desempenho dos modelos, devendo ser ajustados conforme o objetivo da análise e a natureza dos dados.

Referências

- ADEGOYE, G. A., Olorunwa, O. J., Alsajri, F. A., Walne, C. H., Wijewardana, C., Kethireddy, S. R., ... & Reddy, K. R. (2023). Waterlogging effects on soybean physiology and hyperspectral reflectance during the reproductive stage. **Agriculture**, Suíça, v. 13, n4, p. 844. <https://doi.org/10.3390/agriculture13040844>
- ALVES, T. M., Moon, R. D., MacRae, I. V., & Koch, R. L. (2019). Optimizing band selection for spectral detection of *Aphis glycines* Matsumura in soybean. **Pest management science**, Reino Unido, v. 75, n. 4, p. 942-949. <https://doi.org/10.1002/ps.5198>
- AHMAD, U., Alvino, A., & Marino, S. (2021). A review of crop water stress assessment using remote sensing. **Remote Sensing**, Suíça, v. 13, n. 20, p.4155. <https://doi.org/10.3390/rs13204155>

BADZMIEROWSKI, M. J., McCall, D. S., & Evanylo, G. (2019). Using hyperspectral and multispectral indices to detect water stress for an urban turfgrass system. **Agronomy**, Suíça, v. 9, n. 8, p. 439. <https://doi.org/10.3390/agronomy9080439>

BEEGUM, S., Hassan, M. A., Ramamoorthy, P., Bheemanahalli, R., Reddy, K. N., Reddy, V., & Reddy, K. R. (2024). Hyperspectral reflectance-based high throughput phenotyping to assess water-use efficiency in cotton. **Agriculture**, Suíça, v. 14, n. 7, p. 1054. <https://doi.org/10.3390/agriculture14071054>

BELGIU, M.; DRĂGUT, L. Random Forest in remote sensing: a review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Oxford, England, v. 114, p. 24–31, 2016. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>

BREWER, K., Clulow, A., Sibanda, M., Gokool, S., Odindi, J., Mutanga, O., ... & Mabhaudhi, T. (2022). Estimation of maize foliar temperature and stomatal conductance as indicators of water stress based on optical and thermal imagery acquired using an unmanned aerial vehicle (UAV) platform. **Drones**, Suíça, v. 6, n. 7, p. 169. <https://doi.org/10.3390/drones6070169>

BUCHAILLOT, M. L., Soba, D., Shu, T., Liu, J., Aranjuelo, I., Araus, J. L., ... & Sanz-Saez, A. (2022). Estimating peanut and soybean photosynthetic traits using leaf spectral reflectance and advance regression models. **Plant**, Suíça, v. 255, n. 4, p. 93. <https://doi.org/10.1007/s00425-022-03867-6>

CORRALES, D. C., Schoving, C., Raynal, H., Debaeke, P., Journet, E. P., & Constantin, J. (2022). A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France. **Computers and Electronics in Agriculture**, Oxford, England, v. 192, p. 106578. <https://doi.org/10.1016/j.compag.2021.106578>

CHANDRASEKAR, K., Srikanth, P., Chakraborty, A., Choudhary, K., & Ramana, K. V. (2024). Response of crop water indices to soil wetness and vegetation water content. **Advances in Space Research**, Oxford, England, v. 73, n. 2, p.1316-1330. <https://doi.org/10.1016/j.asr.2022.11.019>

CHESHKOVA, A. F. A review of hyperspectral image analysis techniques for plant disease detection and identification. **Vavilov Journal of Genetics and Breeding**, Novosibirsk, Russia, v. 26, n. 2, p. 202, 2022. <https://doi.org/10.18699/VJGB-22-25>

CRUSIOL, L. G. T., Nanni, M. R., Furlanetto, R. H., Sibaldelli, R. N. R., Cezar, E., Sun, L., ... & Farias, J. R. B. (2021). Yield prediction in soybean crop grown under different levels of water availability using reflectance spectroscopy and partial least squares regression. **Remote Sensing**, Oxford, England, v. 13, p. 977. <https://doi.org/10.3390/rs13050977>

DA SILVA JUNIOR, C. A., Nanni, M. R., Shakir, M., Teodoro, P. E., de Oliveira-Júnior, J. F., Cezar, E., ... & Shiratsuchi, L. S. (2018). Soybean varieties discrimination using non-imaging hyperspectral sensor. **Infrared Physics & Technology**, Oxford, England, v. 89, p. 338-350. <https://doi.org/10.1016/j.infrared.2018.01.027>

DE OLIVEIRA, J. F., de Alcântara, J. F., Santana, D. C., Teodoro, L. P. R., Baio, F. H. R., Coradi, P. C., ... & Teodoro, P. E. (2023). Spectral variables as criteria for selection of soybean genotypes at different vegetative stages. **Remote Sensing Applications: Society and Environment**, Oxford, England, v. 32, p. 101026. <https://doi.org/10.1016/j.rsase.2023.101026>

- DHANYA, V. G., Subeesh, A., Susmita, C., Amaresh, Saji, S. J., Dilsha, C., ... & Kumar, S. (2024). High throughput phenotyping using hyperspectral imaging for seed quality assurance coupled with machine learning methods: principles and way forward. **Plant Physiology Reports**, Germany, v. 29, n. 4, p. 749-768. <https://doi.org/10.1007/s40502-024-00839-8>
- FALCIONI, R., Moriwaki, T., Pattaro, M., Furlanetto, R. H., Nanni, M. R., & Antunes, W. C. (2020). High resolution leaf spectral signature as a tool for foliar pigment estimation displaying potential for species differentiation. **Journal of plant physiology**, Oxford, England, v. 249, p. 153161. <https://doi.org/10.1016/j.jplph.2020.153161>
- FAN, L., Zhao, J., Xu, X., Liang, D., Yang, G., Feng, H., ... & Wei, P. (2019). Hyperspectral-based estimation of leaf nitrogen content in corn using optimal selection of multiple spectral variables. **Sensors**, Suíça, v. 19, n. 13, p. 2898. <https://doi.org/10.3390/s19132898>
- FENG, H., Chen, Y., Song, J., Lu, B., Shu, C., Qiao, J., ... & Yang, W. (2024). Maturity classification of rapeseed using hyperspectral image combined with machine learning. **Plant Phenomics**, Oxford, England, v. 6, p. 0139. <https://doi.org/10.34133/plantphenomics.0139>
- FLORES, M. D. S., PASCHOALETE, W. M., Rojo BAILO, F. H., Silva CAMPOS, C. N., de Andréa PANTALEÃO, A., Ribeiro TEODORO, L. P., ... & TEODORO, P. E. (2020). Relationship Between Vegetation Indices And Agronomic Performance Of Maize Varieties Under Different Nitrogen Rates. **Bioscience Journal**, Uberlândia, v. 36, n. 5. <https://doi.org/10.14393/BJ-v36n5a2020-47993>
- FU, P., Montes, C. M., Siebers, M. H., Gomez-Casanovas, N., McGrath, J. M., Ainsworth, E. A., & Bernacchi, C. J. (2022). Advances in field-based high-throughput photosynthetic phenotyping. **Journal of Experimental Botany**, Oxford, England, v. 73, n. 10, p.3157-3172. <https://doi.org/10.1093/jxb/erac077>
- GAO, B. C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. **Remote sensing of environment**, Oxford, England, v. 58, n.3, p.257-266. [https://doi.org/10.1016/S0034-4257\(96\)00067-3](https://doi.org/10.1016/S0034-4257(96)00067-3)
- GAMON, J. A., Peñuelas, J. & Field, C. B. (1992). A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. **Remote Sensing of Environment**, Oxford, England, v. 41, p. 35–44. [https://doi.org/10.1016/0034-4257\(92\)90059-S](https://doi.org/10.1016/0034-4257(92)90059-S)
- GONZÁLEZ-FLOR, C., Serrano, L., & Gorchs, G. (2019). Use of reflectance indices to assess vine water status under mild to moderate water deficits. **Agronomy**, Suiça, v. 9, n. 7, p. 346. <https://doi.org/10.3390/agronomy9070346>
- GU, Q., Sheng, L., Zhang, T., Lu, Y., Zhang, Z., Zheng, K., ... & Zhou, H. (2019). Early detection of tomato spotted wilt virus infection in tobacco using the hyperspectral imaging technique and machine learning algorithms. **Computers and Electronics in Agriculture**, Oxford, England, v. 167, p.105066. <https://doi.org/10.1016/j.compag.2019.105066>
- GUI, J., Fei, J., Wu, Z., Fu, X., & Diakite, A. (2021). Grading method of soybean mosaic disease based on hyperspectral imaging technology. **Information Processing in Agriculture**, Oxford, England, v. 8, n. 3, p. 380-385. <https://doi.org/10.1016/j.inpa.2020.10.006>

- JEONG, S. W., Lyu, J. I., Jeong, H., Baek, J., Moon, J. K., Lee, C., ... & Park, Y. I. (2024). SUnSeT: spectral unmixing of hyperspectral images for phenotyping soybean seed traits. **Plant Cell Reports**, Germany, v. 43, n. 7, p. 164. <https://doi.org/10.1007/s00299-024-03249-0>
- KAKATI, J. P., Fallen, B., Bridges, W., & Narayanan, S. (2022). Characterization of a Soybean (*Glycine max* L. Merr.) population for germination and seedling root traits under Water Stress. **Agronomy**, Suíça, v. 12, n. 8, p. 1944. <https://doi.org/10.3390/agronomy12081944>
- LI, Q., Wong, F. K. K., & Fung, T. (2021). Mapping multi-layered mangroves from multispectral, hyperspectral, and LiDAR data. **Remote Sensing of Environment**, Oxford, England, v. 258, p. 112403. <https://doi.org/10.1016/j.rse.2021.112403>
- LI, F., Li, D., Elsayed, S., Hu, Y., & Schmidhalter, U. (2021). Using optimized three-band spectral indices to assess canopy N uptake in corn and wheat. **European Journal of Agronomy**, Oxford, England, v. 127, p.126286. <https://doi.org/10.1016/j.eja.2021.126286>
- LIAKOS, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. **Sensors**, Suíça, v. 18, n. 8, p.2674. <https://doi.org/10.3390/s18082674>
- LUO, B., Sun, H., Zhang, L., Chen, F., & Wu, K. (2024). Advances in the tea plants phenotyping using hyperspectral imaging technology. **Frontiers in Plant Science**, Suíça, v. 15, p. 1442225. <https://doi.org/10.3389/fpls.2024.1442225>
- MÉRIDA-GARCÍA, R., Gálvez, S., Solís, I., Martínez-Moreno, F., Camino, C., Soriano, J. M., ... & Hernandez, P. (2024). High-throughput phenotyping using hyperspectral indicators supports the genetic dissection of yield in durum wheat grown under heat and drought stress. **Frontiers in Plant Science**, Suíça, v. 15, p. 1470520. <https://doi.org/10.3389/fpls.2024.1470520>
- PANG, Y., Huang, Y., Zhou, Y., Xu, J., & Wu, Y. (2020). Identifying spectral features of characteristics of Sphagnum to assess the remote sensing potential of peatlands: A case study in China. **Mires and Peat**, Germany, v. 26, n. 25. <https://doi.org/10.19189/MaP.2019.OMB.StA.1834>
- QUEMADA, C., Pérez-Escudero, J. M., Gonzalo, R., Ederra, I., Santesteban, L. G., Torres, N., & Iriarte, J. C. (2021). Remote sensing for plant water content monitoring: A review. **Remote Sensing**, Oxford, England, v. 13, v. 11, p. 2088. <https://doi.org/10.3390/rs13112088>
- SANTANA, D. C., de Oliveira Cunha, M. P., Dos Santos, R. G., Cotrim, M. F., Teodoro, L. P. R., da Silva Junior, C. A., ... & Teodoro, P. E. (2022). High-throughput phenotyping allows the selection of soybean genotypes for earliness and high grain yield. **Plant Methods**, [S.l.], v. 18, n. 1, p. 13. <https://doi.org/10.1186/s13007-022-00848-4>
- SARIĆ, R., Nguyen, V. D., Burge, T., Berkowitz, O., Trtílek, M., Whelan, J., ... & Čustović, E. (2022). Applications of hyperspectral imaging in plant phenotyping. **Trends in plant science**, v. 27, n. 3, p. 301-315. <https://doi.org/10.1016/j.tplants.2021.12.003>
- SERRANO, L., Ustin, S. L., Roberts, D. A., Gamon, J. A., & Penuelas, J. (2000). Deriving water content of chaparral vegetation from AVIRIS data. **Remote sensing of Environment**, Oxford, England, v. 74, n. 3, p. 570-581. [https://doi.org/10.1016/S0034-4257\(00\)00147-4](https://doi.org/10.1016/S0034-4257(00)00147-4)

SINGH, R., Kaur, S., Singh, R., Katoch, K., Zotarelli, L., Singh, H., ... & Sharma, L. K. (2025). Advancing prediction of biophysical and biochemical traits in potatoes using hyperspectral data and artificial intelligence. **Agronomy Journal**, [S.l.], v. 117, n. 5, p. e70172. <https://doi.org/10.1002/agj2.70172>

SUKHOVA, E., Yudina, L., Kior, A., Kior, D., Popova, A., Zolin, Y., ... & Sukhov, V. (2022). Modified photochemical reflectance indices as new tool for revealing influence of drought and heat on pea and wheat plants. **Plants**, Suíça, v. 11, n. 10, p. 1308. <https://doi.org/10.3390/plants11101308>

TAVARES, C. J., Ribeiro Junior, W. Q., Ramos, M. L. G., Pereira, L. F., Casari, R. A. D. C. N., Pereira, A. F., ... & Mertz-Henning, L. M. (2022). Water stress alters morphophysiological, grain quality and vegetation indices of soybean cultivars. **Plants**, Suíça, v. 11, n. 4, p. 559. <https://doi.org/10.3390/plants11040559>

TEODORO, P. E., Teodoro, L. P., Baio, F. H., Silva Junior, C. A., Santana, D. C., & Bhering, L. L. (2024). High-throughput phenotyping in maize and soybean genotypes using vegetation indices and computational intelligence. **Plant Methods**, [S.l.], v. 20, n. 1, p. 164. <https://doi.org/10.1186/s13007-024-01294-0>

TORTOSA, I., Escalona, J. M., Opazo, I., Douthe, C., & Medrano, H. (2022). Genotype variations in water use efficiency correspond with photosynthetic traits in Tempranillo grapevine clones. **Agronomy**, Suíça, v. 12, n. 8, p. 1874. <https://doi.org/10.3390/agronomy12081874>

VILLACRÉS, J., Fuentes, A., Reszka, P., & Cheein, F. A. (2021). Retrieval of vegetation indices related to leaf water content from a single index: A case study of Eucalyptus globulus (Labill.) and Pinus radiata (D. Don.). **Plants**, Suíça, v. 10, n. 4, p. 697. <https://doi.org/10.3390/plants10040697>

WANG, Y., Ding, Z., Song, J., Ge, Z., Deng, Z., Liu, Z., ... & Yang, C. (2023). Peanut defect identification based on multispectral image and deep learning. **Agronomy**, Suíça, v. 13, n. 4, p. 1158. <https://doi.org/10.3390/agronomy13041158>

WEN, T., Li, J. H., Wang, Q., Gao, Y. Y., Hao, G. F., & Song, B. A. (2023). Thermal imaging: The digital eye facilitates high-throughput phenotyping traits of plant growth and stress responses. **Science of The Total Environment**, Oxford, England, v. 899, p. 165626. <https://doi.org/10.1016/j.scitotenv.2023.165626>

WU, Q., Xu, L., Zou, Z., Wang, J., Zeng, Q., Wang, Q., ... & Zhou, M. (2022). Rapid nondestructive detection of peanut varieties and peanut mildew based on hyperspectral imaging and stacked machine learning models. **Frontiers in Plant Science**, Suíça, v. 13, p. 1047479. <https://doi.org/10.3389/fpls.2022.1047479>

YOUSEFZADEH-NAJAFABADI, M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. **Frontiers In Plant Science**, Suíça, v. 11, p. 624273. <https://doi.org/10.3389/fpls.2020.624273>

ZAHRAN, S. A. E. S., Saeed, R. A. H., & Elazizy, I. M. (2022). Remote sensing-based water resources and agriculture spatial indicators system. **The Egyptian Journal of Remote Sensing and Space Science**, Cairo, v. 25, n. 2, p. 515-527. <https://doi.org/10.1016/j.ejrs.2022.02.002>

CAPÍTULO 1 - CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO AO DESEMPENHO FISIOLÓGICO USANDO SENSOR VIS-NIR

RESUMO

A soja é uma das culturas agrícolas de maior importância tanto para a economia quanto para a manutenção da oferta de alimentos a nível mundial, por isso é alvo de muitos programas de melhoramento genético, levando a ser uma cultura com grande diversidade de genótipos, a classificação segundo seu comportamento fisiológico pode auxiliar os programas de melhoramento. O objetivo do trabalho foi comparar algoritmos de aprendizagem de máquina e conjuntos de dados espectrais com melhor desempenho na classificação de genótipos de soja com base em características fisiológicas. O experimento foi conduzido a campo, em blocos ao acaso, com três repetições e 32 genótipos de soja. Os dados de refletância obtidos foram agrupados em 20 bandas com intervalos representativos. Os caracteres fisiológicos analisados: fotossíntese líquida (A), concentração interna de CO₂ (Ci), condutância estomática (gs), e eficiência do uso da água (WUE). O agrupamento dos genótipos baseado em seu desempenho fisiológico foi realizado utilizando o algoritmo k-means e a análise de componentes principais, formando dois grupos; após isso, os agrupamentos formados foram utilizados como variáveis de saída dos modelos de aprendizado de máquina, enquanto as variáveis de entrada testadas foram: comprimentos de onda e as médias espectrais das bandas selecionadas. Os algoritmos de aprendizado de máquina testados foram: rede neural artificial, algoritmos de árvore de decisão J48, REPTree, máquina de vetor de suporte (SVM), floresta aleatória e regressão logística (SL), usada como controle. As métricas estatísticas utilizadas para avaliar a acurácia dos modelos foram: porcentagem de classificação correta e F-score. Uma rede de correlações foi gerada para expressar as correlações entre os caracteres fisiológicos e as bandas espectrais, separadas por cluster. Os algoritmos de aprendizado de máquina com melhor desempenho na classificação dos genótipos, utilizando bandas espectrais como variáveis preditoras, foram a máquina de vetor suporte e a regressão logística. Quando a variável preditora utilizada são os comprimentos de onda, a regressão logística apresentou resultado superior.

Palavras-chave: Aprendizado de máquinas; bandas espectrais; fisiologia; melhoramento genético; sensor espectral.

ABSTRACT

Soybean is one of the most important agricultural crops, both for the economy and for maintaining the global food supply. As such, it is the focus of many genetic improvement programs. Given the wide diversity of genotypes in this crop, classification based on physiological behavior can aid in breeding programs. Thus, the objective of this study was to identify machine learning algorithms and inputs with the best performance for classifying soybean genotypes based on physiological characteristics. The field experiment was implemented in a randomized block design with three replications, evaluating 32 soybean genotypes. Multi- and hyperspectral reflectance data were obtained and grouped into 20 representative intervals in bands. The physiological traits analyzed were: net photosynthesis, internal CO₂ concentration, stomatal conductance, and transpiration. Genotype clustering based on their physiological performance was performed using the k-means algorithm and principal component analysis, resulting in two groups. These clusters were then used as output variables in the machine learning models, while the input variables tested included wavelengths and spectral band averages of the selected bands. The machine learning algorithms tested were: artificial neural network, J48 decision tree algorithm, REPTree, support vector machine (SVM), random forest, and logistic regression (LR), which was used as a control. The statistical metrics used to evaluate model accuracy were correct classification percentage and F-score. A correlation network was generated to express the correlations between physiological traits and spectral bands, separated by cluster. The machine learning algorithms that performed best in classifying the genotypes were SVM and LR. Using spectral bands as predictor variables, both algorithms (SVM and LR) showed similar results. When wavelengths were used as the predictor variable, LR showed superior results.

Keywords: Machine learning; spectral bands; physiology; genetic improvement; spectral sensor.

1. INTRODUÇÃO

A soja (*Glycine max* (L.) Merrill) é uma das culturas agrícolas de maior importância tanto para a economia quanto para a manutenção da oferta de alimento a nível mundial. O grão é uma fonte rica em proteínas e óleos, o que a torna indispensável na alimentação humana (WALLY-VALLIM et al., 2014), estando presente na mesa de milhares de pessoas todos os dias, além de sua participação na produção de ração animal entre outros subprodutos (KOUR, 2023; ZHANG et al., 2019). Devido seu teor de proteína elevado e potencial benéfico para a saúde além a ampla gama de produtos que tem em sua composição derivados da soja, sua produção e consumo tem aumentado mundialmente (DENG, 2021). Levando os programas de melhoramento genético a buscar cada vez mais por genótipos com respostas fisiológicas adaptadas aos diferentes ambientes aos quais são expostos.

As características fisiológicas da cultura da soja são vitais na determinação da qualidade e produtividade dos grãos, essas variáveis devem ser consideradas desde a determinação do genótipo a ser utilizado sob diferentes ambientes, considerando que os processos fisiológicos estão intimamente ligados aos fatores genéticos e condições ambientais (BATISTA et al., 2022; SILVA et al., 2017). Diversos fatores abióticos influenciam nos processos moleculares e fisiológicos das plantas, podendo impactar negativamente no crescimento e produtividade da cultura (RASHEED et al., 2022; ZHANG et al., 2020; NGUYEN, 2024). A resposta fisiológica da planta exposta ao estresse abiótico envolve alguns mecanismos genéticos complexos, com genes específicos desempenhando um importante papel na adaptação do genótipo ao ambiente (LESTARI et al., 2014).

A fotossíntese líquida, concentração interna de CO₂, condutância estomática e transpiração são processos fisiológicos importantes e influentes no desenvolvimento e produtividade da soja. A condutância estomática regula a troca gasosa entre a planta e o ambiente, influenciando a eficiência da fotossíntese e a transpiração (CHEN et al., 2022; TANAKA et al., 2010), e está ligada a diversos processos essenciais para o crescimento e desenvolvimento das plantas. Sua atuação sofre interferência de diversos fatores como intensidade luminosa e características genéticas, que podem impactar na eficiência das trocas gasosas e taxas fotossintéticas nas folhas da cultura da soja (CHEN et al., 2022; TANAKA et al., 2010).

Os estômatos são pequenas aberturas nas superfícies vegetais, estando em maior número na face abaxial das folhas, esses poros controlam as trocas gasosas, como a entrada de dióxidos de

carbono (CO_2) utilizado na fotossíntese, liberação de gás oxigênio (O_2) e a liberação de vapor de água durante a transpiração (SHEKOOFA et al., 2023). A condutância estomática determina como ocorrerá esse movimento do gás através dos poros, impactando nas taxas em que essas trocas e perda de água nas plantas (SHEKOOFA et al., 2023; CHEN et al., 2022).

Mensurar essas características pode fornecer importantes informações e contribuir para uma tomada de decisão assertivas durante as fases iniciais dos programas de melhoramento genético de soja. Porém para se obter esses dados é necessário se ter acesso a equipamentos com alto investimento inicial, além da exigência de mão de obra especializada para operar esses equipamentos em campo (SUN et al., 2023; HERRITT & FRITSCHI, 2020), o que implica em maiores custos e demanda tempo para conclusão da tarefa. Uma opção viável para respostas mais rápidas e confiáveis é a utilização de sensoriamento remoto para obtenção de dados espectrais.

O uso de sensores espectrais para captar informações sobre a reflectância da luz em diferentes comprimentos de onda integrados a outras tecnologias fornecem uma ampla base de dados, essas informações podem ser utilizadas para avaliar o estado fisiológico e eficiência do uso da água por exemplo (COTRIM et al., 2021; RAHMAN et al., 2017). Devido ao grande volume de dados obtidos pelo sensoriamento remoto a análise dessas informações pode se tornar uma tarefa difícil para os modelos estatísticos convencionais, sendo assim combinar inteligência computacional e análise espectral, pode melhorar significativamente a capacidade de classificação e predição de caracteres fisiológicos da soja (HOU, et al. 2021).

A utilização das técnicas de aprendizado de máquinas na agricultura está crescendo graças a sua capacidade de analisar grandes conjuntos de dados (LIAKOS et al., 2018), e seu uso tem sido empregado em diversas áreas, como estimativa de produção, identificação de fitopatógenos, reconhecimento de plantas daninhas e fenotipagem de alta precisão (ANG & SENG, 2021; MAHMOOD et al., 2022; SANTANA, 2023). Essas aplicações têm demonstrado a versatilidade da utilização da inteligência computacional para melhorar os resultados da produção agrícola.

Sabendo que os genótipos têm diferentes respostas fisiológicas ao ambiente, a classificação desses materiais segundo sua fisiologia, utilizando suas características espectrais submetidas a algoritmos de aprendizado de máquinas pode auxiliar os melhoristas a selecionarem de maneira rápida e assertivas os genótipos com maior potencial para atender requisitos fisiológicos alvos dos programas de melhoramento. Sendo assim, a hipótese de pesquisa é de que é possível classificar genótipos de soja quanto a sua resposta fisiológica, utilizando informações espectrais submetidas

ao processamento de dados por diferentes algoritmos de aprendizado de máquinas. O objetivo desse estudo foi encontrar a combinação entre diferentes algoritmos de aprendizado de máquinas e configurações de dados espectrais utilizados como variáveis de entrada para os modelos, oferece maior acurácia na classificação dos genótipos de soja com base em suas respostas fisiológicas.

2. MATERIAL E MÉTODOS

O ensaio foi conduzido na Universidade Federal de Mato Grosso do Sul, campus Chapadão do Sul (18°41'33''S, 52°40'45''W, com 810 m de altitude). O clima da região segundo classificação de Köppen e Geiger é Aw (Tropical de Savana). O solo é classificado como Latossolo Vermelho Distrófico Argiloso, com as características na camada de 0 – 0,20 m: pH (H₂O) = 6,20; Al trocável (cmolc dm⁻³) = 0,00; Ca+Mg (cmolc dm⁻³) = 4,31; P (mg dm⁻³) = 41,30; K (cmolc dm⁻³) = 0,20; Matéria orgânica (g dm⁻³) = 19,74; V (%) = 45; m (%) = 0,00; Soma de bases (cmolc dm⁻³) = 2,30; CTC (cmolc dm⁻³) = 5,10.

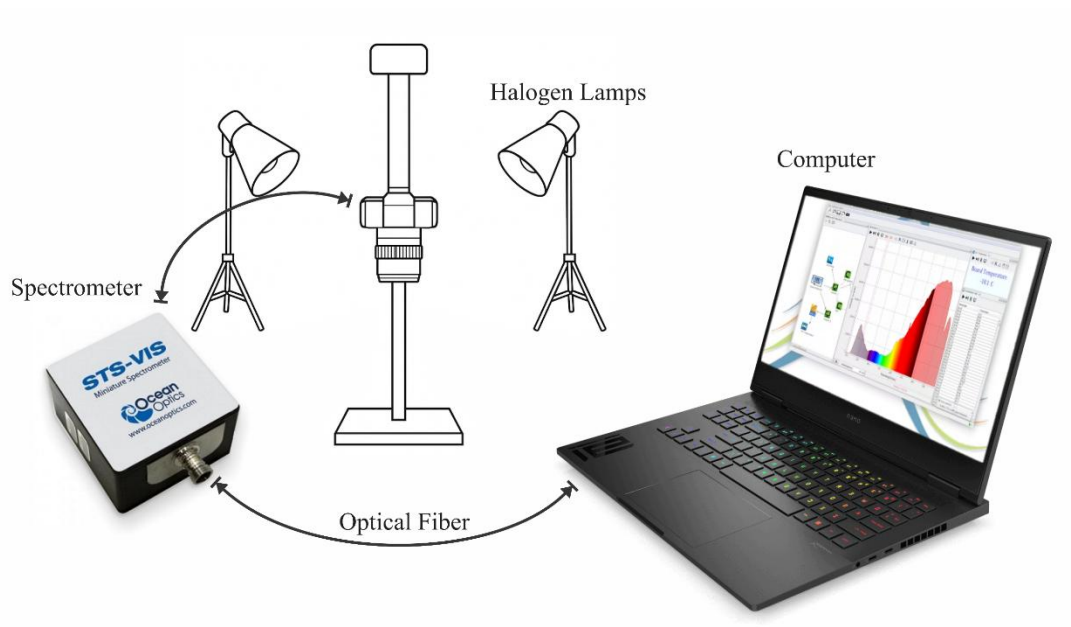
Foram utilizados 32 genótipos F₃ de soja, obtidos a partir de autofecundações sucessivas de populações F₂, a semeadura foi realizada na safra 2021/22 (outubro de 2021), o delineamento utilizado foi o delineamento de blocos casualizados (DBC), com três repetições para cada genótipo. As parcelas foram formadas por cinco linhas com espaçamento de 0,45 m, com um metro de comprimento. Antes da semeadura a área recebeu preparo de solo do tipo convencional (aração e gradagem niveladora).

Na data de semeadura as sementes de soja tratadas utilizando o fungicida (Metil Tiofanato + Piraclostrobina), o inseticida utilizado foi o Fipronil, todos os químicos aplicados na dose de 0,2 L do produto a cada 100 kg de sementes. Foi realizada a inoculação com a bactéria *Bradyrhizobium* para a fixação biológica de nitrogênio, seguindo a orientação da bula, com a dose de 0,2 L de inoculante líquido concentrado a cada 100 kg de sementes. Os sulcos de plantio foram abertos utilizando uma semeadora de 5 linhas, sem deposição de adubação, na sequência as sementes foram distribuídas manualmente nos sulcos com a densidade de 15 sementes por metro. Os tratos culturais utilizados durante o ciclo da cultura, com aplicação de fungicidas, inseticidas e herbicidas foram realizados conforme a necessidade após sucessivas avaliações in loco.

As avaliações iniciaram aos 60 dias após a emergência (DAE), com as análises espectrais de três amostras foliares escolhidas ao acaso da parte superior do dossel da planta de cada parcela. Foi utilizado o sensor hiperespectral Ocean Optics modelo STS-VIS-L-50-400-SMA. Foi capturada a

reflectância nos comprimentos de onda entre 450 e 824 nm em laboratório com iluminação artificial, utilizando duas lâmpadas halógenas de 120 W com faixa de lux convergindo no ponto de leitura a 50 centímetros do alvo. A calibração do sensor foi realizada antes da leitura das amostras de cada bloco, utilizando-se uma placa de sulfato de bário para a calibração do branco e o isolamento do sensor para a calibração do escuro. O sensor atua conectado a um computador que registra as leituras utilizando o software específico do sensor, chamado OceanView (Figura 1), com resolução espectral de 1.5 nm.

Figura 1. Representação esquemática do sistema laboratorial utilizado para a aquisição dos dados espectrais.



Fonte: elaboração própria

Os comprimentos de onda (WL), foram agrupados em médias de intervalos representativos em bandas seguindo a metodologia adaptada de Da Silva Junior et al., (2018), utilizando os valores de comprimentos de onda médios livres de ruído disponibilizados pelo sensor (Tabela 1).

Tabela 1. Intervalos de Faixa espectral (nm) e comprimento de onda médio (nm) usados no estabelecimento de 20 bandas espectrais

Bandas (nº)	Faixa espectral (nm)	Comprimento de onda médio (nm)
B3	481-500	490,47
B4	501-530	515,34
B5	531-539	534,89
B6	540	540,00
B7	541-649	594,80
B8	650	650,00
B9	661-670	665,54
B10	675	675,00
B11	676-684	680,05
B12	685-689	686,83
B13	690-700	695,07
B14	701-709	705,03
B15	710	710,00
B16	711-730	720,60
B17	675	675,00
B18	676-684	680,00
B19	685-689	687,00
B20	690-700	695,00
B21	701-709	705,00
B22	710	710,00
B23	711-730	720,50

Fonte: elaboração própria

Também aos 60 DAE, as características fisiológicas foram avaliadas em cinco plantas por parcela utilizando um analisador de fotossíntese portátil (Infrared Gas Analyzer - IRGA), modelo Li- 6400XT (LiCor Inc., Lincoln, Nebraska, EUA), segundo a metodologia proposta por Teodoro et al. (2019). As medições foram realizadas entre as 8:00 e 11:00 horas da manhã, temperatura variando entre 26 °C e 26,5 °C e umidade relativa do ar entre 50 e 80%. Foi utilizado fluxo de

fótons fotossinteticamente ativos de $1044 \mu\text{mol m}^{-2} \text{s}^{-1}$ e concentrações de CO_2 no ambiente de $372 \pm 10 \text{ mol m}^{-2} \text{s}^{-1}$. As características fisiológicas analisadas foram: fotossíntese líquida (photo, $\mu\text{mol CO}_2 \text{ m}^{-2}\text{s}^{-1}$), condutância estomática (cond, $\text{mol H}_2\text{O m}^{-2} \text{s}^{-1}$), concentração interna de CO_2 (C_i , $\mu\text{mol CO}_2 \text{ mol air}^{-1}$) e transpiração (Trmmol, $\text{mmol H}_2\text{O m}^{-2} \text{s}^{-1}$).

Os dados espectrais e fisiológicos foram submetidos a análises estatísticas de análise de variância (ANOVA) e teste de comparação de médias (skott-Knott). A primeira análise realizada foi o agrupamento dos genótipos aplicando o algoritmo k-means. Esse algoritmo depende de uma determinação prévia do valor de k, sendo esse o número de grupos (cluster's), criados (AHMED et al., 2020), no estudo em questão o valor que melhor representou a separação dos grupos de acordo com as características fisiológicas foi $k = 2$, separando os genótipos em dois grupos, denominados Cluster 1 e Cluster 2.

O k-means determina os valores para os centroides com base no valor de k, após isso ele realiza comparações individuais entre os valores das amostras e o quão distante estão dos centroides definidos anteriormente. Cada amostra pertence a seu respectivo grupo de acordo com a distância dos centroides, sendo consideradas similares entre si os valores próximos e distintos dos valores distantes, que pertencem ao outro grupo. Os clusters definidos pelo algoritmo k-means foram plotados em um gráfico biplot aplicando a análise de componentes principais (PCA), para expressar visualmente a divisão dos genótipos nos clusters.

Após a definição dos clusters, foram construídos boxplot comparativos entre as médias dos valores obtidos para as características fisiológicas e das bandas espectrais (SB), permitindo a observação dos diferentes valores médios separados por cluster formado. Também foi realizado análise de correlação de Pearson entre as variáveis e as bandas selecionadas expressas por um gráfico de calor para cada cluster.

Posteriormente os dados foram submetidos à aprendizagem de máquinas, utilizando seis algoritmos (Tabela 2), sendo a regressão logística (SL), utilizada como modelo controle por ser um algoritmo tradicional de regressão, estando o sucesso dos demais algoritmos relacionados a capacidade de obter melhores resultados quando comparados a SL. Para a entrada nos modelos, foram utilizados dois grupos de dados, as variáveis espectrais (WL), obtidas pelo sensor sem o ruído (450-828nm), e as médias espectrais das bandas selecionadas (SB), (450,0-720,6 nm), com o intuito de que os algoritmos distinguíssem os clusters de acordo com os dados espectrais dos

genótipos pertencentes a cada um deles. Os dois clusters formados foram utilizados como variáveis de saída nos modelos de aprendizado de máquinas.

Tabela 2. Relação dos modelos de aprendizagem de máquinas utilizados na classificação.

Sigla	Modelo de aprendizagem de máquinas	Referência
RT	REPTree	(Snousy et al., 2011)
J48	Árvores de decisão J48	(Quinlan, 2014)
RF	Floresta aleatória	(Belgiu; Dragut, 2016)
RNA	Redes neurais artificiais	(Egmont-Petersen et al, 2002)
SL	Regressão logística	(Štepanovsky et al., 2017)
SVM	Máquina de vetor suporte	(Nalepa and Kawulok, 2019)

Fonte: elaboração própria

Para todos os modelos foram utilizados os parâmetros das configurações padrão do software Weka 3.8.5, com exceção das Redes Neurais Artificiais, onde foram aplicadas duas camadas intermediárias com 10 neurônios em cada. A classificação dos clusters pelos algoritmos utilizou a validação cruzada estratificada, com k-fold = 10 e dez repetições (100 execuções por modelo).

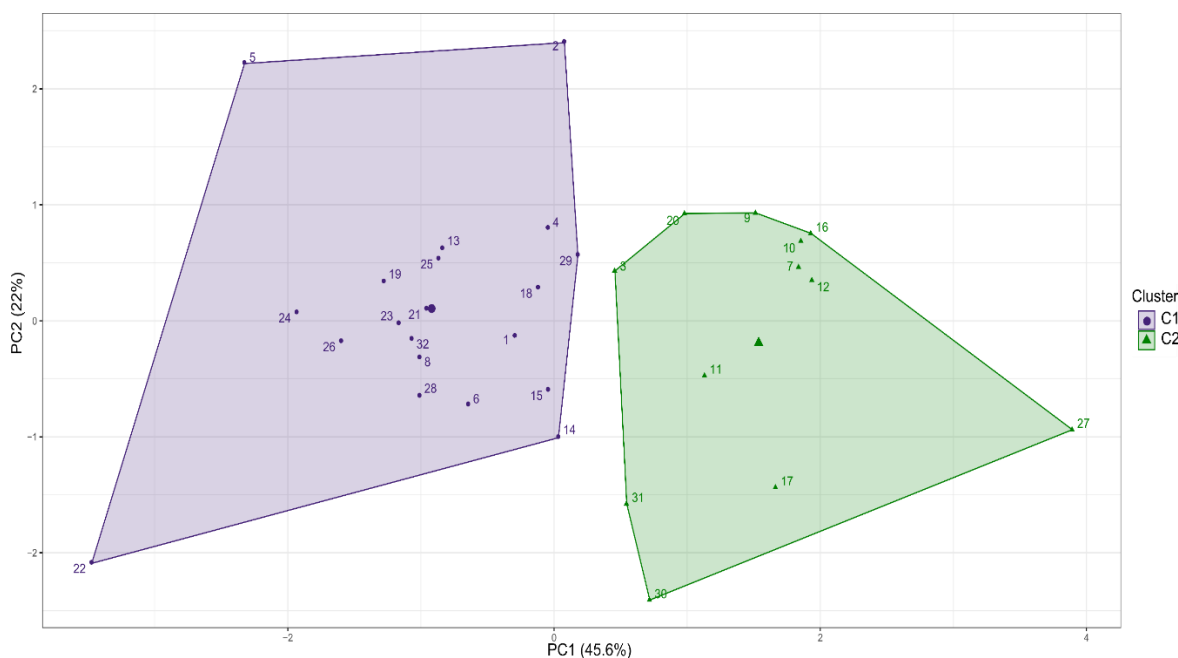
Para avaliar o desempenho dos algoritmos os resultados foram submetidos a duas métricas de acurácia: porcentagem de classificações corretas (CC), e F-score. Foram construídos boxplots para expressar os resultados de acurácia obtidos por cada um dos modelos e dos inputs para ambas as métricas utilizadas. Os valores de CC e F-score foram submetidos uma análise de variância (ANOVA), e teste de médias quando necessário, utilizando o teste de Scott-Knott (SCOTT; KNOTT, 1974), a 5% de significância. A análise de PCA-Kmeans, as correlações de Pearson, os gráficos e os testes de médias foram realizados utilizando o software R (TEAM, 2013), aplicando os pacotes ggplot2, ExpDes.pt e Metan.

3. RESULTADOS

Os 32 genótipos de soja avaliados foram agrupados em dois grupos (cluster), de acordo com suas similaridades relacionadas aos resultados médios para as variáveis fisiológicas: fotossíntese líquida (photo, $\mu\text{mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$), condutância estomática (cond, $\text{mol H}_2\text{O m}^{-2} \text{ s}^{-1}$), concentração interna de CO_2 (C_i , $\mu\text{mol CO}_2 \text{ mol air}^{-1}$) e transpiração (Trmmol, $\text{mmol H}_2\text{O m}^{-2} \text{ s}^{-1}$). O primeiro

grupo denominado Cluster 1 (C1), conta com 20 dos genótipos (PC1 45,6%), sendo os 12 (PC2 22%) restantes agrupados no Cluster 2 (C2). A Figura 2 apresenta um gráfico de componentes principais (PCA), onde é possível observar a dimensionalidade dos dados dos genótipos agrupados em seus respectivos clusters.

Figura 2. Clusters formados pelo algoritmo K-means dividindo os 32 genótipos de soja de acordo com a sua similaridade em variáveis fisiológicas: fotossíntese líquida (photo, $\mu\text{mol CO}_2 \text{ m}^{-2}\text{s}^{-1}$), condutância estomática (cond, $\text{mol H}_2\text{O m}^{-2} \text{ s}^{-1}$), concentração interna de CO_2 (C_i , $\mu\text{mol CO}_2 \text{ mol air}^{-1}$) e transpiração (Trmmol, $\text{mmol H}_2\text{O m}^{-2} \text{ s}^{-1}$).

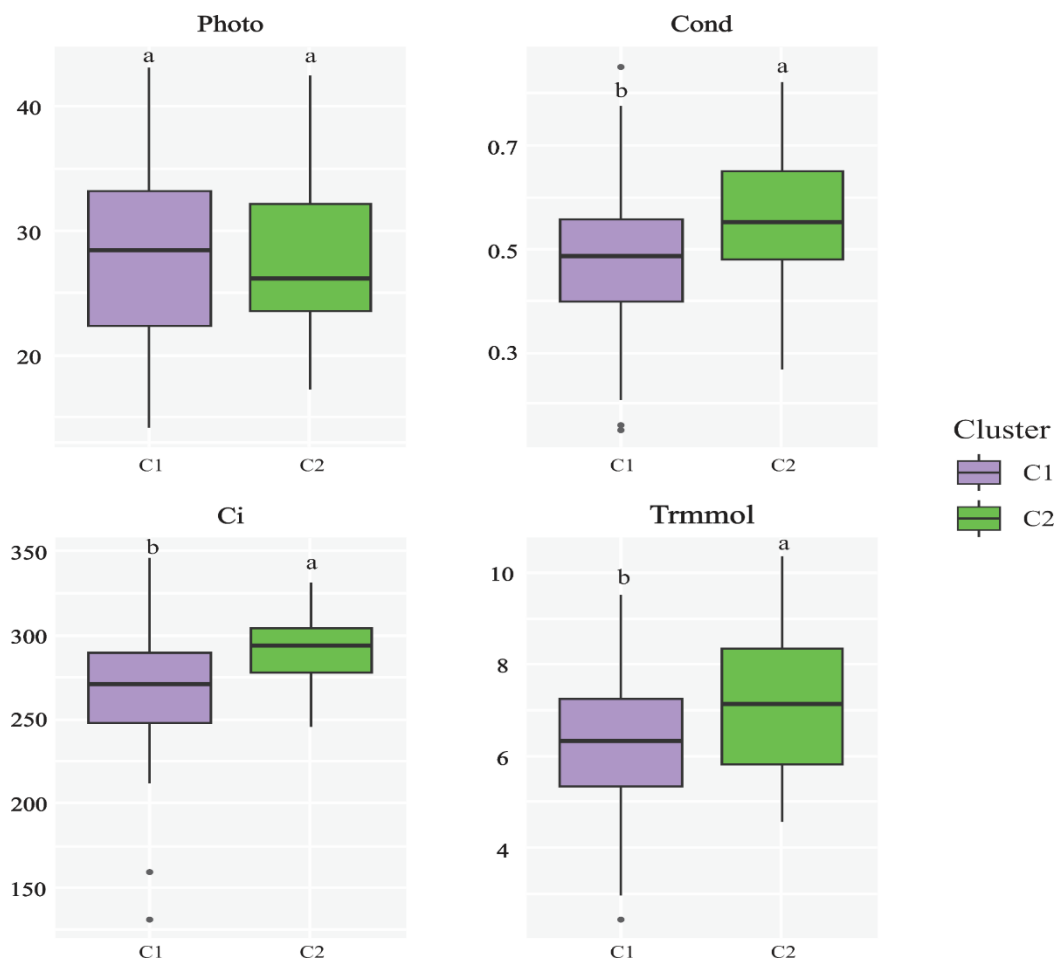


Fonte: elaboração própria

Os valores médios para as variáveis analisadas obtidos pelos grupos estão apresentados nos boxplots da Figura 3. Houve diferença estatística entre os dois clusters para as variáveis condutância estomática, concentração interna de CO_2 e transpiração, sendo o C2 o grupo com melhor média nos três caracteres. Para fotossíntese líquida não houve diferença estatística significativa entre os grupos ainda que a média para essa variável tenha sido superior no C1.

Figura 3. Gráficos de boxplot mostrando a comparação de médias entre os cluster separados por variáveis fisiológicas: fotossíntese líquida (photo, $\mu\text{mol CO}_2 \text{ m}^{-2}\text{s}^{-1}$), condutância estomática

(cond, mol H₂O m⁻² s⁻¹), concentração interna de CO₂ (*C_i*, μmol CO₂ mol air⁻¹) e transpiração (Trmmol, mmol H₂O m⁻² s⁻¹).

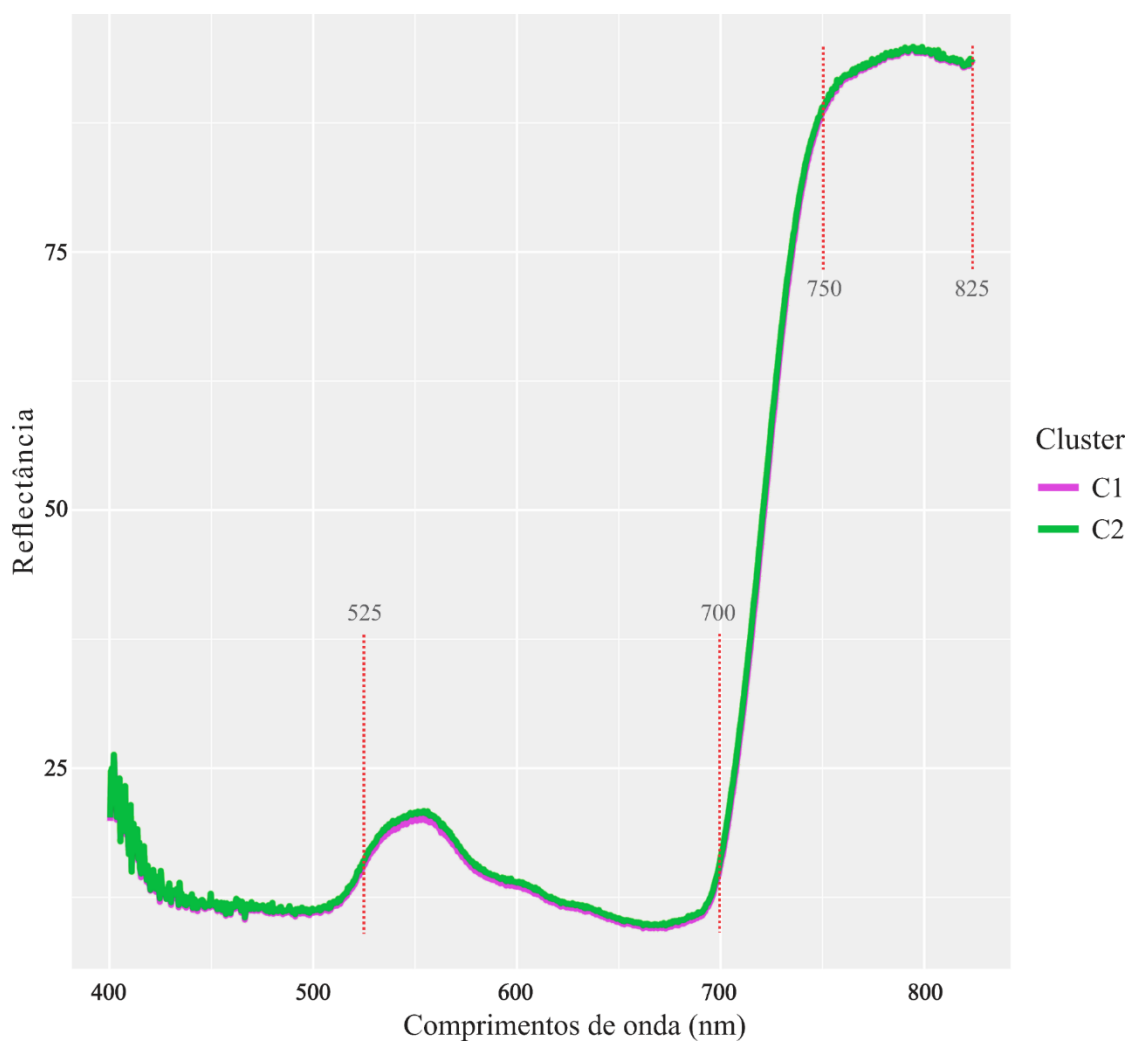


Letras iguais sobre os boxplot's separados por variável não diferenciam entre si pelo teste T a 5% de probabilidade para cada variável.

Fonte: elaboração própria

Com os valores médios de refletância dos comprimentos de ondas dos genótipos alocados em seus respectivos grupos foi traçada a curva espectral dos dois clusters (comprimentos de onda entre 400 e 825nm). Os comprimentos de onda observados no C2 apresentam valores maiores que os do C1, o que fica mais perceptível entre 525 e 700 nm (Figura 4). Entre 750 e 825 nm houve novamente uma diferenciação entre os grupos, sendo a média do C2 superior ao resultado observado no C1.

Figura 4. Assinatura hiperespectral VIS/NIRs dos clusters formados. As linhas verticais indicam o início e o fim da faixa de comprimentos de onda com maior diferenciação entre os cluster C1 e C2.



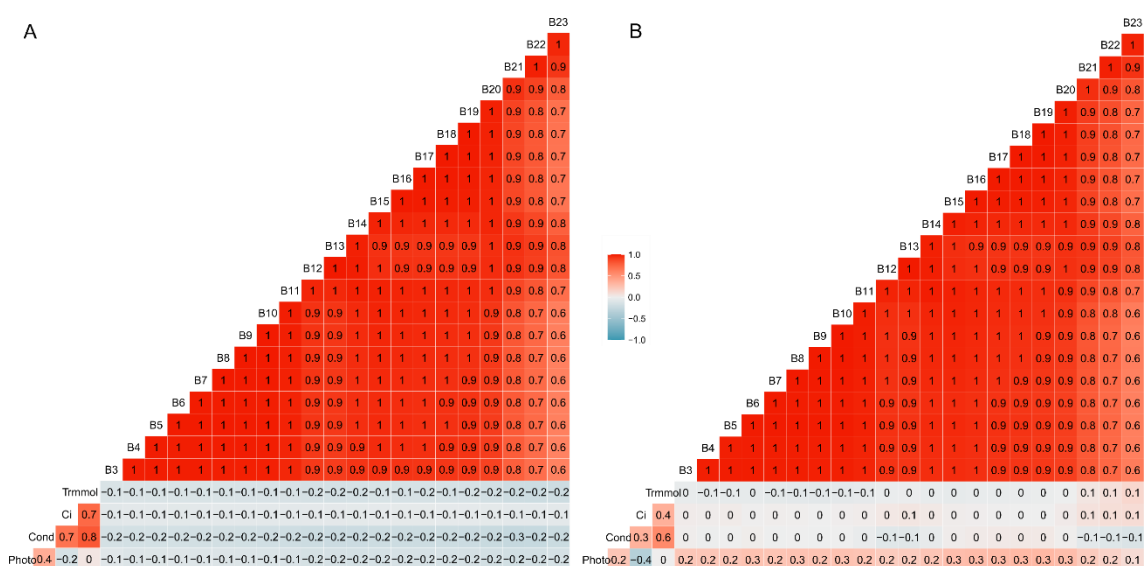
Fonte: elaboração própria

Para expressar a relação linear entre as variáveis fisiológicas e as bandas espectrais da Tabela 1 foi realizada uma análise de correlação de Pearson. Por se tratar de dados com alta similaridade, as bandas espectrais apresentaram de moderada a alta correlação positiva entre si nos dois grupos formados (Figura 4).

Entre as variáveis fisiológicas houve uma diferenciação entre as interações para os dois grupos, com destaque para a correlação alta e positiva entre C_i , cond e Trmmol no C1, e correlação

moderada entre Photo e cond para o mesmo grupo, sendo ainda a interação de Photo fraca e negativa com Ci e nula com Trmmol (Figura 5 A). O gráfico de correlação de Pearson para o C2 (Figura 5 B), mostrou uma interação semelhante as observadas no C1 entre todas as variáveis, porém com menor intensidade, apresentando entre baixa e moderada correlação.

Figura 5. Correlação de Pearson as variáveis fisiológicas: fotossíntese líquida (photo, $\mu\text{mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$), condutância estomática (cond, $\text{mol H}_2\text{O m}^{-2} \text{ s}^{-1}$), concentração interna de CO_2 (Ci , $\mu\text{mol CO}_2 \text{ mol air}^{-1}$) e transpiração (Trmmol, $\text{mmol H}_2\text{O m}^{-2} \text{ s}^{-1}$) para o cluster 1 (A) e cluster 2 (B).



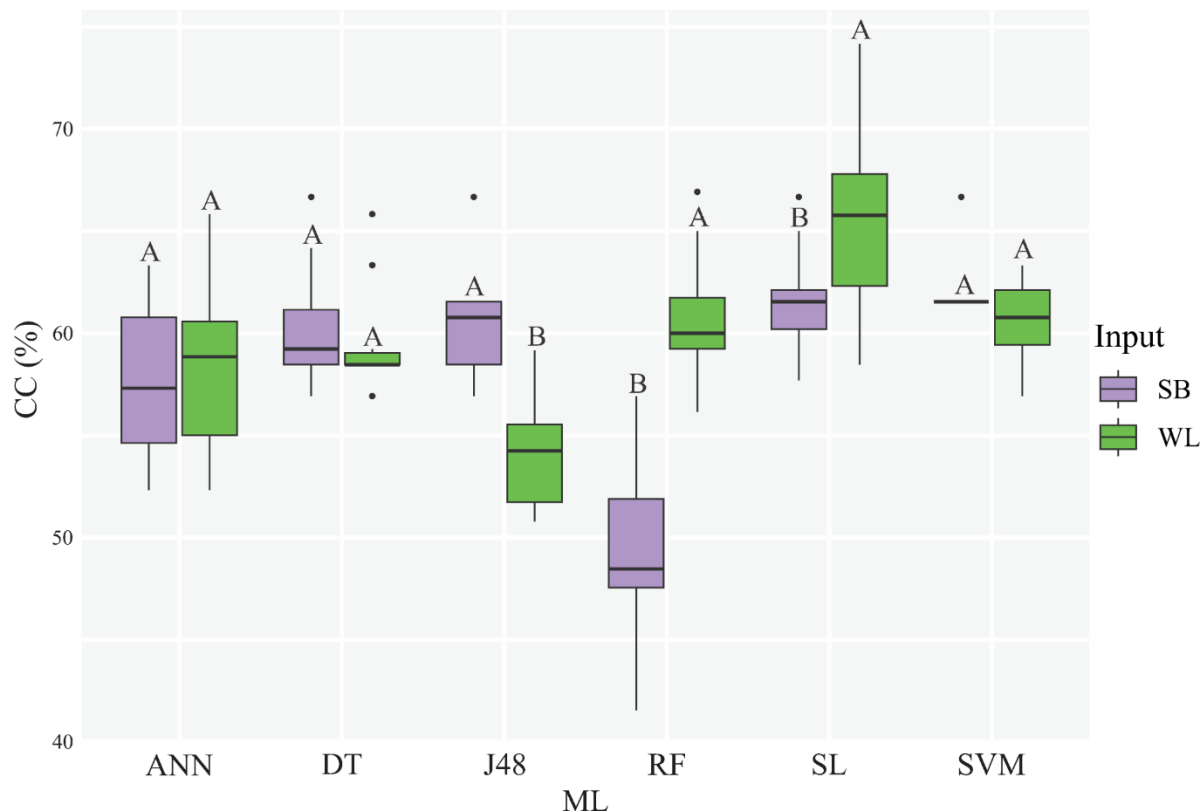
Fonte: elaboração própria

Os dois clusters apresentaram interações diferentes entre as variáveis fisiológicas e as bandas analisadas, sendo todas as interações no C1 negativas e fracas, enquanto no C2 as interações foram majoritariamente nulas, com 14% das interações foram negativas, com destaque para as interações entre a variável Photo que apresentou correlação positiva, ainda que baixa, com todas as bandas testadas.

Essa variação no comportamento dos grupos dificulta a aplicação de métodos estatísticos tradicionais para relacionar e agrupar os genótipos com base em suas respostas fisiológicas, por isso os dados foram submetidos a análise de aprendizagem de máquinas (ML). Os resultados apresentaram interação estatisticamente significativa para as duas métricas de acurácia, entre os modelos de ML utilizados e as entradas. Para a métrica de acurácia Classificações corretas (CC),

as entradas bandas espectrais (SB) e comprimento de ondas (WL), não apresentaram diferença estatística quando submetidos aos algoritmos redes neurais artificiais (ANN), árvore de decisão (DT) ou máquina de vetor de suporte (SVM). Utilizando os algoritmos random forest (RF) e a regressão logística (SL), o melhor percentual de classificações corretas ocorreu quando se utilizou o WL como entrada, sendo a SL o algoritmo com melhor desempenho entre os modelos testados (Figura 6). Quando se utilizou SB como entrada o algoritmo J48 apresentou melhor desempenho, com média estatisticamente superior aos resultados obtidos quando se utiliza o WL como input para esse mesmo algoritmo.

Figura 6. Boxplot para porcentagem de classificação correta (CC) considerando a interação significativa entre os algoritmos de ML e as entradas testadas.



Médias seguidas pelas mesmas letras maiúsculas para as diferentes entradas e algoritmos de ML não diferem pelo teste Scott-Knott a 5% de probabilidade.

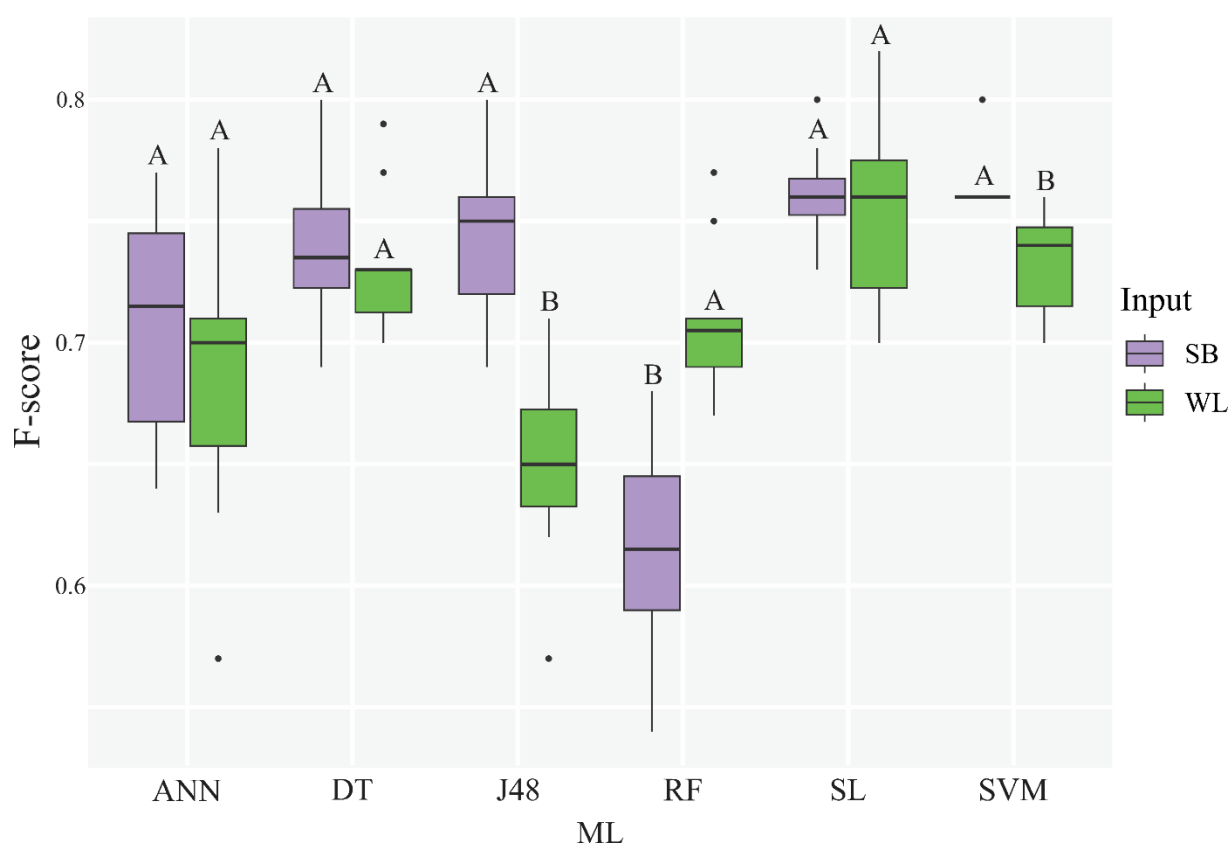
Fonte: elaboração própria

Os resultados mostram que ANN, DT e SVM tiveram desempenho semelhante quando submetidos às duas diferentes entradas. RF e SL tem melhores percentuais de CC quando utilizado

o WL como entrada. Enquanto o algoritmo J48 apresenta maior capacidade de classificações corretas quando o SB é utilizado como entrada.

Para a métrica F-score, os algoritmos ANN, DT e SL não apresentaram diferença significativa em seu desempenho independente da entrada utilizada. Por outro lado os modelos J48 e SVM apresentaram desempenho superiores quando foi utilizado o SB como entrada, resposta oposta ao observado nos resultados do RF, onde o WL foi a entrada que levou o algoritmo a apresentar melhor desempenho na classificação (Figura 7).

Figura 7. Boxplot para F-score considerando a interação significativa entre os algoritmos de ML e as entradas testadas.



Letras maiúsculas sobre os boxplot's das diferentes entradas divididos por algoritmos de ML não diferem entre si pelo teste Scott-Knott a 5% de probabilidade.

Fonte: elaboração própria

Observando os resultados obtidos para as duas métricas é possível concluir que ANN, DT, J48 e RF são constantes em seu desempenho, tendo resposta semelhante para ambas as métricas de

acuracia. O algoritmo SL apresentou as maiores médias em ambos os métodos utilizados como critério de acuracia, quando a entrada utilizada foi o WL e que o SVM apresentou a menor amplitude nos resultados utilizando o SB como entrada tanto para o CC quanto para o F-score, indicando alta acuracia nos resultados. DT apresentou resultados consistentes independente da entrada ou métrica de acuracia utilizado. SL e SVM apresentaram alternância na superioridade de desempenho conforme a entrada e a métrica de acuracia utilizada. As diferenças observadas entre os clusters destacam a necessidade de estratégias personalizadas para o manejo de diferentes genótipos, visando maximizar a produtividade e a sustentabilidade da cultura da soja.

4. DISCUSSÃO

Conhecer a resposta fisiológica dos genótipos ao ambiente, principalmente durante as fases iniciais do programa de melhoramento pode economizar tempo, investimento e esforços dos melhoristas, pois com base em estudo prévio do ambiente em que será inserida a variedade alvo do programa os genótipos podem ser classificados e selecionados para continuar ou não nas fases seguintes de seleção. Essas informações poderiam levar os programas de melhoramento mais rapidamente a variedades com resistência à seca, estresse térmico e tolerância a altos teores salinos por exemplo (BELLALLOUI et al., 2013; VAN DER LAAN, 2024; XU et al., 2019; TOUM et al., 2022).

Os resultados obtidos nesse estudo reforçam a complexidade das interações entre as variáveis fisiológicas e as características espectrais dos grupos formados, e submetidos a diferentes algoritmos de aprendizagem de máquinas. O agrupamento foi possível devido às diferentes respostas dos genótipos ao ambiente, reforçando o que foi observado por Steiner et. al. (2022), que observaram que é possível fazer a discriminação de cultivares de soja com base na eficiência da absorção de nutrientes. Essas diferenças refletem a variabilidade genética e fisiológica presente nos genótipos de soja. Essa variação pode ser observada na anatomia e concentração de pigmentos presentes na folha de diversas culturas e sua interação com a luz em diferentes genótipos conforme observado no estudo de Falcione et. al. (2023).

O fato de não haver diferença estatística entre os grupos para a variável fotossíntese líquida, ainda que as médias tenham sido superiores no Cluster 1, corroboram com o estudo conduzido por Koester et. al. (2016). Esses autores observaram que a concentração de fotossíntese líquida da soja

não sofreu uma variação significativa ao longo do tempo, mesmo com o trabalho dos programas de melhoramento que buscam aumentar o potencial fotossintético da cultura.

A análise espectral apontou que os genótipos do Cluster 2 obtiveram valores de reflectância maiores do que os encontrados no Cluster 1 em diversas faixas espectrais, especialmente entre 525 e 700 nm, e novamente entre 750 e 825 nm. Essas diferenças na reflectância podem estar associadas a variações na estrutura foliar, conteúdo de clorofila e outros pigmentos (CRUSSIOL et al., 2013). Podem ainda estar ligadas a capacidade dos genótipos a resposta ao ambiente conforme descrito no estudo conduzido por Crusiol et al., (2021) que classificou os genótipos de soja sob diferentes níveis de disponibilidade hídrica e estágio fenológico utilizando os dados de reflectância hiperespectral.

A correlação de Pearson indicou que as interações entre variáveis fisiológicas e bandas espectrais são complexas e diferem entre os grupos. No Cluster 1, houve correlações positivas de moderadas a altas entre as variáveis fisiológicas, enquanto no Cluster 2 essas correlações foram majoritariamente baixas ou inexistentes. Sugerindo que os genótipos do Cluster 2 apresentam mecanismos fisiológicos e bioquímicos diferentes dos encontrados nos genótipos do Cluster 1 (ZUBAIDAH et. al., 2023), influenciando e diferindo sua resposta espectral. A complexidade dessas interações faz com que os métodos tradicionais não sejam os mais indicados para demonstrar as nuances que separam os genótipos, uma opção é fazer uso do aprendizado de máquinas para ultrapassar essa barreira.

A utilização de diferentes algoritmos de aprendizagem de máquinas revelou diferenças significativas na capacidade de classificação dos modelos com base nos genótipos e suas especificidades, corroborando com estudos que indicam que algoritmos de aprendizados de máquina podem classificar com eficiência genótipos de soja baseado em diferentes caracteres, oferecendo informações relevantes para os programas de melhoramento genético da soja. Gava et al., (2022) utilizaram técnicas de sensoriamento remoto associados a algoritmos de aprendizado de máquinas para identificar com sucesso diferentes cultivares de soja. O algoritmo de regressão logística (SL) e o random forest (RF) apresentaram melhor desempenho quando o input utilizado foi o comprimento de onda completo (WL).

Esses modelos são eficazes quando se utiliza dados com alta dimensionalidade e complexidade de maneira rápida e confiável, aumentando a precisão de forma não subjetiva. (YOOSEFZADEH-NAJAFABADI et al. 2021; DE MEDEIROS et al., 2020; KHAN et al., 2022). Por outro lado, o

algoritmo J48 mostrou melhor desempenho com as médias espectrais das bandas selecionadas (SB), indicando sua capacidade de simplificação e interpretação de dados reduzidos. Santana (2024) classificou com sucesso genótipos de soja com base em sua eficiência na absorção de nutrientes utilizando modelos de aprendizados de máquinas, destacando a precisão alcançada quando se utiliza algoritmos como máquina de vetores de suporte e J48.

A utilização de métricas de acurácia, como porcentagem de classificações corretas (CC) e F-score, proporcionou uma avaliação robusta do desempenho dos modelos de ML, elas ajudam a entender e identificar os pontos fortes e fracos dos modelos. Os resultados observados demonstraram que alguns algoritmos, como ANN e DT, apresentam desempenho constante independentemente da entrada, outros, como SL e SVM são mais sensíveis a seleção da métrica de desempenho e input utilizado (AWAD & KHANNA, 2015). Este comportamento reforça a importância de selecionar os algoritmos adequados ao conjunto de dados que se busca trabalhar e as entradas com melhor desempenho para tarefas específicas de classificação (VENUGOPAL et al., 2022; SALIH & ABDULAZEEZ, 2021; ADIO, 2024).

Em pesquisas futuras poderiam ser exploradas as aplicações de outros métodos de aprendizagem de máquinas, que atuando com grande número de dados coletados a campo têm mostrado resultados promissores em outras áreas da agricultura de precisão (RIBEIRO et.al., 2024; ED-DAOURI et. al., 2023;). Estudos adicionais poderiam ainda investigar a influência dos fatores ambientais, como estresse hídrico e nutricional, sobre a interação entre as variáveis fisiológicas e espectrais dos genótipos de soja, haja visto que o estudo conduzido por Monteiro et. al., (2021) aponta que esses fatores impactam fortemente a qualidade fisiológica das sementes de soja.

5. CONCLUSÃO

Os resultados deste estudo demonstraram que a aplicação de algoritmos de aprendizado de máquina permite distinguir e classificar genótipos de soja com base em suas respostas fisiológicas, a partir de características espectrais. Entre os modelos testados, a Máquina de Vetor de Suporte (SVM) e a Regressão Logística (SL) apresentaram as maiores precisões de classificação. Além disso, não foram observadas diferenças significativas entre as diferentes formas de entrada dos dados (comprimentos de onda e bandas espectrais), indicando que ambas as abordagens preservam informações relevantes para a modelagem.

REFERÊNCIAS

- AHMED, Mohiuddin; SERAJ, Raihan; ISLAM, Syed Mohammed Shamsul. The k-means algorithm: A comprehensive survey and performance evaluation. **Electronics**, Suíça v. 9, n. 8, p. 1295, 2020. <https://doi.org/10.3390/electronics9081295>
- ANG, Kenneth Li-Minn; SENG, Jasmine Kah Phooi. Big data and machine learning with hyperspectral information in agriculture. **IEEE Access**, United States, v. 9, p. 36699-36718, 2021. <https://doi.org/10.1109/ACCESS.2021.3051196>
- ADIO, M. O., Ogunwuyi, O. J., Oyediran, M. O., Esan, A. O., & Adedapo, O. A. (2024). Performance Evaluation of Some Selected Classification Algorithms in a Facial Recognition System. **ABUAD Journal of Engineering Research and Development**, Nigéria, v. 7, n.1, p.169-177. <https://doi.org/10.53982/ajerd>
- BATISTA, E. C., Villela, G. M., Pires, R. M. D. O., Santos, H. O. D., Carvalho, E. R., & Bruzi, A. T. (2022). Physiological quality of soybean seeds and the influence of maturity group. **Journal of Seed Science**, Londrina, v. 44, p. e202244026. <https://doi.org/10.1590/2317-1545v44261325>
- BELLALLOUI, N., Gillen, A. M., Mengistu, A., Kebede, H., Fisher, D. K., Smith, J. R., & Reddy, K. N. (2013). Responses of nitrogen metabolism and seed nutrition to drought stress in soybean genotypes differing in slow-wilting phenotype. **Frontiers in Plant Science**, Suíça, v. 4, p. 498. <https://doi.org/10.3389/fpls.2013.00498>
- Chen, J., Gao, J., Wang, Q., Tan, X., Li, S., Chen, P., ... & Yang, W. (2022). Blue-light-dependent stomatal density and specific leaf weight coordinate to promote gas exchange of soybean leaves. **Agriculture**, Suíça, v.13, n. 1, p. 119. <https://doi.org/10.3390/agriculture13010119>
- COTRIM, M. F., Gava, R., Campos, C. N. S., de David, C. H. O., Reis, I. D. A., Teodoro, L. P. R., & Teodoro, P. E. (2021). Physiological performance of soybean genotypes grown under irrigated and rainfed conditions. **Journal of Agronomy and Crop Science**, [S.l.], v. 207, n. 1, p. 34-43. <https://doi.org/10.1111/jac.12448>
- COMPANHIA NACIONAL DE ABASTECIMENTO – CONAB. **Acompanhamento da safra brasileira: grãos, v. 11 - safra 2023/2024, n. 6 - sexto levantamento**. Disponível em: <https://www.conab.gov.br/info-agro/safras>. Acesso em: 14 ago. 2024.
- CRUSIOL, L. G. T.; NEIVERTH, W.; RIO, A.; SIBALDELLI, R. N. R.; FERREIRA, L. C.; CARVALHO, J. F. C.; NEPOMUCENO, A. L.; NEUMAIER, N.; FARIAS, J. R. B. NDVI de estádios de desenvolvimento da soja BRS 284 em condições de campo. In: **Jornada Acadêmica Da Embrapa Soja**, 8., 2014, Londrina. Anais... Londrina: Embrapa Soja, 2014. p. 88–91. Disponível em: <https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/964365/1/NDVIdeestadiosdedesenvolvimentoedasojabrs284emcondicoesdecampo.pdf>. Acesso em: 14 de agosto de 2024.

CRUSIOL, L. G. T., Nanni, M. R., Furlanetto, R. H., Sibaldelli, R. N. R., Cezar, E., Sun, L., ... & Farias, J. R. B. (2021). Classification of soybean genotypes assessed under different water availability and at different phenological stages using leaf-based hyperspectral reflectance. **Remote Sensing**, Suíça, v. 13, n. 2, p. 172. <https://doi.org/10.3390/rs13020172>

DA SILVA JUNIOR, C. A., Nanni, M. R., Shakir, M., Teodoro, P. E., de Oliveira-Júnior, J. F., Cezar, E., ... & Shiratsuchi, L. S. (2018). Soybean varieties discrimination using non-imaging hyperspectral sensor. **Infrared Physics & Technology**, Oxford, Engalnd, v. 89, p. 338-350. <https://doi.org/10.1016/j.infrared.2018.01.027>

DE MEDEIROS, A. D., Capobianco, N. P., da Silva, J. M., da Silva, L. J., da Silva, C. B., & dos Santos Dias, D. C. F. (2020). Interactive machine learning for soybean seed and seedling quality classification. **Scientific reports**. [S.l], v. 10, n. 1, p. 11267. <https://doi.org/10.1038/s41598-020-68273-y>

DENG, Lingli. Current progress in the utilization of soy-based emulsifiers in food applications—A Review. **Foods**, Suíça, v. 10, n. 6, p. 1354, 2021. <https://doi.org/10.3390/foods10061354>

EGMONT-PETERSEN, M., de Ridder, D., Handels, H., 2002. Image processing with neural networks—a review. **Pattern Recogn.** Oxford, Engalnd, v. 35, p. 2279–2301. [https://doi.org/10.1016/S0031-3203\(01\)00178-9](https://doi.org/10.1016/S0031-3203(01)00178-9).

ED-DAOUDI, R., Alaoui, A., Ettaki, B., & Zerouaoui, J. (2023). A machine learning approach to identify optimal cultivation practices for sustainable apple production in precision agriculture in Morocco. In E3S Web of Conferences (Vol. 469, p. 00052). **EDP Sciences**. Paris. <https://doi.org/10.1051/e3sconf/202346900052>

FALCIONI, R., Moriwaki, T., Pattaro, M., Furlanetto, R. H., Nanni, M. R., & Antunes, W. C. (2020). High resolution leaf spectral signature as a tool for foliar pigment estimation displaying potential for species differentiation. **Journal of plant physiology**, Oxford, Engalnd, v. 249, p. 153161. <https://doi.org/10.1016/j.jplph.2020.153161>

GAVA, R., Santana, D. C., Cotrim, M. F., Rossi, F. S., Teodoro, L. P. R., da Silva Junior, C. A., & Teodoro, P. E. (2022). Soybean cultivars identification using remotely sensed image and machine learning models. **Sustainability**, Suíça, p. 14, n. 12, p. 7125. <https://doi.org/10.3390/su14127125>

HERRITT, Matthew T.; FRITSCHI, Felix B. Characterization of photosynthetic phenotypes and chloroplast ultrastructural changes of soybean (*Glycine max*) in response to elevated air temperatures. **Frontiers in plant science**, Suíça, v. 11, p. 153, 2020. <https://doi.org/10.3389/fpls.2020.00153>

HOU, M., Tian, F., Ortega-Farias, S., Riveros-Burgos, C., Zhang, T., & Lin, A. (2021). Estimation of crop transpiration and its scale effect based on ground and UAV thermal infrared remote sensing images. **European Journal of Agronomy**, Oxford, Engalnd, v. 131, p. 126389. <https://doi.org/10.1016/j.eja.2021.126389>

HUDSON, Karen. Soybean protein and oil variants identified through a forward genetic screen for seed composition. **Plants**, Suíça, v. 11, n. 21, p. 2966, 2022. <https://doi.org/10.3390/plants11212966>

KALMEGH, S. (2015). Analysis of weka data mining algorithm reptime, simple cart and randomtree for classification of indian news. **Engineering & Technology**, Suíça v. 2, n. 2, p. 438-446. ISSN: 2348-7968

KHAN, A., Vibhute, A. D., Mali, S., & Patil, C. H. (2022). A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications. **Ecological Informatics**, Oxford, Engalnd, v. 69, p. 101678. <https://doi.org/10.1016/j.ecoinf.2022.101678>

Koester, R. P., Nohl, B. M., Diers, B. W., & Ainsworth, E. A. (2016). Has photosynthetic capacity increased with 80 years of soybean breeding? An examination of historical soybean cultivars. **Plant, cell & environment**, United States, v. 39, n. 5, p. 1058-1067. <https://doi.org/10.1111/pce.12675>

KOUR, A.; KAUR, N.; DHADWAL, A. Comprehensive review on soy protein: Health benefits and utilization in food industry. **The Pharma Innovation Journal**, United States, v. 12, p. 2469-2475, 2023. <https://doi.org/10.22271/tpi.2023.v12.i5ad.20225>

LESTARI, Puji; SUTRISNO, Sutrisno; TASMA, I. Made. QTL study to reveal soybean response on abiotic and biotic stresses. **Journal AgroBiogen**, v. 10, n. 3, p. 109-114, 2014. <https://doi.org/10.21082/jbio.v10n3.2014.p109-114>

LIU, Bin; GAO, Pengxiang; ZHANG, Shuoxin. Photosynthesis and Fine Root Growth Dynamics of Soybean in Walnut-Soybean Agroforestry System. **Computational Intelligence and Neuroscience**, France, v. 2022, n. 1, p. 2246824, 2022. <https://doi.org/10.1155/2022/2246824>

LIAKOS, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. **Sensors**, Suíça, v. 18, n. 8, p. 2674. <https://doi.org/10.3390/s18082674>

MAHMOOD, A., Tiwari, A. K., Singh, S. K., & Udmale, S. S. (2022). Contemporary machine learning applications in agriculture: Quo Vadis?. **Concurrency and Computation: Practice and Experience**. [S.l.], v. 34, n.15, p. e6940. <https://doi.org/10.1002/cpe.6940>

MONTEIRO, F. F., Bruzi, A. T., Silva, K. B., Pulcinelli, C. E., Soares, I. O., & Carvalho, M. L. M. (2021). Breeding for yield and seed quality in soybean. **Euphytica**, Suíça, v. 217, n. 12, p. 212. <https://doi.org/10.1007/s10681-021-02943-4>

NALEPA, J., Kawulok, M., 2019. Selecting training sets for support vector machines: a review. **Artif. Intell. Holland**, v. 52, p. 857–900. <https://doi.org/10.1007/s10462-017-9611-1>

NGUYEN, T. C., Jo, H., Tran, H. A., Lee, J., Lee, J. D., Kim, J. H., ... & Song, J. T. (2024). Assessment of drought responses of wild soybean accessions at different growth stages. **Agronomy**, Suíça, v. 14, n. 3, p. 471. <https://doi.org/10.3390/agronomy14030471>

QUINLAN, J. Ross. C4. 5: programs for machine learning. **Elsevier**, Suíça, p. 2014. ISBN: 1-55860-238-0

RAHMAN, T., Ye, L., Liu, X., Iqbal, N., Du, J., Gao, R., ... & Yang, W. (2017). Water use efficiency and water distribution response to different planting patterns in maize–soybean relay strip intercropping systems. **Experimental Agriculture**, England, v. 53, n. 2, p. 159-177. <https://doi.org/10.1017/S0014479716000260>

RASHEED, A., Raza, A., Jie, H., Mahmood, A., Ma, Y., Zhao, L., ... & Jie, Y. (2022). Molecular tools and their applications in developing salt-tolerant soybean (*Glycine max* L.) cultivars. **Bioengineering**, Suíça, v. 9, n. 10, p. 495. <https://doi.org/10.3390/bioengineering9100495>

Ribeiro, J., Mileski, M., Junior, J. L. S., Carvalho, L. F., & Mantovani, R. G. (2024). Image Classification for Precision Agriculture: A Coffee Study Case. **Anais do Computer on the Beach**, 15, 088-095. <https://doi.org/10.14210/cotb.v15.p088-095>

TEAM, R.C. R: **A Language and Environment for Statistical Computing**. 2013.

SANTANA, Dthenifer Cordeiro., de Oliveira, I. C., Cavalheiro, S. B., das Chagas, P. H. M., Teixeira Filho, M. C. M., Della-Silva, J. L., ... & Teodoro, P. E. (2024). Classification of Soybean Genotypes as to Calcium, Magnesium, and Sulfur Content Using Machine Learning Models and UAV–Multispectral Sensor. **AgriEngineering**, Suíça, v. 6, n. 2, p. 1581-1593. <https://doi.org/10.3390/agriengineering6020090>

SANTANA, Dthenifer Cordeiro. Fenotipagem de alta precisão de caracteres industriais de soja usando aprendizagem de máquina e diferentes sensores espectrais. Ilha solteira: **Universidade Estadual Paulista (Unesp)**, 4 ago. 2023.

SALIH, Azar Abid; ABDULAZEEZ, Adnan Mohsin. Evaluation of classification algorithms for intrusion detection system: A review. **Journal of Soft Computing and Data Mining**, Malesia, v. 2, n. 1, p. 31-40, 2021. <https://doi.org/10.30880/jscdm.2021.02.01.004>

SHEKOOFA, A., Moser, V., Dhakal, K., Poudel, I., & Pantalone, V. (2023). Inheritance of early stomatal closure trait in soybean: Ellis× N09-13890 population. **Plants**, suíça, v. 12, n. 18, p. 3227. <https://doi.org/10.3390/plants12183227>

SILVA, K. B., Bruzi, A. T., Zambiazzi, E. V., Soares, I. O., Pereira, J. L. A. R., & Carvalho, M. L. M. (2017). Adaptability and stability of soybean cultivars for grain yield and seed quality. **Genetics and Molecular Research**, Ribeirão Preto, v. 16, n. 2, p. 1-15. <http://dx.doi.org/10.4238/gmr16029646>

SIEBERS, M. H., Yendrek, C. R., Drag, D., Locke, A. M., Rios Acosta, L., Leakey, A. D., ... & Ort, D. R. (2015). Heat waves imposed during early pod development in soybean (*Glycine max*) cause significant yield loss despite a rapid recovery from oxidative stress. **Global change biology**, Reino Unido, v. 21, n. 8, p. 3114-3125. <https://doi.org/10.1111/gcb.12935>

SUN, Y. W., Wang, X. Y., Liu, L., Zhang, Q., Xi, Y. J., & Wang, P. W. (2023). Cloning and functional study of GmRPI2, which is the critical gene of photosynthesis in soybean. **Breeding Science**, 73(3), 290-299. <https://doi.org/10.1270/jsbbs.23002>

STEINER, F., da Rosa, L., Zuffo, A. M., & Bardivieso, D. M. (2022). Classificação de genótipos de soja quanto a eficiência e resposta ao uso de fósforo em solo arenoso do cerrado brasileiro. **Revista de Ciências Agrárias**, Portugal, v. 45, n. 3, p. 94-104. <https://doi.org/10.19084/rca.27469>

SNOUSY, M.B. al, El-Deeb, H.M., Badran, K., Khilil, I.A. al, 2011. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egypt. Inf. J.** Egypt, v. 12, p. 73–82. <https://doi.org/10.1016/j.eij.2011.04.003>.

ŠTEPANOVSÝ, M., Ibrov, A., Buk, Z., Velemnsk, J., 2017. Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. **Forensic Science International (FSI)**. Holland, v. 279, p. 72–82. <https://doi.org/10.1016/j.forsciint.2017.08.005>.

SCOTT, A.J.; KNOTT, M.A.A. Cluster analysis method for grouping means in the analysis of variance. **Biometrics**, United States, v.30, p.507-512, 1974. <https://doi.org/10.2307/2529204>

TANAKA, Y.; FUJII, K.; SHIRAIWA, T. Variability of leaf morphology and stomatal conductance in soybean [Glycine max (L.) Merr.] cultivars. **Crop Science**, v. 50, n. 6, p. 2525-2532, 2010. <https://doi.org/10.2135/cropsci2010.02.0058>

TEAM, R. C. R: **A language and environment for statistical computing**. 2013.

TEODORO, L. P. R., Bhering, L. L., Gomes, B. E. L., Campos, C. N. S., Baio, F. H. R., Gava, R., ... & Teodoro, P. E. (2019). Understanding the combining ability for physiological traits in soybean. **PloS one**, United States, v.14, n. 12, p. e0226523. <https://doi.org/10.1371/journal.pone.0226523>

TOUM, L., Perez-Borroto, L. S., Pena-Malavera, A. N., Luque, C., Welin, B., Berenstein, A., ... & Pardo, E. M. (2022). Selecting putative drought-tolerance markers in two contrasting soybeans. **Scientific Reports**, Reino Unido, v. 12, n. 1, p. 10872. <https://doi.org/10.1038/s41598-022-14334-3>

UNITED STATES DEPARTMENT OF AGRICULTURE – USDA. **Annual Crop Production Report: 2024**. Disponvel em: <https://www.fas.usda.gov>. Acesso em: 14 ago. 2024.

VAN DER LAAN, Liza; PEIXOTO, Leonardo de Azevedo; SINGH, Asheesh K. Genetic Dissection of Heat Stress Tolerance in Soybean through Genome-Wide Association Studies and the Use of Genomic Prediction to Enhance Breeding Applications. **bioRxiv**, p. 2024.04.27.591454, 2024. <https://doi.org/10.1038/s44383-025-00010-8>

VAN DER LAAN, Liza., Parmley, K., Saadati, M., Pacin, H. T., Panthulugiri, S., Sarkar, S., ... & Singh, A. K. (2025). Genomic and phenomic prediction for soybean seed yield, protein, and oil. **The Plant Genome**, 18(1), e70002. <https://doi.org/10.1002/tpg2.70002>

VENUGOPAL, V. K., Takhar, R., Gupta, S., & Mahajan, V. (2022). Clinical Explainability Failure (CEF) & Explainability Failure Ratio (EFR)—changing the way we validate classification algorithms. **Journal of Medical Systems**, United States, v. 46, n. 4, p. 20. <https://doi.org/10.1007/s10916-022-01806-2>

WALLY-VALLIM, A. P., Vanier, N. L., Zavareze, E. D. R., Zambiasi, R. C., de Castro, L. A. S., Schirmer, M. A., & Elias, M. C. (2014). Isoflavone aglycone content and the thermal, functional, and structural properties of soy protein isolates prepared from hydrothermally treated soybeans. **Journal of Food Science**, United States, v. 79, n. 7, p. E1351-E1358. <https://doi.org/10.1111/1750-3841.12506>

XU, C., Xia, Z., Huang, Z., Xia, C., Huang, J., Zha, M., ... & Zhang, C. (2019). Understanding the physiological and transcriptional mechanism of reproductive stage soybean in response to heat stress. **Crop Breeding, Genetics and Genomics**, China, v. 2, n. 1. <https://doi.org/10.20900/cbgg20200004>

ZHANG, M., Liu, Y., Cai, H., Guo, M., Chai, M., She, Z., ... & Qin, Y. (2020). The bZIP transcription factor GmbZIP15 negatively regulates salt-and drought-stress responses in soybean. **International journal of molecular sciences**, Suíça, v. 21, n. 20, p. 7778. <https://doi.org/10.3390/ijms21207778>

ZUBAIDAH, SITI; MUJTAHIDA, IMAMA; KUSWANTORO, HERU. Response of morphological, anatomical, and agronomic characteristics of soybean genotypes to Cowpea Mild Mottle Virus. **Biodiversitas Journal of Biological Diversity**, Indonésia, v. 24, n. 7, 2023. <https://doi.org/10.13057/biodiv/d240739>

YOUSEFZADEH-NAJAFABADI, M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. **Frontiers in plant science**, Suíça, v. 11, p. 624273. <https://doi.org/10.3389/fpls.2020.624273>

CAPÍTULO 2 - USO DE SENSOR VIS-NIR-SWIR PARA CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO AO SEU DESEMPENHO FISIOLÓGICO

Resumo

A soja (*Glycine max* (L.) Merrill) é uma cultura de grande relevância para a segurança alimentar e a economia agrícola, destacando-se pela ampla variabilidade genética e pela capacidade de adaptação a diferentes ambientes. Essa diversidade resulta em respostas fisiológicas distintas entre os genótipos, dificultando a seleção precoce de materiais com potencial agrônomo. Neste estudo, foram avaliadas variáveis fisiológicas e espectrais de 32 populações F₃ de soja, em dois anos (2023/23 e 2023/24), cultivadas em campo, com o objetivo de classificar os genótipos com base em sua resposta fisiológica utilizando algoritmos de aprendizado de máquina. As variáveis fisiológicas fotossíntese líquida (A), concentração interna de CO₂ (C_i), condutância estomática (g_s), eficiência do uso da água (WUE) e transpiração (E) e transpiração (E), foram mensuradas aos 60 dias após a emergência com analisador portátil de fotossíntese (IRGA), enquanto as variáveis espectrais foram obtidas por espectrorradiômetro na faixa de 350 a 2500 nm. As bandas espectrais foram definidas com base em critérios fisiológicos e referências da literatura, totalizando doze intervalos não sobrepostos. Os dados foram submetidos à classificação por seis algoritmos de aprendizado de máquina: redes neurais artificiais (ANN), máquina de vetor suporte (SVM), J48, REPTree (RT), floresta aleatória (RF) e regressão logística (LR). A predição foi realizada por validação cruzada estratificada (k-fold = 10, com dez repetições), e os modelos foram avaliados pelas métricas de acurácia (%CC), F-score e coeficiente Kappa. Os resultados indicaram que o espectro contínuo apresentou desempenho superior às bandas agregadas, e que os algoritmos J48 e REPTree foram os mais eficientes na classificação dos genótipos, seguidos por ANN e SVM. Os modelos RF e LR apresentaram menor desempenho em todas as métricas. A abordagem proposta demonstrou ser eficaz na discriminação de genótipos com diferentes estratégias fisiológicas, evidenciando o potencial da espectroscopia hiperespectral associada ao aprendizado de máquina como ferramenta de suporte à fenotipagem e à seleção em programas de melhoramento genético da soja.

Palavras-Chave: *Glycine max* (L.) Merrill; espectroscopia hiperespectral; fenotipagem fisiológica; aprendizado de máquina; classificação de genótipos.

ABSTRACT

Soybean is a crop of major importance for global food security and agricultural economy, standing out for its broad genetic variability and adaptability to diverse environments. This diversity leads to distinct physiological responses among genotypes, making early selection of promising materials a challenge. In this study, physiological and spectral variables were evaluated in 32 F₃ soybean populations grown under field conditions, aiming to classify genotypes based on their physiological responses using machine learning algorithms. Physiological traits were measured at 60 days after emergence using a portable photosynthesis analyzer (IRGA), while spectral data were collected with a spectroradiometer covering the 350–2500 nm range. Spectral bands were defined based on physiological criteria and literature references, resulting in twelve non-overlapping intervals. Data were classified using six machine learning algorithms: ANN, SVM, J48, REPTree, M5P, and RF. Predictions were performed using stratified k-fold cross-validation ($k = 10$, with ten repetitions), and model performance was evaluated using accuracy (%CC), F-score, and Kappa coefficient. Results showed that the full spectrum input outperformed aggregated bands, and that J48 and REPTree were the most effective algorithms, followed by ANN and SVM. RF and LR showed the lowest performance across all metrics. The proposed approach proved effective in distinguishing genotypes with contrasting physiological strategies, highlighting the potential of hyperspectral reflectance combined with machine learning as a support tool for phenotyping and selection in soybean breeding programs.

Keywords: *Glycine max* (L.) Merrill; hyperspectral reflectance; physiological phenotyping; machine learning; genotype classification.

1. INTRODUÇÃO

A soja (*Glycine max* (L.) Merrill) é uma das principais culturas agrícolas do mundo, desempenhando papel estratégico na segurança alimentar global e na economia de diversos países (CONAB, 2024). Sua ampla variabilidade genética e elevada capacidade de adaptação a diferentes condições edafoclimáticas tornam o cultivo versátil, porém desafiador no que diz respeito à seleção precoce de genótipos com potencial agrônomo. Essa diversidade genética resulta em respostas fisiológicas distintas entre os materiais, dificultando a identificação eficiente de genótipos adaptados a ambientes específicos (YU et al., 2022; SARIC et al., 2022).

Nesse contexto, tecnologias que permitam caracterizar com precisão e rapidez os atributos fisiológicos das plantas têm ganhado relevância. A espectroscopia hiperespectral destaca-se como uma dessas ferramentas, por sua capacidade de gerar dados em tempo reduzido, de forma não destrutiva e com crescente acessibilidade no campo da agricultura de precisão (FALCIONI et al., 2020; DALE et al., 2013). Essa técnica permite mensurar características bioquímicas e estruturais das folhas, como teor de clorofila, conteúdo hídrico e integridade celular, especialmente nas faixas do visível (VIS: 400–700 nm) e do infravermelho próximo (NIR: 700–1100 nm), que são sensíveis a variações fisiológicas (SHI et al., 2023; ADEGOYE et al., 2023).

Entretanto, o grande volume de dados gerados por sensores hiperespectrais exige métodos analíticos robustos, capazes de lidar com alta dimensionalidade e complexidade. As abordagens tradicionais de estatística multivariada, embora úteis, apresentam limitações frente à variabilidade espectral e fisiológica observada em campo. Nesse cenário, algoritmos de aprendizado de máquina (ML) têm se mostrado eficazes na classificação de genótipos e na fenotipagem de alto rendimento, utilizando dados espectrais como base (LIAKOS et al., 2018; YOOSEFZADEH-NAJAFABADI et al., 2021).

Na cultura da soja, algoritmos como Support Vector Machines (SVM), Regressão Logística (LR), Árvores de Decisão (J48, M5P), Random Forest (RF) e Redes Neurais Artificiais (ANN) têm sido aplicados com sucesso em tarefas de classificação e predição, apresentando desempenho variável conforme os dados de entrada utilizados (SNOUSY et al., 2011; BELGIU; DRĂGUȚ, 2016; SANTANA et al., 2023). A escolha entre espectro contínuo ou bandas espectrais agregadas influencia diretamente a acurácia dos modelos, assim como a seleção das métricas de avaliação, como F-score e coeficiente Kappa, que oferecem uma visão mais abrangente da performance dos algoritmos (BENOS et al., 2021; NAIDU et al., 2023).

A espectrometria aplicada à classificação de genótipos tem se consolidado como alternativa promissora para reduzir a subjetividade das avaliações, acelerar os processos de triagem e ampliar a precisão na seleção de materiais genéticos (SANTANA et al., 2022). Diante disso, este estudo parte da hipótese de que é possível classificar genótipos de soja com base em sua resposta fisiológica, utilizando dados espectrais e algoritmos de aprendizado de máquina.

O objetivo deste trabalho foi identificar os algoritmos de ML e as configurações de entrada que proporcionam maior acurácia na classificação de genótipos de soja, considerando suas características fisiológicas mensuradas em condições de campo.

2. MATERIAL E MÉTODOS

O experimento foi realizado nas safras 2023/2024 e 2024/2025, na área experimental da Universidade Federal de Mato Grosso do Sul, campus Chapadão do Sul (18°41'33''S, 52°40'45''W, altitude de 810 m). O clima da região é classificado como Tropical de Savana (Aw), e o solo corresponde a um Latossolo Vermelho Distrófico de textura argilosa. As características químicas da camada de 0–0,20 m foram: pH (H₂O) = 6,20; Al³⁺ trocável = 0,00 cmolc dm⁻³; Ca²⁺ + Mg²⁺ = 4,31 cmolc dm⁻³; P = 41,30 mg dm⁻³; K⁺ = 0,20 cmolc dm⁻³; matéria orgânica = 19,74 g dm⁻³; V = 45%; m = 0%; soma de bases = 2,30 cmolc dm⁻³; CTC = 5,10 cmolc dm⁻³.

O delineamento experimental foi em blocos casualizados, com quatro repetições e 32 populações F₃ de soja. Cada parcela continha cinco linhas de 1,5 m de comprimento, espaçadas em 0,45 m, com densidade de 15 plantas por metro linear. A semeadura ocorreu em outubro de 2023 e 2024, após preparo convencional do solo com aração e gradagem. As sementes foram tratadas com fungicida (Piraclostrobina + Metil Tiofanato) e inseticida (Fipronil), na proporção de 200 mL por 100 kg de sementes. Para a inoculação biológica, utilizou-se *Bradyrhizobium* spp. na mesma proporção. Os tratos culturais foram realizados conforme as exigências da cultura.

As avaliações foram realizadas aos 60 dias após a emergência (DAE), utilizando analisador portátil de fotossíntese por infravermelho (IRGA), modelo Li-6400XT (LiCor Inc., Lincoln, Nebraska, EUA). As medições ocorreram sob fluxo de fótons fotossinteticamente ativos de 1000 $\mu\text{mol m}^{-2} \text{s}^{-1}$.

As variáveis analisadas foram: fotossíntese líquida (A), transpiração (E), condutância estomática (gs), concentração interna de CO₂ (Ci), e eficiência no uso da água (WUE = A/E). As medições foram realizadas entre 8h e 11h, em três plantas por parcela, em dias com céu claro,

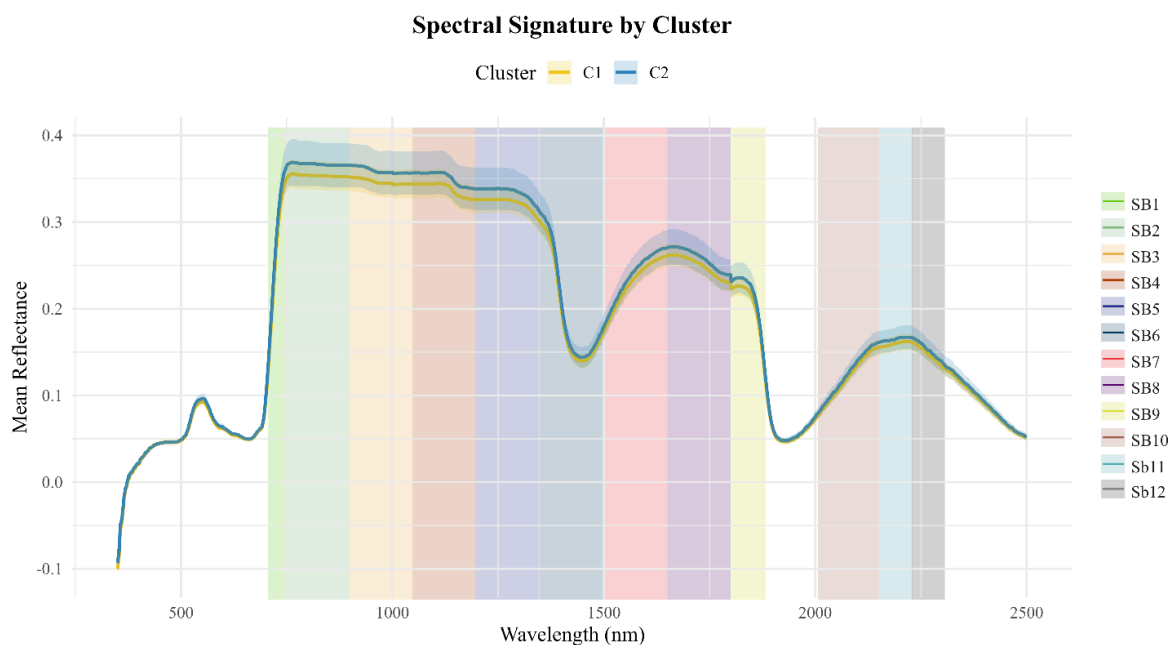
temperatura entre 26,0 e 26,5°C e umidade relativa entre 50% e 80%, conforme Teodoro et al. (2019). A folha utilizada foi a terceira completamente expandida a partir do ápice, considerada padrão para diagnósticos nutricionais.

As mesmas folhas utilizadas nas análises fisiológicas foram avaliadas espectralmente. Foram coletadas três amostras por parcela, medidas com espectrorradiômetro FieldSpec 4 Jr (Analytical Spectral Devices, Boulder, EUA), com faixa de leitura entre 350 e 2500 nm. Utilizou-se o acessório ASD Plant Probe, que permite medições de contato sem interferência da luz ambiente. A calibração foi feita com placa branca de sulfato de bário, repetida ao final das leituras de cada bloco. As leituras foram registradas via software RS3 e exportadas em arquivo com extensão .txt pelo ViewSpectroPro para posterior análise estatística.

Com os dados de comprimentos de onda foi realizada a seleção das bandas espectrais a serem utilizadas foi guiada por critérios fisiológicos, fundamentados nas regiões do espectro eletromagnético que apresentaram maior separação entre os clusters, conforme observado na assinatura espectral média com intervalo de confiança. Adicionalmente, os intervalos foram definidos com base em referências consolidadas da literatura, como Im & Jensen (2009) e da Silva Jr. et al. (2018), que descrevem faixas associadas a componentes bioquímicos e estruturais das folhas. As bandas espectrais foram agrupadas considerando a diferença média suavizada (superior a 0,004), largura mínima de 10 nm e máxima de 150 nm para preservar a capacidade de representatividade do espectro.

A Figura 1 apresenta os intervalos utilizados para cálculo de média de refletância das doze bandas. A segmentação foi baseada em transições conhecidas de reflectância associadas a características fisiológicas das plantas, como a região do Red Edge (701–750 nm), a resposta do mesófilo no infravermelho próximo (750–1050 nm), regiões de absorção de água (1050–1500 nm), e zonas de absorção estrutural e química no SWIR (acima de 1500 nm). A média da reflectância em cada banda foi utilizada como variável de entrada nos modelos de classificação supervisionada.

Figura 1. Assinatura hiperespectral de ambos os clusters formados. Destaque para os comprimentos de onda com diferentes cores mostrando o intervalo contido em cada banda formada.



Fonte: elaboração própria.

Foram identificados doze intervalos espectrais com diferença média suavizada superior a 0,004 e largura mínima de 10 nm: 701–1885 nm e 2068–2308 nm (Tabela 1). Esses intervalos foram divididos em doze bandas espectrais não sobrepostas, respeitando um limite máximo de 150 nm por banda.

Tabela 1. Intervalos de faixa espectral (nm) associados a processos fisiológicos utilizados no estabelecimento das doze bandas espectrais.

BANDA	INÍCIO (NM)	FIM (NM)	RESPOSTA FISIOLÓGICA
B1	701	750	Clorofila (Red Edge)
B2	751	900	Estrutura celular (NIR)
B3	901	1050	Mesófilo
B4	1051	1200	Água (inicial SWIR)

B5	1201	1350	Água + lignina
B6	1351	1500	Água (absorção secundária)
B7	1501	1650	Celulose, lignina
B8	1651	1800	Estrutura vegetal
B9	1801	1885	Transição SWIR
B10	2068	2150	Água e proteína
B11	2151	2230	Lignina e celulose
B12	2231	2308	Estrutura vegetal seca

Os dados foram submetidos à análise por algoritmos de aprendizado de máquina, tendo como variável de saída (output) os atributos fisiológicos. A predição foi realizada por validação cruzada estratificada com k-fold = 10 e dez repetições, totalizando 100 execuções por modelo.

Todos os parâmetros foram mantidos conforme configuração padrão do software Weka 3.8.6. Os algoritmos utilizados foram: redes neurais artificiais (ANN) (EGMONT-PETERSEN et al., 2002), árvore de decisão REPTree (SNOUSY et al., 2011), árvore M5P (BLAIFI et al., 2018), floresta aleatória (RF; BELGIU & DRĂGUȚ, 2016) e máquina de vetor suporte (SVM; NALEPA & KAWULOK, 2019).

Para avaliação dos modelos, foram utilizadas as métricas de acurácia (%CC), F-score e coeficiente Kappa. A significância dos algoritmos e das entradas foi verificada por análise de variância. Quando significativa, os resultados foram agrupados pelo teste de Scott-Knott (SCOTT & KNOTT, 1974), ao nível de 5% de probabilidade. Os gráficos e agrupamentos foram gerados com os pacotes ggplot2 e ExpDes.pt no software RStudio (RSTUDIO TEAM, 2023). As etapas da pesquisa descritas na metodologia estão representadas na Figura 2.

Figura 2. Etapas da pesquisa: A. Obtenção dos valores dos caracteres fisiológicos de 32 genótipos de soja; B. Obtenção dos valores espectrais de amostras dos 32 genótipos F₃ de soja; C. Classificação dos genótipos de soja com base no seu comportamento fisiológico; D. Utilização de algoritmos de aprendizado de máquinas para a classificação dos genótipos utilizando os dados espectrais como entrada e os grupos de genótipos como saída.

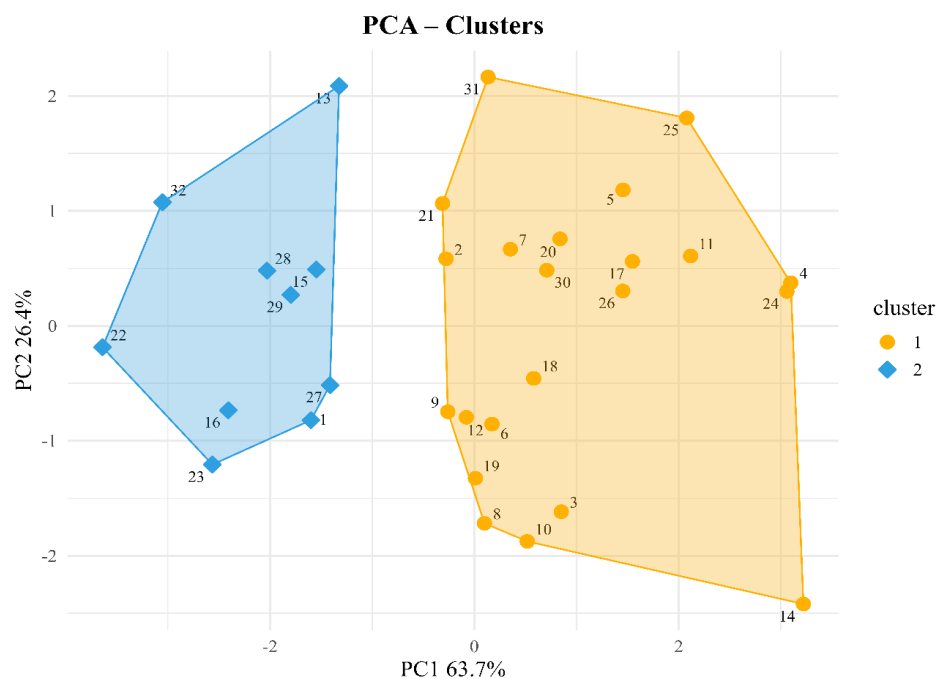


Fonte: elaboração própria.

3. RESULTADOS

A Figura 3 apresenta a projeção bidimensional dos genótipos de soja agrupados em dois grupos (C1 e C2), com base nas variáveis fisiológicas selecionadas. A separação foi obtida por meio da Análise de Componentes Principais (PCA) associada ao algoritmo de agrupamento K-means.

Figura 3. Análise de Componentes Principais (PCA), apresentando a representação bidimensional dos clusters C1 e C2 formados pelo algoritmo k-means.

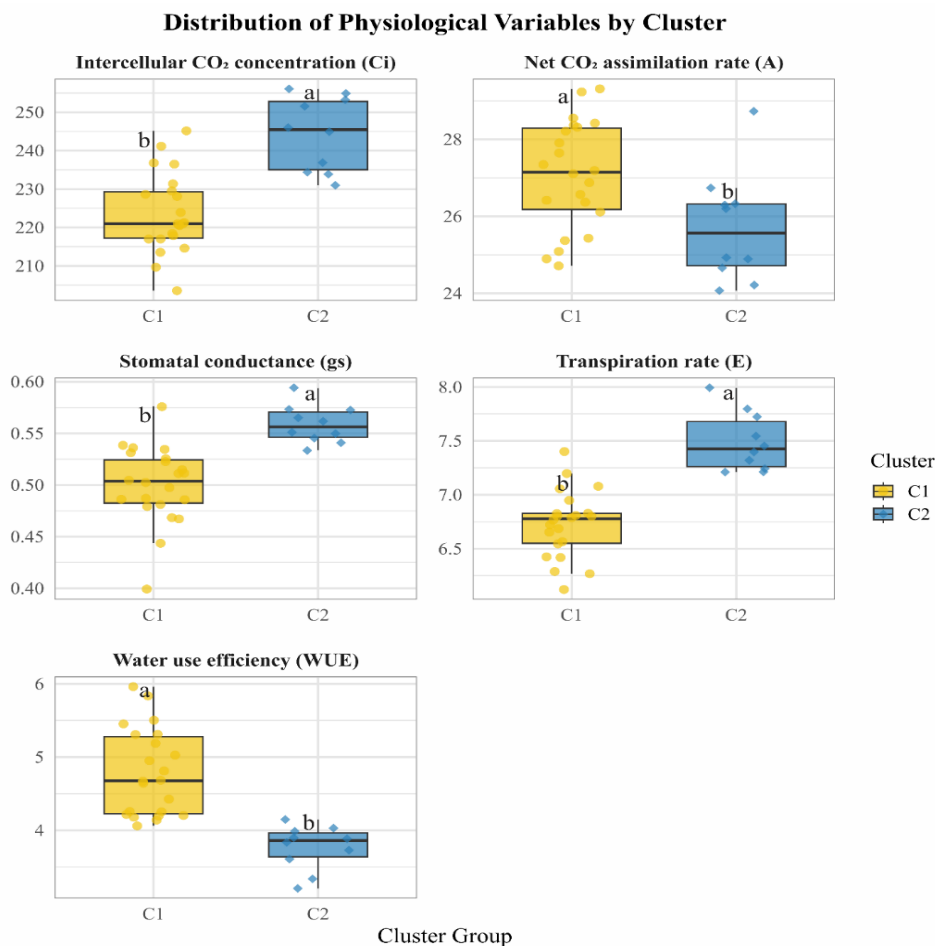


Fonte: elaboração própria.

A formação dos grupos revela padrões fisiológicos distintos. O cluster C1 apresentou maior dispersão ao longo do primeiro componente (PC1 63,7%), sugerindo variabilidade em atributos como fotossíntese líquida (A) e eficiência no uso da água (WUE). Por outro lado, C2 (PC2 26,4%), concentrou-se em uma faixa mais restrita, com destaque para condutância estomática (gs) e taxa de transpiração (E). Esse comportamento demonstram a eficácia da PCA na discriminação de genótipos de soja com desempenho fisiológico contrastante sob estresse moderado.

A Figura 4 mostra os boxplots das variáveis fisiológicas avaliadas por cluster. As diferenças entre os grupos foram estatisticamente significativas para todas as variáveis. O cluster C2 apresentou maiores valores de concentração intercelular de CO₂ (Ci), condutância estomática (gs) e transpiração (E), indicando maior abertura estomática e difusão de gases. Em contraste, C1 destacou-se pela maior taxa de assimilação líquida de CO₂ (A) e maior eficiência no uso da água (WUE), sugerindo uma estratégia voltada à conservação hídrica e ao aproveitamento mais eficiente do carbono fixado.

Figura 4. Distribuição das variáveis fisiológicas C_i , A , g_s , E e WUE entre os grupos de cluster (C1 e C2). Cada boxplot representa a variação dentro de cada variável fisiológica, e as comparações de médias referem-se exclusivamente aos dois clusters dentro de cada gráfico. Os pontos individuais foram deslocados horizontalmente (jitter) para evitar sobreposição e permitir a visualização da densidade de observações em cada grupo. Letras diferentes indicam diferenças estatísticas significativas entre os clusters para a variável correspondente.

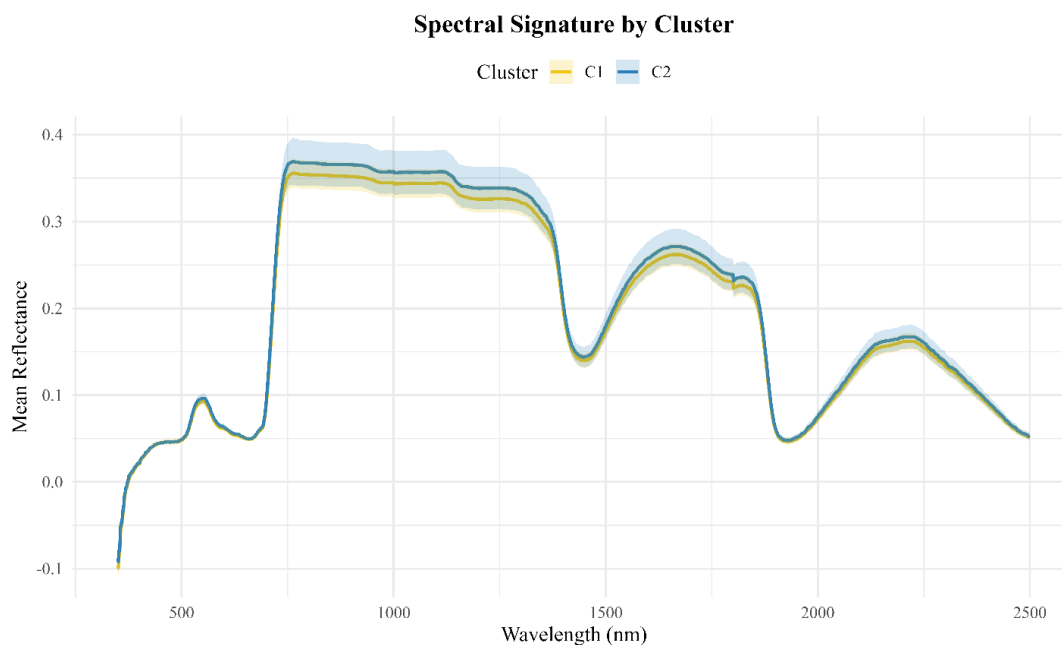


Fonte: elaboração própria.

Esse padrão evidencia um equilíbrio entre intensidade e eficiência no uso dos recursos. Enquanto C2 prioriza trocas gasosas mais intensas, C1 adota uma resposta mais conservadora, com menor transpiração e maior eficiência hídrica. Essa diferenciação tem implicações diretas na seleção de genótipos para programas de melhoramento, especialmente em cenários de estresse hídrico ou ambientes com alta disponibilidade de água.

A Figura 5 apresenta as curvas espectrais médias dos genótipos classificados nos clusters C1 e C2, com bandas de erro ($\pm$ DP), abrangendo a faixa de 350 a 2500 nm. Na região do visível (400–700 nm), observam-se variações consistentes entre os grupos, especialmente entre 540 e 680 nm. A depressão em torno de 680 nm, típica da absorção da clorofila-a, foi mais acentuada em C1, indicando maior concentração de pigmentos fotossintéticos.

Figura 5. Assinaturas espectrais médias dos clusters C1 e C2 ao longo do intervalo VIS-NIR-SWIR (350–2500 nm). As linhas representam a reflectância média de cada cluster em cada comprimento de onda, enquanto as faixas coloridas indicam a variabilidade interna ($\pm$ DP), entre os indivíduos dentro de cada grupo.



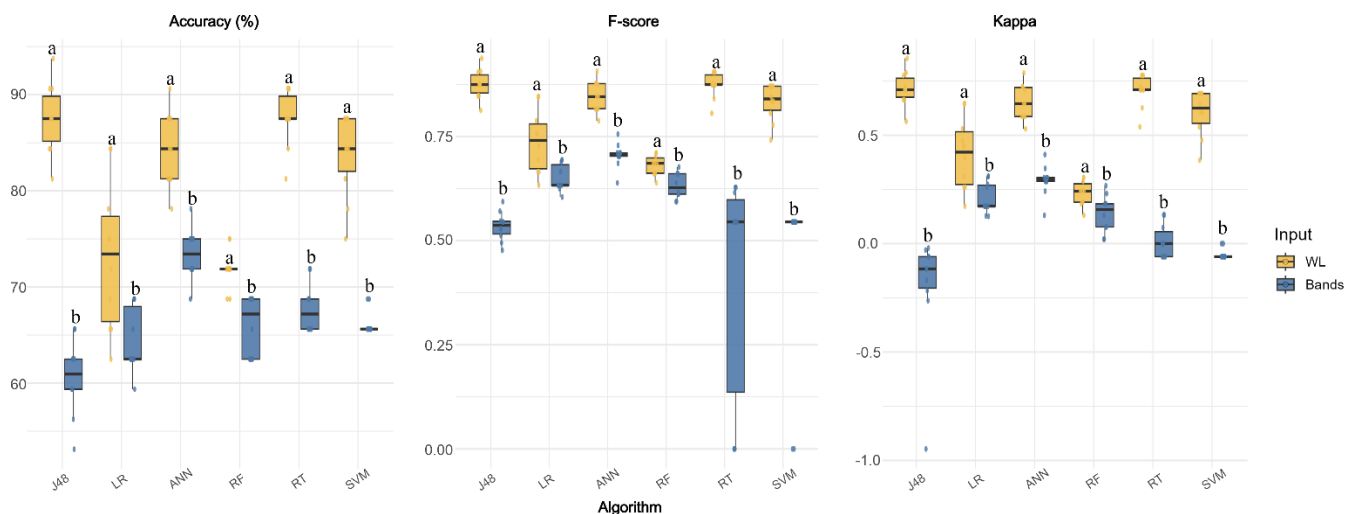
Fonte: elaboração própria.

Na faixa do infravermelho próximo (NIR: 750–1300 nm), sensível à estrutura celular, os valores foram ligeiramente superiores em C2, sugerindo diferenças anatômicas foliares, como maior espessura do mesofilo ou integridade estrutural. Já no infravermelho de ondas curtas (SWIR: 1300–2500 nm), as diferenças foram mais evidentes em bandas associadas à absorção por água e lignina (ex.: 1450, 1950, 2200–2300 nm), indicando variações no teor de água foliar e na

composição estrutural dos tecidos. A análise conjunta dessas regiões espectrais confirma que as diferenças fisiológicas entre os grupos também se expressam ao longo do espectro.

Após o agrupamento, foram realizadas tarefas de classificação utilizando seis algoritmos de aprendizado de máquina (ML): J48, REPTree (RT), máquina de vetor suporte (SVM), redes neurais artificiais (ANN), Regressão Logística (LR) e Random Forest (RF), com duas configurações de entrada: espectro contínuo (WL: 350–2500 nm) e bandas espectrais agregadas (Bands). A Figura 6 apresenta os boxplots comparando a acurácia (CC%) entre os modelos. Observou-se diferença estatística entre as entradas, com desempenho superior para WL em todos os algoritmos. As maiores medianas foram observadas em J48 e RT, seguidos por ANN e SVM. LR e RF apresentaram as menores médias e maior variabilidade.

Figura 6. Comparação do desempenho de algoritmos de ML utilizando espectro contínuo e bandas espectrais.

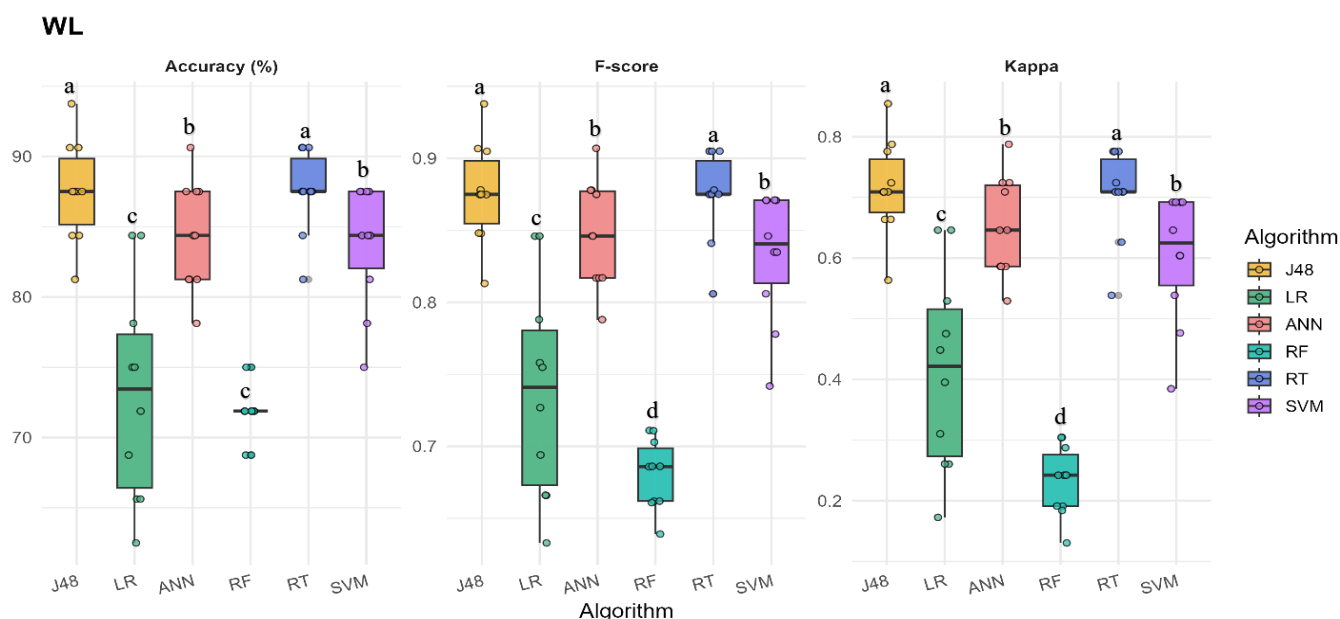


Fonte: elaboração própria.

A agregação do espectro em 12 bandas resultou em perda de desempenho, embora ANN e SVM tenham mantido medianas superiores às dos demais. LR e RF permaneceram com os piores resultados. Esses dados indicam que o espectro contínuo conserva mais informação discriminativa, resultando em maior acurácia e estabilidade nos modelos testados.

Considerando apenas o espectro completo (WL), os boxplots revelam um gradiente claro de desempenho entre os algoritmos nas três métricas avaliadas (Figura 7). Em termos de acurácia, J48 e RT apresentaram os maiores valores médios (87,5%), seguidos por ANN (84,37%) e SVM (83,75%). LR (73,13%) e RF (71,88%) compuseram o grupo com menor desempenho. Esses resultados indicam que a escolha do algoritmo influencia diretamente a taxa de classificações corretas.

Figura 7. Boxplots comparativos de acurácia, F-score e Kappa para algoritmos de classificação com espectro completo (WL).



Fonte: elaboração própria.

Na métrica F-score, J48 (0,88) e RT (0,87) novamente se destacaram, seguidos por ANN (0,85) e SVM (0,83). RF apresentou o menor valor (0,68), indicando baixa capacidade de generalização. Quanto ao coeficiente Kappa, J48 (0,71) e RT (0,70) mantiveram os melhores desempenhos, enquanto ANN (0,65) e SVM (0,60) apresentaram valores intermediários. RF (0,23) foi novamente o pior, e LR (0,41) ficou abaixo dos modelos mais robustos.

Esses resultados reforçam que algoritmos baseados em árvores de decisão e margens, como J48, RT, ANN e SVM, apresentam maior consistência e precisão na classificação de genótipos de

soja com base em dados espectrais, especialmente quando se utiliza o espectro contínuo como entrada.

4. DISCUSSÃO

Os dois primeiros componentes principais explicaram 90,1% da variância total dos dados, valor superior ao observado por Regazzi e Cruz (2020), que indicaram cerca de 70% como referência em estudos similares. Essa abordagem tem sido amplamente utilizada para discriminar genótipos com base em características espectrais e fisiológicas (CRUSIOL et al., 2021; SERON et al., 2025; BUCHAILLOT et al., 2022). A formação dos clusters reflete padrões fisiológicos distintos: o grupo C1 apresentou maior dispersão em atributos como fotossíntese líquida (A) e eficiência no uso da água (WUE), enquanto C2 concentrou-se em condutância estomática (gs) e transpiração (E), o que é coerente com os achados de LU et al. (2020) e WIJEWARDANA et al. (2019), que destacam a sensibilidade dessas variáveis à disponibilidade hídrica.

As curvas espectrais médias dos clusters revelaram diferenças marcantes nas regiões VIS, NIR e SWIR. No VIS (400–700 nm), a depressão em torno de 680 nm, associada à absorção da clorofila-a, foi mais acentuada em C1, sugerindo maior conteúdo de pigmentos fotossintéticos (SHI et al., 2022; SONOBE et al., 2020). No NIR (700–1300 nm), C2 apresentou maior reflectância, possivelmente relacionada à estrutura celular foliar, como maior espessura do mesófilo ou maior integridade anatômica (ADEGOYE et al., 2023). Já no SWIR (1300–2500 nm), as diferenças nas bandas de 1450, 1950 e 2200–2300 nm indicam variações no teor de água e lignina, reforçando a utilidade da espectroscopia como ferramenta de fenotipagem fisiológica (GROßKINSKY et al., 2015; GILL et al., 2022).

A etapa de classificação com algoritmos de aprendizado de máquina revelou que o espectro contínuo (WL) supera as bandas agregadas (Bands) em todos os modelos testados. Essa superioridade está relacionada à preservação da variabilidade espectral, como destacado por TEJASREE & AGILANDEESWARI (2024) e YOOSEFZADEH-NAJAFABADI et al. (2021), que apontam que a redução de dimensionalidade pode comprometer a acurácia da classificação em imagens hiperespectrais.

Entre os algoritmos, J48 e REPTree (RT) apresentaram os melhores desempenhos em acurácia, F-score e coeficiente Kappa. Esses modelos baseados em árvores de decisão são reconhecidos por sua robustez em conjuntos de dados com estrutura hierárquica e por lidarem bem

com variáveis categóricas e contínuas (AHMED et al., 2020). A utilização do algoritmo J48 pode melhorar a precisão da classificação, pois elimina a subjetividade e reduz erros nas respostas (KHATRI et al., 2018; DEMIR et al., 2022). As redes neurais artificiais (ANN) também apresentaram desempenho consistente, especialmente em cenários com maior complexidade espectral. Segundo ABDOLRASOL et al. (2021), a performance das ANN pode ser otimizada por meio de técnicas como PSO e GA, que ajustam parâmetros como número de neurônios e taxa de aprendizado. ANN também tem sido aplicada com sucesso na predição de rendimento de culturas e características químicas (AKBAR et al., 2018; HARA et al., 2021).

O algoritmo SVM mostrou-se eficiente na classificação de genótipos com espectros contínuos, corroborando os achados de NAIDU et al. (2023) e HASSANIJALILIAN et al. (2020), que destacam a estabilidade do SVM em tarefas de classificação binária e multiclasse, especialmente quando avaliado por métricas como F-score e Kappa. SVM também tem sido utilizado com sucesso na predição de produtividade e na detecção de doenças em soja (NAGASUBRAMANIAN et al., 2018; NOI & KAPPAS, 2018).

Por outro lado, algoritmos como Random Forest (RF) e Regressão Logística (LR) apresentaram desempenho inferior. RF, embora eficaz em muitos contextos, pode sofrer com sobreajuste em espectros contínuos quando o número de variáveis é elevado em relação ao número de amostras (HASSAN et al., 2018). LR, por sua vez, é mais sensível à linearidade dos dados e tende a apresentar menor acurácia em cenários com alta complexidade espectral e interações não lineares (RIZAL et al., 2022).

Esses resultados indicam que a escolha do algoritmo de ML deve considerar não apenas a acurácia, mas também a natureza dos dados, o volume amostral e a complexidade espectral. A integração entre espectroscopia hiperespectral e algoritmos de ML representa uma abordagem promissora para a fenotipagem fisiológica de genótipos de soja, permitindo a identificação de grupos com estratégias adaptativas distintas e oferecendo suporte valioso para programas de melhoramento genético voltados à tolerância a estresses ambientais (SANTANA et al., 2022; BOLAÑOS et al., 2023).

5. CONCLUSÃO

Este estudo evidenciou que a espectroscopia combinada à algoritmos de aprendizado de máquina, constitui uma estratégia eficaz para a classificação de genótipos de soja com base em

suas respostas fisiológicas. Os melhores resultados foram obtidos a partir do espectro contínuo (350 a 2500 nm), que possibilitou a preservação da riqueza informacional dos dados espectrais. Entre os algoritmos avaliados, J48 e REPTree destacaram-se por apresentarem maiores valores de acurácia, F-score e coeficiente Kappa, seguidos por máquina de vetor suporte (SVM) e redes neurais artificiais (ANN). Em contrapartida, modelos como Random Forest e Regressão Logística apresentaram desempenho inferior, sugerindo menor capacidade de generalização nesse contexto.

REFERÊNCIAS

- ABDOLRASOL, M. G. M., Hussain, S. M. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., ... & Milad, A. (2021). Artificial Neural Networks Based Optimization Techniques: A Review. **Electronics**, United States, v. 10, p. 2689. DOI: 10.3390/electronics10212689.
- ADEGOYE, G. A., Olorunwa, O. J., Alsajri, F. A., Walne, C. H., Wijewardana, C., Kethireddy, S. R., ... & Reddy, K. R. (2023). Waterlogging effects on soybean physiology and hyperspectral reflectance during the reproductive stage. **Agriculture**, Suíça, v. 13, n. 4, p. 844. 10.3390/agriculture13040844.
- AHMED, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. **Electronics**, United States, v. 9, n. 8, p. 1295. 10.3390/electronics9081295.
- AKBAR, A., Kuanar, A., Patnaik, J., Mishra, A., & Nayak, S. (2018). Application of artificial neural network modeling for optimization and prediction of essential oil yield in turmeric (*Curcuma longa* L.). **Computers and Electronics in Agriculture**, Oxford, England, v. 148, p. 160-178. <https://doi.org/10.1016/j.compag.2018.03.002>
- AWAD, M., & Khanna, R. (2015). Support vector machines for classification. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 39-66). Berkeley, CA: **Apress**. 10.1007/978-1-4302-5990-9_3.
- BELGIU, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. **ISPRS journal of photogrammetry and remote sensing**, Oxford, England, v. 114, p. 24-31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- BENOS, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. **Sensors**, Oxford, England, v. 11, 3758. <https://doi.org/10.3390/s21113758>
- BLAIFI, S. A., Moulahoum, S., Benkercha, R., Taghezouit, B., & Saim, A. (2018). M5P model tree based fast fuzzy maximum power point tracker. **Solar Energy**, [S.l.], v. 163, p.405-424. <https://doi.org/10.1016/j.solener.2018.01.071>
- BUCHAILLOT, M. L., Soba, D., Shu, T., Liu, J., Aranjuelo, I., Araus, J. L., ... & Sanz-Saez, A. (2022). Estimating peanut and soybean photosynthetic traits using leaf spectral reflectance and

advance regression models. **Plants**, Suíça, v. 255, n. 4, p. 93. <https://doi.org/10.1007/s00425-022-03867-6>

BOLAÑOS, J., Corrales, J. C., & Campo, L. V. (2023). Feasibility of early yield prediction per coffee tree based on multispectral aerial imagery: Case of arabica coffee crops in Cauca-Colombia. **Remote Sensing**, Oxford, England, v. 15, n. 1, p. 282. <https://doi.org/10.3390/rs15010282>

CONAB. *Acompanhamento da safra brasileira de grãos 2023/2024*. Brasília: **Companhia Nacional de Abastecimento**, 2024.

Crusiol, L. G. T., Nanni, M. R., Furlanetto, R. H., Sibaldelli, R. N. R., Cezar, E., Sun, L., ... & Farias, J. R. B. (2021). Classification of soybean genotypes assessed under different water availability and at different phenological stages using leaf-based hyperspectral reflectance. **Remote Sensing**, Oxford, England, v. 13, n. 2, p. 172. <https://doi.org/10.3390/rs13020172>

DA SILVA JUNIOR, C. A., Nanni, M. R., Shakir, M., Teodoro, P. E., de Oliveira-Júnior, J. F., Cezar, E., ... & Shiratsuchi, L. S. (2018). Soybean varieties discrimination using non-imaging hyperspectral sensor. **Infrared Physics & Technology**, Oxford, England, v. 89, p. 338-350. <https://doi.org/10.1016/j.infrared.2018.01.027>

DALE, L. M., Thewis, A., Boudry, C., Rotar, I., Dardenne, P., Baeten, V., & Pierna, J. A. F. (2013). Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review. **Applied Spectroscopy Reviews**, [S.l.], v. 48, n. 2, p. 142-159. <https://doi.org/10.1080/05704928.2012.705800>

Demir, S.; Başayığit, L. Classification of Some Biochemical Properties with J48 Classification Tree Algorithms in Hyperspectral Data. *Veri Bilimi*, v.5, n.2, p.20-28, 2022. Disponível em: <https://dergipark.org.tr/en/pub/veri/issue/74116/1178263>

EGMONT-PETERSEN, M.; RIDDER, D.; HANCZAR, B. Image processing with neural networks: A review. **Pattern Recognition**, Reino Unido, v. 35, n. 10, p. 2279–2300, 2002. [https://doi.org/10.1016/S0031-3203\(01\)00178-9](https://doi.org/10.1016/S0031-3203(01)00178-9)

Falcioni, R., Gonçalves, J. V. F., Oliveira, K. M. D., Oliveira, C. A. D., Demattê, J. A., Antunes, W. C., & Nanni, M. R. (2023). Enhancing pigment phenotyping and classification in lettuce through the integration of reflectance spectroscopy and AI algorithms. **Plants**, Suíça, v. 12, n. 6, p. 1333. <https://doi.org/10.3390/plants12061333>

GILL, T., Gill, S. K., Saini, D. K., Chopra, Y., de Koff, J. P., & Sandhu, K. S. (2022). A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. **Phenomics**, 2(3), 156-183. <https://doi.org/10.1007/s43657-022-00048-z>

GROßKINSKY, D. K., Svensgaard, J., Christensen, S., & Roitsch, T. (2015). Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. **Journal of experimental botany**, 66(18), 5429-5440. <https://doi.org/10.1093/jxb/erv345>

HARA, P., Piekutowska, M., & Niedbała, G. (2021). Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. **Land**, 10(6), 609. <https://doi.org/10.3390/land10060609>

HASSAN, Ch Anwar Ul; KHAN, Muhammad Sufyan; SHAH, Munam Ali. Comparison of machine learning algorithms in data classification. In: 2018 **24th International Conference on Automation and Computing (ICAC)**. IEEE, 2018. p. 1-6.

IM, J., & Jensen, J. R. (2008). Hyperspectral remote sensing of vegetation. **Geography Compass**, 2(6), 1943-1961. <https://doi.org/10.1111/j.1749-8198.2008.00182.x>

JEONG, S. W. et al. SUnSeT: spectral unmixing of hyperspectral images for phenotyping soybean seed traits. **Plant Cell Reports**, v. 43, 2024. DOI: 10.1007/s00299-024-03249-0.

LIAKOS, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: **A review. Sensors**, Suiça, v. 18, n. 8, p. 2674. <https://doi.org/10.3390/s18082674>

KHATRI, Sabita; ARORA, Deepak; KUMAR, Anil. Enhancing decision tree classification accuracy through genetically programmed attributes for wart treatment method identification. **Procedia computer science**, v. 132, p. 1685-1694, 2018. <https://doi.org/10.1016/j.procs.2018.05.141>

LU, B., Dao, P. D., Liu, J., He, Y., & Shang, J. (2020). Recent advances of hyperspectral imaging technology and applications in agriculture. **Remote Sensing**, 12(16), 2659. DOI: 10.3390/rs12162659.

NAIDU, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In Computer science on-line conference (pp. 15-25). Cham: **Springer International Publishing**. DOI: 10.1007/978-3-031-35314-7_2.

NAGASUBRAMANIAN, K., Jones, S., Sarkar, S., Singh, A. K., Singh, A., & Ganapathysubramanian, B. (2018). Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. **Plant methods**, 14(1), 86. <https://doi.org/10.1186/s13007-018-0349-9>

NALEPA, J., & KAWULOK, M. (2019). Selecting training sets for support vector machines: a review. **Artificial Intelligence Review**, 52(2), 857-900. <https://doi.org/10.1007/s10462-017-9611-1>

NOI, P. T.; KAPPAS, M. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. **Sensors**, v. 18, n. 1, 2018. DOI: 10.3390/s18010018.

R CORE TEAM. *R: A language and environment for statistical computing*. Vienna: **R Foundation for Statistical Computing**, 2023. Disponível em: <https://www.r-project.org/>.

RIZAL, N. N. M. et al. Comparison between Regression Models, Support Vector Machine (SVM), and Artificial Neural Network (ANN) in River Water Quality Prediction. **Processes**, v. 10, n. 8, 2022. DOI: 10.3390/pr10081652.

RSTUDIO TEAM. *RStudio: Integrated Development Environment for R*. Boston: **RStudio**, PBC, 2023. Disponível em: <https://www.rstudio.com/>.

SANTANA, D. C., Seron, A. C. C., Teodoro, L. P. R., de Oliveira, I. C., da Silva Junior, C. A., Baio, F. H. R., ... & Teodoro, P. E. (2024). Classification of soybean groups for grain yield and

industrial traits using Vnir-Swir spectroscopy. **Infrared Physics & Technology**, 139, 105326. <https://doi.org/10.1016/j.infrared.2024.105326>

SANTANA, D. C., Fenotipagem de alta precisão de caracteres industriais de soja usando aprendizagem de máquina e diferentes sensores espectrais. Ilha Solteira: **Universidade Estadual Paulista (Unesp)**, 4 ago. 2023.

SANTANA, D. C., Teixeira Filho, M. C. M., da Silva, M. R., Chagas, P. H. M. D., de Oliveira, J. L. G., Baio, F. H. R., ... & Shiratsuchi, L. S. (2023). Machine learning in the classification of soybean genotypes for primary macronutrients' content using UAV–multispectral sensor. **Remote Sensing**, 15(5), 1457. <https://doi.org/10.3390/rs15051457>

Sarić, R., Nguyen, V. D., Burge, T., Berkowitz, O., Trtílek, M., Whelan, J., ... & Čustović, E. (2022). Applications of hyperspectral imaging in plant phenotyping. **Trends in plant science**, 27(3), 301-315. <https://doi.org/10.1016/j.tplants.2021.12.003>

SHI, S., Xu, L., Gong, W., Chen, B., Chen, B., Qu, F., ... & Yang, J. (2022). A convolution neural network for forest leaf chlorophyll and carotenoid estimation using hyperspectral reflectance. **International Journal of Applied Earth Observation and Geoinformation**, 108, 102719. <https://doi.org/10.1016/j.jag.2022.102719>

SCOTT, Andrew Jhon; KNOTT, Martin. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, p. 507-512, 1974. <https://doi.org/10.2307/2529204>

SERON, A. C. da S. C. et al. Relationship Between Hyperspectral Data and Amino Acid Composition in Soybean Genotypes. **AgriEngineering**, Suiça, v. 7, n. 8, 2025. DOI: 10.3390/agriengineering7080265.

SONOBE, R., Yamashita, H., Mihara, H., Morita, A., & Ikka, T. (2020). Estimation of leaf chlorophyll a, b and carotenoid contents and their ratios using hyperspectral reflectance. **Remote Sensing**, Reino Unido, v. 12, n. 19, p. 3265. <https://doi.org/10.3390/rs12193265>

SNOUSY, M.B. al, El-Deeb, H.M., Badran, K., Khlil, I.A. al, 2011. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egypt. Inf. J.** Egypt v. 12, p. 73–82. <https://doi.org/10.1016/j.eij.2011.04.003>.

TEODORO, L. P. R., Bhering, L. L., Gomes, B. E. L., Campos, C. N. S., Baio, F. H. R., Gava, R., ... & Teodoro, P. E. (2019). Understanding the combining ability for physiological traits in soybean. **PloS one**, United States, v. 14, n. 12, p. e0226523. <https://doi.org/10.1371/journal.pone.0226523>

TEJASREE, G.; AGILANDEESWARI, L. An extensive review of hyperspectral image classification and prediction: techniques and challenges. **Multimedia Tools and Applications**, v. 83, p. 80941–81038, 2024. DOI: 10.1007/s11042-024-18562-9.

WIJEWARDANA, C., Alsajri, F. A., Irby, J. T., Krutz, L. J., Golden, B., Henry, W. B., ... & Reddy, K. R. (2019). Physiological assessment of water deficit in soybean using midday leaf water potential and spectral features. **Journal of Plant Interactions**, [S.l.] v. 14, n. 1, p. 533-543. <https://doi.org/10.1080/17429145.2019.1662499>

YU, H., Kong, B., Hou, Y., Xu, X., Chen, T., & Liu, X. (2022). A critical review on applications of hyperspectral remote sensing in crop monitoring. **Experimental Agriculture**, England, v. 58, e26. <https://doi.org/10.1017/S0014479722000278>

YOOSEFZADEH-NAJAFABADI, M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. **Frontiers in plant science**, Suíça, v. 11, p. 624273. <https://doi.org/10.3389/fpls.2020.624273>

ZHANG, Q. Y., Fan, K. J., Tian, Z., Guo, K., & Su, W. H. (2024). High-Precision automated soybean phenotypic feature extraction based on deep learning and computer vision. **Plants**, Suíça, v. 13, n. 18, p. 2613. 10.3390/plants13182613.

SHI, H., Guo, J., An, J., Tang, Z., Wang, X., Li, W., ... & Zhang, F. (2023). Estimation of chlorophyll content in soybean crop at different growth stages based on optimal spectral index. **Agronomy**, Suíça, v. 13, n. 3, p. 663. DOI: 10.3390/agronomy13030663.