Classificação Binária Preditiva de Imagens de Folhas de Soja por Algoritmos de Aprendizado de Máquina

Francieli Martins Rocha[1]

Maurício Duarte[2]

Renata Bruna dos Santos Coscolin Favan[2]

RESUMO

Este projeto tem como foco o desenvolvimento e avaliação de desempenho dos modelos de aprendizado aplicados em classificação binária preditiva de doenças em folhas de soja. Foi utilizado um subconjunto de dados SoyNet, que continha anteriormente 1055 imagens de folhas doentes e 139 imagens de folhas saudáveis. Para melhor generalização dos modelos, foi aplicado a técnica de *data augmentation*, aumentando a quantidade de folhas saudáveis para aproximadamente mil imagens e resultando no balanceamento do *dataset.* As imagens apresentam um padrão de escala de cinza e redimensionadas para 64x64. Os algoritmos estudados são *Random Forest, Support Vector Machine (SVM)* e Convolutional Neural Network

(CNN). Para realização do comparativo de resultados, foram considerados métricas como acurácia, precisão, *recall, F1-score* e análise de matrizes de confusão. Foi observado que todos os modelos apresentaram uma acurácia satisfatória superiores a 94%. O *Random Forest* destacou-se em ter uma sensibilidade maior à classe doente, enquanto o *CNN* demonstrou uma alta precisão e um baixo índice de falsos positivos. Conclui-se que os algoritmos *Random Forest* e *CNN* são eficazes para a tarefa de classificação, entretanto a escolha entre ambos dependerá da prioridade prática da atividade.

Palavras-chave: Classificação preditiva. Folhas de soja. Aprendizado de máquina.

1. INTRODUÇÃO

A soja é um produto importantíssimo na exploração agrícola, sendo responsável por uma parcela significativa da agricultura nacional. De acordo com a Companhia Nacional de Abastecimento (CONAB), o Brasil é o maior produtor e exportador mundial de soja, estimado em mais de 150 milhões de toneladas na safra de 2023/2024. A maior parte da produção nacional está localizada nos estados de Mato Grosso, Paraná, Rio Grande do Sul, Goiás e Mato Grosso do Sul, considerados os maiores produtores do país (CONAB, 2025). Com esta grande demanda, o cultivo da soja tem passado por um processo contínuo de modernização, incorporando tecnologias como sensores remotos para irrigação automatizada e o uso de visão computacional. Essa última tem se destacado no agronegócio por permitir a detecção precoce de doenças nas plantas, o que contribui significativamente para o aumento da produtividade e para a redução de riscos à lavoura (EMBRAPA, 2019). Por meio de modelos de aprendizado de máquina, torna-se possível treinar algoritmos com bases de imagens de folhas de soja, identificando padrões visuais associados a doenças, o que viabiliza análises preditivas e decisões mais rápidas e assertivas no manejo agrícola (FERENTINOS, 2018).

O crescimento contínuo de área cultivada em solo brasileiro é uma realidade incontestável, tornando o grão um dos produtos mais relevantes para o mercado econômico nacional. Segundo o 7° Levantamento da Safra 2024/2025 da Conab, foi alcançado 45,3 milhões de hectares da área plantada, o que representa um aumento de 3% comparado com a safra anterior. O crescimento relacionado com sua expansão se dá devido a demanda internacional e por sua rentabilidade da oleaginosa. É estimado que para a safra atual seja possível estabelecer um recorde histórico para o país, conquistando possivelmente 168,3 milhões de toneladas. Os valores surpreendentes conquistados pela agricultura Brasileira afirmam ainda mais a sua consistência em posição de líder mundial na produção de soja.

Por ser uma cultura relevante para o mercado brasileiro há uma preocupação em relação à qualidade e produtividade do cultivo, que pode ser afetada pelo ataque de doenças no plantio causando a diminuição do aproveitamento total da colheita e, consequentemente, tendo um impacto negativo na aquisição do agricultor: o dinheiro e tempo investido na preparação de um ambiente e solo favorável para o plantio pode ser perdido com o atraso na identificação de ataque de doenças na lavoura, resultando em perdas significativas e à necessidade de novas estratégias para mitigar esses

riscos (STOLLER DO BRASIL, 2025). Dito isso, a tecnologia entra como uma solução excelente e inovadora para os desafios enfrentados pelos agricultores.

A soja é um produto de grande relevância na economia nacional, apresentando um papel de destaque nas exportações brasileiras. A implementação de ferramentas tecnológicas, otimizando o manejo agrícola e reduzindo perdas na produção. Para a realização do treinamento dos algoritmos, foi utilizada uma base de dados pré-processada com imagens de folhas de soja convertidas em escala de cinza. A partir dessa base, foram implementados os algoritmos Random Forest, *Support Vector Machine* (SVM) e *Convolutional Neural Network* (CNN). O objetivo do projeto é avaliar o desempenho de diferentes algoritmos de aprendizado na atividade de classificação preditiva binária de imagens, identificando quais folhas são saudáveis e quais estão doentes.

2. APRENDIZADO DE MÁQUINA

Vivencia-se um momento histórico: uma espécie de revolução industrial que vai além de maquinários, sendo possível não só automatizar atividades que necessitam de um trabalho físico como aquelas que dependem de um certo tipo de “raciocínio” e “inteligência” (SciELO Brasil 2021). Apesar de haver estudos e conceitos de inteligência artificial (IA) desde o século passado, o seu recente desenvolvimento só se tornou viável devido à disponibilidade massiva de dados e ao aumento do poder computacional, disponibilizando material o suficiente para realizar o aprendizado de máquina com eficiência. Diferentemente dos humanos que ao longo da vida vão adquirindo um conhecimento diverso cotidianamente, as máquinas necessitam de conjuntos de dados estruturados que servem como base para o treinamento de modelos de aprendizado de máquina para que consiga obter experiências acumuladas através de soluções bem-sucedidas de problemas anteriores, para assim refletir e solucionar o problema proposto.

O objetivo do Aprendizado de Máquina é a construção de programas que melhorem seu desempenho por meio de exemplos, de maneira indutiva. Esses exemplos, que compõem o *dataset (Conjunto de dados)* de aprendizado, podem ser numéricos, textuais ou visuais, como imagens. É preciso uma quantidade significativa de exemplos para que o computador consiga gerar hipóteses através dos dados, fazendo com que o programa consiga aprender e executar sua tarefa baseando-se na sua própria experiência (FACELI et al., 2011). Esta área de Inteligência Artificial

engloba teoria da complexidade computacional, probabilidades estatísticas, teoria da informação, filosofia, psicologia, neurobiologia, entre outros (R. Cerri e A. C. P. de L. F. de Carvalho., 2017).

É possível analisar uma hierarquia de aprendizado dentro dos treinamentos dos modelos, e sua categorização vai partir de se o sistema utilizará informações externamente coletadas, vulgo aprendizado não supervisionado, ou de classes que são rotuladas e conhecidas, denominadas de aprendizado supervisionado. O aprendizado de um algoritmo supervisionado é a criação de um classificador, responsável por determinar corretamente a classe de futuros atributos que ainda não foram rotulados, lógica que será aplicada no estudo presente. Agora pensando no aprendizado não supervisionado, haverá a análise dos exemplos fornecidos e a tentativa de perceber se podem ou não ser agrupados e a identificação do que cada agrupamento significa no contexto da problemática que o sistema lidará (ALPAYDIN, 2020).

Além dos tipos de aprendizado já citados acima, é possível aplicar uma outra abordagem denominada aprendizagem semi supervisionada, na qual uma parcela dos dados é rotulada (com nomes ou categorias) e uma grande parte permanece sem rótulo. Essa estratégia é útil quando a rotulagem manual de todos os dados seria inviável, seja por tempo ou custo, tornando o processo de desenvolvimento muito extenso ou economicamente desfavorável. Ao aplicar a aprendizagem semi supervisionada, é possível alcançar um padrão de aprendizado eficiente mesmo com dados parcialmente anotados.

Os algoritmos de aprendizado de máquina podem ser aplicados em diversas tarefas computacionais. Entre as principais, destacam-se: classificação, usada neste projeto para distinguir folhas saudáveis e doentes; regressão, que prevê valores contínuos como produtividade; sumarização, que gera resumos informativos dos dados; agrupamento, que identifica padrões em dados não rotulados; e associação, que detecta relações entre variáveis, como a ligação entre pragas e épocas do ano.

A figura 1 mostra a representação estrutural dos principais tipos de aprendizado de máquina e suas subcategorias utilizadas para atividades de regressão e classificação, embora apenas a classificação supervisionada foi utilizada neste trabalho.

Figura 1 - Os diferentes tipos de aprendizado.

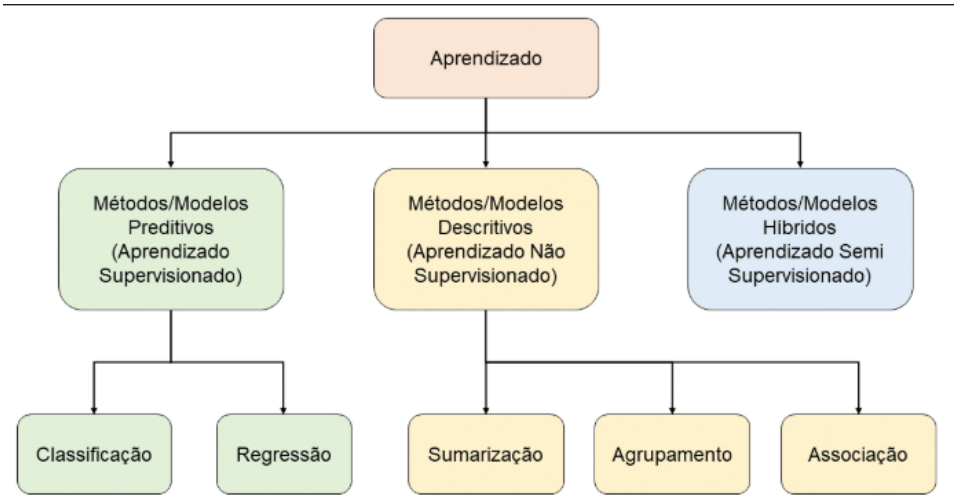


Figura: Embarcados (2020).

A Inteligência Artificial (IA) tem sido amplamente explorada por diversas áreas, como a saúde pública e a agricultura de precisão, revelando-se cada vez mais essencial na resolução de problemas complexos e na análise de grandes volumes de dados. Dentro desse contexto, destaca-se o aprendizado de máquina, um dos principais campos da IA, que permite que sistemas computacionais realizem previsões e tomem decisões com base em padrões aprendidos a partir de dados históricos (HARVARD MEDICAL SCHOOL, 2023).

No campo da agricultura de precisão, o aprendizado de máquina pode ser aplicado a diferentes tarefas, como a análise de dados relacionados à coleta, plantio e fertilização. Essas tarefas envolvem a identificação de padrões que impactam diretamente a qualidade dos grãos, otimizando o desempenho produtivo e reduzindo o tempo necessário para tomadas de decisão estratégicas (KAMILARIS; PRENAFETA-BOLDÚ, 2018).

O aprendizado de máquina tem se mostrado uma ferramenta valiosa no contexto agrícola, especialmente por sua capacidade de lidar com grandes volumes de dados gerados em campo. Entre suas técnicas mais avançadas estão as redes neurais artificiais, compostas por unidades chamadas de neurônios, que aplicam funções matemáticas a partir de entradas e retornam saídas. O sistema “aprende” ajustando os pesos dessas conexões durante o processo de treinamento (ALPAYDIN, 2020).

Na cultura da soja, o aprendizado de máquina pode ser aplicado tanto na análise dos grãos quanto na das folhas. No caso dos grãos, modelos podem ser treinados para identificar características ideais para consumo e exportação,

descartando automaticamente os que não atendem aos critérios. Já a análise da folhagem permite detectar precocemente doenças na planta, evitando perdas e preservando a qualidade da colheita. Neste projeto, optou-se pela análise das folhas, por ser uma abordagem mais eficaz para antecipar danos e garantir a excelência do produto.

Para que os algoritmos de aprendizado de máquina apresentem um bom desempenho, é essencial que os dados utilizados sejam de alta qualidade. Isso se deve ao fato de que esses métodos constroem o conhecimento exclusivamente a partir dos dados fornecidos, sem o apoio de regras explícitas ou informações externas (BATISTA, 2003). Nesse sentido, um pré-processamento inadequado pode comprometer significativamente a acurácia dos modelos, tornando o resultado menos confiável e eficiente.

3. MATERIAL E MÉTODOS

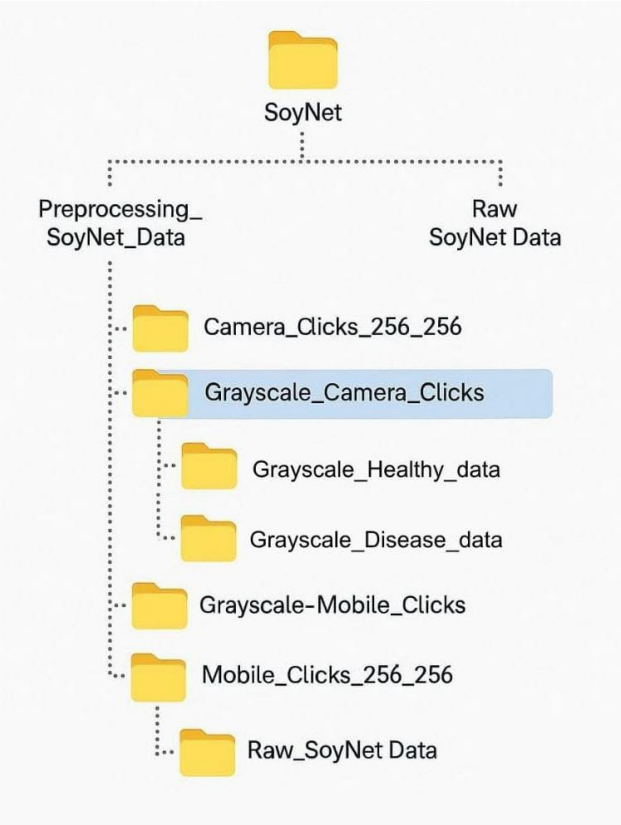
O ambiente escolhido para o desenvolvimento do projeto foi a plataforma *Google Collaboratory (Colab)*, visando um melhor aproveitamento da disponibilidade de manipulação de dados em grande volume. A linguagem de programação utilizada na inteligência artificial foi *Python* 3, preferencialmente escolhida por conta da vasta possibilidade de bibliotecas que é oferecida ao usuário principalmente voltado ao aprendizado de máquina.

As bibliotecas necessárias para a o desenvolvimento do trabalho foram a Numpy, OpenCV, Scikit-learn, TensorFlow e outras para manipulação de dados, visualização e implementação dos algoritmos."

A base de dados utilizada neste projeto foi o SoyNet, elaborado para fornecer imagens de alta qualidade de folhas de soja classificadas como doentes e saudáveis. Este dataset tem como objetivo garantir a confiabilidade das informações, oferecendo imagens adequadas para um melhor treinamento dos modelos de aprendizado de máquina e, consequentemente, proporcionando diagnósticos e classificações mais eficazes em um ambiente real da agricultura. Foi preciso equilibrar o número de amostras entre as classes por meio da técnica de *data augmentation*, a fim de evitar o viés de aprendizado decorrente do desbalanceamento de dados.

Para a obter das imagens utilizadas neste trabalho, foram selecionadas exclusivamente aquelas capturadas por uma câmera digital, com o objetivo de garantir alta resolução e qualidade. Embora o conjunto de dados original também disponibilize

imagens feitas com câmeras de celular (mobile), optou-se por utilizar apenas as imagens da câmera digital, pois são as únicas que se encontram pré-processadas. Os dados utilizados estão organizados em duas categorias: os dados brutos (*Raw_SoyNet_Data*) e os dados pré-processados (*Preprocessing_SoyNet_Data*), conforme ilustrado na Figura 2, que apresenta a estrutura de pastas do projeto.
Figura 2 - Estrutura de diretórios do *dataset* SoyNet.



Fonte: Elaborado pelo autor (2025).

O *dataset SoyNet* apresenta aproximadamente 9 mil imagens para uso. Entretanto para esta análise de desempenho dos algoritmos foi utilizada uma parcela específica da base, focando nas categorias de folhas saudáveis e folhas doentes. Foram selecionadas 1055 imagens de folhas doentes e 139 imagens de folhas saudáveis, e para evitar o desbalanceamento do algoritmo, foi preciso equilibrar o número de amostras entre as classes através da técnica de *data augmentation*, que consiste na geração de novas imagens a partir de transformações nas já existentes, como rotações, espelhamentos e deslocamentos. Com isto, foi possível ampliar o número de imagens saudáveis para aproximadamente 1000 amostras, tornando o *dataset* mais equilibrado e adequado para o treinamento dos modelos de

classificação. A padronização das imagens em escala de cinza e o redimensionamento para 64x64 pixels também foram etapas fundamentais para garantir consistência nos experimentos.

Para que houvesse a avaliação do comportamento de um mesmo *dataset* com diferentes algoritmos, foram escolhidos três modelos supervisionados: *Random Forest, Support Vector Machine (SVM)* e o *Convolutional Neural Network* (*CNN)*. Estes algoritmos possuem características distintas tanto em estrutura como no seu processo de aprendizado, entregando uma diversidade de resultados e análises do porquê de seu desempenho.

Segundo *Breiman* (2001, tradução feita pelo autor) “O *Random Forest* é uma combinação de preditores de árvores, de modo de que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores da floresta”. O algoritmo de *Random Forest* é uma técnica utilizada no aprendizado de máquina que visa fazer a combinação de diferentes tipos de árvores de decisão das quais têm uma amostra aleatória dos dados de treinamento. As decisões conclusivas tomadas pelo algoritmo têm como base as previsões de árvores individuais.

Dentro do funcionamento do *Random Forest* pode-se encontrar um processo de etapas, como a seleção de amostras, onde se é feita a extração de informações do conjunto original com reposição, e então aplicadas para treinamento de uma árvore de decisão. Após isso, para determinar uma melhor divisão, haverá a construção da árvore de decisão, onde seus subconjuntos aleatórios de *features* são selecionados em cada nó da árvore, fazendo com que seja introduzida diversidade entre as árvores e reduz a correlação que possa existir entre elas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Por fim, será possível criar uma combinação entre as previsões das árvores para criar uma previsão final do aprendizado.

Para este projeto o *Random Forest* foi escolhido devido a sua robustez e capacidade de trabalhar com um conjunto de dados complexos, o que se casa muito bem com o *dataset* de imagens de soja em diferentes condições. Em um projeto de classificação binária, a capacidade do algoritmo de lidar com um considerável volume de dados e informações correlacionadas o torna um dos preferidos para esse tipo de atividade, como em um estudo do Mateletto (2022) onde o algoritmo *Random Forest* demonstrou satisfatoriamente a eficácia em realização de tarefas envolvendo seleção de atributos e detecção de tendências.

Os cruciais parâmetros aplicados no *Random Forest* foi uma variável para definição de número de árvores geradas, nomeada de "*n_estimators*" e limitada ao valor 100, e um variável para garantir reprodutibilidade de resultados, denominada de "*random_state=42*". Os demais parâmetros foram mantidos com o valor padrão previsto na biblioteca Scikit-learn.

O algoritmo de Rede Neural Convolucional (CNN) é uma técnica de aprendizado profundo especialmente eficaz para o processamento de dados com estrutura em grade, como imagens. Esse modelo utiliza camadas convolucionais que extraem automaticamente atributos relevantes dos dados de entrada, reduzindo a necessidade de pré-processamento manual (LECUN et al., 1998).

Para seu funcionamento, acontece a aplicação de filtros convolucionais sobre a imagem de entrada, fazendo com que seja gerado mapas de atributos que são gradualmente refinados ao longo de suas camadas de rede. Após a aplicação destas camadas, é utilizado as camadas mais densas para que seja possível consolidar a informação obtida e realizar a classificação de maneira eficiente (GU et al.,2018).

A escolha deste algoritmo para o projeto levou em conta sua capacidade de captura de padrões visuais complexos em imagens, o que o torna eficiente para este projeto de classificação. Além disso, este algoritmo mostra-se relevante inclusive na agricultura de precisão e tem sido adotado em aplicações agrícolas que envolvem reconhecimento de doenças, graças a sua alta precisão nestes casos (FERENTINOS, 2018).

Os principais parâmetros para o *CNN* no trabalho foram as bibliotecas *TensorFlow* e *Keras*, contendo as camadas convolucionais e camadas densamente conectadas. Foi utilizado funções de ativação *ReLU* nas camadas intermediárias e colocado a função *softMax* na camada de saída para classificação binária e uma função de perda *binary_crossentropy*. Para o treinamento, foi utilizado o otimizador Adam visando o trabalho de classificação de imagens.

Support Vector Machine (SVM) é um tipo de aprendizado supervisionado que visa separar os dados em diferentes classes através de um hiperplano separador. Ele busca encontrar a linha que tem a maior margem possível, estando mais distante dos pontos mais próximos de cada classe, denominados vetores de suporte. Caso aconteça de os dados não serem linearmente separáveis, é utilizado um *kernel* para que possa criar uma dimensão onde esta realidade seja possível, criando uma

dimensão mais alta. Geralmente, o *kernel* utilizado é o *Radial Basis Function (RBF)* que traz uma ótima solução para classificações não lineares (Chang e Lin, 2011).

O *SVM* foi escolhido pela sua capacidade de lidar com dados complexos que apresentam uma alta dimensionalidade, como o *dataset* de imagens que será utilizado no projeto. Graças a sua grande robustez e eficácia em problemas de classificação, é possível capturar as relações não lineares das imagens.

Na implementação do código em *Python*, foi utilizada a classe SVC() da biblioteca *scikit-learn*, com seus parâmetros padrão, incluindo o *kernel='rbf'*, que utiliza o *Radial Basis Function* para realizar a separação dos dados. Também foi mantido o parâmetro de regularização, responsável por controlar o equilíbrio entre a margem e os erros de classificação, além da definição automática do parâmetro *gamma*, baseada na variância dos dados, o que controla a influência de cada ponto na decisão final. Os valores padrão foram mantidos com o objetivo de avaliar o desempenho do *SVM* sem a necessidade de ajustes manuais.

Concluída a definição do ambiente e dos algoritmos de classificação, foi utilizado o dataset de folhas de soja contendo imagens de plantas saudáveis e doentes. As imagens foram divididas em 80% para treinamento e 20% para teste. O próximo capítulo apresenta os resultados obtidos com cada técnica aplicada, permitindo uma análise comparativa entre os modelos testados. Dessa forma, é possível avaliar a eficiência de cada abordagem e identificar aquela que apresentou o melhor desempenho no contexto proposto.

4. RESULTADOS E DISCUSSÃO

A avaliação dos modelos inicialmente utilizou parâmetros amplamente consolidados no aprendizado de máquina, sendo eles:

- **Acurácia**: Indicador do percentual de acertos obtidos no treinamento.
- **Precisão**: Proporção de acertos relacionados à classe correta. No contexto do projeto, servirá para identificar quantas folhas doentes realmente foram classificadas como doentes.
- ***Recall***: Medidor da capacidade do algoritmo identificar amostras positivas, responsável por apresentar seu nível de eficácia.

- **F1-score**: Média entre precisão e recall, interessante para considerar tanto os falsos positivos quanto os falsos negativos e assim identificar enganos do modelo.
- **Matriz de confusão**: Representação visual do desempenho do modelo, onde é possível visualizar claramente os erros e acertos do algoritmo. Com ela, é possível realizar uma análise mais minuciosa em relação aos resultados.

Todos os modelos testados apresentaram excelente acurácia. O Random Forest obteve a maior acurácia, com 96%, seguido pelo SVM com 95% e, por fim, a CNN com 94%. Esse alto percentual de acertos pode ser atribuído à qualidade da base de dados utilizada, que apresenta boa resolução de imagem e padronização. Além da acurácia, foram analisadas outras métricas quantitativas dos algoritmos. Os valores obtidos estão detalhados na Tabela 1.

Tabela 1 - Análise dos valores de precisão, *recall* e *F1-Score* para os algoritmos testados.

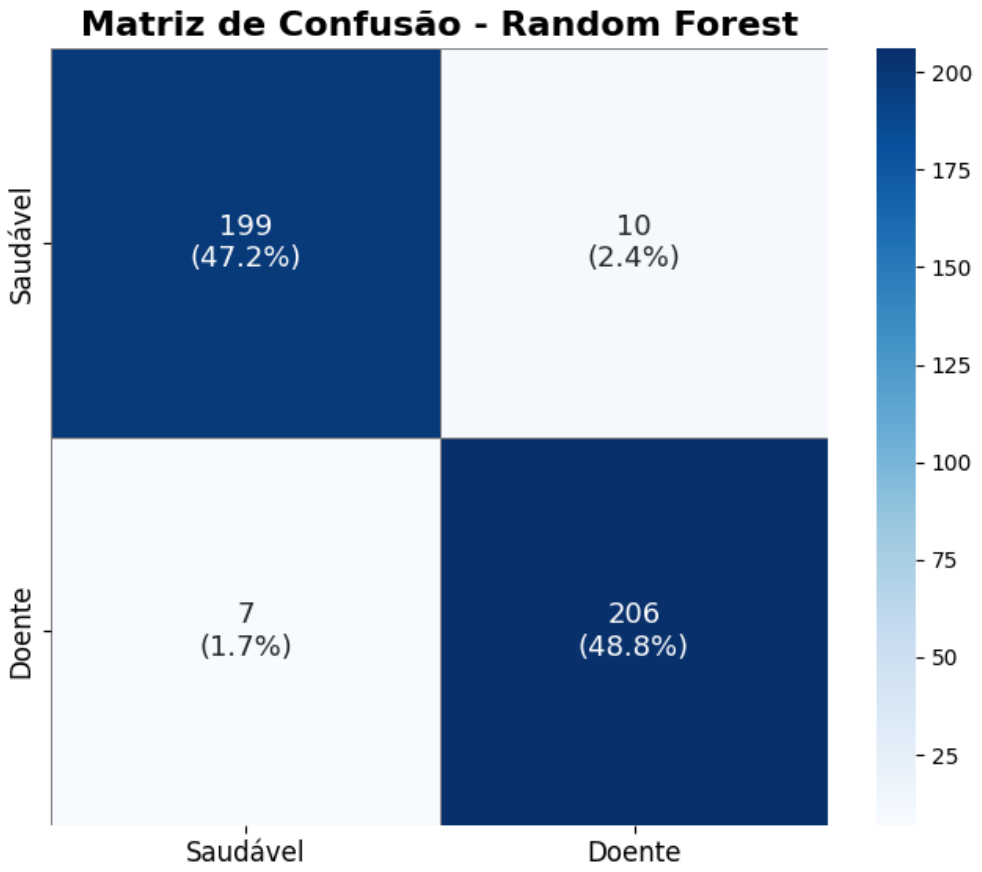
Modelo	Acurácia	Precisão	Recall	F1-Score
Random Forest	0,96	0,96	0,96	0,96
SVM	0,95	0,94	0,94	0,94
CNN	0,94	0,95	0,95	0,95

Fonte: Elaborada pelo autor.

Para melhor compreensão de acertos e erros dos três algoritmos, foi realizada a análise das matrizes de confusão. As matrizes auxiliam a visualização do comportamento preditivo em relação as suas duas classes (folhas de soja saudáveis e folhas de soja doentes), revelando o número exato de verdadeiros negativos e verdadeiros positivos (acertos), como os falsos negativos e falsos positivos (erros).

A matriz de confusão do *Random Forest* traz um resultado muito positivo: 213 folhas doentes classificadas corretamente e apenas 7 classificadas incorretamente. A classe saudável também teve valores favoráveis de 199 acertos, e 10 erros. É notável que o algoritmo conseguiu diferenciar as classes apresentando um baixo índice de erros (falsos negativos), sendo ótimo em um contexto de detecção precoce de doenças em plantas.

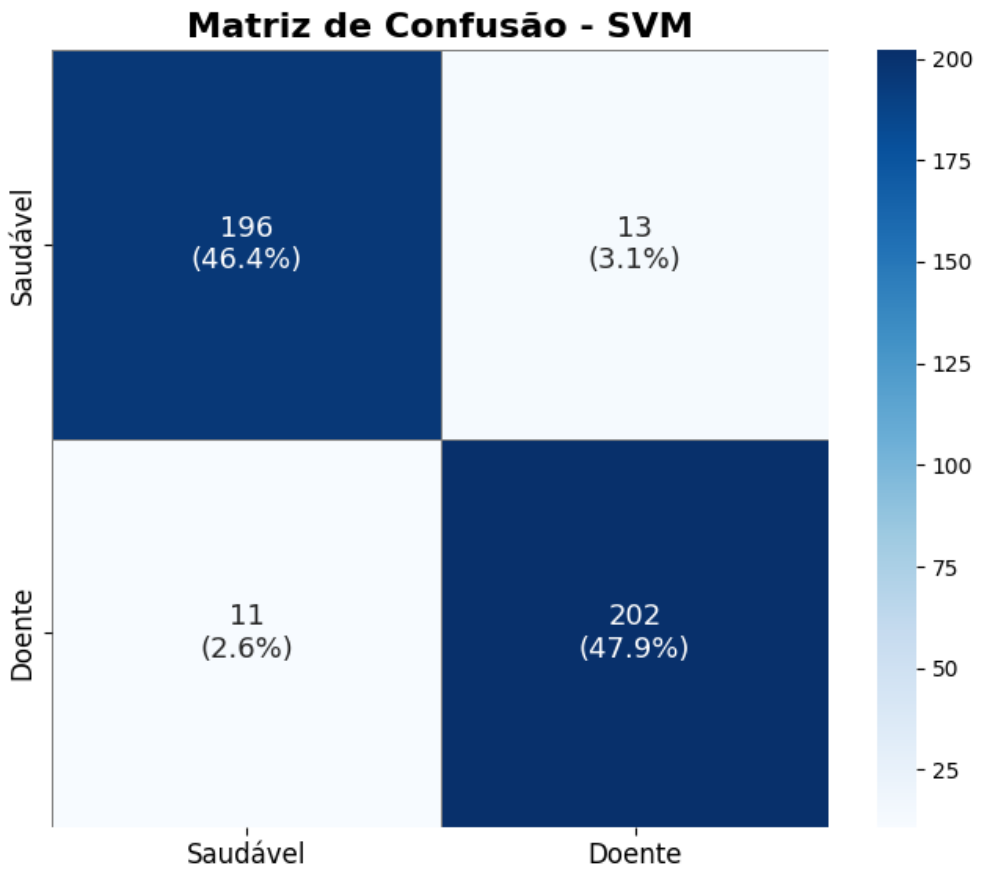
Figura 3 - Matriz de Confusão *Random Forest*.



Fonte: Elaborada pelo Autor (2025).

O *SVM*, assim como o *Random Forest*, apresentou valores positivos: 202 folhas foram corretamente classificas como doentes e 11 incorretamente. Em relação a classe saudável, sua leitura foi ótima apresentando 196 acertos e 13 erros. Observando de maneira geral, o desempenho foi satisfatório, porém apresentou uma taxa de falsos negativo (11) ligeiramente maior em relação ao *Random Forest* (7), o que pode ser uma informação relevante em um contexto em que não detectar a doença existente traga mais riscos do que uma identificação errônea.

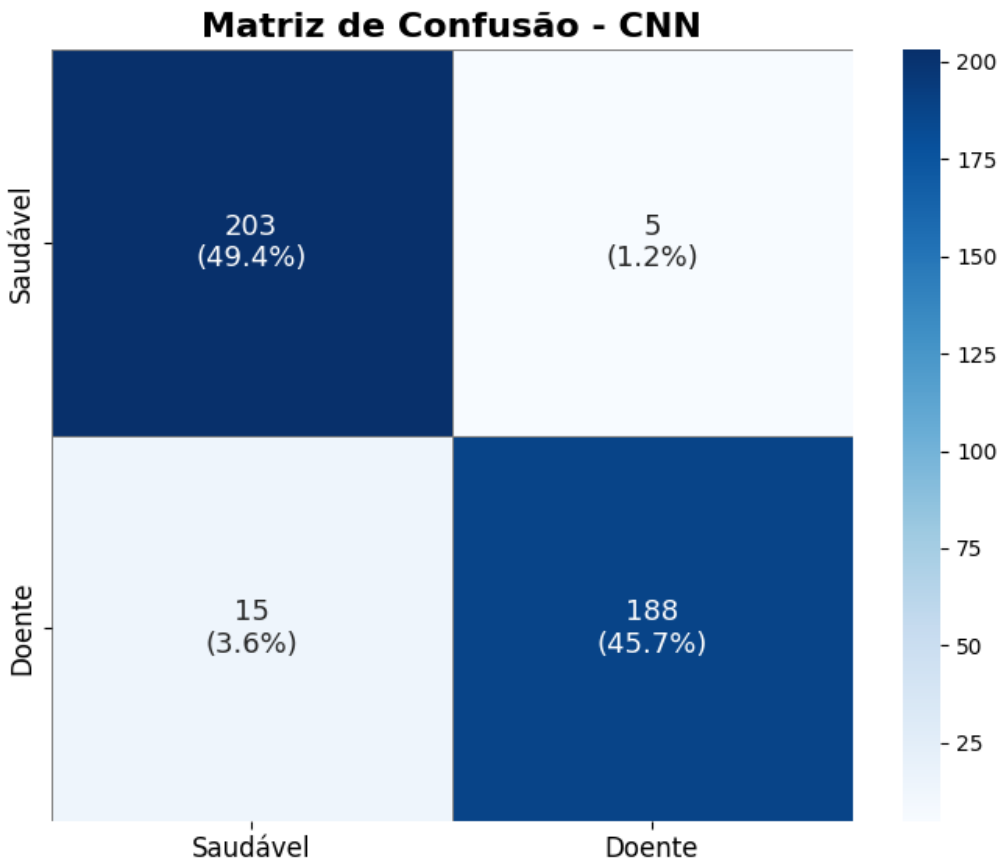
Figura 4 - Matriz de confusão *SVM*.



Fonte: Elaborada pelo autor (2025).

O *CNN* apresentou um resultado equilibrado e robusto. Foram classificados corretamente 198 folhas doentes, tendo uma taxa de erro de 15 folhas confundidas com as saudáveis, sendo um número relativamente maior que os outros dois algoritmos. Na classe saudável houve 204 acertos e apenas 5 erros. Neste modelo é notável sua baixa quantidade de falsos positivos, mostrando-se eficiente na classificação de folhas saudáveis e evitando equívocos futuros.

Figura 5 - Matriz de confusão *CNN*.



Fonte: Elaborada pelo autor (2025).

Apesar do *CNN* ter entregado um número maior de falsos negativos comparado aos outros algoritmos, o modelo apresentou o menor número de falsos positivos (5) e entregou um desempenho equilibrado entre as classes de folhas doentes e folhas saudáveis. A precisão relacionada a classe doente foi mais alta, afirmando sua capacidade de generalização. Entretanto o *Random Forest* também apresentou resultados notáveis, como o modelo com maior sensibilidade à classe “doente”, entregando o menor número de erros.

As diferenças entre os modelos estão relacionadas às estratégias de aprendizagem de cada um. O Random Forest responde melhor a padrões estruturais mais claros (BREIMAN, 2001), enquanto a CNN é capaz de extrair representações visuais complexas por meio de hierarquias (LECUN et al., 1998; GU et al., 2018), o que a torna mais criteriosa nas decisões. O SVM se destaca em contextos com baixa variabilidade visual, mas pode ser sensível a ruídos sutis. Esses comportamentos foram observados durante os testes com o conjunto de dados de teste. Assim, a escolha do modelo depende do objetivo: o Random Forest é mais indicado para evitar falhas, enquanto a CNN oferece melhor equilíbrio entre sensibilidade e precisão.

5. CONSIDERAÇÕES FINAIS

Este trabalho avaliou o desempenho dos algoritmos Random Forest, SVM e CNN na classificação binária de folhas de soja saudáveis e infectadas, com base em imagens em escala de cinza. Os modelos foram treinados com dados balanceados por *data augmentation*, alcançando acurácia acima de 94% e F1-Score consistente. O SVM apresentou bom desempenho, embora com menor destaque. A CNN obteve menos falsos positivos, enquanto o Random Forest se sobressaiu pela menor taxa de falsos negativos — o que é relevante para aplicações na agricultura de precisão, ao permitir a detecção mais eficaz de folhas doentes.

Foi notado que Random Forest quanto a CNN são modelos viáveis, e a escolha depende do contexto de aplicação: quando o objetivo for maior precisão e equilíbrio, a CNN tende a oferecer melhores resultados; já o Random Forest mostra-se mais eficiente na minimização de falsos negativos, o que pode ser crucial na detecção precoce de doenças em lavouras.

Como trabalhos futuros, sugere-se a ampliação do dataset com folhas em diferentes estágios de desenvolvimento, a remoção do fundo das imagens, a aplicação de validação cruzada para maior robustez estatística, e a adoção de redes neurais mais avançadas, como a EfficientNet. Além disso, recomenda-se a consideração do tipo de lesão ou injúria nas folhas, o que pode refinar ainda mais a assertividade do modelo em cenários reais.

REFERÊNCIAS:

ALPAYDIN, Ethem. *Introduction to machine learning*. 4. ed. Cambridge: MIT Press, 2020.

BATISTA, Gustavo Enrique de Almeida Prado Alves. *Pré-processamento de dados em aprendizado de máquina supervisionado*. 2003. 157 f. Dissertação (Mestrado em Ciência da Computação) – Universidade de São Paulo, São Carlos, 2003. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/pt-br.php. Acesso em: 31 mar. 2025.

BREIMAN, Leo. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324. Acesso em: 21 abr. 2025.

CHANG, Chih-Chung; LIN, Chih-Jen. *A practical guide to support vector classification*. 2011. Disponível em: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. Acesso em: 22 abr. 2025.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). *Acompanhamento da Safra Brasileira de Grãos – Safra 2024/2025 – 7º levantamento*. Brasília: Conab, maio 2025. Disponível em: https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos. Acesso em: 16 maio 2025.

EMBRAPA. *Monitoramento de pragas da soja utilizando aprendizado de máquina*. Londrina: Embrapa Soja, 2019. (Circular Técnica, 169). Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/223151/1/Circ-Tec-169.pdf. Acesso em: 31 mar. 2025.

EMBRAPA. *Sistema de produção para a cultura da soja na Região Central do Brasil*. Londrina: Embrapa Soja, 2000. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/95489/1/SP-16-online.pdf. Acesso em: 31 mar. 2025.

FERENTINOS, Konstantinos P. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, v. 145, p. 311–318, 2018.
GU, J. et al. Recent advances in convolutional neural networks. *Pattern Recognition*, v. 77, p. 354–377, 2018.

HARVARD MEDICAL SCHOOL. *Artificial intelligence in medicine and healthcare: Applications and challenges*. Harvard University, 2023.
HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The elements of statistical learning: data mining, inference, and prediction*. 2. ed. New York: Springer, 2009. Disponível em: https://link.springer.com/book/10.1007/978-0-387-84858-7. Acesso em: 21 abr. 2025.

INSTITUTO FEDERAL GOIANO. *Aplicações de aprendizado de máquina na agricultura de precisão*. 2024.

KAMILARIS, A.; PRENAFETA-BOLDÚ, F. X. Deep learning in agriculture: a survey. *Computers and Electronics in Agriculture*, v. 147, p. 70–90, 2018.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998.

MOHAMMED, Sedir et al. The effects of data quality on machine learning performance. *arXiv*, 2022. Disponível em: https://arxiv.org/pdf/2207.14529. Acesso em: 31 mar. 2025.

OLIVEIRA, Luiz H. M.; SILVA, João A. Aplicações do aprendizado de máquina na agricultura brasileira. *Estudos Avançados*, v. 34, n. 98, p. 129–147, 2020. Disponível em: https://www.scielo.br/j/ea/a/wXBdv8yHBV9xHz8qG5RCgZd. Acesso em: 31 mar. 2025.

SILVA, J. A. O. S. *Aplicação de modelos de aprendizagem profunda para detecção e segmentação de plantas daninhas em imagens RGB capturadas por UAV*. 2024. Dissertação (Mestrado em Ciência da Computação) – Instituto Federal Goiano, 2024. Disponível em: https://repositorio.ifgoiano.edu.br/bitstream/prefix/4716/5/Disserta%C3%A7%C3%A3o_JOSEF%20AUGUSTO%20OBERDAN%20SOUZA%20SILVA.pdf. Acesso em: 14 abr. 2025.

STOLLER DO BRASIL. *Doenças na soja: o impacto na produtividade da cultura*. Disponível em: https://www.stoller.com.br/blog/o-impacto-das-doencas-da-soja-na-produtividade/. Acesso em: 14 abr. 2025.