Universidade Estadual de Londrina

MURILO CAMINOTTO BARBOSA

ABORDAGEM BASEADA NA EXTRAÇÃO DE ATRIBUTOS DO PIXEL PARA QUANTIFICAR A SEVERIDADE DA FERRUGEM ASIÁTICA EM IMAGENS DE FOLHA DE SOJA

Londrina
2021

MURILO CAMINOTTO BARBOSA

ABORDAGEM BASEADA NA EXTRAÇÃO DE ATRIBUTOS DO PIXEL PARA QUANTIFICAR A SEVERIDADE DA FERRUGEM ASIÁTICA EM IMAGENS DE FOLHA DE SOJA

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

Orientador: Dr. Alan Salvany Felinto

Londrina
2021

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Barbosa, Murilo .
Abordagem baseada na extração de atributos do pixel para quantificar a severidade da ferrugem asiática em imagens de folha de soja / Murilo Barbosa. - Londrina, 2021.
49 f. : il.

Orientador: Alan Salvany Felinto.
Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Ciência da Computação, 2021.
Inclui bibliografia.

1. Processamento digital de imagens - Tese. 2. Aprendizado de Máquina - Tese. 3. Ferrugem Asiática - Tese. I. Salvany Felinto, Alan. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU 519

MURILO CAMINOTTO BARBOSA

ABORDAGEM BASEADA NA EXTRAÇÃO DE ATRIBUTOS DO PIXEL PARA QUANTIFICAR A SEVERIDADE DA FERRUGEM ASIÁTICA EM IMAGENS DE FOLHA DE SOJA

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

BANCA EXAMINADORA

_____

Orientador: Dr. Alan Salvany Felinto
Universidade Estadual de Londrina – UEL

_____

Prof. Dr. Sylvio Barbon Junior
Universidade Estadual de Londrina – UEL

_____

Prof. Dr. Marcelo Giovanneti Canteri
Universidade Estadual de Londrina – UEL

Londrina, 16 de junho de 2021.

Se você não gosta do seu destino, não o aceite. Em vez disso, tenha a coragem para transformá-lo naquilo que você quer que ele seja.

AGRADECIMENTOS

Os agradecimentos principais são direcionados a minha família, principalmente meus pais, Flávio e Rosângela, que tem se esforçado ao máximo para manter um teto em nossas cabeças e comida gostosa em nossas barrigas. Agradeço aos meus avós, Haroldo e Leonor, e meus tios que tem se preocupado comigo mesmo durante este período de pandemia onde o grupo de maior risco são eles e não eu. Agradecimentos especiais são direcionados a minha amada namorada, Barbara Oliveira, que suportou de perto minhas crises de ansiedade durante a escrita e manutenção deste projeto, e que foi capaz de botar minha cabeça em ordem. Quero agradecer ao meu orientador Alan Salvany Felinto por ter me ensinado a escrever uma dissertação de mestrado. Agradecer ao Marcelo Canteri e ao Lucas Fantin por terem tido a paciência de me ensinar os termos agronômicos que não conhecia, e pelo suporte financeiro na publicação do artigo.

Sem qualquer um de vocês eu tenho certeza que este trabalho seria mil vezes mais difícil.

Quem confere fere confere será conferido

BARBOSA, M. C. **Abordagem baseada na extração de atributos do pixel para quantificar a severidade da ferrugem asiática em imagens de folha de soja.** 2021. 49 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2021.

RESUMO

A soja é um produto global, servindo como alimento tanto para humanos quanto para animais e possui subprodutos fabricados na industria, como óleo de soja, Shoyu, missô, Tofu, lecitina de soja (emulsificante para produtos como sorvete, achocolatados, salsichas, barras de cereais, e produtos congelados), com demanda na sociedade e portanto valor comercial agregado. A Ferrugem Asiática é uma doença de soja que ocorre principalmente em regiões tropicais e subtropicais, que causa prejuízos para o Brasil na ordem de bilhões de dólares anuais. A eficiência do controle da doença, assim como as tomadas de decisões relacionadas ao manejo do cultivo são avaliadas de acordo com nível da severidade. Quantificar a severidade da Ferrugem Asiática é uma das chaves para o manejo e controle a base de fungicida. A quantificação do nível de infecção da doença conta com o auxílio da inspeção visual de especialistas que utilizam escalas diagramáticas para estimarem a severidade da Ferrugem Asiática na plantação. A estimativa por análise visual, por especialistas, possuem erros inerentes a condição humana, e por este motivo, análises de imagem e estimativas feitas por modelos criados, por métodos de aprendizado de máquina, tem sido aplicados na área da agronomia. Este trabalho, utiliza técnicas de Processamento de Imagens e aprendizado de máquina para quantificar o grau da Severidade da Ferrugem Asiática. O diferencial deste trabalho é a extração de descritores baseado em pixeis, fundamentado nas cores apresentadas em cada estágio do fungo causador da doença, sendo que todos os estágios podem ser encontrados em uma mesma folha. A abordagem é pautada na classificação dos pixeis entre as classes ‘sadio’ e ‘patogênico’, com um total de 160 mil amostras de pixeis retiradas de 7 (sete) folhas, sendo estas uma de cada nível de infecção atestado pelo especialista. Ao classificar cada pixel a área afetada pela doença é automaticamente segmentada durante o processo de classificação, consequentemente, não há necessidade de outro custo computacional para realizar a segmentação da área afetada pelo fungo. A divisão entre o total de pixeis classificados como patogênicos pelo total dos pixeis da folha resulta no nível de infecção da ferrugem Asiática na folha da soja. Ao término da análise dos resultados, o melhor modelo foi criado pelo algoritmo CART, obtendo uma acurácia superior a 97% na classificação entre pixeis sadios e patogênicos.

Palavras-chave: processamento digital de imagens; aprendizado de máquina; ferrugem asiática.

BARBOSA, M. C. **Pixel attribute extraction-based approach to quantify Asian rust severity in soybean leaf images.** 2021. 49 p. Master's Thesis (Master in Science in Computer Science) – State University of Londrina, Londrina, 2021.

ABSTRACT

Soy is an global product, serving as food for humans and animals and has by-products manufactured in the industry, such as soybean oil, Shoyu, miso, Tofu, soy lecithin (emulsifier for products such as ice cream, chocolate, sausages, cereal bars, and frozen products), with demand in society and therefore added commercial value. Asian Rust is one of the most devastating soy diseases in tropical and subtropical regions, causing losses to Brazil in the order of billions of dollars annually. The efficiency of disease control, as well as making decisions related to crop management, are evaluated according to the level of severity. Quantifying the severity of Asian Rust is one of the keys to fungicide-based management and control. The quantification of the level of infection of the disease has the aid of visual inspection by specialists who use diagrammatic scales to estimate the severity of Asian Rust in the plantation. Estimation by visual analysis, by specialists, has errors inherent to the human condition, therefore, image analysis and estimates made by models created by machine learning methods have been applied in the area of agronomy. This work uses Image Processing techniques and machine learning to quantify the level of Severity of Asian Rust. The differential of this work is the extraction of descriptors from pixels, from the colors presented in each stage of the fungus that causes the disease. The approach is based on the classification of pixels between the 'healthy' and 'pathogenic' classes, with a total of 160 thousand pixel samples taken from 7 (seven) leaves, one of each level of infection attested by the specialist. The area affected by the disease is automatically segmented during the classification process,consequently, there is no need for another computational cost to perform the segmentation of the area affected by the fungus. The division between the total number of pixels classified as pathogenic by the total number of pixels on the leaf results in the level of Asian rust infection in the soybean leaf. At the end of the analysis of the results, the best model was created by the CART algorithm, obtaining an accuracy greater than 97% in the classification between healthy and pathogenic pixels.

Keywords: digital image processing; machine learning; asian rust.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARFF	Arquivo no formato Relação-Atributo
ASSESS	Image Analysis Software for Plant Disease Quantification
AUC	Área sob a Curva
CV	Cross-Validation
FN	Falso Negativo
FP	Falso Positivo
Gimp	GNU Image Manipulation program
ML	Machine Learning
NDVI	Índice de Vegetação com diferença Normalizada
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SVM	Máquinas de Vetor de Suporte
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo
UEL	Universidade Estadual de Londrina

SUMÁRIO

1 INTRODUÇÃO

A Ferrugem Asiática, que é causada pelo fungo *Phakopsora pachyrhizi* Syd & P. Syd, é uma doença de soja que ocorre principalmente em regiões tropicais e subtropicais, causando apenas no Brasil, perdas anuais de aproximadamente dois bilhões de dólares [4]. A Ferrugem Asiática primeiro ocorreu no Brasil em 2001 [5], e foi reportada na Argentina em 2003 [6] e nos Estados Unidos em 2004 [7]. A redução na produtividade causada por esta doença está diretamente relacionada com sua progressão e severidade, que estão ligados a fatores bióticos e abióticos. Juntamente com estes fatores, condições ambientais podem afetar a doença, como temperatura e umidade [8]. Estudos apontam que Brasil, Argentina e Estados Unidos têm maiores chances de epidemias desta doença devido às condições climáticas [9]. Outra situação que contribui para a manifestação da doença é o plantio de soja no período entre-safras [10].

A soja é um importante produto global, servindo como alimento para animais e tendo diversos usos para seus subprodutos na indústria, como cosméticos, fibras e farmacêuticos. Argentina, Brasil e Estados Unidos são os maiores produtores, sendo o Brasil o líder de exportação contribuindo com cerca de 40% do suprimento global [11].

Estratégias de controle são compostas de métodos integrados, como escolha da data e ciclo de cultivo [12], implementação de períodos-livre de soja [13], nutrição do plantio [14], utilização de cultivos resistentes [15], monitoramento do campo [16, 17, 18] e aplicação de fungicidas [19, 20, 21, 22]. A eficiência do controle dessas estratégias, assim como as tomadas de decisões relacionadas ao manejo do cultivo são avaliadas de acordo com nível da severidade. O nível da severidade geralmente é quantificado através de estimativas visuais, e tem migrando para análise de imagens no espectro visível, hyperspectral e multispectral [23].

A automação de detecção e classificação de doenças na agricultura é encorajada e estudada por pesquisadores para todos os tipos de plantas e frutas devido a eficiência da aplicação [24, 25]. Alguns dos problemas quando se lida com detecção de doenças em plantas são: extração dos descritores devido a variação da luminosidade e principalmente o tamanho do banco de imagens necessário para o treinamento de modelos de aprendizado de máquina [26]. Devido a dificuldade em obter bancos de imagens para o desenvolvimento de novos modelos, pesquisadores têm se dedicado a criação destes tipos de banco de imagens para diferentes tipos de doenças, de forma que possam ser utilizados em trabalhos futuros [27].

A acurácia na quantificação do nível de infecção de Ferrugem Asiática melhora a tomada de decisão relacionada ao manejo e controle desta doença, principalmente devido

a eficácia da aplicação de fungicidas ser influenciada por este fator. Além disso, a aplicação incorreta do fungicida pode levar a ocorrência de resistência ao fungicida [28, 29].

Neste contexto, a quantificação da Ferrugem Asiática em seus estágios iniciais é uma das chaves para o manejo e controle a base de fungicida [17]. A quantificação da Ferrugem Asiática já é um problema conhecido, e trabalhos como [30] já foram feitos para resolver este problema, porém esta dissertação traz uma nova abordagem para encontrar a solução. Esta nova abordagem consiste em obter vantagem no fato de que todos os estágios da Ferrugem Asiática, com toda a sua amplitude de cores, pode ser encontrada em uma única folha [31]. Métodos que utilizam aprendizagem profunda para a classificação de doenças em folhas de plantas (como CNN por exemplo) recebem como entrada exemplos de imagens da folha, de forma que cada camada possa extrair diferentes tipos de informações e combiná-las no final para encontrar a classe correspondente. Este processo requer uma quantidade de imagens na casa dos milhares para que o modelo tenha uma acurácia satisfatória. Em contraste, a abordagem proposta, os atributos são extraídos diretamente de cada pixel, portanto o número de amostras é definido pela quantidade de pixeis e não mais pelo número de imagens de folhas.

Extraindo informações de cada pixel, esta abordagem visa obter maior quantidade de informação de cada imagem de folha, o que resolve um dos problemas quando se lida com doenças em folhas [32, 26]. O objetivo é medir o nível de ferrugem asiática através da classificação dos pixeis entre sadio e patogênico, pois o nível da infecção pode ser encontrado dividindo o total de pixeis classificados como patogênicos pelo total de pixeis da folha.

Utilizando esta abordagem, uma acurácia superior a 97% foi obtida na classificação entre pixeis sadios e patogênicos. O comportamento do modelo foi comparado com um total de três especialistas humanos que se voluntariaram para classificar o nível da Ferrugem Asiática de acordo com os métodos convencionais, e o software ASSESS que é utilizado na área de agronomia para avaliação de doenças.

As principais contribuições deste trabalho são:

- Extrair informações a nível de pixel, reduzindo a quantidade de imagens necessárias para treinar um modelo de aprendizado de máquina.
- Reduzir erro associado a quantificação do nível de Ferrugem Asiática, uma vez que métodos visuais possuem grande erro associado.
- Reduzir a necessidade de utilizar algoritmos para segmentação de regiões patologicamente diagnosticadas, uma vez que tem-se todos os pixeis classificados, a doença estará segmentada.

A estrutura do trabalho segue da seguinte maneira: Capítulo 2 contem a fundamentação teórica com explicações sobre descritores, seleção de atributos destes descritores, aprendizado de máquina, métricas e trabalhos correlatos. No Capítulo 3 é abordado o procedimento metodológico com análise e explicações sobre as etapas de aquisição de imagens, identificação da doença e extração e seleção dos descritores. Capítulo 4 aborda a análise de resultados contendo os ajustes, resultados experimentais e comparações com outras métricas. Por fim o Capítulo 5 contendo a conclusão deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Para o entendimento teórico desta dissertação, abordou-se os tópicos como descritores, seleção de atributos presentes nestes descritores, algoritmos de classificação e métricas de avaliação, além de um estudo dos trabalhos correlatos na área de diagnósticos de doenças em folhas.

2.1 Descritores

Descritores são informações retiradas de um objeto ou problema que sejam capazes de nos informar algo sobre o objeto ou problema em questão. Se quantificar o peso de um objeto, por exemplo, este descritor é capaz de informar a força necessária (através de uma medida padronizada) para elevar esse dado objeto acima do nível do solo, por outro lado, quando extraída a cor de um objeto, este descritor informa qual é o comprimento de onda refletido por ele, e assim por diante. Existem uma ampla variedade de descritores, podendo ser agrupados de acordo com o que descrevem, como por exemplo: descritores de cor, textura, forma, etc.

Se tratando de processamento digital de imagens e detecção de doenças em plantas, sistemas de cores, como HSV, RGB YCbCr, e LAB são os descritores mais comumente usados. O sistema LAB é o tido como o mais poderoso no quesito de remoção de ruídos adicionados por câmeras e perturbações de fundo, como é discutido em [33]. As técnicas que mais utilizam estes sistemas são Otsu e *K-means* como verificado em [34], cujo objetivo era pesquisar diferentes técnicas para detecção de doenças em folhas. Entretanto, aprendizado de máquina é comumente aplicado para detecção de doenças usando estes mesmos sistemas de cores; por exemplo, SVM e PSO foram utilizados juntamente com os sistemas HSI e LAB para criar modelos que alcançaram 98% de acurácia na detecção de doenças em girassóis [35], e 88,9% para a detecção míldio como descrito em [36].

O sistema RGB é o sistema mais utilizado na computação gráfica, onde R indica Vermelho (Red), G indica Verde (Green) e B indica Azul (Blue). Essas cores são chamadas de cores aditivas, e suas combinações podem criar todas as outras [37].

O sistema HSI é baseado na percepção humana das cores, onde H indica matiz (Hue) que descreve a cor pura, S indica a saturação (Saturation) e I indica a intensidade (Intensity) [33]. Este sistema pode ser derivado do RGB com as seguinte fórmulas:

$$H = \begin{cases} \theta, & \text{if } B \leq G \\ 360 - \theta, & \text{caso contrário} \end{cases} \tag{2.1}$$

com

$$\theta = \cos^{-1}\left(\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{\frac{1}{2}}}\right) \tag{2.2}$$

onde $360 - \theta$ é o ângulo em relação ao eixo vertical sendo este último correspondente a luminosidade

$$S = 1 - \frac{3 * \min(R,G,B)}{(R+G+B)} \tag{2.3}$$

$$I = \frac{1}{3}(R+G+B) \tag{2.4}$$

No sistema Lab, L descreve o brilho, 'a' descreve a cor do verde até o vermelho e 'b' descreve a cor do azul até o amarelo [33]. Este sistema pode ser derivado do RGB com as seguintes fórmulas:

$$L = 0.2126 * R + 0.7152 * G + 0.0722 * B \tag{2.5}$$

$$a = 1.4749 * (0.2213 * R - 0.3390 * G + 0.1177 * B) + 128 \tag{2.6}$$

$$b = 0.6245 * (0.1949 * R + 0.6057 * G - 0.8006 * B) + 128 \tag{2.7}$$

2.2 Seleção de Atributos

O processo de seleção de atributos tem como objetivo melhorar os resultados obtidos com o conjunto de atributos. Um atributo irrelevante pode aumentar a complexidade do modelo ou causar distúrbios nos resultados. Uma grande quantidade de atributos não significa necessariamente a criação de um modelo mais robusto ou melhor, podendo aumentar o tempo necessário para que haja o treinamento do modelo. Por estes e outros motivos, o processo de seleção de atributos é utilizado no processo de aprendizado de máquina [38, 39].

Os métodos de seleção de atributos podem ser divididos entre:

- Métodos de Filtragem:

  Métodos de filtragem são métodos que atribuem pontuações aos atributos analisados através de medidas estatísticas, normalmente univariadas, que consideram a independência do atributo com relação a variável alvo. Este tipo de método costuma ser aplicado em bancos de dados com grande número de atributos, uma vez que, o poder computacional exigido para aplicar um métodos de filtragem é consideravelmente menor que outros métodos de seleção de atributos [40]. O ranqueamento é feito tendo por base a capacidade de um atributo conter informações relevantes sobre diversas

classes, sendo este comportamento caracterizado como relevância do atributo [41]. Devido ao fato do método de filtragem criar um ranqueamento dos atributos, um parâmetro contendo a quantidade de atributos que devem ser selecionados, geralmente é utilizado para o retorno destes atributos. Um ponto importante para ressaltar é que um subconjunto ótimo de atributos não é necessariamente único, uma vez que para um mesmo classificador, pode-se alcançar a mesma acurácia utilizando diversos subconjuntos [42].

O coeficiente de correlação de Pearson [41] é um dos métodos de filtragem mais utilizados atualmente no processamento de imagens [43], geralmente aplicado na forma de uma matriz, chamada de matriz de correlação. Este método é capaz de detectar dependências lineares entre a variável e o alvo, retornando valores entre -1 e 1, onde +-1 significa máximo de correlação e 0 indica que as variáveis não são correlacionadas. É utilizado para calcular o grau de redundância de um atributo perante outros.

- Métodos de Conjuntos:

  Métodos de Conjuntos são métodos que utilizam modelos preditivos para atribuir pontuações baseadas na acurácia que um determinado conjunto de atributos alcançou no modelo escolhido. Este método visa criar um ranqueamento, porém baseado em conjuntos distintos de atributos e não em atributos isolados. Devido ao número de subconjuntos crescer exponencialmente, este método se torna custoso para bancos de dados com um grande número de atributos. Em geral tende-se a aplicar métodos heurísticos ou algoritmos evolucionários como algoritmos genéticos [44] ou optimização de enxame de partículas [41] para a escolha dos subconjuntos visando diminuir o custo operacional desta abordagem, uma vez que abordagens exaustivas se tornam impraticáveis [42].

  Um algoritmo utilizado é o algoritmo de Remoção de Atributo por Recursividade ou RFE (*Recursive Feature Elimination*) [45, 46]. O algoritmo RFE recebe como parâmetro um modelo preditivo que será utilizado para testar o subconjunto de atributos criado a cada iteração. O primeiro subconjunto de atributos contém todos os atributos a serem analisadas (conjunto completo) e então, a cada nova iteração um atributo é removido do subconjunto e o teste é repetido. Ao final do algoritmo tem-se um ranking contendo todas as possibilidades de subconjuntos e seus respectivos valores alcançados de acurácia com o modelo preditivo proposto.

- Métodos embutidos:

  Métodos Embutidos são métodos que aprendem quais atributos melhor contribuem para o resultado final do modelo durante a criação do próprio modelo [47, 48]. A principal motivação para estes métodos é reduzir o tempo computacional gasto du-

rante a reclassificação dos subconjuntos, como é feito nos métodos de conjuntos. A ideia é incorporar a seleção de atributos durante a etapa de treinamento do modelo classificador. Este tipo de abordagem é utilizado principalmente em algoritmos de árvores de decisão como por exemplo o *RandomForest*. O algoritmo *RandomForest* possui internamente (embutido) um algoritmo responsável por atribuir valores aos atributos de acordo com sua importância [41], selecionando assim os melhores atributos para criar seu modelo final. Desta forma, é possível utilizar algoritmos como este para realizar a seleção de atributos como um pré-processamento para uma aplicação em outro modelo [49].

2.3 Escala diagramática

Escalas diagramáticas são ferramentas utilizadas para avaliar a severidades de doenças de plantas. Para a criação de uma escala diagramática, um grupo de imagens de folha apresentando níveis variados de sintomas são coletados e cada imagem tem seu nível avaliado [50, 51, 52]. Após a etapa de avaliação, uma escala de níveis é proposta, geralmente obedecendo a lei de Webber Fechner [53]. Para a validação, um grupo de imagens com os níveis de infecção conhecidos são distribuídos para avaliadores inferirem seus níveis utilizando como ferramenta a escala recém-crida. Caso as avaliações feitas pelos avaliadores possuam um grau de acurácia elevado e uma coerência entre os avaliadores, pode-se dizer que a escala foi validada [54].

2.4 Aprendizado de Máquina

Inteligência Artificial (IA) possui um amplo escopo e tende a ser definida como uma ciência que tenta mimetizar habilidades humanas. Dentro do amplo escopo de IA, tem-se um ramo específico chamado de Aprendizado de Máquina (ML - do inglês *Machine Learning*) dedicado a criação de algoritmos capazes de aprender com dados, gerando modelos que possuem um poder preditivo relacionado aos objeto de estudo. ML é um método de análise de dados que automatiza a construção de modelos analíticos, identificando padrões e tomando decisões com o mínimo de interferência humana [55].

Com o constante aumento da disponibilidade de informações, barateamento de armazenamento de dados digitais e poder computacional, este tipo de ciência tem sido aplicada nas mais diversas áreas, de saúde a marketing e vendas. No ramo da agricultura não poderia ser diferente, com as mais variadas aplicações para detecção e classificação de doenças em frutas, verduras, grãos e folhas.

Alguns dos métodos mais populares de aprendizado dentro de ML são:

- **Aprendizado supervisionado**: é treinado através de exemplos rotulados, onde

cada entrada possui a saída desejada já conhecida. Este tipo de aprendizado utiliza padrões para prever os rótulos de dados adicionais não rotulados. Comumente utilizado em aplicações cujos dados históricos preveem eventos futuros.

- **Aprendizado não-supervisionado**: é utilizado em conjuntos de dados não rotulados, onde o objetivo é encontrar padrões ou estruturas dentro do conjunto analisado. O resultado deste tipo de método são agrupamento dos dados. Comumente utilizado para recomendar produtos ou itens para conjuntos de clientes.

- **Aprendizado semi-supervisionado**: possui as mesmas aplicações do aprendizado supervisionado, com o diferencial que parte dos exemplos (em geral, a maioria deles) não está rotulada. É comumente aplicado em casos em que o custo associado a rotular um exemplo é elevado.

2.5 Classificação

Se tratando de aprendizado de máquina, um modelo de classificação tem por objetivo receber um conjunto de dados (atributos) como entrada e retornar como saída uma das possíveis classes atribuídas ao problema em questão. Estas classes, em geral, são categóricas, ou seja, um rótulo definido por um humano especialista no problema em questão. Um algoritmo de classificação induz um modelo através de análises estatísticas do conjunto de dados de treinamento de forma que ao ser inserido um novo exemplo, o modelo seja capaz de predizer a classe a qual pertence.

Com relação ao tipo de treinamento, o escolhido para este trabalho foi o treinamento supervisionado, que consiste em ter todos exemplos utilizados no treino já classificados por um especialista. Os classificadores utilizados neste trabalho foram: Árvores de Decisão e Máquina de Vetor de Suporte, principalmente devido a sua ampla utilização em trabalhos correlatos e facilidade de aplicação com o método de treinamento escolhido.

2.5.1 Árvore de Decisão

Uma árvore de decisão permite que um indivíduo (ou máquina) compare possíveis ações baseado em seus custos e benefícios. Este método é comumente utilizado em sistemas classificadores por meio da avaliação de atributos em um conjunto de dados [56].

Árvores de Decisão são constituídas no mínimo de:

1. Nós: são a representação dos atributos, contém sua descrição e valor. O primeiro nó de uma árvore é chamado de raiz; dois nós diretamente ligados por uma aresta constituem um nó pai, sendo este o nó do nível superior, e o nó filho, sendo este o nó do nível inferior.

2. Arestas: são "linhas"que tem origem de um nó superior e se conectam com um nó de um nível abaixo; possui um valor associado que representa o "corte"feito no plano do atributo de origem.

3. Nós folha: são nós que não possuem arestas ligando-os a níveis mais abaixo, encerrando a profundidade de um determinado "galho"da árvore. Os nós folhas determinam a classe.

4. Galho: Subconjunto de nós de uma árvore, contendo um nó inicial (raiz) e seu conjunto de ramificações (nós filhos). Um galho também pode ser um conjunto bem definido de nós interligados, ignorando algumas de suas ramificações.

Dada uma representação gráfica de uma árvore de decisão pode-se seguir suas arestas verificando, a cada novo nó, qual aresta deve-se seguir baseado no valor que existe no atributo sendo analisado atualmente, chegando assim a uma das folhas da árvore e consequentemente a uma das classes. O caminho da raiz (primeiro nó de uma árvore) até a folha, chamado de regra de decisão. A Figura 1 ilustra uma árvore de decisão e os respectivos cortes no plano.

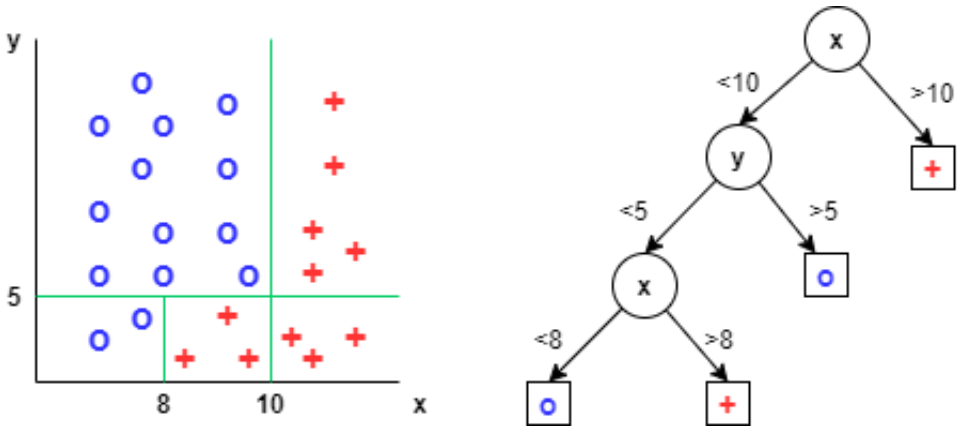

Figura 1 – Representação de uma árvore de decisão e os respectivos cortes no plano; Imagem adaptada de: Monard e Baranauskas[1]

Algoritmos de árvore de decisão, em geral utilizam uma função de mérito que define o melhor atributo para realizar um novo "corte"e criar novas arestas.

Uma função de mérito é um algoritmo capaz de avaliar a contribuição individual de cada atributo para predizer uma determinada classe. Dentre os algoritmos mais utilizados para esta função está o *Ganho de informação* (será abordado na seção 2.6.1)

Para evitar que uma árvore cresça indefinidamente, regras de poda costumam ser aplicadas. Uma poda evita que um novo galho seja criado, possivelmente diminuindo a profundidade da árvore. Regras de podas costumam ser aplicadas dado o fato de que uma árvore de decisão com uma grande profundidade tende ao super-ajuste, o que prejudica o

processo de classificação de novas instâncias. Duas das principais regras de poda utilizadas são: um número máximo de níveis; um limiar de erro a se alcançar. Algumas vantagens de árvores de decisão são:

1. Flexibilidade: Árvores não assumem nenhuma distribuição para os dados e criam decisões para todo o espaço de busca.
2. Robustez: Possui uma sensibilidade a valores *outliers* (fora do padrão) reduzida.
3. Seleção de atributos: Durante a construção da árvore são selecionados os atributos mais relevantes através da função de mérito.

Algumas desvantagens:

1. atributos contínuos: vários autores recomendam discretizar este valores devido ao pré-processamento necessário para utilização da função de mérito.
2. Valores ausentes: uma vez que o valor não é conhecido não é possível realizar sua predição.
3. Instabilidade: pequenas mudanças no conjunto de treino podem significar grandes alterações na árvore resultante.

Os modelos baseados em árvores de decisão utilizados neste trabalho foram:

- *Random Forest*:

  É um algoritmo que cria um conjunto de árvores de decisão no momento de treinamento, ao invés de uma única árvore. Ao indicar uma instância a ser classificada, o retorno é a classe que mais foi "votada"por cada uma de suas árvores individuais. Devido às múltiplas árvores criadas serem treinadas com diferentes partes do conjunto de treino, este método tende a evitar o super-ajuste, e reduzir a variância, porém perde completamente a capacidade de ser interpretável por um humano [57].

- *Random Tree*:

  É uma árvore de decisão que é criada usando partes aleatórias do conjunto de atributos disponíveis e que não utiliza regras de poda. Este método é aplicado com intuito de evitar o super-ajuste. [58].

- CART:

  É algoritmo de árvore de decisão binária muito semelhante ao C4.5 que é uma evolução do algoritmo pioneiro em indução de árvores ID3 [59].

A árvore é construída de forma recursiva e baseada em busca gulosa, procurando, sobre um conjunto de atributos, aqueles que “melhor” dividem os exemplos, gerando sub-árvores [41].

- RepTree:

  possui como objetivo um rápido aprendizado, criando novos ramos através do método de ganho de informação e realizando podas através do método de redução de erro que consiste em remover nós iterativamente sendo estes escolhidos de forma que maior aumente a acurácia da árvore de decisão, atribuindo a classe de maior dominância do ramo removido a este nó. Este método tende a gerar árvores menores e de fácil interpretação [60].

2.5.2 Máquina de Vetor de Suporte

Máquinas de Vetor de Suporte (SVM) se baseiam nos princípios da teoria do aprendizado estatístico. Uma SVM é uma implementação de um mapeamento não-linear dos dados de entrada para um espaço de descritores de alta-dimensão, onde é criado um hiperplano ótimo para separar os dados linearmente em duas classes [41]. Este mapeamento é executado por um kernel bem definido escolhido anteriormente.

O kernel é a função responsável por separar inicialmente as classes, podendo ser:

- Linear: divide o plano de forma linear
- Polinomial: aplica uma função polinomial para a divisão do plano, exigindo como parte do parâmetro um grau para este polinômio.
- RBF: divide o plano tendo por base o crescimento de uma função de base radial, necessitando do parâmetro "gamma", que indica a importância de uma única instância no conjunto.
- Sigmoid: divide o plano através de uma tangente hiperbólica. Este Kernel tem sua origem de redes neurais artificias.

Se os dados de entrada forem linearmente separáveis, o hiperplano considerado ótimo é aquele que apresenta a máxima margem de separação. Para a criação desta margem, primeiro a SVM classifica as classes corretamente e depois, em função desta restrição, calcula a melhor distância entre as margens (Figura 2). Os elementos de bordas utilizados para definir a margem são chamados de vetores de suporte.

Outro parâmetro que pode ser manipulado, afetando diretamente o resultado é o chamado "termo de regularização"(também conhecido como "custo"). Este custo é penalidade associada as instâncias que são classificada erroneamente. Com um custo elevados,

tem-se a tendência de evitar erros durante a etapa de classificação, o que pode levar a superajustes. Custos mais baixos tendem a criar separações mais suaves entre as classes.

A utilização de uma SVM deve ser feita com cuidado, uma vez que o poder computacional necessário para sua execução pode crescer exponencialmente para bases muito grandes, uma vez que este algoritmo exige a inversão de matrizes. Elementos que fogem muito do padrão, e/ou ruídos podem causar grandes distúrbios na construção do modelo.

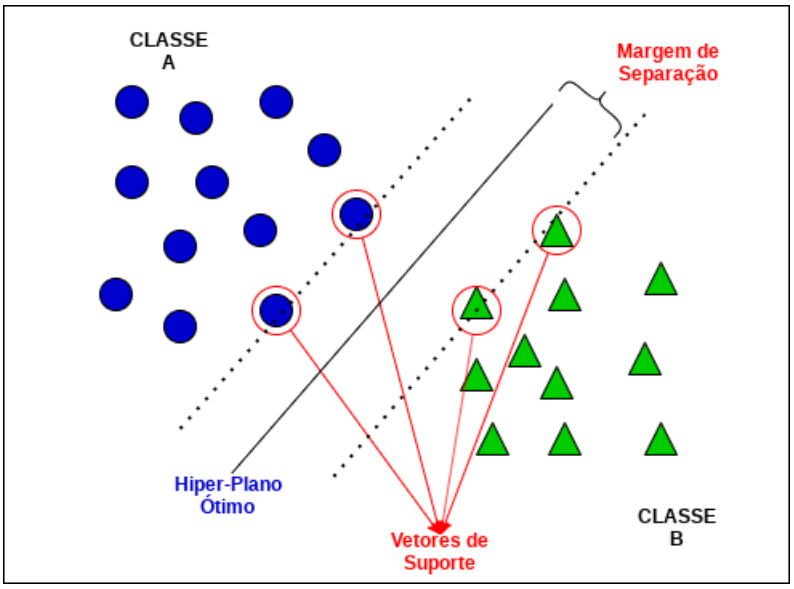

Figura 2 – Representação da separação de classes criada pelo algoritmo SVM e o hiperplano ótimo com os vetores de suporte utilizado para sua criação; Imagem adaptada de: Lorena e Carvalho[2]

2.6 Métricas

No contexto de classificação por aprendizado de máquina, métricas são funções ou medidas capazes de trazer informações úteis a respeito do desempenho do modelo criado pela técnica de aprendizado de máquina aplicada. Em geral, mais de uma métrica é utilizada para afirmar se um modelo foi bem sucedido ou não. Estas funções ou medidas podem ser aplicadas para compreender comportamentos específicos, como por exemplo saber se um modelo está acertando mais os casos positivos ou negativos.

Métricas também podem ser aplicadas em outras etapas do processo de classificação, como por exemplo na averiguação da importância dos descritores. É de extrema importância a utilização de métricas para a análise da importância dos descritores para um bom processo de seleção destes descritores.

Para avaliar o modelo como um todo, diversas métricas podem ser utilizadas. As medidas mais utilizadas são as taxas de: Verdadeiro Positivo (TP), Verdadeiro Negativo

(TN), Falso Positivo (FP) e Falso Negativo (FN). Estas medidas podem ser aplicadas em conjunto para criar outras métricas de avaliação. As métricas e funções utilizadas nestes trabalhos são: Ganho de informação, Acurácia, Precisão, Sensibilidade, Area sob Curva, *F-score* (a média entre as classes) e o Coeficiente de correlação de Pearson.

2.6.1 Ganho de informação

O *Ganho de informação* é frequentemente aplicado como função para medir a relevância de um atributo, medindo a quantidade de presença ou abstenção deste atributo no conjunto de dados utilizado para treino. Seu cerne é a medida de Entropia, que mede a aleatoriedade dos dados, calculada da seguinte forma:

$$H(A) = -\sum_{i} Pi * \log_2 Pi \tag{2.8}$$

onde

- P: é a probabilidade de se observar um valor A

Durante a aplicação desta função em uma árvore de decisão, a cada nó , o atributo que mais reduz a aleatoriedade da classe será escolhido para criar uma nova aresta, sendo o resultado do Ganho de informação justamente esta medida de aleatoriedade. Sua fórmula é a seguinte:

$$IG(A, p, q) = I(p, q) = E(A, p, q) \tag{2.9}$$

sendo

- 'p' e 'q': o número de objetos de duas classes diferentes.

2.6.2 Acurácia

A acurácia é um bom indicativo da performance do modelo. Ela mede a frequência de acerto com relação ao total de amostras, ou seja, indica a quantidade de acertos obtidos pelo modelo. Uma desvantagem desta métrica é que ela pode ser enganosa quando lidando com bases de dados desbalanceados. Sua fórmula é como segue:

$$Acuracia = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

2.6.3 Precisão

A precisão indica o quão bem o modelo é capaz de acertar os casos positivos (Verdadeiros Positivos). A precisão não leva em consideração os casos negativos, mesmo que estes sejam falsos negativos. Geralmente, esta métrica é utilizada quando um Falso Positivo pode causar graves consequências. Sua fórmula é como segue:

$$Precisao = \frac{TP}{TP + FP} \tag{2.11}$$

2.6.4 Sensibilidade

A sensibilidade indica o quão bem o modelo é capaz de acertar os casos negativos (Verdadeiros Negativos). A sensibilidade não leva em consideração os casos positivos, mesmo que sejam falsos positivos. Geralmente, esta métrica é utilizada quando um Falso Negativo pode causar graves consequências. Sua fórmula é como segue:

$$Sensibilidade = \frac{TP}{TP + FN} \tag{2.12}$$

2.6.5 Curva ROC

Uma vez que o resultado de sistemas de classificação geralmente estão situados dentro de um intervalo contínuo, é necessário definir um limiar de decisão para se classificar e contabilizar o número de predições positivas e negativas. Para cada limiar são calculados valores de sensibilidade e especificidade, que são dispostos em um gráfico denominado curva ROC (*Receiver Operating Characteristic*), apresentando no eixo das ordenadas os valores de sensibilidade e nas abscissas o seu complemento.

É uma abordagem gráfica para exibir a troca entre a taxa Verdadeiro Positivo (TP) e a taxa de Falsos Positivos (FP) de um classificador. TP é traçado ao longo do eixo Y e FP é traçado ao longo do eixo X. O desempenho de cada classificador é representado como um ponto na curva.

Um classificador perfeito possuiria uma linha horizontal no topo do gráfico, mas em geral as linhas consideradas "boas"se encontram entre a diagonal e a linha perfeita. Uma das medidas para comparação de sistemas é justamente a área sobre a Curva (AUC), uma vez tendo a curva ROC, basta aplicar um método de integral numérica, sendo o número obtido um discriminante da qualidade do sistema, quanto maior melhor [61].

2.6.6 *F-score*

Em análises estatísticas de classificação binária o *F-score* (ou *F-measure*) é a medida de acurácia, considerando tanto a precisão quanto a sensibilidade (também chamado de *recall*.

F-score é a média harmônica da precisão e sensibilidade, onde o melhor valor possível é 1 e o pior é 0 [41].

$$F = \frac{2PR}{P + R} \tag{2.13}$$

onde:

1. P: Precisão
2. R: Sensibilidade (*Recall*)

Em casos de bases com ambas as classes com a mesma importância, porem com números de amostras desbalanceados para uma das classes, tende-se a escolher modelos com maior *F-score* para ambas as classes. Em bases com uma das classes mais importante que a outra, tende-se a escolher o modelo com maior *F-score* para a esta classe em específico.

2.6.7 Coeficiente de correlação de Pearson

O coeficiente de correlação de Pearson, mede o grau da correlação entre duas variáveis. Geralmente representado por ρ, o coeficiente de correlação de Pearson assume valores entre -1 e 1, indicando se a correlação foi positiva, negativa ou se não houve correlação conforme ilustrado na Figura 3.

Figura 3 – Coeficiente de correlação de Pearson com resultados positivos (esquerda), negativos (centro) e sem correlação (direita) ; Adaptado de: Paranhos et al.[3]

Examinando o valor de ρ obtém-se 3 casos possíveis:

1. $\rho > 0$: Correlação positiva entre duas variáveis
2. $\rho < 0$: Correlação negativa entre duas variáveis, ou seja, uma variável é inversamente proporcional a outra.
3. $\rho = 0$: Não existe correlação linear entre as duas variáveis

Entretanto, a verdadeira importância do coeficiente de correlação de Pearson é descobrir se as variáveis estão ou não correlacionadas entre si, seja uma correlação positiva ou negativa. Por esse motivo, a análise de correlação é feita com o $|\rho|$, sendo interpretado da seguinte forma:

- Muito fraco: 0 - 0,19
- Fraco: 0,20 - 0,39
- Moderado: 0,40 - 0,59
- Forte: 0,60 - 0,79

- Muito forte: 0,80 - 1

2.6.8 Validação Cruzada

Ao aplicar métodos de ML, é fundamental avaliar a estabilidade do modelo criado, ou seja, tentar avaliar como este modelo vai se comportar com amostras desconhecidas. Em outras palavras, deve-se avaliar se o modelo criado foi capaz de generalizar os dados a partir do conjunto de treinamento, ou se captou muito ruido ou se super-ajustou ao conjunto de treinamento.

Uma das técnicas comumente utilizadas para avaliar a capacidade de generalização de um modelo cujo objetivo é a predição é a chamada Validação Cruzada (CV - do inglês, *Cross-Validation*). O objetivo da validação cruzada é estimar o quão preciso é um modelo na prática, ou seja, em um novo conjunto de dados. O ideal é que o modelo seja avaliado em amostras que não foram utilizadas para treino ou ajuste, de forma que exista um senso imparcial de eficácia do modelo [62].

O método de validação cruzada utilizado neste trabalho consiste em separar o conjunto de dados entre um conjunto de treino (contendo a maior parte das amostras) e um conjunto de teste que será utilizado apenas após a criação do modelo para avaliar seu desempenho. Este processo irá se repetir um total de 'K' vezes, onde o 'K' é um valor inteiro pré definido, que irá dividir o conjunto de dados em 'K' partes iguais, onde $(K-1)/K$ partes serão utilizadas para treino e o restante para teste e, a cada iteração tanto o conjunto de teste quanto o de treino serão diferentes. Ao criar um total de 'K' modelos construídos e testados com 'K' conjuntos distintos de amostras, é possível ter uma média da eficácia do modelo criado de forma imparcial.

2.7 Trabalhos Correlatos

Melo, (2015) [30] desenvolveu um sistema capaz de inferir o nível de ferrugem asiática em plantações de soja utilizando dispositivos móveis. Para tal trabalho, cerca de três mil exemplares de folhas, entre contaminadas e sadias, foram coletadas diretamente do campo e guardadas em sacos plásticos para prevenir a perda de umidade. Posteriormente, estas mesmas imagens foram submetidas a uma caixa preta para obtenção de imagens sem que houvesse perturbação de luzes externas. Após uma etapa de pré-processamento e segmentação, o fundo das imagens foi removido e o resultado foi utilizado para alimentar uma Rede Neural Convolucional. Como resultado 78,86% de acurácia foi obtido com relação a inferência do nível de severidade de ferrugem asiática.

Marcos (2019) [63] utilizou máscaras de kernel convolucional para realçar as cores presentes nas feridas das plantas, com o intuito de tornar o processo de segmentação por limite (threshold) mais viável. O tamanho do kernel pode variar amplamente, dependendo

das características da doença e seus padrões sintomáticos. Um problema que deve ser apontado por este tipo de abordagem, está relacionado a dificuldade de segmentação das regiões periféricas da lesão, o que pode causar distúrbios na avaliação e diminuir a confiabilidade do modelo. Tendo isto em vista, algoritmos genéticos foram empregados para descobrir o valor a ser utilizado no kernel, porém, o valor ótimo não é garantido de ser encontrado.

Khan (2018) [32] propôs um método para segmentar e reconhecer doenças em plantios de frutas baseado em coeficientes de correlação e descritores de aprendizado profundo. Seu método utiliza de uma técnica conhecida como *contrast stretching* que tem por objetivo aumentar o contraste entre a lesão causada pela doença e o fundo. Após esta etapa, um algoritmo genético é aplicado para apontar as características mais relevantes, que são utilizadas como entrada para uma Máquina de Vetor de suporte que é treinada para distinguir entre diferentes tipos de doenças. Este modelo foi capaz de alcançar 98,60% de acurácia, porém milhares de imagens de cada tipo de doença foram necessárias para treinar o algoritmo.

Khan (2019) [64] propôs um método de pré-processamento para segmentação seguido de classificação para doenças de maçãs . O método se baseia no conjunto de filtro de caixa 3D, filtro Gaussiano 3D, filtro da mediana 3D e decorrelação para realçar a ferida, e então a segmentação é feita através de um método de correlação forte. Após o realce e a segmentação, a seleção de descritores e classificação são feitas por um algoritmo genético e uma SVM Multi Classe respectivamente. O método proposto atingiu 97,20% de acurácia.

Ainda no tópico de pré-processamento, Adeel (2019) [65] propôs uma nova sequência de passos para aumento de contraste e segmentação de doenças em folhas de uva. Primeiro, a redução de borramento de contraste local (LCHR- do inglês "Local Contrast Haze Reduction") foi aplicado para aumentar o contraste local, então o melhor canal da transformação Lab é selecionado, baseando-se em informações de pixel. Após esta etapa, os descritores de cor, textura e geometria são combinados através da abordagem de análise de correlação canônica. A classificação foi feita através de uma SVM multiclasses, e obteve uma acurácia acima de 92%.

Adeel (2020) [66] desenvolveu um novo framework para seleção de descritores de aprendizagem profunda para o reconhecimento de doenças em folhas de uva. Neste trabalho, os modelos pré treinados AlexNet and ResNet101 são utilizados para a extração dos descritores através de técnicas de transferência de aprendizado. Os métodos de entropia Yager e a formulação de curtose são utilizados para seleção de descritores. Os descritores selecionados são inseridos em uma SVM para a classificação com o método de validação cruzada. Como resultado o modelo foi capaz de atingir 99% de acurácia no banco de dados *Plant Village*, porém o grande número de imagens necessárias para o aprendizado pode se tornar um problema quando lidando com novos tipos de dados.

A abordagem baseada em pixel proposta nesta pesquisa não irá necessitar de segmentação da doença, reduzindo assim a etapa de pré-processamento. Algoritmos baseados em pixel, são comumente utilizados para processamento e classificação de imagens, consistindo em agrupar ou apenas classificar pixeis individualmente. Na abordagem proposta por esta pesquisa, a classificação individual de pixeis é utilizada para um melhor aproveitamento das informações contidas em imagens de folha de soja. Uma vez que o objetivo é classificar um pixel entre patogênico e sadio, ao término da classificação, os conjuntos de pixeis contaminados que formam a lesão da doença já estarão segmentados.

Outro ponto contrastante entre os métodos apresentados e a proposta desta pesquisa é o número de amostras necessárias para o treinamento do modelo. Conforme explicado, em cada trabalho citado, milhares de imagens das folhas contendo as doenças estudadas foram necessárias para que os algoritmos fossem capazes de induzir um modelo com alto poder preditivo. Na abordagem baseada em pixel aqui proposta, estas amostras se tornam os pixeis contidos em uma imagem e não mais a própria imagem, dessa forma, reduzindo a quantidade de imagens de folhas necessárias para induzir um modelo com alta acurácia.

3 PROCEDIMENTO METODOLÓGICO

A metodologia do trabalho consiste numa etapa de aquisição de imagens, onde um banco de imagens foi fornecido pela equipe "Agro informática- Agricultura Digital e Tecnológica", seguida pela identificação da doença com o auxílio de um especialista, uma etapa de extração dos descritores seguida pela seleção dos seus atributos através da função de ganho de informação em conjunto com a correlação de Pearson, onde após análise dos atributos, os algoritmos *Random Forest*, *Random Tree*, CART, *RepTree* e SVM foram testados, tendo acurácia, precisão, sensibilidade, área sob a curva ROC, *F-score* medidos e comparados. Por fim, realizou-se comparações entre o modelo criado, a ferramenta APS ASSESS, e três avaliadores humanos.

3.1 Aquisição de imagens

As folhas foram coletadas em um campo de experimentos de plantas de soja, naturalmente infectadas pelo fungo, de forma aleatória de diferentes plantas. Após a coleta, foram anexadas a uma chapa de vidro transparente usando fita adesiva também transparente. As imagens foram coletadas com as folhas sobre um fundo contrastante, utilizando um telefone celular modelo *Iphone 6 Plus* e com uma distância aproximada de 30 cm entre a lente da câmera e folha. Uma limitação notada nesta abordagem é a variação de intensidade de luz, que pode acabar por distorcer os resultados, portanto um processo de normalização poderia ser aplicado na fase de pré processamento, porém as imagens foram obtidas com um mesmo padrão de luz, o que evita a influência da variação da iluminação nos resultados obtidos. Cada imagens possui uma versão com fundo e outra sem, possuindo 582 pixeis de largura e 870 pixeis de altura, conforme Figura 4. Ao total setenta e sete (77) imagens foram utilizadas nesta dissertação.

3.2 Identificação da doença

Para identificar as áreas específicas contaminadas pela doença e seus arredores, a ferramenta Gimp (*GNU Image Manipulation program*) foi utilizada para identificar as áreas infectadas com a coloração rosa. A segmentação da doença foi validada por um especialista e está exemplificada pela Figura 5.

A divisão dos níveis de severidade foi feita baseando-se na lógica das escalas diagramáticas, que dividem os níveis de tal modo que a visão humana é capaz de distingui-las, explicada pela lei de Weber-Fechner [50, 51, 52]. A escala diagramática utilizada na avaliação da Ferrugem Asiática é pautada neste princípio [67]. Os níveis de severidades definidos

Figura 4 – Exemplo de imagem de folha de soja do banco de imagens fornecida pelo grupo Agro informática- Agricultura Digital e Tecnológica com o fundo(esquerda) e mesma imagem sem o fundo (direita).

Figura 5 – Folha apresentando sintomas de Ferrugem Asiática (esquerda) e sintomas da ferrugem asiática segmentados em rosa e aprovados pelo especialista.

pelo especialista e as respectivas quantidades de amostras de folhas de soja classificadas por nível estão descritas na Tabela 1

3.3 Extração dos Descritores

Considerando que um fungo possui quatro estágios [68] sendo eles :

1. Estágio latente: neste estágio o fungo não é visível, porém já se encontra inserido na folha.

2. Clorose ou amarelamento: este estágio ocorre logo após o primeiro e causa a clorose

(aparecimento de regiões cor de palha) na folha. Nenhuma estrutura do fungo é visível.

3. Esporulação: neste estágio, o uredosporo é visível com a coloração de ferrugem. A estrutura do fungo é totalmente visível e o processo de infecção se repete.

4. Necrose: após a esporulação a urédia morre, deixando apenas uma área escura de tons pretos/marrom escuro.

Todos estes estágios podem ser encontrados em uma única folha [69], como demonstrado pela Figura 6. Com a análise sendo feita em nível de pixel, todas as cores apresentadas por cada estágio do fungo podem ser extraídas e analisadas durante o processo de construção do modelo.

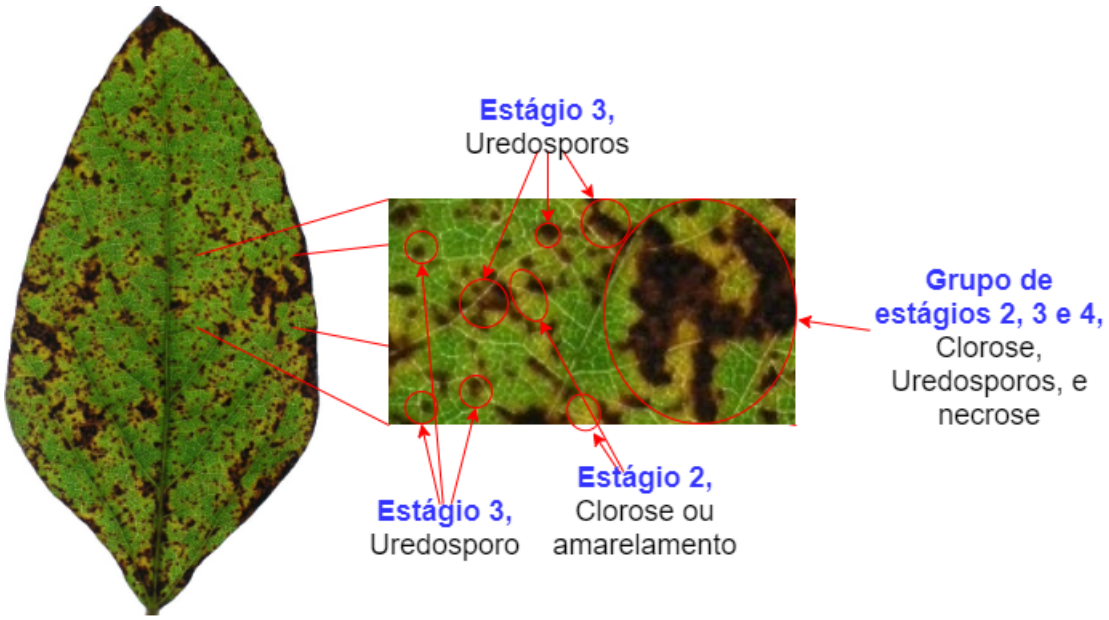

Figura 6 – Exemplo de imagem de folha de soja contaminada pelo fungo causador da Ferrugem Asiática e seus 4 estágios circulados em vermelho (Imagem criada pelo autor)

Nível	Qtd. de Amostras
0-2%	11
2-5%	11
5-10%	11
10-25%	11
25-50%	11
50-75%	11
75-100%	11

Tabela 1 – Níveis de severidades definidos pelo especialista e as respectivas quantidades de amostras de folhas de soja classificadas por nível.

Neste trabalho, 12 atributos são testados: os canais dos sistemas de cores RGB, LaB e HSI e os valores absolutos da diferença entre os canais RGB. A extração dos atributos foi feita com ajuda da biblioteca OpenCV [70] utilizando a linguagem de programação Python na versão 3.0. Todas as imagens estavam no formato JPG. Cada uma das imagens inseridas sofrem o processo de transformação de cores para HSI e Lab. Após este passo, todos os pixeis que foram definidos como patogênicos são extraídos e, todos os atributos de cor são computados e escritos em um arquivo do tipo ARFF para serem analisados posteriormente. Os pixeis considerados não patogênicos são extraídos de forma aleatória até que se tenha a mesma quantidade dos pixeis patogênicos, garantindo assim uma base balanceada.

Para esta pesquisa, foram utilizadas sete imagens, uma para cada nível de ferrugem conforme ilustrado pela imagem 7, garantindo que exemplares do fungo em todos os seus estágios estejam presentes na base de dados. Ao todo, 160 mil pixel foram coletados sendo 80 mil patogênicos e 80 mil sadios. Uma base de dados balanceada foi criada com intuito de que não haja enviesamento do modelo por existir classe majoritária. Com relação ao super-ajuste, que é quando o modelo é apenas capaz de classificar entre o conjunto de dados utilizado para o treinamento, este problema foi minimizado garantindo que na base de 160 mil amostras, todos os estágios do fungo foram adicionados.

Figura 7 – As sete (7) imagens de folhas de soja contaminadas pelo fungo *P. pachyhizi* que foram utilizadas para a coleta de pixeis. Os níveis categorizado pelo especialista para cada imagem são 0-2%, 2-5%, 5-10%, 10-25%, 25-50%, 50-75%, 75-100% da esquerda para a direita de cima para baixo respectivamente.

3.4 Seleção dos atributos

O próximo passo foi a seleção dos melhores atributos para o problema, que neste caso é classificar entre pixeis sadios e patogênicos.

No caso do estudo aqui apresentado, por razões de otimização, a função *ganho de informação* foi aplicada e, logo na sequência foi criada uma matriz de correlação dos atributos, o que possibilitou a escolha de um menor número de atributos que contribuam com a maior quantidade de informação com a menor sobreposição de informação. A função *ganho de informação* (descrita pela equação 2.8) é uma função de análise estatística, cujo retorno é um ranking dos melhores atributos. A matriz de correlação indica quão correlacionados dois atributos estão, em outras palavras, indica o quanto dois atributos descrevem porções semelhantes do problema, sendo 1 ou -1 indicando o máximo possível de correlação diretamente e indiretamente, respectivamente e 0 o mínimo. A Tabela 2 ilustra a matriz de correlação e a Tabela 3 ilustra o resultado da função *ganho de informação*.

Para selecionar os melhores atributos, começando pelo topo do ranking cada atributo foi comparado com o atributo atual e, caso possua uma correlação forte (>0.7) [71] este atributo é removido do conjunto, restando apenas o mais bem colocado no ranking. Ao final deste procedimento, os atributos selecionados foram ‘a’, ‘H’, ‘B-R’ e ‘S’.

Um dado interessante e digno de discussão é o fato de que embora o atribuo ‘a’, sendo este o atributo melhor colocado no ranking, que descreve a cor variando do verde ao vermelho, obteve elevada correlação (-0.73) com o atribuo ‘b’ que descreve a cor do azul ao amarelo. Acredita-se que esta correlação se deve ao fato de que os tons de verdes encontrados na folha são tons mais escuros, o que indica forte presença de combinação da cor azul para a criação destas tonalidades.

Após a seleção dos atributos, o próximo passo é realizar os testes com a aplicação dos atributos encontrados com os algoritmos selecionados.

	R	G	B	L	a	b	H	S	V	|B-G|	|B-R|	|G-R|
R	1	0.83	0.80	0.88	-0.27	0.72	0.02	-0.35	0.88	0.63	0.77	0.23
G	0.83	1	0.79	0.99	-0.76	0.88	0.35	-0.27	0.99	0.89	0.51	0.65
B	0.80	0.79	1	0.79	-0.37	0.42	0.20	-0.68	0.80	0.42	0.23	0.32
L	0.88	0.99	0.79	1	-0.70	0.89	0.31	-0.28	1	0.88	0.58	0.61
a	-0.27	-0.76	-0.37	-0.70	1	-0.73	-0.58	-0.01	-0.68	-0.84	-0.04	-0.85
b	0.72	0.88	0.42	0.89	-0.73	1	0.27	0.11	0.88	0.98	0.71	0.63
H	0.02	0.35	0.20	0.31	-0.58	0.27	1	-0.13	0.28	0.38	-0.18	0.43
S	-0.35	-0.27	-0.68	-0.28	-0.01	0.11	-0.13	1	-0.27	0.09	0.15	0.05
V	0.88	0.99	0.80	1	-0.68	0.88	0.28	-0.27	1	0.87	0.58	0.64
|B-G|	0.63	0.89	0.42	0.88	-0.84	0.98	0.38	0.09	0.87	1	0.58	0.72
|B-R|	0.77	0.51	0.23	0.58	-0.04	0.71	-0.18	0.15	0.58	0.58	1	0.02
|G-R|	0.23	0.65	0.32	0.61	-0.85	0.63	0.43	0.05	0.64	0.72	0.02	1

Tabela 2 – Matriz de correlação dos atributos extraídos das imagens de folha de soja infectadas por ferrugem asiática sendo que: R indica Vermelho (Red), G indica Verde (Green) e B indica Azul (Blue), H indica matiz (Hue), S indica a saturação, I indica intensidade, L descreve o brilho, 'a' descreve cor do verde até o vermelho e 'b' descreve cor do azul até o amarelo

Ranking	Atributos
0.5745	a
0.5487	|G-R|
0.5291	H
0.3912	|B-R|
0.3333	|B-G|
0.2968	G
0.2566	L
0.2546	V
0.2453	b
0.1888	R
0.1156	B
0.0696	S

Tabela 3 – Ranking resultante da função ganho de informação aplicada sobre os atributos utilizados para descrever a doença Ferrugem Asiática. O atributo R indica Vermelho (Red), G indica Verde (Green) e B indica Azul (Blue), H indica matiz (Hue), S indica a saturação, I indica intensidade, L descreve o brilho, 'a' descreve cor do verde até o vermelho e 'b' descreve cor do azul até o amarelo

4 ANÁLISE E RESULTADOS

Após alguns testes, notou-se que a nervura central e suas áreas adjacentes estavam sendo consideradas patogênicas devido a sua proximidade de cores com o terceiro estágio do fungo (vide seção 3.3), portanto uma nova segmentação foi feita onde a nervura central e suas proximidades foram assinaladas como não-patogênico. A marcação foi feita com coloração azul (Figura 8). Esta nova segmentação permite ao algoritmo de extração de atributos identificar e marcar a nervura central e suas áreas adjacentes corretamente, primeiro incluindo os pixeis em azul e depois selecionando aleatoriamente até balancear a base. A segmentação foi feita com a mesma estratégia e programas descritos na seção 3.2.

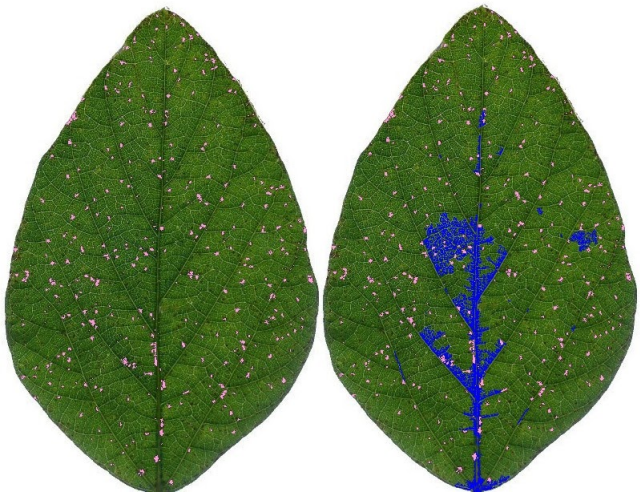

Figura 8 – Folha contaminada pelo fungo *Phakopsora pachyrhizi* com sintomas segmentados em rosa (esquerda) e segmentação adicional da nervura central e suas áreas adjacentes em azul (direita)

4.1 Resultados experimentais

Tanto para os algoritmos de arvore de decisão quanto para Máquina de Vetor de Suporte, o método de validação cruzada com o valor de 'K' = 10 foi aplicado. Todos os testes foram realizados com auxílio da biblioteca *scikit learn* em Python (versão 3.0). Para o algoritmo *Random Forest* foi alterado o número de árvores que são utilizadas para estimar o resultado, e para o SVM foram alterados os tipos de kernel entre RBF, linear e sigmoide, com intuito de melhorar suas performances. A Tabela 4 ilustra os parâmetros utilizados em cada um dos algoritmos e seus resultados. Na sequência, a Tabela 5 ilustra os melhores resultados de cada algoritmo.

Algoritmo	Parâmetro	Acurácia	Precisão	Sensibilidade	*F-Score*	AUC
Random Forest	N° árvores: 100	97.20	0.97	0.97	0.97	0.99
Random Forest	N° árvores: 200	97.21	0.97	0.97	0.97	0.99
Random Forest	N° árvores: 300	97.23	0.97	0.97	0.97	0.99
SVM	Kernel: RBF	97.22	0.97	0.97	0.97	0.98
SVM	Kernel: Sigmoid	49.76	0.24	0.50	0.33	0.50
SVM	Kernel: Linear	97.17	0.97	0.97	0.97	0.99

Tabela 4 – Variação do número de árvores do algoritmo *Random Forest* e seus resultados; Variação do kernel do algoritmo SVM e seus resultados.

Algoritmo	Acurácia	Precisão	Sensibilidade	*F-Score*	AUC
CART	97.24	0.97	0.97	0.97	0.97
Random Tree	91.13	0.92	0.91	0.91	0.93
RepTree	92.90	0.95	0.91	0.93	0.97
Random Forest 300 Estimators	97.23	0.97	0.97	0.97	0.99
SVM Kernel RBF	97.22	0.97	0.97	0.97	0.98

Tabela 5 – Melhores resultados de cada algoritmo

Como o modelo criado deve ser capaz de quantificar baixos níveis de ferrugem asiática, sua precisão e sensibilidade devem possuir um valor elevado, e para realizar a comparação entre modelos, a área sob a curva e o *F-Score* (seções 2.6.5 e 2.6.6) foram utilizados, sendo que o *F-Score* utilizado foi a média entre ambas as classes. Obtendo a maior acurácia dentro todos os modelos criados e com um valor de precisão, sensibilidade e *F-score* de 0.97 que indicam que o modelo foi capaz de diferenciar entre ambas as classes e com a área sob a curva indicando um bom desempenho, o modelo criado pelo algoritmo CART foi determinado como sendo o melhor para o estudo.

Para a validação do modelo, setenta imagens (dez para cada nível descrito na Tabela 1) foram aplicadas e o seu nível de severidade foi calculado. O calculo da severidade é baseado na divisão entre o total de pixeis pelo total de pixeis patogênicos. A severidade calculada pelo modelo foi então comparada com duas outras técnicas: APS Assess software [72], que é uma ferramenta utilizada para quantificação de doenças em plantas, e três especialistas humanos utilizando escalas diagramáticas desenvolvidas por [67] para auxiliar na estimativa da severidade da doença. A Figura 9 ilustra a segmentação gerada após a classificação de cada pixel pelo modelo criado.

Conforme ilustrado pela Figura 10, as medidas de severidade obtidas pelo modelo estão de acordo com o método de avaliação APS Assess. As colunas em preto na Figura 10 representam os sete intervalos de níveis de severidade, e as linhas vermelhas são o máximo e o mínimo esperado em cada intervalo de acordo com o especialista (Tabela 1). O modelo desenvolvido seguiu os mesmo padrões gerados pela ferramenta APS Assess, porém ambos não foram capazes de se encaixar exatamente na área esperada, o que pode

Figura 9 – Imagem de folha de soja apresentando sintoma de Ferrugem Asiática (esquerda) ao lado da mesma imagem após ter todos os seus pixeis classificado (direita), onde os pixeis considerados patogênicos foram pintados de vermelho e os sadios deixados como são.

indicar um erro de classificação por parte do especialista.

Além do APS Assess, três especialistas humanos foram convidados a avaliar o mesmo conjunto de imagens (setenta imagens) utilizando escalas diagramáticas. A Figura 11 ilustra a dispersão dos resultados quando comparados com o modelo aceito APS Assess (linha vermelha).

Os avaliadores humanos tendem a superestimar a severidade da ferrugem asiática acima de 50%. A Tabela 6 exibe a média de erro e o desvio padrão de dos avaliadores e do novo método proposto.

X	Erro Médio	Desvio Padrão
Novo Modelo	2.8672	2.5839
Avaliador 1	10.9685	13.2347
Avaliador 2	14.1508	15.7334
Avaliador 3	9.6128	9.2763

Tabela 6 – Erro Médio e desvio padrão dos avaliadores humanos e de novo modelo quando comparados a ferramenta APS ASSESS com total de setenta (70) imagens distintas de folhas de soja.

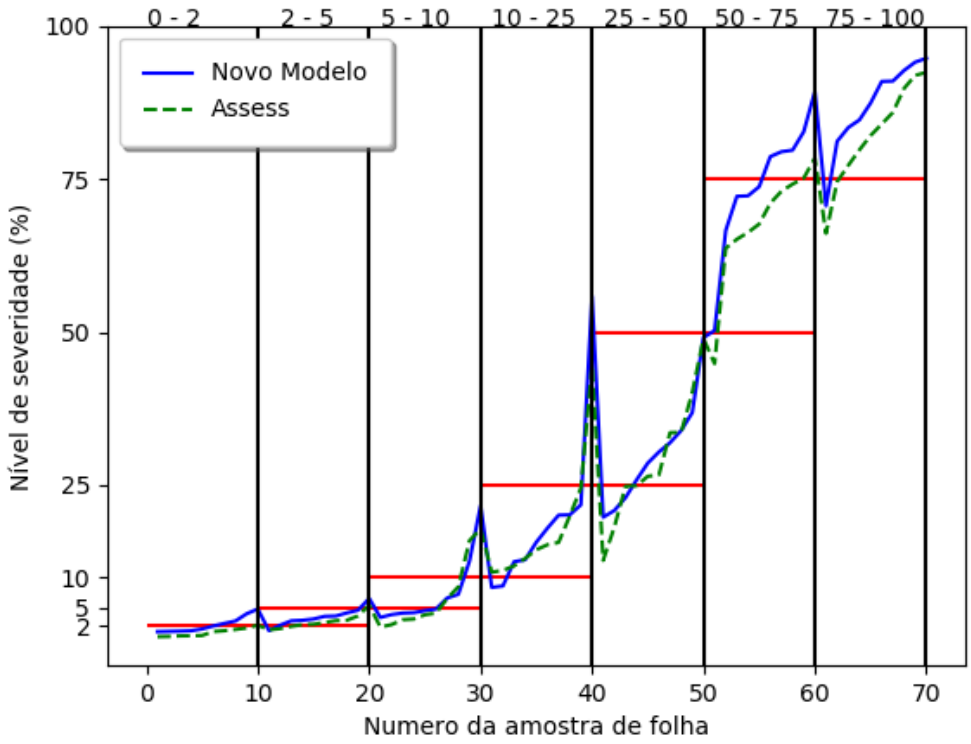

Figura 10 – Comparação dos resultados obtidos pelo novo modelo e pela ferramenta APS ASSESS na quantificação do nível de severidade de Ferrugem Asiática em setenta (70) imagens distintas de folhas de soja em intervalos definidos por especialista

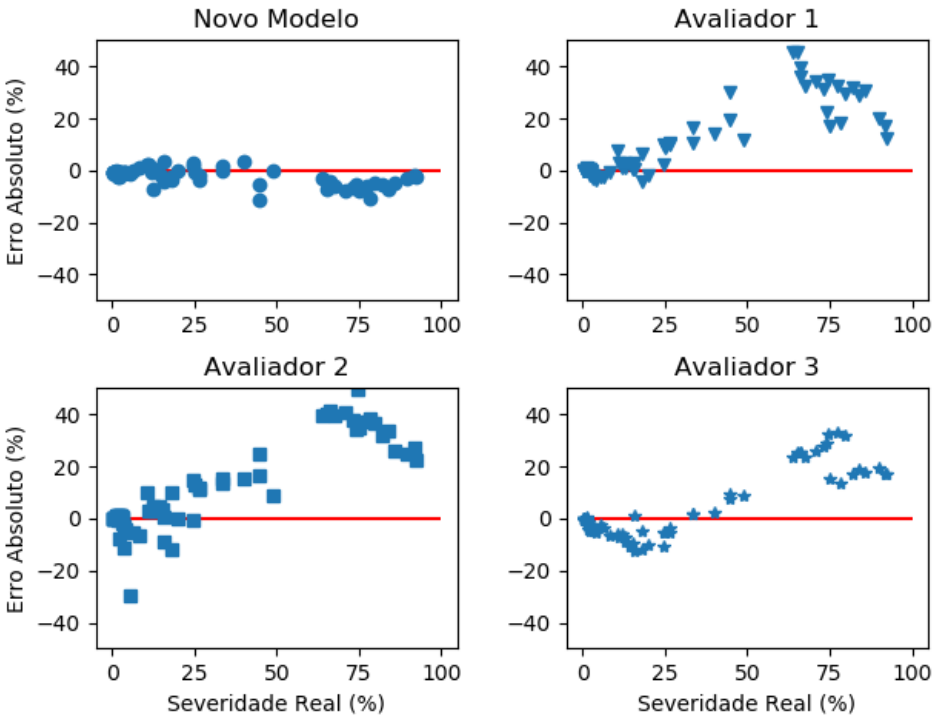

Figura 11 – Erro Absoluto do novo modelo e dos avaliadores humanos comparado com a ferramenta APS ASSESS com total de setenta (70) imagens distintas de folhas de soja.

5 CONCLUSÃO

A soja é um importante produto de uso industrial, variando suas aplicações de óleo de soja para frituras, passando por emulsificantes para diversos produtos, até ração para animais de criação. Essa grande variedade de aplicações gera um valor elevado de mercado, portanto a sanidade do plantio é um ponto fundamental para que se possa ter uma boa produção.

A Ferrugem Asiática é uma doença causada por um fungo que atinge principalmente países tropicais, causando danos as folhas fazendo-as cair, diminuindo assim a produção de grãos por parte da planta. O seu principal tratamento é a aplicação de fungicida, que depende da medição do nível de severidade da doença. Esta medição por sua vez, é feita através de escalas diagramáticas, sendo estas, ferramentas que auxiliam um profissional humano na avaliação do nível de infecção de Ferrugem Asiática. A taxa de erro e variação por parte do humano, pode ter impactos negativos na sanidade do plantio.

Neste trabalho, foi desenvolvida abordagem baseada em pixel para avaliar o nível de infecção por Ferrugem Asiática em folhas de soja, usando um banco de imagens fornecido pelo grupo "Agro-informática - Agricultura Digital e Tecnológica"da Universidade Estadual de Londrina do departamento de Agronomia. Múltiplos descritores de cor visíveis como RGB, Lab E HSV foram testados usando os algoritmos *Random Forest*, *Random Tree*, CART, RepTree e SVM. Processo de seleção de atributos foi aplicado com intuito de melhorar os resultados obtidos. Os resultados obtiveram acurácia acima de 91% com o melhor resultado obtido pelo algoritmo CART com 97.24% de acurácia no modelo gerado.

O diferencial deste trabalho se encontra na abordagem que reduz o problema a nível de pixel através de uma classificação binária entre ‘sadio’ e ‘patogênico’. Esta abordagem só foi possível devido a condição natural do fungo causador da ferrugem asiática ser dividida em 4 (quatro) estágios, sendo destes 3 (três) visíveis a olho nu, e portanto descritíveis através de atributos de cor, além disso, todos os estágios podem ser encontrados em um mesmo exemplar de folha contaminada. Através desta abordagem foi possível reduzir consideravelmente a quantidade de imagens de folhas necessárias para treinar o modelo, e ainda assim treina-lo com 160 mil amostras (de pixel) em uma base robusta e com ampla variedade de exemplos. Além disto, devido a classificação ser a nível de pixel, após cada pixel da folha ter sido classificado, as regiões patogênicas são automaticamente segmentadas, não necessitando de outro processo para que haja o destaque da doença.

O modelo criado foi comparado com 3 (três) avaliadores humanos utilizando escalas diagramáticas, assim como um software de medição de nível de severidade de doenças em

folhas conhecido e aceito no mundo agronômico chamado de ASSESS. As comparações se deram utilizando um total de 70 (setenta) imagens de folhas contaminadas pelo fungo e os resultados obtidos pelo ASSESS foram utilizados como o padrão ouro durante a comparação com os avaliadores humanos. O modelo criado obteve considerável redução no erro médio quando comparado com a média dos avaliadores humanos, sendo esta redução de 75,23% , assim como uma redução do desvio padrão de 79,73%, o que indica que a abordagem proposta foi capaz de quantificar o nível de ferrugem asiática em folhas de soja com elevada confiança.

REFERÊNCIAS

[1] MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. *Sistemas Inteligentes-Fundamentos e Aplicações*, v. 1, p. 115–139, 2003.

[2] LORENA, A. C.; CARVALHO, A. C. de. Introduçaoas máquinas de vetores suporte. *Relatório Técnico do Instituto de Ciências Matemáticas e de Computaçao (USP/Sao Carlos)*, v. 192, p. 11, 2003.

[3] PARANHOS, R. et al. Desvendando os mistérios do coeficiente de correlação de pearson: o retorno. *Leviathan (São Paulo)*, n. 8, p. 66–95, 2014.

[4] GODOY, C. V. et al. Asian soybean rust in brazil: past, present, and future. *Pesquisa Agropecuária Brasileira*, SciELO Brasil, v. 51, n. 5, p. 407–421, 2016.

[5] YORINORI, J. et al. Epidemics of soybean rust (phakopsora pachyrhizi) in brazil and paraguay from 2001 to 2003. *Plant Disease*, Am Phytopath Society, v. 89, n. 6, p. 675–677, 2005.

[6] IVANCOVICH, A. Soybean rust in argentina. *Plant Disease*, Am Phytopath Society, v. 89, n. 6, p. 667–668, 2005.

[7] SCHNEIDER, R. et al. First report of soybean rust caused by phakopsora pachyrhizi in the continental united states. *Plant disease*, Am Phytopath Society, v. 89, n. 7, p. 774–774, 2005.

[8] TSUKAHARA, R. Y.; HIKISHIMA, M.; CANTERI, M. G. Relationship between climate and the progress of the asian soybean rust (phakopsora pachyrhizi) in two micro-regions of paraná state. *Semina: Ciências Agrárias*, v. 29, n. 1, p. 47–52, 2008.

[9] ROSA, C.; SPEHAR, C.; LIU, J. Asian soybean rust resistance: an overview. *J Plant Pathol Microb*, v. 6, n. 307, p. 2, 2015.

[10] MINCHIO, C. A. et al. Epidemias de ferrugem asiática no rio grande do sul explicadas pelo fenômeno enos e pela incidência da doença na entressafra. *Summa Phytopathologica*, SciELO Brasil, v. 42, n. 4, p. 321–326, 2016.

[11] OLIVEIRA, G. d. lt (2016). the geopolitics of brazilian soybeans. *The Journal of Peasant Studies*, v. 43, n. 2, p. 348–372.

[12] MOREIRA, E. et al. Temporal dynamics of soybean rust associated with leaf area index in soybean cultivars of different maturity groups. *Plant disease*, Am Phytopath Society, v. 99, n. 9, p. 1216–1226, 2015.

[13] MINCHIO, C. et al. Predicting asian soybean rust epidemics based on off-season occurrence and el niño southern oscillation phenomenon in paraná and mato grosso states, brazil. *Journal of Agricultural Science*, v. 10, n. 11, p. 562, 2018.

[14] GASPAR, G. G. et al. Balance among calcium, magnesium and potassium levels affecting asian soybean rust severity. *Agronomy science and Biotechnology*, v. 1, n. 1, p. 39–39, 2015.

[15] ISHIWATA, Y. I.; FURUYA, J. Evaluating the contribution of soybean rust-resistant cultivars to soybean production and the soybean market in brazil: A supply and demand model analysis. *Sustainability*, Multidisciplinary Digital Publishing Institute, v. 12, n. 4, p. 1422, 2020.

[16] BRAGA, K. et al. Sensitivity of populations of phakopsora pachyrhizi to the fungicide prothioconazole. *Summa Phytopathologica*, SciELO Brasil, v. 46, n. 2, p. 150–154, 2020.

[17] FANTIN, L. H. et al. Spectral characterization and quantification of phakopsora pachyrhizi urediniospores by fourier transformed infrared with attenuated total reflectance. *European Journal of Plant Pathology*, Springer, v. 154, n. 4, p. 1149–1157, 2019.

[18] XAVIER, S. A. et al. Older leaf tissues in younger plants are more susceptible to soybean rust. *Acta Scientiarum. Agronomy*, SciELO Brasil, v. 39, n. 1, p. 17–24, 2017.

[19] CARMONA, M. et al. Development and validation of a fungicide scoring system for management of late season soybean diseases in argentina. *Crop Protection*, Elsevier, v. 70, p. 83–91, 2015.

[20] FIGUEIREDO, G. V. C. et al. A bayesian probability model can simulate the knowledge of soybean rust researchers to optimize the application of fungicides. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, IGI Global, v. 10, n. 4, p. 37–51, 2019.

[21] GODOY, C. et al. Eficiência de fungicidas para o controle da ferrugem-asiática da soja, phakopsora pachyrhizi, na safra 2018/19: Resultados sumarizados dos ensaios cooperativos. *Embrapa Soja-Circular Técnica (INFOTECA-E)*, Londrina: Embrapa Soja., 2019.

[22] REIS, E. M.; ZANATTA, M.; REIS, A. C. Performance of chlorothalonil levels and spraying intervals on asian rust control and soybean grain yield. *Summa Phytopathologica*, SciELO Brasil, v. 45, n. 3, p. 261–264, 2019.

[23] BOCK, C. H. et al. From visual estimates to fully automated sensor-based measurements of plant disease severity: status and challenges for improving accuracy. *Phytopathology Research*, Springer, v. 2, p. 1–30, 2020.

[24] SHARIF, M. et al. Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. *Computers and electronics in agriculture*, Elsevier, v. 150, p. 220–234, 2018.

[25] SAFDAR, A. et al. Intelligent microscopic approach for identification and recognition of citrus deformities. *Microscopy research and technique*, Wiley Online Library, v. 82, n. 9, p. 1542–1556, 2019.

[26] IQBAL, Z. et al. An automated detection and classification of citrus plant diseases using image processing techniques: A review. *Computers and electronics in agriculture*, Elsevier, v. 153, p. 12–32, 2018.

[27] RAUF, H. T. et al. A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning. *Data in brief*, Elsevier, v. 26, p. 104340, 2019.

[28] JØRGENSEN, L. N. et al. Targeting fungicide inputs according to need. *Annual review of phytopathology*, Annual Reviews, v. 55, p. 181–203, 2017.

[29] BOSCH, F. v. d. et al. Governing principles can guide fungicide-resistance management tactics. *Annual Review of Phytopathology*, Annual Reviews, v. 52, p. 175–195, 2014.

[30] MELO, G. d. A. et al. Utilização de processamento digital de imagens e redes neurais artificiais para o reconhecimento de índices de severidade da ferrugem asiática da soja. UNIVERSIDADE ESTADUAL DE PONTA GROSSA, 2015.

[31] GOELLNER, K. et al. Phakopsora pachyrhizi, the causal agent of asian soybean rust. *Molecular plant pathology*, Wiley Online Library, v. 11, n. 2, p. 169–177, 2010.

[32] KHAN, M. A. et al. Ccdf: Automatic system for segmentation and recognition of fruit crops diseases based on correlation coefficient and deep cnn features. *Computers and electronics in agriculture*, Elsevier, v. 155, p. 220–236, 2018.

[33] CHAUDHARY, P. et al. Color transform based approach for disease spot detection on plant leaf. *International journal of computer science and telecommunications*, Citeseer, v. 3, n. 6, p. 65–70, 2012.

[34] RATHOD, A. N.; TANAWAL, B.; SHAH, V. Image processing techniques for detection of leaf disease. *International Journal of Advanced Research in Computer Science and Software Engineering*, v. 3, n. 11, 2013.

[35] SINGH, V. Sunflower leaf diseases detection using image segmentation based on particle swarm optimization. *Artificial Intelligence in Agriculture*, Elsevier, v. 3, p. 62–68, 2019.

[36] PADOL, P. B.; YADAV, A. A. Svm classifier based grape leaf disease detection. In: IEEE. *2016 Conference on advances in signal processing (CASP)*. [S.l.], 2016. p. 175–179.

[37] GONZALEZ, R. C.; WOODS, R. E. Processamento digital de imagem. *Pearson, ISBN-10: 8576054019,*, v. 10, p. 11–27, 2010.

[38] KHAN, M. A. et al. Automated design for recognition of blood cells diseases from hematopathology using classical features selection and elm. *Microscopy Research and Technique*, Wiley Online Library, v. 84, n. 2, p. 202–216, 2021.

[39] NAHEED, N. et al. Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review. *Computer Modeling in Engineering & Sciences*, Tech Science Press, v. 125, n. 1, p. 314–344, 2020.

[40] SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A.; TOMBILLA-SANROMÁN, M. Filter methods for feature selection–a comparative study. In: SPRINGER. *International Conference on Intelligent Data Engineering and Automated Learning*. [S.l.], 2007. p. 178–187.

[41] JOSHI, A. V. *Machine learning and artificial intelligence*. [S.l.]: Springer, 2020.

[42] CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers & Electrical Engineering*, Elsevier, v. 40, n. 1, p. 16–28, 2014.

[43] NETO, A. M. et al. Image processing using pearson's correlation coefficient: Applications on autonomous robotics. In: IEEE. *2013 13th International Conference on Autonomous Robot Systems*. [S.l.], 2013. p. 1–6.

[44] GOLDBERG, D. E. *Genetic algorithms*. [S.l.]: Pearson Education India, 2006.

[45] YAN, K.; ZHANG, D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, Elsevier, v. 212, p. 353–363, 2015.

[46] GRANITTO, P. M. et al. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 83, n. 2, p. 83–90, 2006.

[47] GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003.

[48] BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. A review of feature selection methods on synthetic data. *Knowledge and information systems*, Springer, v. 34, n. 3, p. 483–519, 2013.

[49] LAL, T. N. et al. Embedded methods. In: *Feature extraction*. [S.l.]: Springer, 2006. p. 137–165.

[50] BRAGA, K. et al. Development and validation of a diagrammatic scale for the assessment of the severity of bacterial leaf streak of corn. *European Journal of Plant Pathology*, Springer, v. 157, p. 367–375, 2020.

[51] FANTIN, L. et al. Development and validation of diagrammatic scale to assess target spot severity in cotton. *Australasian Plant Pathology*, Springer, v. 47, n. 5, p. 491–497, 2018.

[52] NUTTER, F. W.; ESKER, P. D. The role of psychophysics in phytopathology: The weber–fechner law revisited. *European Journal of Plant Pathology*, Springer, v. 114, n. 2, p. 199–213, 2006.

[53] PORTUGAL, R.; SVAITER, B. F. Weber-fechner law and the optimality of the logarithmic scale. *Minds and Machines*, Springer, v. 21, n. 1, p. 73–81, 2011.

[54] CUNHA, R. L. d. et al. Desenvolvimento e validação de uma escala diagramática para avaliar a severidade da ferrugem (hemileia vastatrix) do cafeeiro. 2001.

[55] ZHANG, X.-D. Machine learning. In: *A Matrix Algebra Approach to Artificial Intelligence*. [S.l.]: Springer, 2020. p. 223–440.

[56] SIMP, A. X. I. I.; REMOTO, S. Classificação de imagens de sensoriamento remoto pela aprendizagem por árvore de decisão: uma avaliação de desempenho. d, p. 4319–4326, 2005.

[57] BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

[58] LAVALLE, S. M. *Rapidly-Exploring Random Trees: A New Tool for Path Planning*. [S.l.], 1998.

[59] QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Elsevier, 2014.

[60] JAYANTHI, S.; SASIKALA, S. Reptree classifier for identifying link spam in web search engines. *IJSC*, v. 3, n. 2, p. 498–505, 2013.

[61] PRATI, R. et al. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008.

[62] BERRAR, D. Cross-validation. *Encyclopedia of bioinformatics and computational biology*, Academic, v. 1, p. 542–545, 2019.

[63] MARCOS, A. P.; RODOVALHO, N. L. S.; BACKES, A. R. Coffee leaf rust detection using genetic algorithm. In: IEEE. *2019 XV Workshop de Visão Computacional (WVC)*. [S.l.], 2019. p. 16–20.

[64] KHAN, M. A. et al. An optimized method for segmentation and classification of apple diseases based on strong correlation and genetic algorithm based feature selection. *IEEE Access*, IEEE, v. 7, p. 46261–46277, 2019.

[65] ADEEL, A. et al. Diagnosis and recognition of grape leaf diseases: An automated system based on a novel saliency approach and canonical correlation analysis based multiple features fusion. *Sustainable Computing: Informatics and Systems*, Elsevier, v. 24, p. 100349, 2019.

[66] ADEEL, A. et al. Entropy-controlled deep features selection framework for grape leaf diseases recognition. *Expert Systems*, Wiley Online Library, 2020.

[67] GODOY, C.; KOGA, L.; CANTERI, M. Escala diagramática para avaliação da severidade da ferrugem da soja. *Fitopatologia Brasileira*, v. 31, n. 1, p. 63–68, 2006.

[68] KELLY, H. Y. et al. From select agent to an established pathogen: the response to phakopsora pachyrhizi (soybean rust) in north america. *Phytopathology*, Am Phytopath Society, v. 105, n. 7, p. 905–916, 2015.

[69] NAVARRO, B. L. et al. Histopathology of phakopsora euvitis on vitis vinifera. *European Journal of Plant Pathology*, Springer, v. 154, n. 4, p. 1185–1193, 2019.

[70] BRADSKI, G.; KAEHLER, A. *Learning OpenCV: Computer vision with the OpenCV library*. [S.l.]: " O'Reilly Media, Inc.", 2008.

[71] BENESTY, J. et al. Pearson correlation coefficient. In: *Noise reduction in speech processing*. [S.l.]: Springer, 2009. p. 1–4.

[72] LAMARI, L. *Assess: image analysis software for plant disease quantification*. [S.l.]: APS press, 2002.

TRABALHO ACEITO

1. Murilo Caminotto Barbosa, Ana Sottana, Deryk Ribeiro, Alan Salvany Felinto, Lucas Fantin e Marcelo Canteri. **Regression applied to measure normalized difference vegetation index in soybean images with visible color spaces collected by smartphones** aceito no evento 2021 IEEE International Instrumentation & Measurement Technology Conference, apresentado dia 20 de Maio de 2021 (Qualis A3 2021)