UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

DAVID ANTONIO SBRISSA NETO

Diagnóstico de doenças de soja através da emissão de fluorescência das folhas: estudo de caso para Ferrugem asiática e “Haste verde e retenção foliar”

São Carlos

2019

DAVID ANTONIO SBRISSA NETO

Diagnóstico de doenças de soja através da emissão de fluorescência das folhas: estudo de caso para Ferrugem asiática e “Haste verde e retenção foliar”

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.
Área de concentração: Física Aplicada
Opção: Física Biomolecular
Orientador: Profª. Drª. Débora Marcondes Bastos Pereira Milori

Versão Corrigida
(Versão original disponível na Unidade que aloja o programa)

São Carlos

2019

Sbrissa Neto, David Antonio
Diagnóstico de doenças de soja através da emissão de fluorescência das folhas: estudo de caso para Ferrugem asiática e "Haste verde e retenção foliar" / David Antonio Sbrissa Neto; orientadora Débora Marcondes Bastos Pereira Milori - versão corrigida -- São Carlos, 2019.
99 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2019.

1. Soja. 2. Fluorescência. 3. DIagnóstico precoce. 4. Aprendizado de máquina. I. Milori, Débora Marcondes Bastos Pereira, orient. II. Título.

Dedico esse trabalho à ciência e a toda comunidade científica. Que esse trabalho sirva de inspiração aos cientistas novos e antigos, assim como me inspirou, não apenas na busca pela verdade, mas também na maneira de compreender a vida e a sociedade.

AGRADECIMENTOS

Agradeço à minha orientadora, Dra. Débora Marcondes Bastos Pereira, que idealizou esse projeto, ofereceu todas as condições para que eu pudesse desenvolve-lo e sempre me apoiou nos contratempos que a vida nos põe a prova, e não foram poucos. Agradeço por toda confiança em mim depositada. Agradeço também por me mostrar uma visão interdisciplinar da ciência, que muito me inspirou e certamente carregarei na sequência da minha trajetória acadêmica.

Agradeço às Instituições que estive vinculado nos últimos anos: à USP, pelo programa e pela estrutura, à UNIFAP por possibilitar a continuidade do programa, mesmo depois da aprovação no concurso, e também à EMBRAPA Instrumentação, por toda estrutura laboratorial requerida no projeto.

Agradeço às agências de pesquisa e fomento CNPq e FAPESP pelo financiamento dos laboratórios e projetos de pesquisa os quais me inseri durante o doutorado.

Agradeço à minha família, em especial aos meus pais David Antonio Sbrissa Junior e Elaine Aparecida Lucatti Sbrissa, que sempre me apoiaram nas minhas escolhas acadêmicas e deram todo suporte para que eu chegasse até aqui. Sem eles nada disso seria possível.

Agradeço ao Colegiado de Física da UNIFAP, por permitirem minha liberação das atividades acadêmica no ano de 2018 para finalização do trabalho, etapa crucial para a conclusão do trabalho

Agradeço amigos do laboratório e de ciência, Carla, Anielle, Alex, Vitão, Max, Amanda, Ká, Guto Alfredo. Muito obrigado pelo apoio, pelo carinho e pela motivação nos momentos difíceis, e em especial, pelos bandejões obrigatórios, que renderam boas conversas e incontáveis equívocos com sobremesas.

Agradeço aos meus amigos e físicos brilhantes Luis Felipe (Nó), Renato (Thomp) e Tiago (Timo), por compartilhamos bons momentos nesses últimos anos, com destaque aos intermináveis CIV's marcados para as 18h que começavam as 22h.

Agradeço aos amigos do xadrez Vivaldinho, Filipe Guerra, Felipe El Debs, pelas boas histórias e recordações do mundo enxadrístico, que desde sempre guiou meu universo de escolhas.

Agradeço, por fim, e em especial, a minha namorada Ana Lidia Salmazo, por esse ainda curto, porém intenso relacionamento. Agradeço por todas as vezes que me apoiou quando precisei, por todas as vezes que me criticou quanto agi errado, e obrigado por me mostrar que podemos continuamente amadurecer, sem deixar de viver com alegria e amor.

"Tudo é uma questão de manter

A mente quieta

A espinha ereta

E o coração tranquilo"

Walter Franco

RESUMO

SBRISSA NETO, D. A. **Diagnóstico de doenças de soja através da emissão de fluorescência das folhas**: estudo de caso para Ferrugem asiática e "Haste verde e retenção foliar". 2019. 99 p. Tese (Doutorado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2019.

O presente trabalho investiga a emissão de fluorescência de plantas de soja contaminadas por duas importantes doenças de soja para o cenário brasileiro: a Ferrugem Asiática e a Haste Verde e Retenção Foliar – popularmente conhecida como soja louca II. Para tal, foi realizado um experimento nas dependências da EMBRAPA Soja, localizada em Londrina-PR, com a expressão de ambas as doenças em plantas de soja, além de um grupo controle para comparação. Foram obtidas cinco coletas caracterizadas como Assintomáticas para HVRF e três coletas para as Assintomáticas FA. Foram utilizadas três técnicas de fluorescência: o estereomicroscópio comercial SteReo Lumar.V12 com excitação em λ=365 nm (EM-365), a *Laser Induced Fluorescence Spectroscopy* (LIFS) e *Laser Induced Fluorescence Imaging* (LIFI), ambas com excitação em λ=405 nm. Como principais resultados, destaca-se a contribuição das bandas de emissão em 520 nm e 740 nm para caracterização das assintomáticas. Na etapa de classificação, a taxa de acerto na predição dos grupos de teste foi elevada, com mais de 90% de acurácia total da classificação para todas as metodologias de fluorescência. No caso HVRF, a instrumentação LIFI apresentou a maior acurácia de teste, com 98%. No caso FA, as instrumentações LIFI e LIFS apresentaram 93% e 94% de acurácia, respectivamente. Essas informações comprovam a hipótese fundamental do trabalho, sobre a capacidade discriminatória das emissões de fluorescência de plantas assintomáticas com relação às plantas sadias, com ênfase às técnicas LIFS e LIFI, que podem ser adaptadas para instrumentações portáteis. Além disso, foi realizado um teste de classificação para as três classes – Sadia, Assintomática HVRF e Assintomática FA – para todas as instrumentações. Os resultados desse teste foram igualmente consistentes em comparação com os resultados para duas classes, e permitem concluir que os dados de fluorescência obtidos pelas instrumentações são eficientes na caracterização distinta de cada classe.

Palavras-chave: Doenças de soja. Fluorescência. Diagnóstico precoce. Aprendizado de máquina.

ABSTRACT

SBRISSA NETO, D. A**. Diagnosis of soybean diseases through leaf fluorescence emission:** case study for Asian rust and “Green Stem and Leaf Retention”. 2019. 99 p. Tese (Doutorado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2019.

The aim of this work is to analyze the fluorescence emission of soybean plants contaminated by two important diseases for the Brazilian scenario: Asian Rust and Green Stem and Leaf Retention - popularly known as *soja louca II*. For this purpose, an experiment was performed in EMBRAPA Soja, located in Londrina-PR, with the expression of both diseases in soybean plants, as well as a control group for comparison. Five collections characterized as asymptomatic for HVRF and three collections for asymptomatic FA were obtained, and each collection consisted of 32 samples from each group. Three fluorescence techniques were used: The SteReo Lumar.V12 commercial stereomicroscope with excitation at λ = 365 nm (EM-365); the *Laser Induced Fluorescence Spectroscopy* (LIFS) and *Laser Induced Fluorescence Imaging* (LIFI) systems, both with λ=405 nm excitation. The main results of this analysis are the contribution of the emission bands at 520 nm and 740 nm for the characterization of asymptomatic ones. In the classification stage, the prediction accuracy of the test groups was high, with over 90% of total classification accuracy for all fluorescence methodologies. In the HVRF case, the LIFI instrumentation presented the highest test accuracy, with 98%. In the FA, the LIFI and LIFS instrumentations presented 93% and 94% accuracy, respectively. This information confirms the fundamental hypothesis of the work, regarding the discriminatory capacity of fluorescence emissions of asymptomatic plants, compared to healthy plants, with emphasis on LIFS and LIFI techniques, which can be adapted for portable instrumentation. In addition, a classification test was performed with the three classes – Healthy, Asymptomatic HVRF and Asymptomatic FA - for all instrumentations. Their results were equally consistent with the results for two classes, and allow us to conclude that the fluorescence data obtained by the instrumentations are efficient in the distinct characterization of each class.

Keywords: Soybean. Fluorescence. Early diagnosis. Machine learning.

LISTA DE FIGURAS

Figura 1 – Exemplo de um Diagrama de Jablonski..... 28

Figura 2 – Espectro simplificado de absorção luminosa por parte das plantas. Nesse esquema as absorções luminosas das moléculas Ch-a, Ch-b e β-caroteno formam o espectro total absorvido. 30

Figura 3 – (A) Configuração das moléculas de Ch-a e Ch-b. (B) Configuração do β-caroteno. 31

Figura 4 – Esquema do funcionamento do complexo de antena na produção de energia no processo da clorofila. Fonte: Adaptada de TAIZ et al.[20] 32

Figura 5 – Espectro de emissão de fluorescência com excitação em 405 nm obtida com a técnica de fluorescência induzida por laser. 33

Figura 6 – a) Formação das pústulas na região foliar, b) Exemplo de uma urédia, sistema reprodutivo do fungo e c) Lavoura infestada por ferrugem asiática. 35

Figura 7 – (a) Aphelenchoides besseyi extraído de uma planta contaminada com HVRF. (b) Trifólios de plantas sadias (mão direita) e assintomática HVRF com distorções (mão esquerda). (c) Período de colheita de lavoura de soja contaminada por HVRF. 36

Figura 8 – Esquema do aparato experimental LIFS. Um laser diodo de 405 nm de excitação, um espectrômetro Ocean Optics USB 4000 conectados a uma ponteira de prova. 40

Figura 9 – Aparato experimental para obtenção de imagens de fluorescência induzida por laser (LIFI). 41

Figura 10 – Estereomicroscópio (EM-365) utilizado na aquisição de imagens de fluorescência. 42

Figura 11 – A) Imagem original, de dimensão 1296 × 986 pixels. Nota-se um gradiente escurecendo regiões periféricas da imagem. B) Imagem recortada, de dimensão 370 x 370 pixels. 44

Figura 12 – Exemplo de segmentação automática e determinação da máscara. a) Imagem de fluorescência em RGB, b) Imagem resultado da transformação em escala de cinza e suavização gaussiana e c) Máscara final, como resultado da aplicação de filtros morfológicos de eliminação de ruído. 45

Figura 13 – Exemplo de conjuntos (P, R) para o algoritmo LBP..... 48

Figura 14 – Espaço de cor RGB. 49

Figura 15 – Exemplo de decomposição nos canais de cores RGB. A) imagem RGB original, B) Imagem referente ao canal de cor Vermelho (R), C) Imagem referente ao canal de cor Verde (G) e D) Imagem referente ao canal de cor Azul (B). 50

Figura 16 – Espaço de cor HSV. 51

Figura 17 – Exemplo de decomposição nos canais de cores HSV. A) imagem RGB original, B) Imagem referente ao canal da Matriz (H), C) Imagem referente ao canal da Saturação (S) e D) Imagem referente ao canal da Intensidade (V). 51

Figura 18 – Exemplo de decomposição nos canais de cores La*b*. A) imagem RGB original, B) Imagem referente ao canal (L), C) Imagem referente ao canal (a*) e D) Imagem referente ao canal (b*). 52

Figura 19 – Obtenção dos atributos de cor dominante. Os pixels da imagem original sofrem um agrupamento em três principais cores, e suas intensidades RGB e sua frequência relativa na área da folha são utilizados como atributos. 53

Figura 20 – Representação da decomposição em k componentes principais na matriz de *scores* e na matriz de *loadings*: 56

Figura 21 – Exemplo do resultado de uma validação cruzada e sua terminologia. 58

Figura 22 – Fluxograma de análise. 59

Figura 23 – Evolução dos valores médios da área F520 referente às classes Sadia e Assintomática HVRF e o respectivo desvio padrão para cada coleta. 62

Figura 24 – Evolução dos valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF para o sistema LIFS, e o respectivo desvio padrão para cada coleta. 63

Figura 25 – Distribuição dos 92 comprimentos de onda indicados pelos algoritmos de seleção de atributos para a doença HVRF. 65

Figura 26 – Projeção bidimensionais das duas primeiras componentes principais para as classes Sadia e Assintomática HVRF referente à matriz de dados dos espectros LIFS. O gráfico também mostra a área resultante da aplicação do algoritmo k-*Means* para as respectivas classes. 66

Figura 27 – Contribuição de cada comprimento de onda para a direção da primeira componente principal (1ª CP). A figura ainda traz um espectro arbitrário de uma amostra para referência dos comprimentos de onda. 67

Figura 28 – Matrizes de confusão para a classificação das amostras assintomáticas HVRF com dados LIFS. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste. 69

Figura 29 – Projeção bidimensional nas duas componentes principais de maior variância acumulada para as classes Sadia e Assintomática HVRF referente a matriz de dados das imagens EM-365. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes. 70

Figura 30 – Matrizes de confusão para a classificação das amostras assintomáticas HVRF com dados do EM-365. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste. 73

Figura 31 – Projeção bidimensional nas duas componentes principais de maior variância acumulada para as classes Sadia e Assintomática HVRF referente à matriz de dados das imagens LIFI. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes. Além disso, são destacados na figura os centroides de cada *cluster*. 74

Figura 32 – Projeção tridimensional nas três componentes principais de maior variância acumulada para as classes Sadia e Assintomática HVRF referente à matriz de dados das imagens LIFI. 75

Figura 33 – Matrizes de confusão para a classificação das amostras assintomáticas HVRF com dados LIFI. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste. 77

Figura 34 – Evolução dos valores médios da área F520 para as classes Sadia e Assintomática FA através das coletas 4 dai, 7 dai e 11 dai. 78

Figura 35 – Evolução dos valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF para o sistema LIFS e o respectivo desvio padrão para cada coleta. 79

Figura 36 – Distribuição dos 95 comprimentos de onda indicados pelos algoritmos de seleção de atributos para a doença FA. 80

Figura 37 – Projeção bidimensional das duas primeiras componentes principais para as classes Sadia e Assintomática FA referente à matriz de dados dos espectros LIFS. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes..... 81

Figura 38 – Contribuição de cada comprimento de onda para a direção da primeira componente principal (1ª CP). A figura ainda traz um espectro arbitrário de uma amostra para referência dos comprimentos de onda. 82

Figura 39 – Matrizes de confusão para a classificação das amostras assintomáticas FA com dados LIFS. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste..... 83

Figura 40 – Projeção bidimensional das duas primeiras componentes principais para as classes Sadia e Assintomática FA para as imagens EM-365. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes. 84

Figura 41 – Matrizes de confusão para a classificação das amostras assintomáticas FA com dados EM-365. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste. 86

Figura 42 – Projeção bidimensional das duas primeiras componentes principais para as classes Sadia e Assintomática FA para as imagens LIFI. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes. 87

Figura 43 – Matrizes de confusão para a classificação das amostras assintomáticas FA com dados LIFI. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste..... 90

Figura 44 – Matrizes de confusão A) para validação dos modelos de classificação e B) para predição do conjunto de teste das amostras Sadia, Assintomática HVRF e Assintomáticas FA para os espectros do LIFS. 91

Figura 45 – Matrizes de confusão A) para validação dos modelos de classificação e B) para predição do conjunto de teste das amostras Sadia, Assintomática HVRF e Assintomáticas FA para as imagens do EM-365. 91

Figura 46 – Matrizes de confusão A) para validação dos modelos de classificação e B) para predição do conjunto de teste das amostras Sadia, Assintomática HVRF e Assintomáticas FA para as imagens do LIFI. 92

LISTA DE TABELAS

Tabela 1 – Coletas e quantidades de amostras obtidas no experimento. 38

Tabela 2 – Valores médios da área F520 referente às classes Sadia e Assintomática HVRF e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias. 62

Tabela 3 – Valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF, e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias. 63

Tabela 4 – Nome e módulo da contribuição dos 50 principais atributos na direção da primeira componente principal. 71

Tabela 5 – Nome e módulo da contribuição dos 50 principais atributos na direção da primeira componente principal. 76

Tabela 6 – Valores médios da área F520 referente às classes Sadia e Assintomática FA e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias. 78

Tabela 7 – Valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias. 79

Tabela 8 – Nome e módulo da contribuição dos 50 principais atributos na direção da primeira componente principal. 85

Tabela 9 – Nome e módulo da contribuição dos 46 principais atributos na direção da primeira componente principal para os dados LIFI da doença FA. 89

SUMÁRIO

1 INTRODUÇÃO 23
1.1 OBJETIVOS 25
2 REVISÃO BIBLIOGRÁFICA 27
2.1 FLUORESCÊNCIA INDUZIDA POR LASER 27
2.2 PROCESSOS FOTOSSINTÉTICOS DE PLANTAS 29
2.3 FLUORESCÊNCIA DA CLOROFILA E METABÓLITOS SECUNDÁRIOS 32
2.4 FERRUGEM ASIÁTICA 34
2.5 HASTE VERDE E RETENÇÃO FOLIAR (HVRF) 35
3 MATERIAIS E MÉTODOS 37
3.1 REALIZAÇÃO DO EXPERIMENTO E OBTENÇÃO DE AMOSTRAS 37
3.2 SISTEMAS DE AQUISIÇÃO DA FLUORESCÊNCIA 39
3.2.1 Sistema de fluorescência induzida por laser 39
3.2.2 Imagem de fluorescência induzida por LED 40
3.2.3 Estereomicroscópio de fluorescência com excitação 365 nm (EM-365) 41
3.2.4 Processamento de dados de espectroscopia 42
3.2.5 Processamento de imagens digitais 42
3.2.5.1 Pré-processamento 43
3.2.5.2 Segmentação 44
3.2.5.3 Extração de atributos 45
3.2.5.4 Atributos estatísticos 46
3.2.5.5 Canais de cores 49
3.2.5.6 Cor dominante 52
3.2.6 Seleção de atributos 53
3.2.7 Redução de dimensionalidade (PCA e normalização *z*-score) 54
3.2.8 Classificação e validação cruzada 56
3.2.9 Matriz de confusão 58
3.2.10 Fluxograma de análise 59
4 RESULTADOS E DISCUSSÕES 61
4.1 HASTE VERDE E RETENÇÃO FOLIAR: LIFS 61
4.2 HASTE VERDE E RETENÇÃO FOLIAR: EM-365 69
4.3 HASTE VERDE E RETENÇÃO FOLIAR: LIFI 73
4.4 FERRUGEM ASIÁTICA: LIFS 77
4.5 FERRUGEM ASIÁTICA: EM-365 83

4.6 FERRUGEM ASIÁTICA: LIFI 86
4.7 COMPARAÇÃO DAS TRÊS CLASSES 91
5 CONCLUSÕES 93
REFERÊNCIAS 95

1 INTRODUÇÃO

O agronegócio é uma das mais importantes atividades comerciais brasileiras. Dentre os insumos negociados pelo Brasil, a soja e seus subprodutos destacam-se como umas das mais rentáveis. De acordo com o Ministério da Agricultura, Pecuária e Abastecimento (MAPA), o total de exportações com o complexo da soja alcançou aproximadamente U$ 31,7 bilhões no ano de 2018.[1] Essas cifras colocam o Brasil dentre os maiores exportadores de soja do mundo. Esse patamar de exportação deve-se aos investimentos e pesquisas em áreas estratégicas do setor, que têm alavancado a produtividade dos grãos em todas as regiões brasileiras.

Apesar do imenso volume de soja exportado pelo Brasil, grande parte da produção é perdida devido a patógenos nocivos à planta. A Organização Mundial para Alimentação e Agricultura (FAO) estima que tais patógenos sejam responsáveis por 42,1 % das perdas totais em soja no mundo, sendo que 13,3% são causados por fitopatógenos, tais como fungos, bactérias, nematoides e vírus.[2] No Brasil, já são catalogadas cerca de 40 tipos diferentes de doenças que acometem as plantações de soja nas lavouras brasileiras.[3]

Tendo isso em vista, o manejo de agentes patógenos é uma importante estratégia no aumento de produtividade da soja, refletindo também numa maior qualidade do produto, pois possibilita o uso estratégico de defensivos na lavoura. Há necessidade de se identificar rapidamente a presença de patógenos, pois as doenças diagnosticadas em estágio inicial podem ser mais facilmente tratadas, o que também evita sua proliferação. Atualmente no Brasil, o manejo e controle de pragas se dá majoritariamente através de uso de defensivos, do controle biológico e da inspeção em campo.[4]

Dentre as principais doenças que afetam os cultivares de soja, destacam-se a Ferrugem Asiática (FA) e a Haste Verde e Retenção Foliar (HVRF). A Ferrugem Asiática existe no Brasil desde a década de 70 e é transmitida pelo fungo *phakopsora pachyrhizi*. A proliferação de uma doença fúngica é rápida, levando aproximadamente cinco dias para completar seu ciclo. Além disso, seus esporos podem permanecer inativos por um longo tempo, até encontrarem condições favoráveis para seu desenvolvimento. A ferrugem asiática, quando não tratada, pode acometer até 80% de toda lavoura.[5] Atualmente, as principais técnicas de manejo para a ferrugem asiática consistem na aplicação de defensivos químicos, utilização de variedades geneticamente resistentes e também através da realização do vazio sanitário – período de descanso do lote de terra, que compreende de dois a três meses do ano. O vazio sanitário garante a eliminação dos esporos da Ferrugem Asiática e também de outros patógenos nocivos à soja.[6] Pelo fato de ser uma doença de rápida proliferação, é necessário que haja um acompanhamento contínuo da lavoura, a

fim de evitar que o fungo se espalhe e, consequentemente, provoque diminuição da produção total. Nesse viés, o diagnóstico precoce é considerado fundamental para que seja feito o manejo apropriado da doença, com aplicações de defensivos nas proporções corretas e fazendo rotatividade dos mesmos.

Já a haste verde e retenção foliar (HVRF) – popularmente conhecida como Soja Louca II – é uma doença causada pelo nematoide *aphelenchoides besseyi*. É uma doença relativamente nova no cenário brasileiro, sendo descoberta recentemente por pesquisadores da EMBRAPA Soja e reconhecida apenas em 2015 pelo MAPA.[7] Com relação ao nematoide causador da doença, sabe-se que o mesmo já é hospedeiro em outras culturas, como algodão e feijão, além de mais quatro plantas daninhas.[8] Por se tratar de uma doença descoberta recentemente, pouco se sabe sobre sua fisiologia, prevenção e combate. O ciclo reprodutivo desse nematoide é de aproximadamente dez dias, ou seja, a HVRF, assim como a ferrugem, também é uma doença de rápida proliferação. Além disso, tem-se observado grandes infestações em regiões de clima mais quente no Brasil, compreendendo majoritariamente os estados das regiões Norte e Nordeste, com condições de temperatura e umidade elevadas. O uso de defensivos nas lavouras, como vermífugos, por exemplo, tem se mostrado ineficientes no combate da HVRF, de tal maneira que o controle mais efetivo se dá através da inspeção manual da lavoura e extração de plantas contaminadas. O sintoma mais marcante da HVRF é a manutenção da planta em um estado vegetativo perene, o que dificulta a inspeção das mesmas, pois as plantas contaminas ficam evidentes por contrate, quando as demais plantas sadias secam. Sendo assim, o diagnóstico precoce da doença é crucial, a fim de que seja realizada a extração dos pés infectados, impedindo que a doença se dissemine pela lavoura.

Nos últimos anos, a agricultura de precisão tem ganhado espaço em diversas linhas de pesquisa, onde projetos e parcerias têm visado melhorias na cadeia produtiva agrícola. A agricultura de precisão é um termo relativamente novo, e engloba toda uma gama de aparatos tecnológicos no auxílio das tomadas de decisões inteligentes. A temática da agricultura de precisão remete desde o preparo e análise do solo, modificações genéticas para obtenção de sementes resistentes, controle biológico de pragas, sensoriamento remoto, até o desenvolvimento de tecnologias inspeção das plantas em campo.[9-10]

No que tange ao desenvolvimento de equipamentos portáteis de inspeção, a utilização de técnicas biofotônicas de interação da luz tem possibilitado novas perspectivas de análise acerca do estado de saúde das plantas. Baseado no fenômeno físico da fluorescência dos metabólitos foliares, é possível determinar variações sutis em plantas e identificar padrões que seriam impossíveis de serem determinados a olho nu. Em artigo publicado em 2012, Milori e

colaboradores utilizaram espectros de fluorescência de folhas de citrus e conseguiram sensibilidade suficiente para diferenciar até dez variedades distintas.[11] No mesmo ano, Pereira e colaboradores utilizaram imagens de fluorescência para identificar a presença de *huanglogbing* – bactéria responsável pela doença de citrus *Greening* –, logo nos primeiros meses de infestação da doença.[12] Ainda na temática do *Greening*, Ranulfi e colaboradores utilizaram espectros de fluorescência na categorização de plantas sadias, sintomáticas e assintomáticas de HBL.[13] Todos os trabalhos acima citados constituem parte das linhas de pesquisas realizadas nas dependências da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) Instrumentação, localizada na cidade de São Carlos, interior de São Paulo.

Baseado na expertise alcançada no grupo e observando as demandas da agricultura nacional, foi proposto para o presente trabalho um estudo sistemático das doenças de soja Ferrugem Asiática e Haste Verde e Retenção Foliar e a avaliação das condições de realização do diagnóstico das mesmas através de técnicas fotônicas de fluorescência. A hipótese original proposta é de que as técnicas espectrais e de imagens de fluorescência revelem alterações físico-químicas nas plantas contaminadas com as referidas doenças, possibilitando a identificação precoce das mesmas, ainda em estágio assintomático. Destaca-se como novidade do trabalho o pioneirismo no estudo das respectivas doenças de soja, através da análise de imagens e espectros de fluorescência das mesmas em estágio assintomático. Além disso, pela primeira vez é reportado um comparativo sistemático de duas técnicas distintas de fluorescência para folhas de soja.

1.1 OBJETIVOS

O principal objetivo do presente trabalho é investigar a evolução das doenças ferrugem asiática e HVRF em plantas de soja através da análise de imagens e espectros de fluorescência das mesmas. Como estratégia de desenvolvimento do projeto, inicialmente foi realizado um ensaio com plantas de sojas inoculadas com as doenças em ambiente controlado, a fim de obter amostras das doenças durante sua evolução.

Foram utilizadas três técnicas de fluorescência. A primeira apresentada é a Espectroscopia de Fluorescência Induzida por Laser (*Laser Induced Fluorescence Spectrum* – LIFS). As outras duas técnicas de fluorescência captam imagens de fluorescência: Um sistema de Imagem de Fluorescência Induzida por Laser (*Laser Induced Fluorescence Image* – LIFI) e um estereomicroscópio comercial da marca Zeiss® Lumar v12 (aqui referenciado como EM-365). A técnica LIFS possui excitação em 405 nm e um espectrômetro com variação de 190 nm a 890 nm.

A técnica LIFI também apresenta excitação em 405 nm e utiliza uma câmera digital comercial. Já o EM-365 utiliza uma fonte de excitação em 365 nm e um microscópio para obtenção de imagens ampliadas de fluorescência. Essa gama de instrumentações foi proposta a fim de entender as características do problema proposto e buscar soluções de diagnóstico em campo, através de imagens e espectros de fluorescência.

2 REVISÃO BIBLIOGRÁFICA

A seguir são apresentados os principais temas e conceitos referentes ao trabalho. Inicia-se uma apresentação dos processos físicos e biológicos que compõem as atividades fotossintéticas das plantas em geral e, em seguida, serão abordadas as doenças de soja exploradas neste trabalho. Este capítulo de revisão se encerra com uma explanação sobre o conceito de fluorescência e as características do espectro resultante da excitação de folhas

2.1 FLUORESCÊNCIA INDUZIDA POR LASER

A interação da luz com a matéria é um processo físico de extrema relevância nos dias de hoje, com infinitas aplicações. Ao interagir com a matéria, a luz pode sofrer efeitos ópticos tais como absorção e reflexão parcial ou total.[14] De acordo com a teoria quântica,[15] a luz é constituída de pequenos pacotes energéticos chamados de fótons. Esses fótons possuem valor de energia discreta que depende linearmente da sua frequência, de acordo com a relação:

$$E = h\nu \tag{1}$$

Onde E é a energia, h é a constante de Planck e ν a frequência do fóton.[15]

No processo de absorção da luz, os fótons podem ser absorvidos por átomos e moléculas, dependendo de sua frequência de excitação. A absorção de um fóton conduz as moléculas a um estado quântico excitado. Baseado no princípio de Franck-Condon[16], o processo de absorção energética ocorre na ordem de 10^{-15} segundos. Para ocorrer a transição de um estado fudamental E_m para um estado quântico excitado E_n através da absorção do fóton, a frequência desse fóton deve ser proporcional à diferença das energias dos estados quânticos, de acordo com:

$$\nu = \frac{E_n - E_m}{h} \tag{2}$$

No estado excitado, as moléculas estão mais instáveis e tendem a retornar ao estado fundamental, por meio de um relaxamento, quando pode ou não haver emissão de um novo fóton. No caso do relaxamento não radioativo, a molécula excitada retorna ao seu estado fundamental através de decaimentos vibracionais moleculares. No caso de relaxamento radioativo, há emissão de um novo fóton e retorno da molécula ao estado fundamental. Esse novo fóton de

luz possui energia menor que o absorvido anteriormente. Esse fenômeno é chamado de Luminescência, que por sua vez pode ser dividido em dois processos distintos: a Fluorescência e a Fosforescência.[17]

Para exemplificar os processos envolvidos na absorção energética, utiliza-se o diagrama de Jablonski, ilustrado na Figura 1.

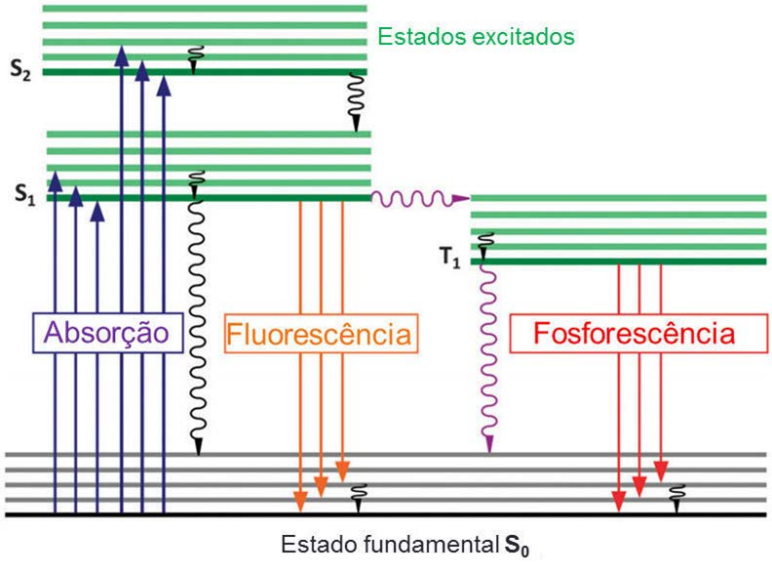

Figura 1 – Exemplo de um Diagrama de Jablonski.
Fonte: Adaptada de HEINE; MÜLLER-BUSCHBAUM.[18]

No diagrama, as letras S e T representam estados singletos e tripletos, respectivamente. Cada um dos estados singletos (S_1 e S_2) ou tripletos (T_1) possui configurações vibracionais diferentes que variam minimamente de um para o outro, representadas pelas linhas acima dos estados. A Relaxação Vibracional é a perda energética que ocorre entre as configurações vibracionais distintas dentro de um mesmo estado. Essa perda acontece por meio de colisões moleculares das espécies excitadas com o solvente, acarretando em um aumento de energia da temperatura do sistema. A relaxação vibracional ocorre em torno de 10^{-12} segundos e não há emissão de fótons. O relaxamento entre estados singletos (S_1 para S_2, por exemplo) é chamado de conversão interna e a principal característica é a manutenção do estado de *spin* da molécula. No caso de uma transição de um estado singleto para outro tripleto, há variação no estado do spin e o fenômeno é chamado de Cruzamento Intersistema. A transição de estados singletos para o estado fundamental com emissão de radiação e sem inversão de *spin* é chamado de Fluorescência. O tempo gasto nessa transição é da ordem de 10^{-8} segundos. Isso implica que o valor de

energia emitida na fluorescência é menor do que o processo de absorção. Sendo assim, a frequência do fóton emitido na fluorescência é menor.

A transição de estados tripletos para o estado fundamental é proibida por dipolos elétricos, de acordo com as regras de seleção quântica, mas podem ocorrer por interações de quadripolos elétricos, porém com probabilidade muito inferior quando comparada com as transições da fluorescência. Esse processo de transição de um estado tripleto para o estado fundamental com emissão de luz é chamado de Fosforescência. Tendo em vista a baixa probabilidade de ocorrência do fenômeno, o tempo de emissão da Fosforescência é da ordem de milissegundos a segundos.

2.2 PROCESSOS FOTOSSINTÉTICOS DE PLANTAS

A fotossíntese é um processo bioquímico que, através da captura da luz solar e síntese da água e dióxido de carbono, gera energia na forma de ATP e glicose para manutenção dos ciclos vitais das plantas.[19] Esse processo ocorre em organelas intracelulares chamadas cloroplastos, que são reservatórios dos pigmentos fotossintetizantes. Dentre os diversos pigmentos responsáveis pela fotossíntese, as clorofilas e os carotenoides são os mais abundantes nas plantas. As clorofilas *a* e *b* (Ch-a e Ch-b) encontram-se nas plantas na proporção de 3:1, sendo a Ch-b responsável por ampliar o espectro de absorção luminosa. O carotenoide β-caroteno soma-se às clorofilas na absorção luminosa e compõe predominantemente o espectro de absorção luminosa, exemplificada na Figura 2. Cada pigmento é responsável por determinado perfil espectral de absorção, que somados, constituem o espectro total utilizado pela planta para realização da fotossíntese. É importante ressaltar que os processos quânticos ocorrem para configurações energéticas bem definidas, e uma alteração no perfil de absorção pode gerar *déficits* nas sínteses de plantas sadias.

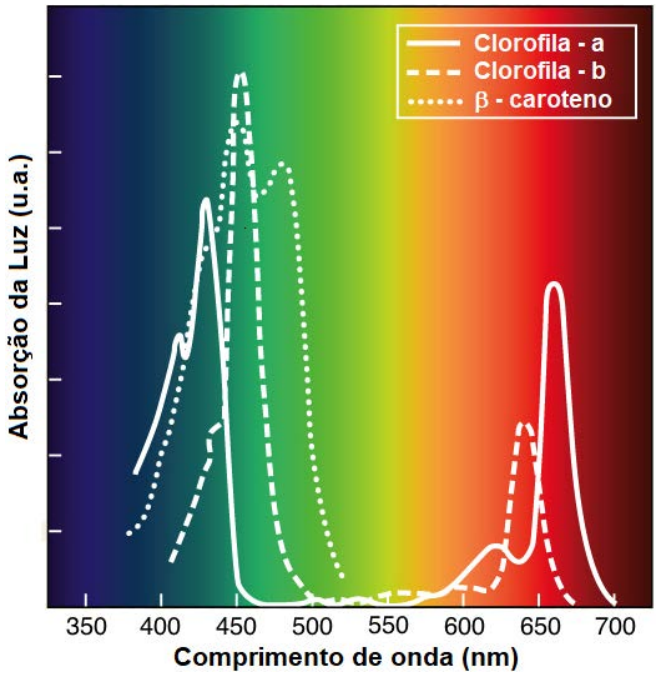

Figura 2 – Espectro simplificado de absorção luminosa por parte das plantas. Nesse esquema as absorções luminosas das moléculas Ch-a, Ch-b e β-caroteno formam o espectro total absorvido.
Fonte: Adaptada de KHAN ACADEMY.[20]

As clorofilas são pigmentos fotossintetizantes produzidos naturalmente pelas plantas, e estão quimicamente relacionadas com os grupos do tipo porfirina (um grupo circular de átomos circundando um íon de magnésio), contendo uma complexa estrutura em anel e normalmente uma longa cauda de hidrocarbonetos de característica hidrofóbica, de acordo com a Figura 3. As clorofilas *a* e *b* são quimicamente similares e diferem estruturalmente apenas no radical do carbono C-3 do anel de porfirina. Na Ch-a, esse carbono está ligado a um grupo metil (-CH3), enquanto que na Ch-b ele se liga a um grupo aldeído (-CHO). Essa diferença, apesar de sutil, confere uma variação espectral entre ambas, crucial para a formação do espectro de absorção característico.

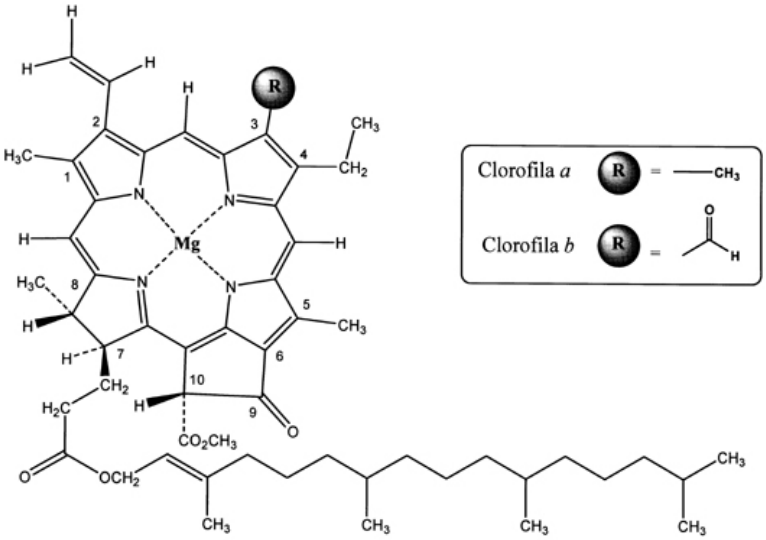

Figura 3 – (A) Configuração das moléculas de Ch-a e Ch-b. (B) Configuração do β-caroteno.
Fonte: STREIT et al.[21]

Os carotenoides também são pigmentos fotossintetizantes naturais caracterizados por biomoléculas lineares com múltiplas ligações duplas conjugadas. Nas folhas, esses pigmentos apresentam coloração alaranjada característica, devido às suas bandas de absorção na região dos 400 a 500 nm. Os carotenoides contribuem com a captação luminosa realizada nos cloroplastos. Além disso, essas moléculas têm papel de regulação de temperatura dentro da planta, ajudando na absorção do excesso de luz solar incidente nas folhas.

Uma única molécula de clorofila não é capaz de absorver toda luz necessária para realização da fotossíntese. É necessário um conjunto de moléculas operando para fornecer a quantidade adequada de captação solar. Nas plantas, o processo de absorção luminosa ocorre em complexos do tipo antena, onde clorofilas e carotenoides fazem a absorção luminosa e transferência de energia por ressonância, molécula a molécula, até o centro de ação das antenas para que clorofilas especiais realizem o processo de transferência eletrônica de um pigmento doador para outro receptor. O modelo de funcionamento do complexo antena está ilustrado da Figura 4.

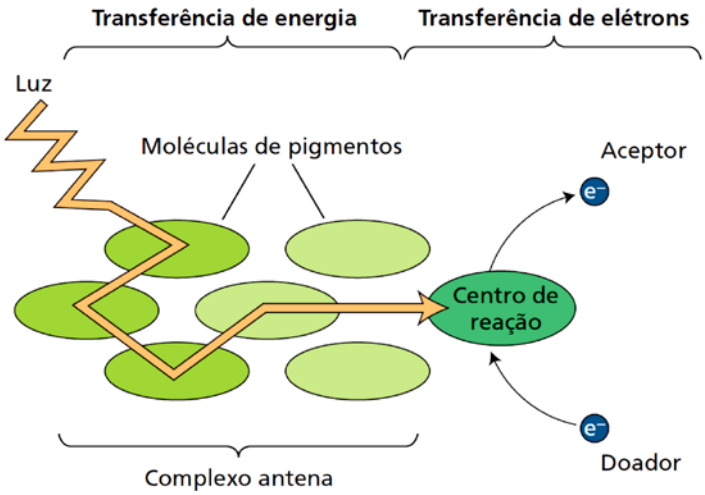

Figura 4 – Esquema do funcionamento do complexo de antena na produção de energia no processo da clorofila. Fonte: Adaptada de TAIZ et al.[19]

2.3 FLUORESCÊNCIA DA CLOROFILA E METABÓLITOS SECUNDÁRIOS

Além de sua importância biológica nos processos fotossintéticos, as clorofilas *a* e *b*, carotenoides e outros metabólitos ditos secundários apresentam características ópticas importantes. A presença dos anéis porfirina fenólicos na composição estrutural da maioria das moléculas confere estabilidade às mesmas, gerando características de fluorescência ao serem excitadas por comprimentos de onda adequados.[16, 22] O resultado da emissão de fluorescência por diversas moléculas proporciona a formação do espectro característico da planta. Esse espectro é a composição da emissão de diversas moléculas e sua correta interpretação é capaz de revelar alterações dos processos metabólicos de uma planta, caracterizando condições de stress bióticos e abióticos, mesmo sem expressar visivelmente seus sintomas.[23] A Figura 5 apresenta o espectro de emissão fluorescência da clorofila para uma excitação em 405 nm de uma folha de soja sadia.

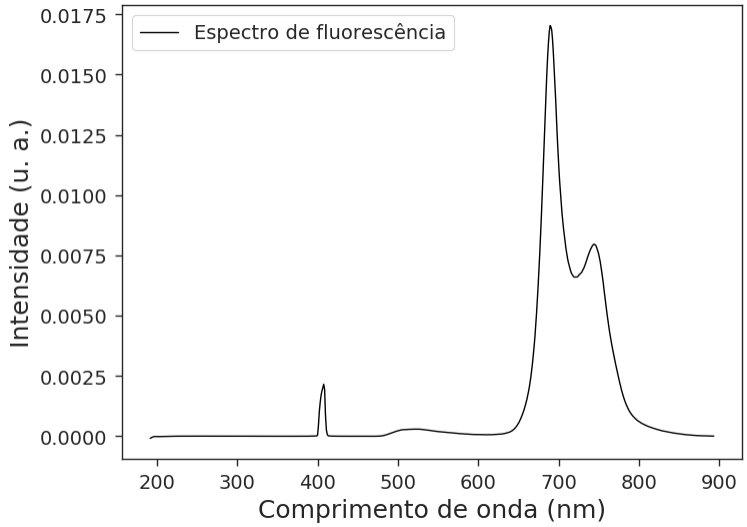

Figura 5 – Espectro de emissão de fluorescência com excitação em 405 nm obtida com a técnica de fluorescência induzida por laser.
Fonte: Elaborada pelo autor

O gráfico apresentado na Figura 5 revela faixas espectrais cruciais à interpretação do estado de saúde de uma planta. Segundo Buschmann[24], o espectro de fluorescência referente à emissão no vermelho / infravermelho, centrada em 690 e 740 nm, é formado majoritariamente pela emissão de fluorescência da Ch-a. A Ch-b e carotenoides, que são responsáveis pela formação do espectro de absorção, contribuem fracamente para a formação do espectro característico nessa faixa. Os dois picos auxiliam na determinação do estado de saúde da planta, pois um desbalanceamento na proporção desses picos sugere uma variação na atividade metabólica da planta. O principal indicativo de stress da planta é bem referenciado pela relação de intensidades máximas da banda de emissão 690 nm pela banda 740 nm.[25-28]

Além da emissão no vermelho-infravermelho, nota-se uma banda de emissão no azul – verde, localizada entre aproximadamente 470 e 600 nm.[29] Nessa região existe uma variedade de moléculas que contribuem para a formação do espectro.[30-31] A maioria delas apresentam anéis fenólicos em sua composição, que são estáveis e possibilitam a emissão de fluorescência, desde que excitados com um comprimento de onda adequado.[27, 32] A constituição dessa banda característica é uma somatória de múltiplas emissões, sendo reconhecida como a assinatura espectral da planta, pois a proporção das moléculas que contribui para a formação do espectro é invariável com o tempo de vida da planta, diferentemente do que ocorre para a emissão no vermelho-infravermelho, onde as quantidades de clorofila aumentam com o desenvolvimento da mesma e influenciam na intensidade do espectro. Meyer et al[33] utilizou essa propriedade para desenvolver uma técnica para mensurar a idade de uma planta através da análise das bandas de emissão da Ch-a. Dentre os inúmeros compostos responsáveis pela fluorescência no azul e

vermelho, apresentam-se coumarinas, ligninas, diversos tipos de fenóis, ácidos ferúlicos ligados à parede celular da planta e flavonoides.[34] Esse último, em especial, é produzido nas folhas como um mecanismo de defesa contra patógenos externos, o que faz dessa banda de emissão um componente importante para o reconhecimento de doenças em estágio assintomático, pois um aumento na produção de flavonoides implica diretamente numa maior emissão da fluorescência dessa banda.[35]

2.4 FERRUGEM ASIÁTICA

A monocultura da soja, bem como a maioria das monoculturas, sofre com diversos tipos de pragas que acometem as lavouras, desde vírus até insetos e plantas daninhas. Tais patógenos externos estão diretamente vinculados a perdas de produção. Uma das principais doenças de soja está associada ao fungo *phakopsora*, causador da doença ferrugem asiática. A ferrugem asiática, alvo deste trabalho, é causada pelo fungo *phakopsora pachyrhizi*, nativo do oriente e presente em praticamente todos os países com plantações de soja.[6]

Os sintomas apresentados pela ferrugem asiática podem ocorrer em qualquer estágio do desenvolvimento da planta. Porém, tem-se observado que a maior parte da sua contaminação ocorre no estágio germinativo e com maior quantidade na parte de baixo da folha.[3] As condições para sua proliferação são a alta umidade e temperaturas na faixa de 15 a 28 °C. Conforme ilustrado na Figura 6, os sintomas iniciais da ferrugem são pequenas lesões nas folhas, de coloração levemente marrom, referente à germinação dos esporos.[4-5] Com o desenvolvimento do fungo, ocorre a formação das ureias, que são as estruturas reprodutivas do fungo, possibilitando a disseminação do fungo e a formação de lesões foliares visíveis. Se não tratada, a ferrugem asiática pode acometer até 80% da lavoura.[37]

O manejo da ferrugem asiática feito atualmente no Brasil é caracterizado pelo uso de defensivos fúngicos e monitoramento da lavoura. Apesar de existirem diversas marcas de fungicidas com elevada efetividade, o uso indiscriminado desses produtos pode acarretar em uma resistência do fungo quanto ao seu uso. Rotatividade de marcas e dosagens adequadas são também condições necessárias para o manejo correto. Nos últimos anos, o Brasil também tem adotado, como estratégia de manejo da ferrugem asiática, o período de vazio sanitário nas lavouras de soja. Esse período corresponde aos meses de agosto a outubro, quando são inutilizadas as porções de terra onde a soja será plantada, garantindo assim a morte dos esporos da ferrugem e também de outros patógenos. Por fim, o monitoramento da lavoura próxima à época de floração

e em períodos de chuva e umidade é fundamental, inclusive para a avaliação sobre o uso de fungicidas.

Figura 6 – a) Formação das pústulas na região foliar, b) Exemplo de uma urédia, sistema reprodutivo do fungo e c) Lavoura infestada por ferrugem asiática.

Fonte Adaptada de ROHÁČEK et al. [34]

2.5 HASTE VERDE E RETENÇÃO FOLIAR (HVRF)

A HVRF, conhecida popularmente por soja louca II, é atualmente a mais nova doença de soja catalogada e reconhecida pelo MAPA em 2015.[38] A síndrome da HVRF é transmitida pelo nematoide *aphelenchoides besseyi*, mesmo organismo responsável por doenças em cultivares de arroz e algodão.[3]

Os principais sintomas apresentados pelas plantas infectadas pela HVRF são mais aparentes nas fases de floração e colheita da soja. Nessa fase, as plantas contaminadas mantêm-se verdes, em constante estado vegetativo. Todas as estruturas da planta (caule, folhas e pecíolos) ficam mal ou subdesenvolvidas, causando diminuição na produção das lavouras através do abortamento das vagens, enquanto as plantas sadias completam sua senescência e produzem seus frutos. Além do prejuízo na produção, a realização da colheita das plantas infectadas é dificultada, pois as plantas verdes podem danificar os maquinários de colheita, que são desenvolvidos exclusivamente para plantas maduras.[39]

Outro problema enfrentado pelos produtores de soja consiste na ineficiência dos defensivos agrícolas contra os nematoides. Estima-se que menos de 30% dos parasitas são eliminados com vermífugos disponíveis atualmente. Sendo assim, o principal manejo realizado pelos produtores é o diagnóstico precoce da doença e extração manual das plantas infectadas. Essa prática apresenta duas grandes dificuldades. A primeira delas é o diagnóstico precoce em si, antes da expressão dos sintomas, pois ainda não é reconhecida uma técnica eficiente para tal. A segunda é a inaplicabilidade desse manejo para grandes lavouras de soja, que chegam a alcançar

milhares de hectares.[39] A Figura 7 apresenta uma imagem ampliada do nematóide, bem como exemplos de infestação da mesma.

Figura 7 – (a) Aphelenchoides besseyi extraído de uma planta contaminada com HVRF. (b) Trifólios de plantas sadias (mão direita) e assintomática HVRF com distorções (mão esquerda). (c) Período de colheita de lavoura de soja contaminada por HVRF.
Fonte: Adaptada de AGROLINK [36]

3 MATERIAIS E MÉTODOS

Neste capítulo serão abordados três assuntos fundamentais para o desenvolvimento do trabalho. Inicialmente são apresentados os detalhes da realização dos ensaios referentes ao desenvolvimento das doenças de soja Haste Verde e Retenção Foliar e Ferrugem Asiática. Em seguida serão abordados os instrumentos de obtenção de fluorescência: *Laser Induced Fluorescence Spectroscopy* (LIFS), *Laser Induced Fluorescence Imaging* (LIFI) e o estereomicroscópio comercial (EM-365). Por fim, são apresentados os conceitos para o processamento e análise dos espectros e imagens digitais de fluorescência obtidos, bem como os métodos computacionais referentes à etapa de classificação dos grupos.

3.1 REALIZAÇÃO DO EXPERIMENTO E OBTENÇÃO DE AMOSTRAS

A hipótese que norteia o presente trabalho é a detecção precoce, em estágio dito assintomático, das doenças Ferrugem Asiática e Haste Verde e Retenção Foliar, as quais serão referenciadas por FA e HVRF, respectivamente. Para tal, foi realizado um ensaio com o desenvolvimento dessas doenças em ambiente controlado para coleta e acompanhamento das plantas. O experimento ocorreu na cidade de Londrina - PR, nas dependências da EMBRAPA Soja, com a coordenação do pesquisador Maurício Meyer. Os experimentos ocorreram em casa de vegetação simulando as condições favoráveis ao desenvolvimento das plantas e das doenças. A expertise do pesquisador definiu os parâmetros de controle do ambiente, com temperatura média em torno de 25 °C e nebulização a cada hora. A condição de nebulização é fundamental para o desenvolvimento de ambas as doenças, que necessitam de umidade elevada.

A variedade utilizada neste experimento foi a Brasmax Apolo RR. Essa variedade é bem adaptada ao clima da região sul do Brasil, e não possui nenhum tipo de resistência genética a fungos e nematoides. Foram preparados 100 vasos, sendo 33 destinados à inoculação da FA, 33 destinados à HVRF e os 34 restantes ao grupo Controle. Cada vaso desenvolveu cinco plantas e durante a fase inicial, as plantas que apresentavam falhas de desenvolvimento eram subtraídas. As amostras foram obtidas coletando uma folha por vaso, selecionando o melhor trifólio mais antigo. É importante ressaltar que cada amostra corresponde a uma folha de cada vaso e não necessariamente da mesma planta. Ou seja, para a próxima coleta, a amostra de determinado vaso não precisamente será extraída da planta anterior, mas de quaisquer outras plantas do mesmo vaso, levando em consideração a expertise do fitopatologista responsável pela coleta

para a escolha do melhor conjunto planta/trifólio. Essa alternativa foi proposta a fim de se obter uma quantidade de amostras estatisticamente significativa e evitar alterações fisiológicas por *stress* nas plantas devido às constantes subtrações de folhas. Tal alternativa gerou uma quantidade amostral de folhas suficiente para a sequência do trabalho, sendo cada coleta composta por 32 das melhores amostras de cada grupo.

O processo de inoculação das doenças seguiu os protocolos tradicionais conhecidos para ambas, com alguns detalhes que serão aqui abordados. A casa de vegetação possuía inóculos da FA, devido a experimentos paralelos realizados no mesmo local. Sendo assim a estratégia adotada para prosseguir com o cronograma foi fazer a semeadura das plantas designadas à inoculação da FA em outra casa de vegetação, e após 60 dias da semeadura, inseri-las na casa de vegetação destinada, marcando assim o início da inoculação da FA. Já os demais grupos Controle e inoculadas com HVRF foram tratados com defensivos fúngicos durante todo o desenvolvimento do ensaio para que os mesmos não fossem contaminados pelos esporos do fungo da FA presentes no ambiente. Os inóculos da HVRF foram multiplicados previamente e a inoculação das plantas com os nematoides foi feito na proporção de 5.000 por planta no 45º dia de desenvolvimento da planta, a contar da semeadura.

A tabela 1 mostra a disposição dos dias das coletas, bem como as quantidades de coletas realizadas para cada grupo.

Tabela 1 – Coletas e quantidades de amostras obtidas no experimento.

	Grupo Controle	**Grupo Ferrugem / (d.a.i)**	**Grupo HVRF / (d.a.i)**
Coleta 01	32	–	32 / (0 dai)
Coleta 02	–	–	– / (3 dai)
Coleta 03	32	–	32 / (7 dai)
Coleta 04	32	–	32 / (10 dai)
Coleta 05	–	–	– / (14 dai)
Coleta 06	–	–/ (0 dai)	– / (17 dai)
Coleta 07	32	32 / (4 dai)	32 / (21 dai)
Coleta 08	32	32 / (7 dai)	32 / (24 dai)
Coleta 09	32	32 / (11 dai)	32 / (28 dai)
Coleta 10	32	32 / (14 dai)	32 / (31 dai)

- dai = dia após inoculação

Fonte: Elaborada pelo autor

Do total de amostras geradas, duas coletas foram descartadas por inconsistência e uma foi descartada por degradação amostral. O ensaio para a HVRF resultou em cinco coletas assintomáticas adequadas para o estudo – coletas 7 dai 11 dai, 21 dai 25 dai e 28 dai –, totalizando 192 amostras para cada classe. Já o desenvolvimento da doença FA gerou apenas três coletas assintomáticas – 4 dai, 7 dai e 11 dai –, com 96 amostras por classe. Cabe ressaltar que as amostras, caracterizadas como assintomáticas, foram acompanhadas até a expressão dos sintomas de cada doença.

3.2 SISTEMAS DE AQUISIÇÃO DA FLUORESCÊNCIA

A seguir serão detalhados os sistemas de aquisição de fluorescência utilizados no projeto. Foram utilizados três sistemas de aquisição, baseados em espectros e imagens de fluorescência. Inicialmente será apresentado o sistema de espectroscopia *Laser Induced Fluorescence Spectrum* (LIFS). Na sequência, serão apresentados o sistema *Laser Induced Fluorescence Imaging* (LIFI) e o estereomicroscópio Zeiss® Lumar v12 (EM-365).

3.2.1 Sistema de fluorescência induzida por laser

O *Laser Induced Fluorescence System* (LIFS) é uma técnica espectroscópica de aquisição de fluorescência induzida por laser. O LIFS é composto de um laser diodo Coherent, sendo o modelo Cube com excitação em 405 nm e Potência 20 mW e um espectrômetro USB2000-Ocean Optics, que resolve na faixa espectral de 194 – 894 nm. O sistema é controlado por um computador acoplado que permite as parametrizações necessárias. Para o presente trabalho utilizou-se Tempo de integração $t_{integ} = 20ms$, Numero de médias $N = 20$ e Boxcar (parâmetro de suavização do espectro) $b = 4$. A excitação das amostras e aquisição da fluorescência são feitas através de um conjunto de fibras ópticas, ilustradas na Figura 8, com a fibra central responsável pela excitação, e as outras seis fibras adjacentes responsáveis pela captação da fluorescência e seu encaminhamento até o espectrômetro.

O padrão de aquisição dos espectros foi feito na face abaxial das folhas, posicionando a fibra ótica adjacente a nervura central na região próxima ao início da folha. Além disso, a aquisição dos espectros para os grupos sadio e assintomático foram intercalados a cada cinco folhas, a fim de evitar possíveis enviesamentos das aquisições. Por exemplo, fazendo todas as medidas de um grupo e posteriormente de outro grupo, podem ocorrer degradações indesejáveis

nas amostras durante esse intervalo e assim haveria distinção dos espectros pela degradação, e não pelas respostas fisiológicas da presença dos patógenos.

As condições de aquisição dos espectros foram padronizadas. A temperatura ambiente do laboratório foi mantida a 22 ºC, bem como a temperatura das amostras, que foram utilizadas após o equilíbrio térmico com o ambiente. O laboratório foi mantido em escuridão no momento da aquisição dos espectros.

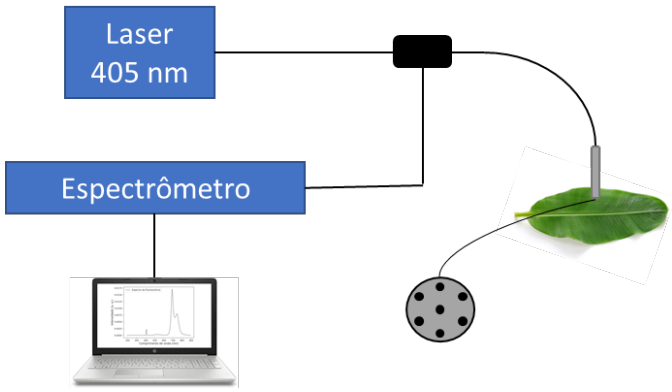

Figura 8 – Esquema do aparato experimental LIFS. Um laser diodo de 405 nm de excitação, um espectrômetro Ocean Optics USB 4000 conectados a uma ponteira de prova.
Fonte: Elaborada pelo autor

3.2.2 Imagem de fluorescência induzida por LED

O *Laser Induced Fluorescence Image* (LIFI) é um aparato de aquisição de imagens de fluorescência induzida por LEDs. O aparato proposto possui como fonte de excitação um conjunto de LEDs, que possibilita a estruturação de um sistema de imagem de fluorescência compacto e adequado para utilização no campo.

Os LEDs utilizados possuem comprimento de onda de 405 nm, mesmo valor do laser do sistema LIFS, o que torna fiel a comparação dos dois sistemas. O sistema utilizado era constituído de oito LEDs de 1W de potência elétrica. Os LEDs inicialmente foram instalados em difusores de temperatura, importantes para manter a estabilidade térmica e não prejudicar a eficiência luminosa dos LEDs. Eles foram dispostos circularmente em uma placa de acrílico, adaptada para acoplar uma câmera fotográfica digital no centro do conjunto. Em frente à câmera foi disposto um filtro passa-alta de corte em 440 nm, a fim de evitar a reflectância da fonte excitadora. Um anteparo para fixação das amostras e captura das imagens foi colocado a 30 centímetros da fonte. Para esse novo sistema LIFI, a irradiância luminosa média incidente nas

amostras foi de 0,4 mW/cm^2. O local de posicionamento das amostras possuía distribuição luminosa homogênea. A Figura 9 mostra a montagem experimental utilizada.

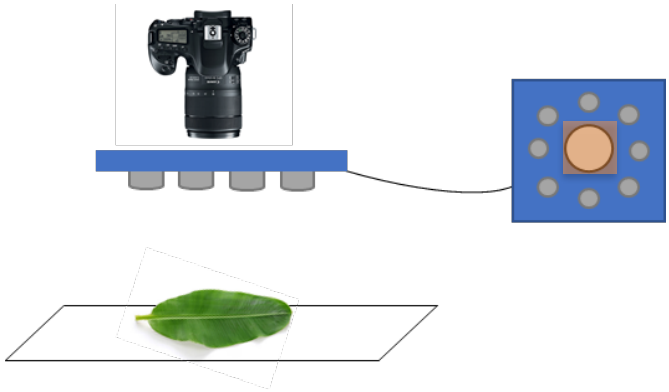

Figura 9 – Aparato experimental para obtenção de imagens de fluorescência induzida por laser (LIFI).
Fonte: Elaborada pelo autor.

A aquisição das imagens foi feita com as condições do laboratório em escuridão com temperatura estabilizada em 22 ºC. Para cada imagem obtida, a fonte de excitação era estabilizada durante 10 segundos de luz incidente, para evitar flutuações da fluorescência e garantir padronização das imagens. Os parâmetros definidos para a câmera foram: tempo de aquisição de 6 segundos, abertura ISO de 200 e distância focal de 28 mm. A focalização das imagens foi definida manualmente para cada captura.

3.2.3 Estereomicroscópio de fluorescência com excitação 365 nm (EM-365)

O equipamento EM-365 é um equipamento comercial de fins científicos. É constituído de uma lâmpada de vapor de mercúrio como fonte de excitação e um filtro passa banda alta em 450 nm. Foi utilizada a linha de emissão Hg em $\lambda_{EM\text{-}365}$ = 365 nm. O sistema de aquisição das imagens para o EM-365 é feito com o software *AxionVision* e exige calibração do sistema e definição de parâmetros referentes à imagem. O sistema passou por uma refinada calibração inicial, que foi mantida durante toda fase de aquisição de imagens. Com relação aos parâmetros para captura das imagens, destaca-se o tempo de exposição. Esse parâmetro é associado ao intervalo de tempo em que os sensores recebem a informação da fluorescência. Deve-se garantir um tempo de exposição padronizado e adequado às características do problema, pois isso afeta diretamente a sensibilidade do detector e consequentemente as características da fluorescência da imagem. Dependendo das conduções diárias do experimento, o tempo de aquisição foi

definido entre 1,7s e 2,2s. Outro fator considerável era o tempo de estabilização da fluorescência, que foi padronizado em cinco segundos, ou seja, cada amostra era excitada durante esse tempo, para depois ser feita a captura da imagem. A calibração do *software* de captura de imagens *Axio Vision* foi 0,50 *Brightness*, 1,00 *Contraste* e 1,00 *gamma*. Foram definidos também os seguintes parâmetros: ampliação de 21 vezes e campo de fundo de 11,0 mm. A resolução das imagens foi de 1290 × 968 pixels, e foram salvas no formato sem compressão TIFF. A Figura 10 traz uma imagem do sistema EM-365 utilizado no trabalho.

Figura 10 – Estereomicroscópio (EM-365) utilizado na aquisição de imagens de fluorescência.
Fonte: Elaborada pelo autor

3.2.4 Processamento de dados de espectroscopia

O espectrômetro utilizado no sistema LIFS apresenta variabilidade de 200 a 900 nm. O tratamento dos valores de fluorescência de cada comprimento de onda foi feito a partir de 400 nm, considerando o pico de refletância centrado em 405 nm. Os comprimentos de onda inferiores a 400 nm foram desconsiderados, por serem inferiores ao comprimento de onda de excitação. O ajuste pela linha de base for feito a partir da média da região entre 390 e 400 nm. Essa é uma etapa fundamental para evitar propagação de erro referente ao ruído branco inerente ao aparelho. Em seguida, cada valor de comprimento de onda foi normalizado pela área total do espectro. Os valores finais desses processos de recorte espectral, remoção de linha de base e normalização pela área foram utilizados para a elaboração de classificadores.[40]

3.2.5 Processamento de imagens digitais

A seguir, é apresentada uma revisão acerca dos conceitos utilizados no trabalho para o processamento digital de imagens. Tendo em vista a ampla gama de algoritmos e processamentos possíveis,[41] optou-se, na presente redação, por abordar apenas os conceitos utilizados diretamente a cada etapa, indicando ao leitor interessado as fontes para aprofundamento de cada tema. Cabe ressaltar que os algoritmos foram implementados na linguagem de programação Python, com a utilização de bibliotecas código aberto (*open-source*), amplamente utilizadas em trabalhos científicos, tais como *numpy*,[42] *scipy*,[43] *sklearn*,[44] *opencv*,[45] dentre outros que serão citados conforme a apresentação da metodologia.

3.2.5.1 Pré-processamento

No pré-processamento, são realizados processos de preparação para a análise das imagens.[46] Processos como recorte das imagens, localização do objeto de estudo, adequação do banco de dados, análise e conversão de formatos de imagens, resolução e qualidade das imagens, dentre outros, são realizados nessa etapa. A finalidade do processo de pré-processamento é facilitar ou viabilizar as etapas posteriores. Tais processos podem ser automatizados ou não, dependendo da necessidade de cada tipo de imagem. No presente projeto, foi realizado o recorte automatizado das imagens do EM-365, garantindo que todas as imagens seriam analisadas na mesma referência. Esse procedimento se deve ao gradiente irregular sombreado que as imagens apresentam. Cabe citar que, durante o tempo do doutorado, foram testadas duas técnicas para minimização desse efeito apresentado nas imagens do EM-365. A primeira delas foi através da correção *gamma*,[46-47] onde a imagem passa por um uma transformação não linear do tipo $f(x) = Ax^{\gamma}$, com $\gamma > 0$. Essa transformação realça as regiões mais escuras da imagem e mantem as regiões mais claras com poucas modificações. A outra foi a utilização da filtragem homomórfica,[49] Ambas as manipulações foram descartadas por não apresentarem resultados satisfatórios para o seguimento do projeto. Sendo assim, optou-se pela seleção de uma área homogênea e seu recorte, sem nenhuma manipulação de pixels, a fim de manter a padronização das análises.

O recorte inicia-se no pixel (330, 230) e termina no pixel (700, 600), resultando numa imagem com dimensão 370 × 370, conforme a Figura 11. A escolha desses pontos foi feita manualmente, de maneira a preservar a maior região possível da figura, sem que os efeitos de gradiente interferissem. Para as imagens do LIFI, não foram realizados pré-processamentos.

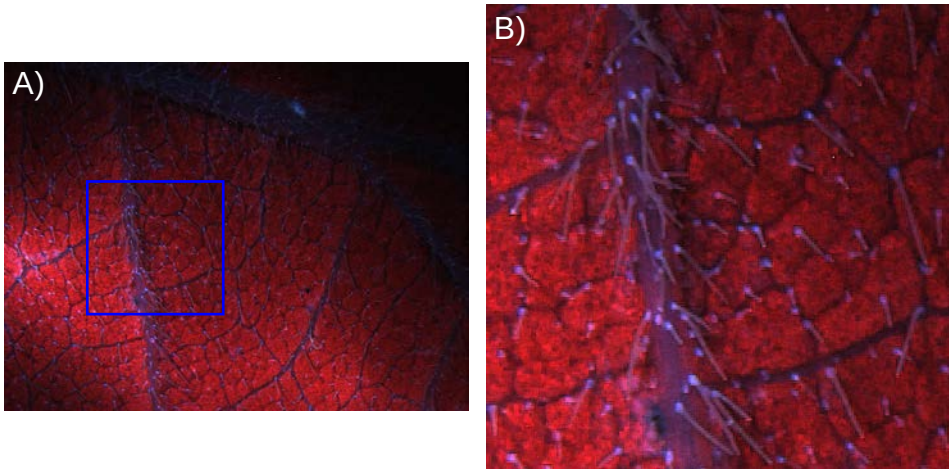

Figura 11 – A) Imagem original, de dimensão 1296 × 986 pixels. Nota-se um gradiente escurecendo regiões periféricas da imagem. B) Imagem recortada, de dimensão 370 x 370 pixels.

Fonte: Elaborada pelo autor

3.2.5.2 Segmentação

De posse da imagem pré-processada, é necessário referenciar o objeto de estudo para posteriormente ser analisado. Para tal, faz-se uso de técnicas de segmentação. A segmentação consiste na diferenciação de uma ou mais regiões de interesse (ROI) do seu respectivo fundo. É desejável que o processo de segmentação seja automatizado, pois evita interferências humanas subjetivas. A literatura reporta uma vasta gama de trabalhos voltados exclusivamente para a segmentação de imagens digitais, tanto no desenvolvimento de algoritmos quanto no objetivo final de localizar os objetos com a maior confiabilidade possível.[50]

No presente projeto, foi realizada a segmentação automatizada das imagens LIFI para a determinação da máscara da folha, utilizada como referência para toda a sessão de extração de atributos. O objetivo da determinação da máscara é eliminar da análise o fundo preto das imagens. Não foram realizadas segmentações nas imagens do EM-365. A Figura 12 apresenta as etapas da segmentação realizada nas imagens do LIFI. Nessas imagens, os pixels referentes à folha devem ser destacados, e aqueles referentes ao anteparo devem ser descartados das análises. Inicialmente a imagem original RGB foi convertida para escala de cinza e suavizada por uma filtragem gaussiana no domínio do espaço.[46,50] A seleção da máscara foi feita através da correção *gamma*, com valores da exponencial elevados. Essa transformação faz um realce dos pixels claros da imagem, destacando facilmente a forma da folha. Em seguida, é aplicado um filtro morfológico para eliminar os ruídos remanescentes da etapa anterior.[52] Finalmente, a imagem é convertida para a forma binária [0, 1] destacando a forma da região de interesse. Essa

imagem final é chamada de máscara e é utilizada como referência para eliminar o fundo da imagem.

Figura 12 – Exemplo de segmentação automática e determinação da máscara. a) Imagem de fluorescência em RGB, b) Imagem resultado da transformação em escala de cinza e suavização gaussiana e c) Máscara final, como resultado da aplicação de filtros morfológicos de eliminação de ruído.

Fonte: Elaborada pelo autor

3.2.5.3 Extração de atributos

Um problema de classificação, tal qual proposto, consiste em diferenciar com a maior confiabilidade possível, duas ou mais categorias de imagens. Conforme abordado no Capítulo 1, o objetivo principal do presente trabalho é elaborar um protocolo para reconhecimento de duas doenças de soja com relação a plantas sadias. Para tal, é necessário que os grupos apresentem características qualitativa e quantitativamente diferenciáveis entre si. A obtenção e análise dessas características são abordadas na etapa de extração e seleção de atributos.

A extração de atributos consiste em realizar operações com os pixels da imagem, a fim de obter um conjunto de informações significativo dessas.[52-53] Tais informações, comparadas grupo a grupo, podem revelar as características de cada tipo de imagem, possibilitando a posterior classificação delas. A classificação adequada dos grupos dependerá da qualidade dos atributos e da capacidade de discriminação deles. A literatura reporta uma vasta diversidade com relação às possibilidades de extração de atributos, variando dos atributos mais simples até os com maior grau de complexidade.[54-57]

No desenvolvimento do presente trabalho, foram implementados diferentes tipos de atributos. A escolha do conjunto de atributos que compõe o trabalho foi feita pelo autor, baseado na capacidade discriminatória de cada atributo e na sua relevância para o problema. Na sequência, serão detalhados os atributos selecionados para comporem a matriz de dados na tarefa de classificação. No total, foram escolhidos 146 atributos para nessa etapa, que fazem referência à informação de textura e coloração presente nas imagens de fluorescência.

3.2.5.4 Atributos estatísticos

Com o intuito de explorar a intensidade dos valores de pixels dos canais de cores das imagens, foram estipulados atributos que fazem referência à distribuição dos valores de pixels na imagem. Sendo uma imagem $I(i,j)$ de dimensão $(m \times n)$ composta por $x_i = 1, 2, \dots, N$ pixels, os atributos estatísticos utilizados foram os seguintes[51]:

- Média dos valores dos pixels ($\bar{x}$)

$$\bar{x} = \frac{1}{N}\sum_i x_i \tag{3}$$

- Variância (σ) e Desvio Padrão (s)

$$s = \sigma^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 \tag{4}$$

- *Skewness*

$$skewness = \sum_i \frac{(x_i - \bar{x})^3}{\sigma^3} \tag{5}$$

- *Kurtosis*

$$kurtosis = \sum_i \frac{(x_i - \bar{x})^4}{\sigma^4} \tag{6}$$

- Entropia

$$entropia = -\sum_i p_i . \log(p_i) \tag{7}$$

onde p_i é a probabilidade de ocorrência do pixel i.

Tendo explorado as intensidades dos pixels de uma imagem, cabe agora extrair informações com relação à vizinhança deles. Para tal, foram implementados e testados atributos de textura. Existem três abordagens clássicas para a descrição da textura de uma imagem: abordagens *estatísticas*, definida por um conjunto de métricas locais; abordagens *estruturais*, baseadas na composição de estruturas periódicas na imagem; e por fim, abordagens *espectrais*, baseados em propriedades de espectros de *Fourier* e *Wavelet* e outras transformações.[58-60] No presente

trabalho, foi utilizada uma abordagem estatística através dos atributos de Haralick. Os atributos de textura de Haralick foram propostos inicialmente em 1979, e até hoje são amplamente utilizados em problemas de classificação. Esses atributos são obtidos a partir da matriz de co-ocorrência $P(i,j,d,\theta)$, onde i e j são pixels vizinhos separados por uma distância d. Essa distância d é analizada através de uma direção θ, que pode ser $[0°, 45°, 90°,\ 135°]$. Foram utilizados para determinação da matriz de co-ocorrência, os parâmetros $d = 1$ e $\theta = 0°$. A teoria proposta por Haralick destaca 14 atributos extraídos da matriz de co-ocorrência. Destes, foram utilizados os seguintes:

- Contraste

$$contraste = \sum_i \sum_j (i-j)^2 \, P(i,j) \tag{8}$$

- Dissimilaridade

$$dissimilaridade = \sum_i \sum_j |i-j| P(i,j) \tag{9}$$

- Homogeneidade

$$homogeneidade = \sum_i \sum_j \left(\frac{1}{1 + \ (i-j)^2}\right) P(i,j) \tag{10}$$

- ASM e Energia

$$ASM = \ energia^2 = \sum_i \sum_j P(i,j)^2 \tag{11}$$

- Correlação

$$correlação = \ \sum_i \sum_j P(i,j) \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y} \tag{12}$$

onde $\mu_x;\ \mu_y$ e $\sigma_x;\ \sigma_y$ são a média e o desvio padrão das imagens i e j, respectivamente.

Além de uma abordagem *estatística*, baseada nos atributos de Haralick, foi utilizada também uma abordagem *estrutural* e *estatística* ao mesmo tempo, podendo ser considerada uma abordagem unificadora: os descritores de padrão local binário (LBP – do inglês *Local*

Binary Pattern). Inspirado num modelo proposto por Wang & He,[62] o algoritmo LBP divide a imagem toda em células de dimensão pré-definida, e aplica uma operação que gera um valor para cada célula, que serão utilizados para a determinação dos atributos LBP. O método LBP proposto por Ojala[63] difere do original em algumas etapas, tornando sua execução mais eficiente. O algoritmo trata uma textura T em uma vizinhança, onde g_c corresponde ao valor do nível de cinza do pixel central, P é a quantidade de pontos ao redor do pixel central, g_p são os valores dos P pixels espaçados por um círculo de raio R. A Figura 13 mostra um conjunto de três configurações distintas do algoritmo LBP.

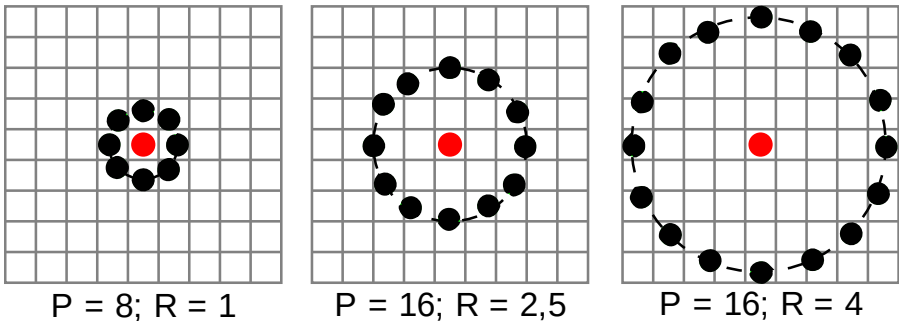

Figura 13 – Exemplo de conjuntos (P, R) para o algoritmo LBP.
Fonte: Adaptada de LOCAL... [64]

O algoritmo LBP faz uma análise de cada célula pré-definida dentro da imagem, chamada código LBP, onde inicialmente compara-se o valor do pixel central com a vizinhança, atribuindo 0 para valores menores que o pixel central e 1 para valores de vizinhança maiores. Feito esse limiar, multiplica-se uma máscara ao redor da vizinhança de 2^n e efetua-se a soma resultante das multiplicações por 0 e 1. A equação que traduz essa operação para cada célula apresenta-se:

$$LBP(P,R) = \sum_{P=1}^{P-1} s(g_P - g_c)2^P$$
$$s = \begin{cases} 1, & se\ x \geq 0 \\ 0, & se\ x < 0 \end{cases} \qquad (13)$$

De posse dos códigos LPB celulares, são calculados os atributos LBP pelo histograma de frequência relativa dos valores que, nesse caso, gera um total de $2^8 = 256$ possibilidades de textura.

No presente trabalho, foram testadas diferentes configurações de P e R que melhor retratassem o problema, e elegeu-se para utilização no trabalho a parametrização $P = 24$ e $R = 8$. Os atributos LBP utilizados foram resultantes do cálculo de histograma para 26 janelas.

3.2.5.5 Canais de cores

Um modelo de cores é, basicamente, a determinação de um sistema de coordenadas e um subespaço adequado a esse sistema, na qual cada cor é bem definida em um ponto único. No processamento de imagens, o modelo de cor mais utilizado na prática é o RGB (Red, Green, Blue – vermelho, verde e azul). A configuração RGB é padrão para monitores e vídeos utilizados no dia-a-dia, e é a maneira mais tradicional de representar uma imagem.[47] Espacialmente, pode-se entender o espaço de cor RGB em geometria cúbica com Preto e Branco ocupando vértices opostos, e as cores primárias Vermelho, Verde e Azul como as dimensões dos eixos do cubo representativos, de acordo com a Figura 14. Essa disposição gera também as cores Magenta, Ciano e Amarelo como uma combinação de valores dos canais principais.[65] Na Figura 15 é apresentada uma decomposição dos canais de cores em questão.

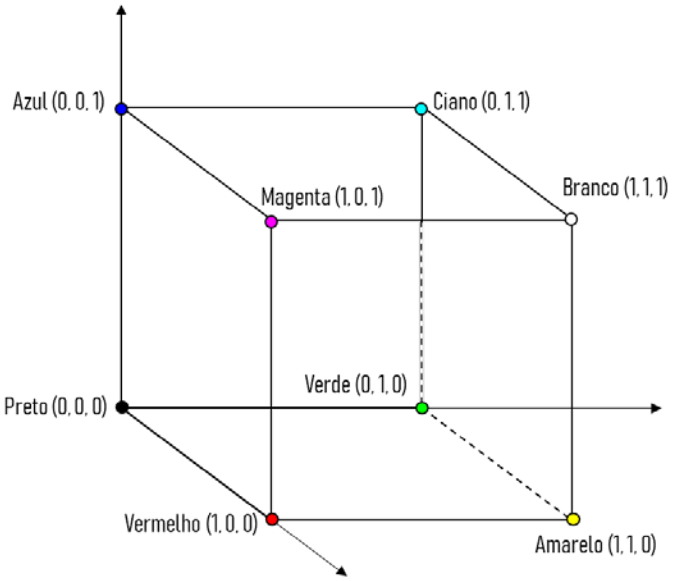

Figura 14 – Espaço de cor RGB.
Fonte: Elaborada pelo autor.

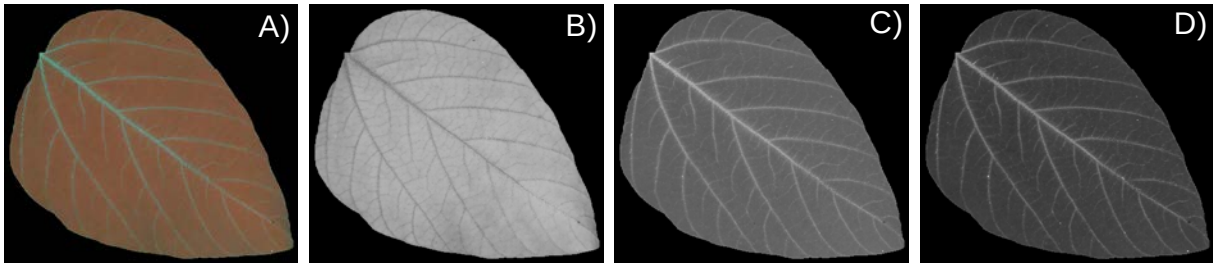

Figura 15 – Exemplo de decomposição nos canais de cores RGB. A) imagem RGB original, B) Imagem referente ao canal de cor Vermelho (R), C) Imagem referente ao canal de cor Verde (G) e D) Imagem referente ao canal de cor Azul (B).
Fonte: Elaborada pelo autor.

Outro modelo de cor utilizado é o HSV (*hue*, *saturation* e *value* ou *intensity*): matiz, saturação e intensidade. Essa representação é bastante adequada à maneira como a visão humana é interpretada, através da matiz de cores, da sua saturação e brilho.[65] Matiz é a métrica que trata de uma cor pura. Já a saturação faz referência à diluição da cor pura pela luz branca e o brilho representa a intensidade. O espaço de cor HSV é definido geometricamente conforme a Figura 16.

A conversão do espaço de cor RGB para HSV é uma manipulação amplamente empregada em técnicas de análise de imagens, pois as transformações podem realçar informações relevantes na discriminação de duas classes. As equações de conversão do espaço RGB cúbico para o espaço HSV cônico é dado por: [65]

$$H = \begin{cases} \theta, & se\ B \leq G \\ 360 - \theta, & se\ B > G \end{cases} \tag{14}$$

onde

$$\theta = \cos^{-1}\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)+(R-B)(G-B)]^{1/2}}\right\} \tag{15}$$

$$S = 1 - \frac{3}{(R+G+B)}[\min(R,G,B)] \tag{16}$$

$$I = \frac{1}{3}(R+G+B) \tag{17}$$

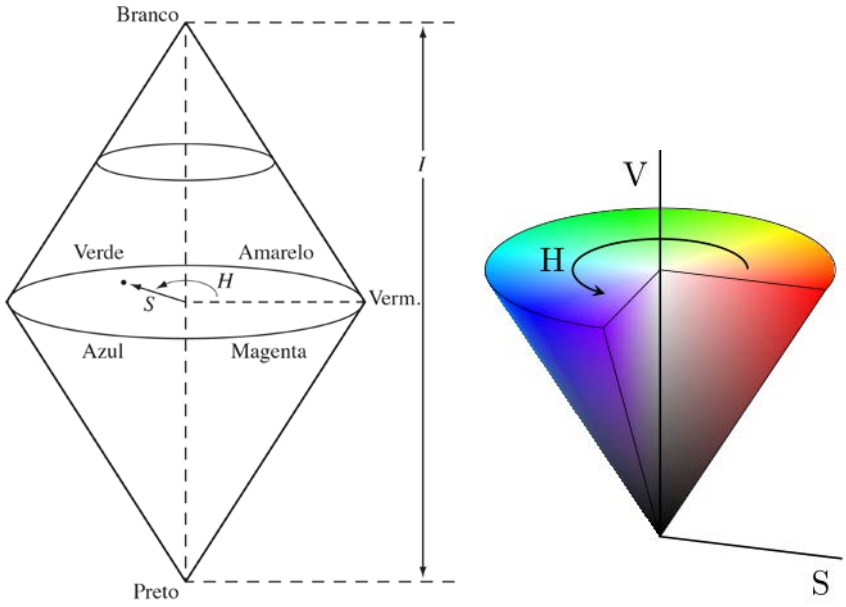

Figura 16 – Espaço de cor HSV.
Fonte: Elaborada pelo autor.

Na Figura 17 é apresentada uma decomposição dos canais de cores em questão.

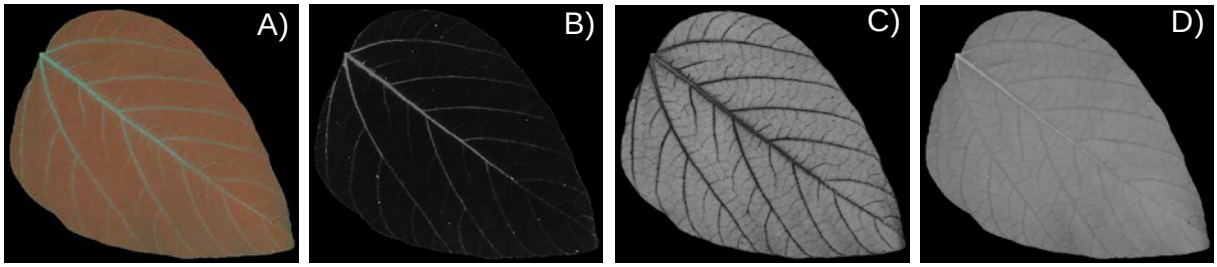

Figura 17 – Exemplo de decomposição nos canais de cores HSV. A) imagem RGB original, B) Imagem referente ao canal da Matriz (H), C) Imagem referente ao canal da Saturação (S) e D) Imagem referente ao canal da Intensidade (V).
Fonte: Elaborada pelo autor.

Por fim, foi utilizada mais uma transformação não linear do espaço de cor RGB, o La*b*. Assim como o HVS, o espaço de cor em questão consegue separar muito bem a informação de intensidade, caracterizada pelo canal *L*, das informações de cores, definida nos canais a* e b*. O modelo utilizado em muitos *sistemas de gerenciamento de cores* (SGC) é o modelo CIE *L*a*b**, *também* chamado de Cielab.[66] Os componentes de cor *L*a*b** são determinados pelas equações a seguir: [67]

$$L = 116.h\left(\frac{Y}{Y_W}\right) - 16 \qquad (18)$$

$$a^* = 500\left[\left(\frac{X}{X_W}\right) - h\left(\frac{Y}{Y_W}\right)\right] \quad (19)$$

$$b^* = 200\left[h\left(\frac{Y}{Y_W}\right) - \left(\frac{Z}{Z_W}\right)\right] \quad (20)$$

sendo

$$h(q) = \begin{cases} \sqrt[3]{q}, & se\ q > 0,008856 \\ 7{,}787q + \dfrac{16}{116}, & se\ q \leq 0,008856 \end{cases} \quad (21)$$

e X_W, Y_W e Z_W são valores de referência do tri estímulo branco – normalmente o branco de um difusor de reflexão perfeita no padrão CIE de iluminação D65. Na Figura 18 é apresentada uma decomposição dos canais de cores em questão.

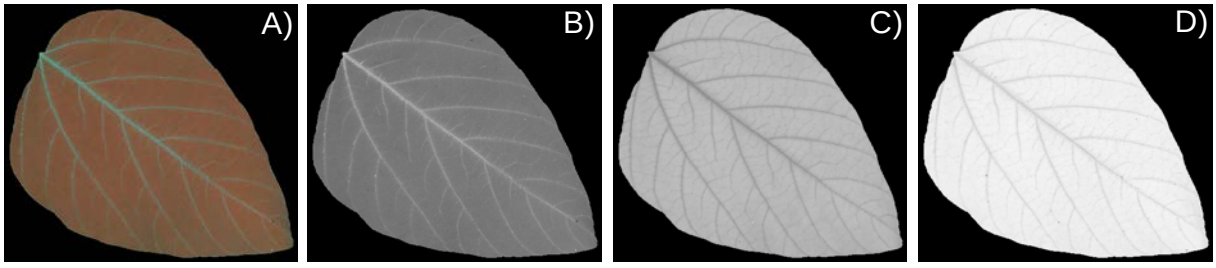

Figura 18 – Exemplo de decomposição nos canais de cores La*b*. A) imagem RGB original, B) Imagem referente ao canal (L), C) Imagem referente ao canal (a*) e D) Imagem referente ao canal (b*).
Fonte: Elaborado pelo autor.

3.2.5.6 Cor dominante

Até o presente momento, foram discutidos conceitos já amplamente utilizados e bem estabelecidos na literatura, com relação à extração de atributos para tarefas de classificação. Neste tópico será apresentado o conceito de cor dominante empregado nas imagens de fluorescência. Esses atributos foram propostos com base nas na evolução da análise das amostras e supre uma necessidade observada de avaliar tonalidades características das imagens de fluorescência e a sua frequência relativa de aparição. Basicamente, é aplicado um algoritmo de agrupamento (*cluster*) aos pixels da imagem em três tonalidades e a frequência relativa dessas cores nas folhas. O conceito de cor dominante é uma nova maneira de avaliar a intensidade dos pixels, além dos atributos estatísticos já citados. O método de agrupamento utilizado foi o clássico *k-Means*. O resultado da determinação das cores dominantes de uma imagem de fluorescência é apresentado na. Figura 19. As imagens foram agrupadas em três cores principais, e as

intensidades dos canais de cor RGB, bem como a frequência relativa de aparição de cada cor na área foliar, foram utilizadas como atributo na matriz de dados.

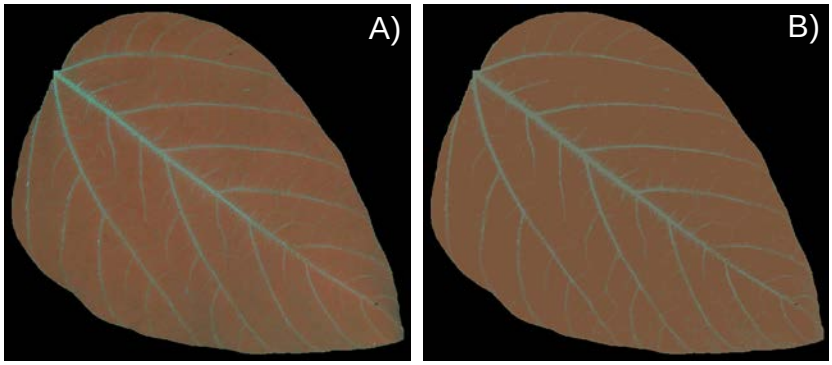

Figura 19 – Obtenção dos atributos de cor dominante. Os pixels da imagem original sofrem um agrupamento em três principais cores, e suas intensidades RGB e sua frequência relativa na área da folha são utilizados como atributos.
Fonte: Elaborada pelo autor.

Os atributos de imagens, utilizados no trabalho, podem ser sumarizados da seguinte maneira: três espaços de cores RGB, HSV e La*b*, com três canais cada. De cada canal, foram extraídos seis atributos estatísticos de intensidade – média, desvio padrão, variância, skewness, kurtosis e entropia – e mais seis atributos estatísticos de textura – contraste, dissimilaridade, homogeneidade, ASM, energia e correlação; 26 atributos de textura LBP e por fim; três cores dominantes com 4 atributos cada, totalizando 12 atributos. No total, são 146 atributos relacionados à intensidade dos canais de cor e textura.

3.2.6 Seleção de atributos

A seleção de atributos é, essencialmente, o processo de escolha dos atributos mais relevantes para a classificação dos grupos. É uma importante etapa para a análise dos dados que contém número de atributos excessivos. Em geral, problemas envolvendo grande quantidade de atributos costumam apresentar problemas de dimensionalidade, como por exemplo aumentar significativamente o tempo de execução dos algoritmos, tornando o modelo extremamente complexo e levando ao clássico problema de *overfitting*, quando o modelo reproduz perfeitamente o conjunto de treino, mas não é capaz de reproduzir os resultados para amostras de teste. Em suma, a seleção de atributos permite execução mais rápida dos algoritmos, reduz a complexidade do problema tornando-o mais fácil para interpretar, e como consequência, melhora a as taxas de acerto do classificador. [68-69]

Existem três abordagens para os algoritmos de seleção de atributos: filtro, *wrapper* e *embedded*.[70] A abordagem filtro utiliza um algoritmo de aprendizado externo para fazer uma filtragem dos atributos com base em critérios de avaliação exatos, tais como distância entre esses atributos, informação, dependência ou consistência. Essa abordagem é comumente utilizada para fazer um pré-processamento dos dados pois, por utilizarem critérios exatos, são independentes de algoritmos de classificação. No presente trabalho, foram utilizados dois critérios para seleção dos atributos: o coeficiente de correlação de Pearson e o teste Chi-Square.

Na abordagem *wrapper*, é necessário um algoritmo de classificação para a seleção de atributos, fazendo uso do critério de avaliação deste para eleger o melhor desempenho. O algoritmo faz uma busca pelo melhor conjunto de atributos baseado nos resultados de classificação, o que garante eficiência do método. Alguns exemplos de utilização da abordagem *wrapper* englobam a) seleção direta, b) eliminação direta, e até mesmo c) combinação entre eliminação e seleção a cada iteração. No presente trabalho, utilizou-se a abordagem Eliminação Recursiva de Atributos (*Recursive Feature Elimination* – RFE), que faz uma pesquisa exaustiva para encontrar o subconjunto de atributos com o melhor desempenho. O algoritmo de classificação utilizado foi o Regressão Logística.[71]

Por fim, além das duas abordagens filtro e de uma *wrapper*, foram utilizadas outras três metodologias da abordagem *embedded*, que assim como a *wrapper*, utiliza algoritmos de classificação e sua respectiva métrica de desempenho para seleção do melhor subconjunto de atributos. A diferença da abordagem *embedded* está na utilização de métodos de regularização, também conhecidos por métodos de penalização, que imputam uma penalidade a cada atributo testado. Isso restringe a otimização dos algoritmos de classificação, tornando-os menos complexos. No presente trabalho, optou-se pela utilização dos seguintes algoritmos para a metodologia *embedded*: Regressão Logística,[71] *Random Forest*[72]e *LigthBGM*,[73] , sobre as quais deixaremos, para o leitor interessado, a opção de consultar as referências para aprofundamento.

3.2.7 Redução de dimensionalidade (PCA e normalização *z*-score)

Conforme apresentado no tópico anterior, a Seleção de Atributos visa reduzir a complexidade do problema através da exclusão de atributos de baixa relevância para o problema. Nessa mesma temática, existem técnicas de redução de dimensionalidade que buscam alterar a representação de um conjunto de dados, de maneira que suas principais componentes representem a informação inerente com o mínimo de perda de informação possível, mas com uma dimensão menor do que a matriz original.

A técnica de redução de dimensionalidade empregada foi a Análise de Componentes Principais (*Principal Component Analysis* – PCA). A PCA é provavelmente o método de redução de dados mais utilizado atualmente e foi proposto por Pearson em 1901.[74] De maneira intuitiva, a PCA produz um novo sistema de coordenadas, de dimensão ($n \times n$) a partir dos dados originais de uma matriz ($p \times n$) de p variáveis e n atributos. Esse novo sistema de coordenadas é chamado de Componentes Principais, e tem como característica fundamental o acúmulo da maior parcela da variância dos dados em poucas componentes, sendo a primeira componente responsável pela maior porcentagem da variância, a segunda componente correspondendo à segunda maior variância e assim por diante.[75-76]

Teoricamente, a ordenação das variâncias é obtida a partir do cálculo da matriz de correlação dos dados originais. Dessa matriz, são extraídos os autovetores, que correspondem à variância dos dados acumulada em seu respectivo autovetor. Para o cálculo da PCA, a primeira operação necessária é a padronização dos dados, a fim de evitar que as altas variâncias dominem as componentes principais, pois, quanto maior a ordem de grandeza dos dados, proporcional será sua variância. A padronização dos dados é feita através da relação:

$$x_{norm} = \frac{x - \bar{x}}{\sigma_x} \tag{22}$$

onde $\bar{x}$ e σ_x correspondem às médias do atributo e o desvio padrão, respectivamente. Em seguida é calculada a matriz de correlação (r_{xy}) para cada par de atributos, através da equação:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \tag{23}$$

onde x e y são atributos da matriz de dados e n o número de amostras. A correlação indica o que ocorre com a variação de um atributo com relação a outro, ou seja, uma correlação positiva indica a mesma direção de variação, enquanto uma correlação negativa implica em variações opostas. Por fim, os autovalores e autovetores da matriz r_{xy} são determinados através da resolução da equação:

$$r_{xy} - \lambda I = 0 \tag{24}$$

onde I é uma matriz quadrada e λ, os autovalores. Os autovalores são ordenados por seu valor e as projeções das componentes principais definidas pela multiplicação da matriz original pelos

autovetores ordenados. Dessa maneira, a primeira componente principal representará a maior variância dos dados, a segunda componente principal, a segunda maior variância, e assim sucessivamente. Em geral, na avaliação de uma projeção com duas classes, busca-se a componente principal que melhor segregue os grupos, e então se avalia quais atributos mais contribuem para formação das Componentes Principais. A avaliação dos pesos de cada componente principal é dada pelo parâmetro *loadings*, que são obtidos diretamente como resultado da decomposição da matriz original $X(n \times p)$dos dados para k componentes principais. O resultado dessa decomposição resulta na matriz de *scores* U e na matriz de *loadings* V^T. A matriz de *scores* é o resultado da projeção das n amostras na direção dos autovetores das k componentes principais e a matriz dos *loadings*, que traz a contribuição relativa de cada p atributo para direção das componentes principais. A Figura 20 apresenta uma representação desse processo.

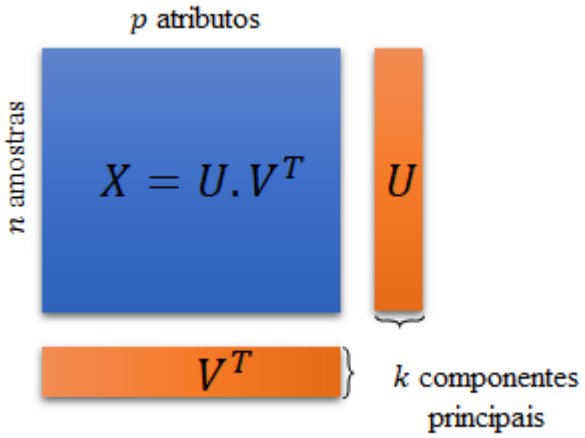

Figura 20 – Representação da decomposição em k componentes principais na matriz de *scores* e na matriz de *loadings*.
Fonte: Elaborada pelo autor

3.2.8 Classificação e validação cruzada

Um processo de classificação binária envolve duas referências: a matriz de dados X, onde costuma-se associar as linhas da matriz às p amostras, e as colunas correspondem aos n atributos. Nessa mesma matriz, estão presentes atributos das duas ou mais categorias a serem preditas. A referência das classes é comumente associada a um vetor $y_i \subset \{1, 0\}$, onde $y_i = 1$ e $y_i = 0$ representam as classes. O processo de classificação é baseado no treinamento e parametrização dos algoritmos para um grupo de treino e posterior predição de um grupo de teste.

Chama-se de validação cruzada à etapa de treinamento e ajuste de parâmetros dos classificadores. Existem três abordagens principais para a validação cruzada.[71,77] O *LeaveOneOut*,

o *Hold-out* e o *k-fold*, sendo esse último o mais robusto dos três. Esse método faz a separação do conjunto de dados em *k* grupos e realiza o treino e teste dos dados *k* vezes. A cada treinamento, seleciona-se um grupo de teste e os demais *k-1* grupos diferentes são utilizados como grupo de treino. Os demais métodos são casos particulares do *k-fold*. Por exemplo, se o número de *k* grupos for igual ao número de amostras, então tem-se o *LeaveOneOut*, ao passo que para o caso específico k = 1 , tem-se o *Hold-out*.

Já o processo de validação do conjunto de teste, aqui chamado de predição, é a técnica mais confiável para avaliar se o comportamento de um classificador é estável e não apresenta *overfitting*. A predição é feita dividindo a matriz de dados *X* e o vetor *y* geralmente na proporção de 70% treino para 30 % teste. Diferentemente da validação cruzada, o grupo de teste não é utilizado em nenhum momento para treinamento dos modelos. Com o grupo de treino, é realizada uma validação cruzada para treinamento e parametrização do modelo, e posterior classificação do grupo de teste. Em geral, espera-se que não haja discrepância entre os resultados da validação cruzada do grupo de treino e a predição do grupo de teste. A indicação desses resultados é a melhor estratégia para avaliação de *overfitting*.

Com relação à escolha dos algoritmos utilizados na predição e validação cruzada, a estratégia inicial foi buscar na literatura os principais algoritmos utilizados para tarefas de classificação. Nesse viés, os métodos *NaiveBayes*, Árvores de Decisão, Regressão Logística, k *Nearest Neighbor* (kNN), Redes Neurais e *Support Vector Machine* (SVM) foram tomados como base para treinamento.[71,75,78-79] A avaliação dos algoritmos de classificação é exaustiva, pois a maioria dos métodos atuais necessitam de ajustes paramétricos. Uma parametrização fora das necessidades do conjunto, ou escolhida de forma aleatória, apresentará resultados de classificação imprecisos. Sendo assim, foi utilizado o algoritmo *GridSerachCV* da biblioteca *scikit-learn*,[44] que faz uma busca exaustiva do melhor conjunto de parâmetros para cada caso. Baseado nas respostas dessa busca exaustiva e nas características biológicas e físicas do problema, optou-se pela utilização do algoritmo *Support Vector Machine* (SVM).

O algoritmo SVM, de maneira global, busca maximizar a distinção dos grupos através da minimização do parâmetro de erro atribuído ao hiperplano responsável pela separação. Essa abordagem é adequada ao presente problema, pois a matriz de dados utilizada apresenta caráter biológico de evolução de assintomáticas, e essa característica gera classes interseccionadas e consequentemente, uma dificuldade para classificar corretamente o ponto. Sabendo do erro inerente ao problema, espera-se que o classificado seja eficiente na minimização desse erro, garantindo estabilidade do resultado.

3.2.9 Matriz de confusão

Os resultados de predição dos grupos de teste são realizados através da validação cruzada, que utiliza o modelo treinado para predizer coletas diferentes. A avaliação do resultado da validação cruzada é expressa na forma de uma matriz de confusão, onde acertos e erros de ambas as classes são contabilizados. A figura 8 traz um modelo da apresentação de uma matriz de confusão e sua terminologia.[77]

As siglas apresentadas na Figura 21, comumente expressas em porcentagens, representam respectivamente: VP (taxa de verdadeiro positivo), a quantidade de amostras da classe Doente classificadas corretamente, enquanto FP (taxa de falso positivo), a quantidade da classe Doente classificadas erroneamente. As siglas FN (taxa de falso negativo) e VN (taxa de verdadeiro negativo) seguem o mesmo padrão para a classe Sadia.

		PREDITO	
		P(Sadia)	P(Doente)
VERDADEIRO	Sadia	VN	FN
	Doente	FP	VP

Figura 21 – Exemplo do resultado de uma validação cruzada e sua terminologia.
Fonte: Elaborada pelo autor.

A avaliação da qualidade da tarefa de classificação é feita através de métricas de desempenho. No presente trabalho, foram utilizadas três métricas: acurácia, sensibilidade e especificidade. A acurácia é medida através da soma dos acertos do classificador, ou seja, a diagonal principal da matriz de confusão, dividido pelo número total de testes. A equação para o cálculo de acurácia é apresentada abaixo.

$$A = \frac{VP + VN}{VN + FP + FN + VN} \tag{25}$$

Já os termos sensibilidade e especificidade referem-se à taxa de acerto para cada classe. A sensibilidade trata da capacidade de diagnóstico da classe contaminada, que no presente trabalho serão Assintomáticas HVRF e FA. Em contrapartida, a especificidade trata da capacidade

de diagnóstico da classe Sadia, que são obtidos do grupo Controle. A equação para o cálculo da sensibilidade (S) e da especificidade (E) é dado por:

$$S = \frac{VP}{VP + FN} \tag{26}$$

$$E = \frac{VN}{VN + FP} \tag{27}$$

3.2.10 Fluxograma de análise

Por fim, a estrutura de análise para o projeto pode ser visualizada na Figura 22.

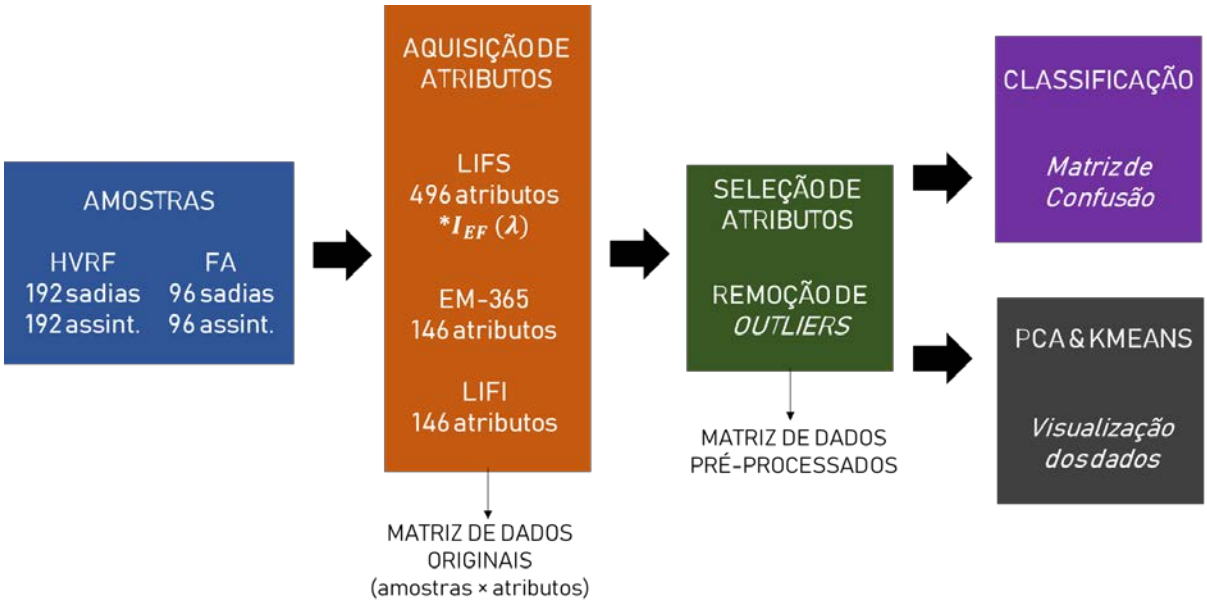

Figura 22 – Fluxograma de análise.

Fonte: Elaborada pelo autor

O fluxo da análise é feita para as 384 amostras referentes a HVRF e também para as amostras FA. Na aquisição de atributos, são extraídos 496 atributos da instrumentação LIFS e mais 146 atributos de imagem tanto para o LIFI quanto para o EM-365. Na sequência, são extraídos os atributos relevantes e excluídos outliers. Por fim, essa matriz pré-processada gera a visualização dos dados e as matrizes de confusão.

4 RESULTADOS E DISCUSSÕES

A seguir serão expostos os principais resultados obtidos dos processos de análise dos sinais de fluorescência e de classificação.

4.1 HASTE VERDE E RETENÇÃO FOLIAR: LIFS

O presente capítulo apresenta a análise dos espectros obtidos pela instrumentação LIFS. O objetivo dessa etapa é avaliar o comportamento e a capacidade de discriminação dos espectros de fluorescência entre as classes Sadia e Assintomática HVRF.

Após os processos de remoção de linha de base e normalização, foram considerados como atributos os comprimentos de onda superiores a 400 nm. Essa operação resultou em 476 comprimentos de onda compreendidos entre 400 e 890 nm. Além disso, a instrumentação LIFS também permite avaliar as intensidades das bandas de emissão F520, F690 e F740, que fazem referência à emissão de fluorescência de moléculas cruciais ao funcionamento metabólico das plantas. A área da banda de emissão F520 (área compreendida entre 478 e 583 nm) está relacionada diretamente à intensidade de fluorescência dos metabólitos de defesa presentes na planta, enquanto as bandas F690 (faixa espectral 690 ± 5 nm) e F740 (faixa espectral 744 ± 5 nm), estão associadas à intensidade de emissão das moléculas localizadas nos cloroplastos – Ch-a, Ch-b e carotenoides –, sendo a relação F690 / F740 um indicativo quanto à taxa metabólica da planta.[24-25] A escolha desses pontos baseou-se nas características do espectro resultante da variedade de soja em questão, em que os centros das faixas espectrais são pontos de máximos locais. Há de se salientar que os valores máximos diferem de maneira sutil entre plantas e até mesmo entre outras variedades de cultivares, pois o espectro é composto da emissão de várias moléculas, e suas proporções alteram a forma e centralidade dos picos.

De posse dos atributos para as coletas 7 dai, 10 dai, 21 dai, 24 dai e 28 dai, foi realizada uma exclusão de *outliers* para cada período. Importante salientar que o presente trabalho trata de uma evolução temporal, e não convém fazer a exclusão dos pontos divergentes levando em consideração todas as coletas, pois elas diferem em parâmetros essenciais, como por exemplo, nas quantidades de clorofila de cada folha. Por isso, há de se implementar a exclusão de *outliers* coleta a coleta. Foram exploradas técnicas de remoção de *outliers* multivariada presentes na biblioteca em Python chamada PyOD,[80] que apresenta uma vasta gama de algoritmos de remoção multivariada. A utilização da biblioteca PyOD requer definição da fração de outliers para a exclusão. A experiência mostrou que quanto maior essa fração, mais precisos são os resultados

de classificação. Porém, limitar a quantidade de amostras torna o algoritmo pouco robusto. Por isso, baseado no conhecimento do tratamento prévio dos dados, optou-se por limitar a quantidade de *outliers* a 10% por coleta para a instrumentação LIFS. Sendo assim, para a classe Sadia, foram detectados 18 das 160 amostras, enquanto que para a classe Assintomática HVRF, foram detectados 18 das 160 amostras, totalizando aproximadamente 11,25% de amostras excluídas.

Na sequência, a Figura 23 sumariza a média e os desvios padrões da intensidade da banda F520 para cada coleta. Finalmente, a Tabela 2 dispõe os valores numéricos das médias, bem como o resultado do teste de hipótese para diferenciação das médias para cada coleta.

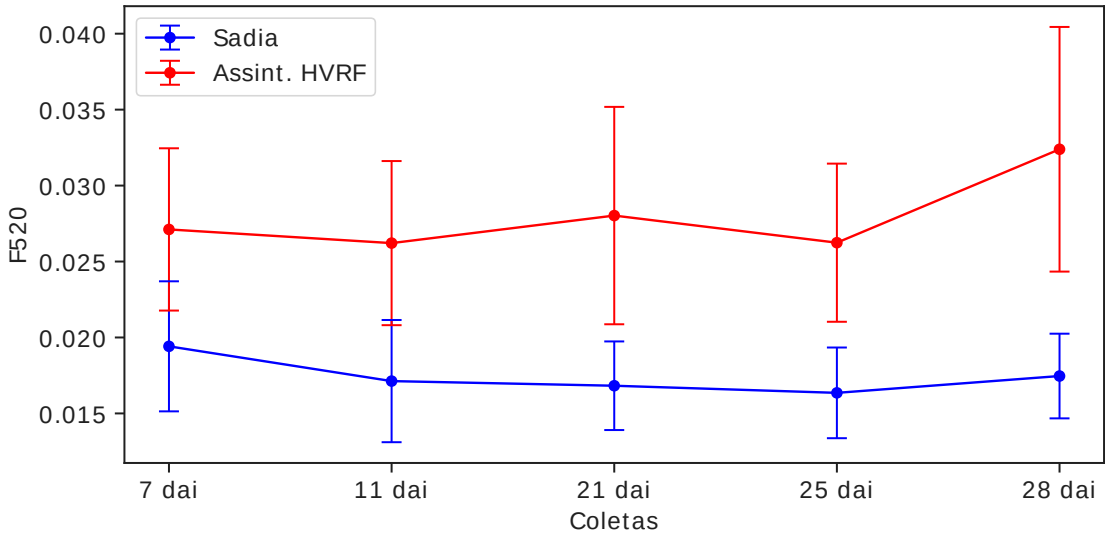

Figura 23 – Evolução dos valores médios da área F520 referente às classes Sadia e Assintomática HVRF e o respectivo desvio padrão para cada coleta.
Fonte: Elaborada pelo autor.

Tabela 2 – Valores médios da área F520 referente às classes Sadia e Assintomática HVRF e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias.

F520	7 dai	11 dai	21 dai	25 dai	28 dai
Sadia	0,018 ± 0,004	0,017 ± 0,003	0,017 ± 0,003	0,016 ± 0,003	0,018 ± 0,003
Assint. HVRF	0,026 ± 0,004	0,025 ± 0,004	0,027 ± 0,006	0,026 ± 0,006	0,031 ± 0,005
p-valor	< 0,05	< 0,05	< 0,05	< 0,05	< 0,05

Fonte: Elaborada pelo autor.

A banda F520, região de emissão dos metabólitos secundários, apresentou resultado consistente, onde era esperada uma constância nas médias para a classe sadia e um aumento gradual de intensidade para a classe Assintomática HVRF, devido a emissões de substâncias defesa da planta, como os já citados flavonoides. Também se destaca a diferenciação

significativa das médias em todas as coletas, indicando que a banda em F520, logo nos primeiros dias de infestação, pode revelar a presença dos patógenos através da análise dessa banda.

Na sequência, a Figura 24 mostra a disposição dos valores médios da relação das áreas F690 / F740 para cada coleta. A Tabela 3 apresenta os valores médios e o desvio padrão da referida relação para cada classe, bem como o resultado do p-valor para comparação da média.

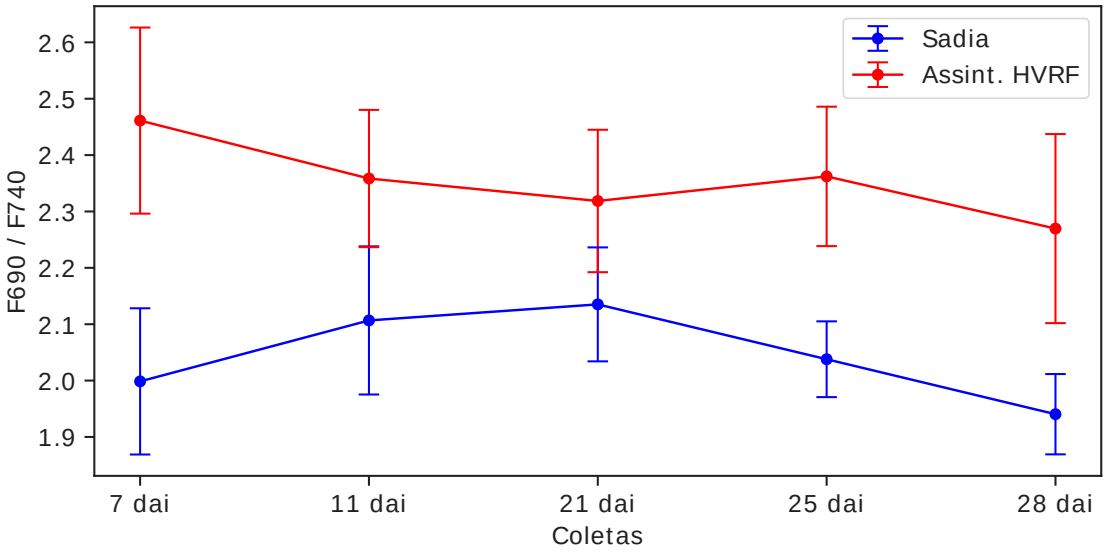

Figura 24 – Evolução dos valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF para o sistema LIFS, e o respectivo desvio padrão para cada coleta.
Fonte: Elaborada pelo autor.

Tabela 3 – Valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF, e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias.

F690 / F740	**7 dai**	**11 dai**	**21 dai**	**25 dai**	**28 dai**
Sadia	2,0 ± 0,1	2,1 ± 0,1	2,1 ± 0,1	2,0 ± 0,1	1,9 ± 0,1
Assint. HVRF	2,4 ± 0,1	2,3 ± 0,1	2,3 ± 0,1	2,3 ± 0,1	2,3 ± 0,1
p-valor	< 0,05	< 0,05	< 0,05	< 0,05	< 0,05

Fonte: Elaborada pelo autor.

Tal qual ocorrido para a banda F520, houve diferenciação significativa para todas as coletas para a relação das bandas F690 / F740, indicando que a referida relação pode também revelar a presença de patógenos nos primeiros dias de infestação. Destaca-se a diferenciação elevada da coleta 7 dai, onde esperava-se que a diferença dos valores médios fosse menor, ou no mínimo, que as coletas seguintes acompanhassem essa proporção. Tal comportamento pode estar associado, dentre outros fatores, a adversidades na realização do experimento e na

aquisição das amostras, tendo em vista todas as particularidades já citadas para a realização do ensaio. Cabe mais uma vez argumentar a dificuldade de se trabalhar com materiais biológicos *in vivo* e a imprevisibilidade de fatores que alteram as condições de uma coleta para outra. Vale lembrar que, das dez coletas programadas no ensaio, apenas metade foi utilizada, pois durante as coletas houve problemas de degradação e inconsistência de dados. Apesar de toda complexidade inerente ao estudo, as coletas assintomáticas apresentam diferenciação estatística da classe Sadia e utiliza-se esse argumento para a sequência do trabalho.

Após verificar a diferenciação estatística entre as classes Sadia e Assintomática HVRF para cada coleta, avaliou-se o comportamento das amostras de maneira única. Essa verificação faz-se necessária pois os modelos de classificação para diagnóstico devem ser independentes do período de desenvolvimento da planta. A elaboração do classificador deve possuir características referentes a todos os estágios de evolução da doença para que o mesmo seja viável. Sendo assim, o desenvolvimento do estudo seguiu com a união de todas as amostras numa matriz única, tanto para avaliação dos atributos como para elaboração dos classificadores.

Algoritmos de seleção de atributos foram implementados para definição dos comprimentos de onda mais relevantes para a classificação. Lembrando que se optou por excluir os comprimentos de onda inferiores a 400 nm por representarem ruídos intrínsecos ao espectrômetro. A banda em 405 nm aparente no espectro característico corresponde à reflectância da luz incidente nas amostras e também foram consideradas como atributos. A informação da reflectância superficial das amostras pode diferir de uma classe para a outra, devido a possíveis alterações fisiológicas e estruturais proporcionados pela infecção patogênica.

Devido à variabilidade de algoritmos disponíveis e com o intuito de obter um resultado robusto, optou-se pela implementação de seis diferentes métodos para seleção dos atributos e cada um selecionou 100 dos 476 comprimentos de onda. Desse montante, foram utilizados para a classificação os atributos escolhidos por pelo menos três algoritmos. As quantidades de atributos selecionados por cada algoritmo (100 / 476) e o critério para eleger os atributos foram escolhidos pelo autor com base no conhecimento dos dados tratados. No total, a seleção de atributos resultou em 91 atributos relevantes para a classificação. Para visualização de tais comprimentos de onda, optou-se por dispô-los graficamente no espectro médio das amostras Sadias e Sintomáticas HVRF, obtidos na coleta 31 dai. Cabe destacar que tal coleta foi utilizada apenas nessa exibição. A representação da Figura 25 elucida as regiões que apresentam maior diferença de valor espectral médio.

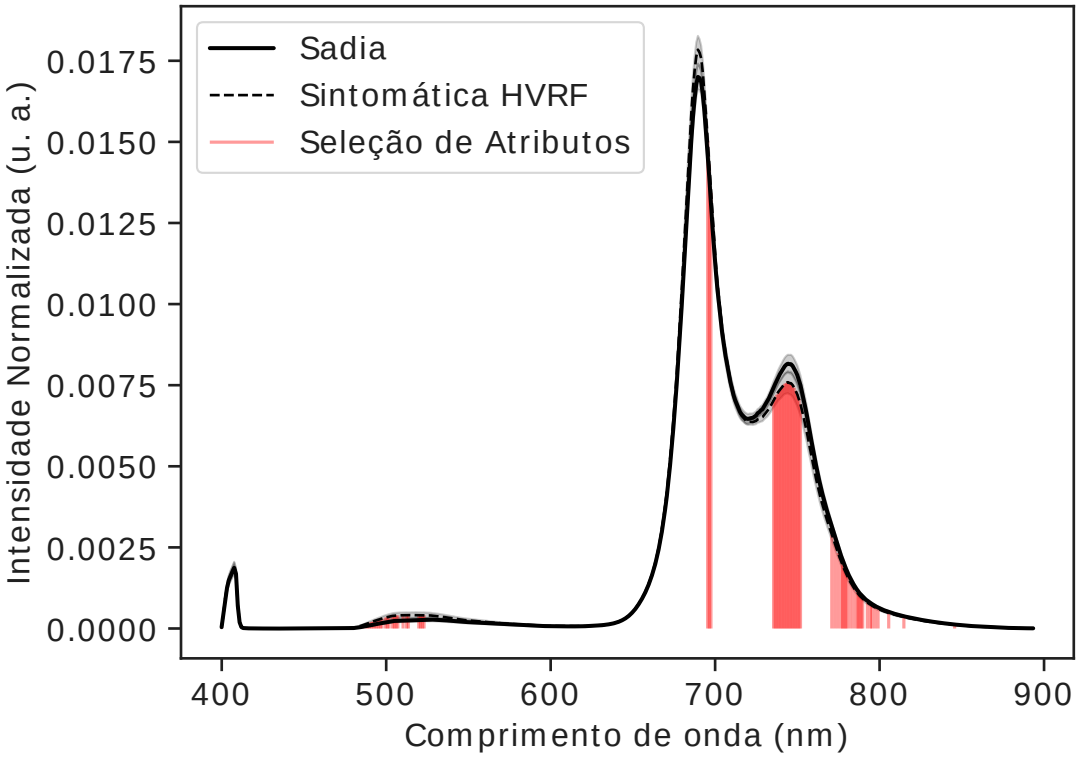

Figura 25 – Distribuição dos 92 comprimentos de onda indicados pelos algoritmos de seleção de atributos para a doença HVRF.
Fonte: Elaborada pelo autor.

No gráfico médio das classes Sadia e Sintomática HVRF, destacam-se as bandas de emissão dos principais metabólitos das folhas. Nota-se um aumento médio das bandas em 520 e 690 nm, bem como uma diminuição no valor médio da banda em 740 nm. Percebe-se também que os atributos selecionados se concentram nas bandas em 520, 740 e 780nm, bem como alguns atributos próximos a 700 nm. Esse resultado é satisfatório, pois revela a importância das bandas de emissão em 520 e 740 nm para a discriminação das classes Sadia e Assintomática HVRF, que estão diretamente associadas à emissão de metabólitos fundamentais para caracterização do estado de saúde da planta. Os atributos selecionados que não estão situados nas bandas características, como por exemplo a faixa em 780 nm, podem estar associados à emissão de outros metabólitos de menor quantidade na folha, ou à variação das bandas de emissão devido a presença do patógeno.

Em seguida, foi proposta uma visualização bidimensional dos dados das classes Sadia e Assintomática HVRF através da *Principal Component Analysis* (PCA). Para o cálculo das componentes principais, foi utilizada a mesma matriz de dados utilizada para a seleção de atributos. Os objetivos dessa etapa consistem em avaliar se as componentes principais são capazes de produzir agrupamentos das classes, e verificar quais faixas espectrais contribuem efetivamente para a discriminação, e qual a relação entre elas. Para a avaliação do agrupamento não

supervisionado, foi aplicado o algoritmo k-*Means* para k = 2, e as regiões delimitadas pelo resultado do algoritmo foram utilizadas como referência para avaliação dos conglomerados. A Figura 26 apresenta a projeção da matriz para duas primeiras componentes principais, que juntas acumulam a maior variância dos dados, bem como o resultado dos agrupamentos.

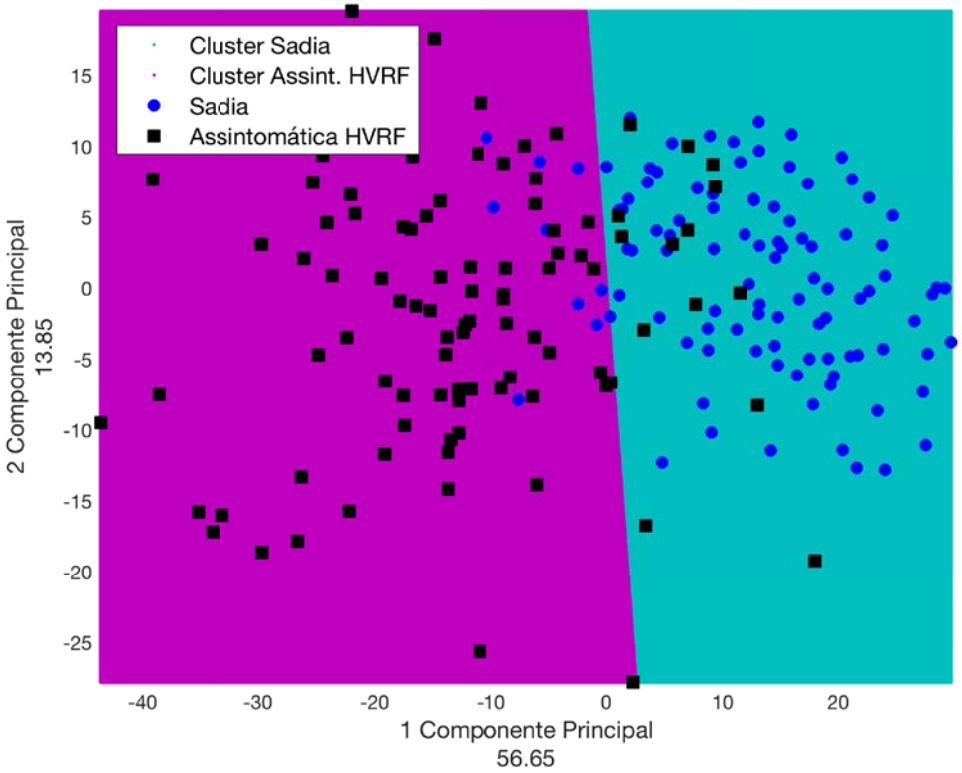

Figura 26 – Projeção bidimensionais das duas primeiras componentes principais para as classes Sadia e Assintomática HVRF referente à matriz de dados dos espectros LIFS. O gráfico também mostra a área resultante da aplicação do algoritmo k-*Means* para as respectivas classes.
Fonte: Elaborada pelo autor.

Nota-se claramente a distinção entre as amostras que compõe as classes Sadia e Assintomática HVRF, e a região dos agrupamentos (*clusters*) deixa nítida a discriminação das classes pela PCA. Essa informação é importante para a sequência dos resultados, pois confirma a tendência de agrupamento das classes. Outra consideração importante é a variância da classe Assintomática HVRF ser maior que a distribuição dos dados da classe Sadia. Esse resultado era esperado, pois sabidamente o grupo contaminado sofreu mudanças em sua composição química, o que proporcionou maior variabilidade dos dados no decorrer da evolução da doença.

A primeira componente principal apresentou uma variância acumulada de aproximadamente 56,65%, enquanto a segunda componente representou 13,85%. Visivelmente, a 1ªCP é a responsável majoritária pela segregação das classes ter sido satisfatória, e a linha divisória do resultado do k-*Means* confirma essa premissa. Baseado nessa informação, é possível estimar a

contribuição de cada comprimento de onda para a formação dessa componente analisando seu fator *loadings*. A Figura 27 apresenta a contribuição de cada comprimento de onda para a formação da componente principal.

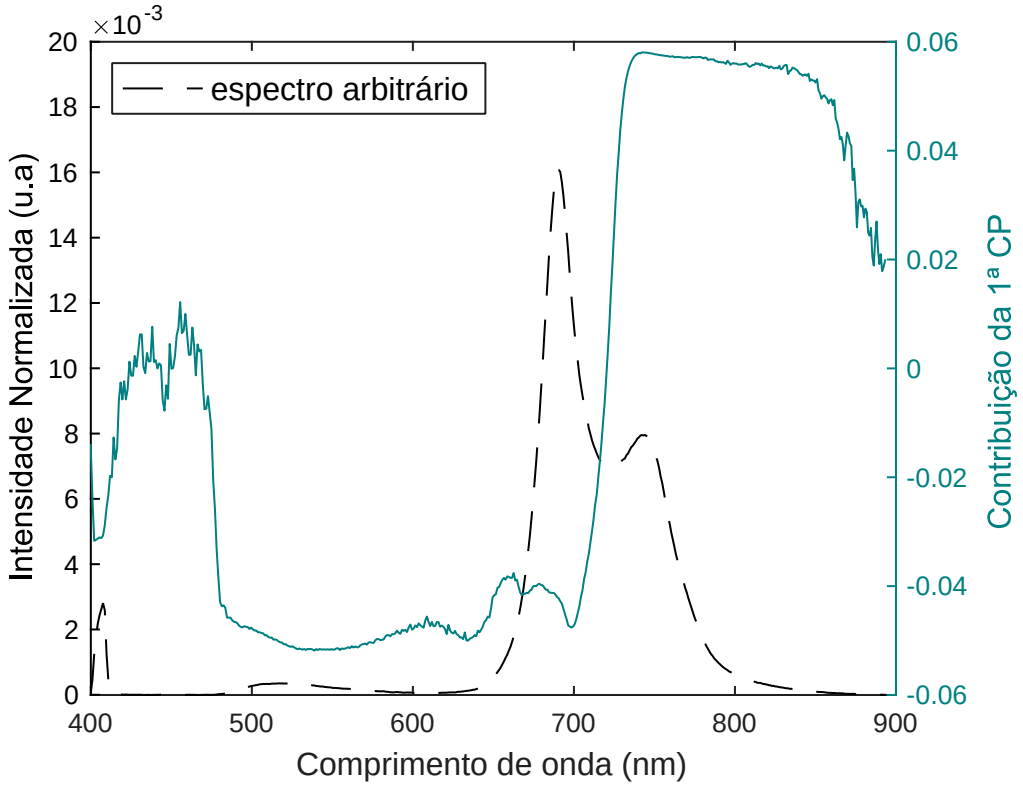

Figura 27 – Contribuição de cada comprimento de onda para a direção da primeira componente principal (1ª CP). A figura ainda traz um espectro arbitrário de uma amostra para referência dos comprimentos de onda.
Fonte: Elaborada pelo autor

As bandas de emissão que mais contribuem, em módulo, para discriminação dos grupos estão nas regiões dos metabólitos secundários em F520 e também na banda de emissão do F740. Tais bandas se relacionam de maneira inversa na formação da componente, reflexo da característica do problema, onde nota-se um aumento médio na intensidade do F520 e uma diminuição da F740 para as classes assintomáticas. Esse resultado está condizente também com as informações obtidas com os algoritmos de seleção de atributos.

Após explorar as características dos espectros de fluorescência gerados pelo sistema LIFS, iniciou-se a etapa de treinamento dos modelos e validação. A escolha de um único algoritmo não é uma tarefa trivial, sendo que, na maioria das vezes, o classificador adequado para determinado problema não mantém sua eficácia para bases de dados distintas. Atualmente, há uma vasta gama de modelos e métodos de classificação disponíveis na literatura, sendo alguns

reconhecidamente consagrados pelos resultados consistentes em diversos artigos. Além disso, diversos classificadores requerem ajustes paramétricos para obterem o maior índice de acerto possível. Tendo em vista esse cenário, foram testados previamente diversos modelos de classificação, com diferentes métodos e características, e uma filtragem dos modelos mais consistentes levou à escolha do algoritmo *Support Vector Machine* (SVM) para a apresentação dos resultados. Todos os classificadores testados foram implementados em linguagem *Python*, com utilização da biblioteca *scikit-learn* para acesso aos algoritmos de classificação, além da utilização das funções de parametrização.

A etapa de classificação foi realizada com a separação da matriz total de dados em dois grupos: um grupo de treino e validação do classificador, correspondendo a 70% do total, e um grupo de teste com os 30% restantes das amostras. Importante destacar que essa divisão foi realizada aleatoriamente pra cada coleta e posteriormente unificada nos grupos de treino e teste. Essa alternativa foi escolhida para garantir que os grupos não apresentassem informações desbalanceadas com relação às coletas, o que certamente enviesaria a classificação. Ademais, essa divisão por coletas é estratégica para avaliação da capacidade discriminatória da técnica com relação ao período de desenvolvimento da doença, ou seja, pode-se utilizar cada grupo de treino separadamente para testes de predição e analisar a evolução das métricas da matriz de confusão.

O treinamento do classificador foi realizado para o grupo de treino, com o método *k-fold* para k = 10. O resultado dessa validação cruzada gerou o modelo de classificação que foi utilizado pra predizer o grupo de teste. A comparação dos dois resultados – validação e predição – é uma boa estratégia para garantir confiabilidade dos resultados, tendo em vista a limitação imposta pelas características do experimento à base de dados. A exibição dos resultados será feita através das matrizes de confusão geradas na validação cruzada e na predição do conjunto de teste.

Voltando a atenção para o caso dos dados espectrais do LIFS para a HVRF, o principal resultado de classificação ocorreu com a utilização dos 97 comprimentos de onda eleitos na seleção de atributos, e o classificador SVM com núcleo polinomial de ordem 2 e padronização dos dados. O treinamento do modelo gerou uma matriz de confusão com 91% de acurácia, mesmo valor para os índices de sensibilidade e especificidade. A predição do grupo de teste, obviamente realizada com os mesmos atributos do treinamento do modelo, gerou um valor e acurácia de 93,98%, com sensibilidade de 95,12% e especificidade de 92,86%. A Figura 28 apresenta as matrizes de confusão para o treinamento do modelo e posterior predição do grupo de teste.

A)

	Sadia	***HVRF***
Sadia	91%	9%
HVRF	9%	91%

B)

	Sadia	***HVRF***
Sadia	95%	7%
HVRF	5%	93%

Figura 28 – Matrizes de confusão para a classificação das amostras assintomáticas HVRF com dados LIFS. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste.
Fonte: Elaborada pelo autor.

Esses resultados podem ser considerados satisfatórios para o presente caso, tendo em vista sua complexidade já detalhada. Ambas as matrizes apresentaram resultados balanceados, sendo possível descartar a possibilidade de *overfitting* para o presente caso. Também se destaca a taxa de aproximadamente 5% de falsos negativos, que trata das amostras Assintomáticas HVRF erroneamente classificadas. Num teste de predição, é desejável que esse valor seja o menor possível, e o presente resultado pode ser considerado aceitável por se tratar de um estudo inicial.

Sendo assim, essa primeira explanação permite concluir que os espectros de emissão obtidos do LIFS são capazes de fornecer dados para a discriminação de classes Sadia e Assintomática HVRF com alto grau de acurácia.

4.2 HASTE VERDE E RETENÇÃO FOLIAR: EM-365

O estereomicroscópio EM-365 apresenta uma fonte de excitação em $\lambda_{excitação}$ = 365 nm, valor este mais energético do que o comprimento de onda da fonte de excitação do LIFS e do LIFI em $\lambda_{excitação}$ = 405 nm. A excitação em 365 nm é próxima do ideal para a fluorescência dos metabólitos secundários, situadas na banda de emissão do azul e verde. Conforme visto no tópico anterior, essa banda é igualmente relevante para a identificação de das amostras assintomáticas HVRF.

Conforme detalhado na seção Materiais e Métodos, os 146 atributos extraídos de cada imagem fazem referência à cor e textura. Inicialmente, buscou-se avaliar a qualidade das amostras obtidas fazendo detecção de outliers. O método de detecção utilizado neste caso foi diferente do caso LIFS, com uma fração de *outliers* de 20% por coleta. Optou-se por essa fração maior para o presente caso devido a dificuldades na aquisição nas imagens por parte do

aparelho. A experiência na manipulação das amostras e do aparelho mostra que a qualidade na aquisição das imagens do EM-365 é altamente variável a fatores intrínsecos ao problema, como por exemplo, o tempo de estabilidade da fluorescência, que diversas vezes inviabilizava a aquisição de algumas imagens e requeria novas tentativas, e até mesmo novos ajustes paramétricos para todas as amostras. Levando essa característica em conta e realizando a detecção dos pontos anômalos, foram excluídos 19 das 160 amostras Sadias e 21 das 160 amostras Assintomáticas HVRF.

Na sequência, foram exploradas a seleção de atributos e a decomposição PCA para visualização dos dados e análise da contribuição dos mesmos na discriminação das classes. Para os atributos de imagem, optou-se por utilizar os mesmos algoritmos já implementados no LIFS, com a diferença que cada um dos seis algoritmos elegeu 50 dos 146 atributos possíveis. Os atributos eleitos por pelo menos três dos algoritmos foram selecionados para elaboração dos classificadores e redução de dimensionalidade. Para a presente base de imagens, foram selecionados 51 atributos, posteriormente utilizados na decomposição PCA e na etapa de classificação. A Figura 29 apresenta a projeção da matriz de dados EM-365 das duas primeiras componentes principais e o resultado do agrupamento das classes Sadia e Assintomática HVRF.

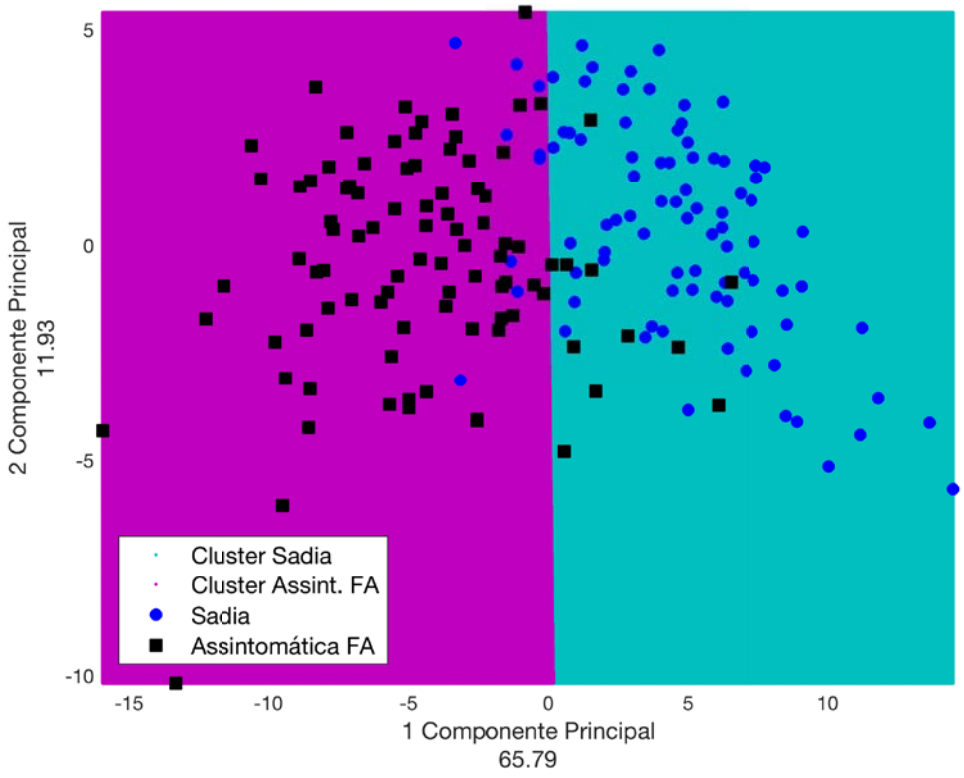

Figura 29 – Projeção bidimensional nas duas componentes principais de maior variância acumulada para as classes Sadia e Assintomática HVRF referente a matriz de dados das imagens EM-365. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes.

Fonte: Elaborada pelo autor.

Tabela 4 – Nome e módulo da contribuição dos 50 principais atributos na direção da primeira componente principal.

	Atributos	|Contribuição| (1º CP)		Atributos	|Contribuição| (1º CP)
1	media_B	0,16	26	lbp_12	0,14
2	energia_H	0,16	27	homogeneidade_S	0,14
3	energia_B	0,16	28	homogeneidade_G	0,14
4	media_s	0,15	29	ASM_S	0,14
5	media_G	0,15	30	lbp_21	0,14
6	media_b*	0,15	31	lbp_13	0,14
7	lbp_9	0,15	32	lbp_7	0,14
8	lbp_8	0,15	33	lbp_19	0,14
9	lbp_11	0,15	34	lbp_17	0,14
10	lbp_10	0,15	35	lbp_16	0,14
11	homogeneidade_B	0,15	36	lbp_15	0,14
12	energia_S	0,15	37	lbp_14	0,14
13	dcolor_G2	0,15	38	kurtosis_S	0,14
14	dcolor_G1	0,15	39	correlação_H	0,13
15	dcolor_B1	0,15	40	desvio_H	0,13
16	ASM_H	0,15	41	skewness_B	0,13
17	energia_G	0,15	42	dissimilaridade_B	0,13
18	dcolor_B2	0,15	43	homogeneidade_b*	0,12
19	ASM_G	0,15	44	kurtosis_B	0,12
20	skewness_S	0,15	45	entropia_S	0,12
21	lbp_20	0,15	46	lbp_1	0,11
22	ASM_B	0,15	47	variância_H	0,11
23	lbp_25	0,14	48	desvio_G	0,1
24	lbp_24	0,14	49	dissimilaridade_b*	0,09
25	lbp_22	0,14	50	variancia_G	0,09

Fonte: Elaborada pelo autor.

O resultado da decomposição PCA apresentou uma segregação de grupos satisfatória. Pode-se notar claramente a formação de agrupamentos para ambas as classes e o resultado da aplicação do algoritmo k-*Means*, e a respectiva demarcação dos *clusters* corrobora com essa informação. Tendo em vista essa distinção, cabe observar que a 1ª CP é responsável majoritária pela segregação obtida, tendo a mesma representação de 65,79 % da variância dos dados. Para avaliar quais atributos têm maior contribuição na direção da 1ª CP, é mostrado na Tabela 4 o valor da contribuição relativa, caracterizada pelos *loadings* desta componente. Tais valores estão apresentados em módulo, pois, para o presente estudo, interessa a magnitude da contribuição de cada atributo.

Devido à homogeneidade dos valores da contribuição relativa, com baixa variabilidade, optou-se por fazer a análise dos atributos relevantes nos seguintes blocos: a) Atributos referentes aos espaços de cores RGB (referência _R, _G, _B), b) HSV (referência _H, _S, _V), c) La*b* (referência _L, _a*, _b*), d) Atributos de textura LBP e por fim e) Atributos das cores dominantes (_dcolor_). Essa mesma organização será apresentada nas demais instrumentações de imagens e doenças que seguem, pois o intuito dessa estruturação é comparar as doenças apresentadas neste trabalho e traçar perspectivas sobre a utilização dos atributos de imagens para caracterização em ambos os casos.

Os atributos de textura LBP foram os mais numerosos para as imagens EM-365 da HVRF, correspondendo a 36% do total de atributos. O segundo bloco de atributos mais relevante foi o espaço de cores RGB, com 26% dos atributos. Por fim, tem-se o espaço de cor HSV com 24%, a cor dominante com 8% e o espaço de cores La*b* com 6%.

De posse dos atributos relevantes, os algoritmos de classificação foram implementados e validados, de modo que o melhor resultado ocorreu para o classificador SVM, com núcleo linear e padronização dos dados. O treinamento do modelo de classificação gerou uma matriz de confusão com acurácia de 93,9%, com aproximadamente 92% de sensibilidade e 93% de especificidade. A predição do grupo de teste apresentou resultados próximos, sendo 91% de acurácia, 89% de sensibilidade e 92 % de especificidade. A Figura 30 apresenta as matrizes de confusão anteriormente citadas.

A)

	Sadia	***HVRF***
Sadia	93%	5%
HVRF	7%	95%

B)

	Sadia	***HVRF***
Sadia	92%	11%
HVRF	8%	89%

Figura 30 – Matrizes de confusão para a classificação das amostras assintomáticas HVRF com dados do EM-365. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste.
Fonte: Elaborada pelo autor.

Esses resultados podem ser considerados satisfatórios e estão próximos às taxas obtidas no LIFS para as mesmas amostras, o que comprova o caráter discriminatório que as imagens de fluorescência do EM-365 podem proporcionar.

Tendo avaliado a viabilidade das imagens de fluorescência para diagnóstico em estágio assintomático da HVRF, cabe agora extrapolar essas características para uma instrumentação capaz de realizar tal diagnóstico de maneira portátil. Sendo assim, as mesmas amostras utilizadas nas instrumentações anteriores foram também avaliadas pelo LIFI e seus resultados são apresentados na sequência.

4.3 HASTE VERDE E RETENÇÃO FOLIAR: LIFI

A última instrumentação apresentada para o caso das amostras Assintomáticas HVRF é o LIFI. Vale lembrar que essa montagem experimental tem por objetivo proporcionar uma configuração plausível de ser adaptada para diagnóstico em campo, assim como já foi alcançado com sucesso para a instrumentação LIFS. Lembra-se também que a instrumentação proposta possui comprimento de onda de excitação em 405 nm.

Inicia-se a apresentação dos resultados obtidos do sistema LIFI com a importante etapa de exclusão de *outliers*. Diferentemente do ocorrido para as imagens do EM-365, não foram encontradas dificuldades para aquisição das imagens LIFI, visto que ela possui uma estruturação mais robusta e menos dependente de parâmetros. Conforme já citado, a fração de outliers por coleta foi definida em 10%, sendo que esta resultou na exclusão de 12 das 160 amostras Sadias e 14 das 160 amostras Assintomáticas HVRF.

A execução dos algoritmos de seleção de atributos resultou na escolha de 50 atributos. Estes foram usados na decomposição PCA e posteriormente na validação dos modelos de

classificação e predição dos grupos de teste. A Figura 31 apresenta a projeção nas duas primeiras componentes principais obtidas da matriz de dados LIFI e o resultado da aplicação do k-NN para agrupamento das classes Sadia e Assintomática HVRF.

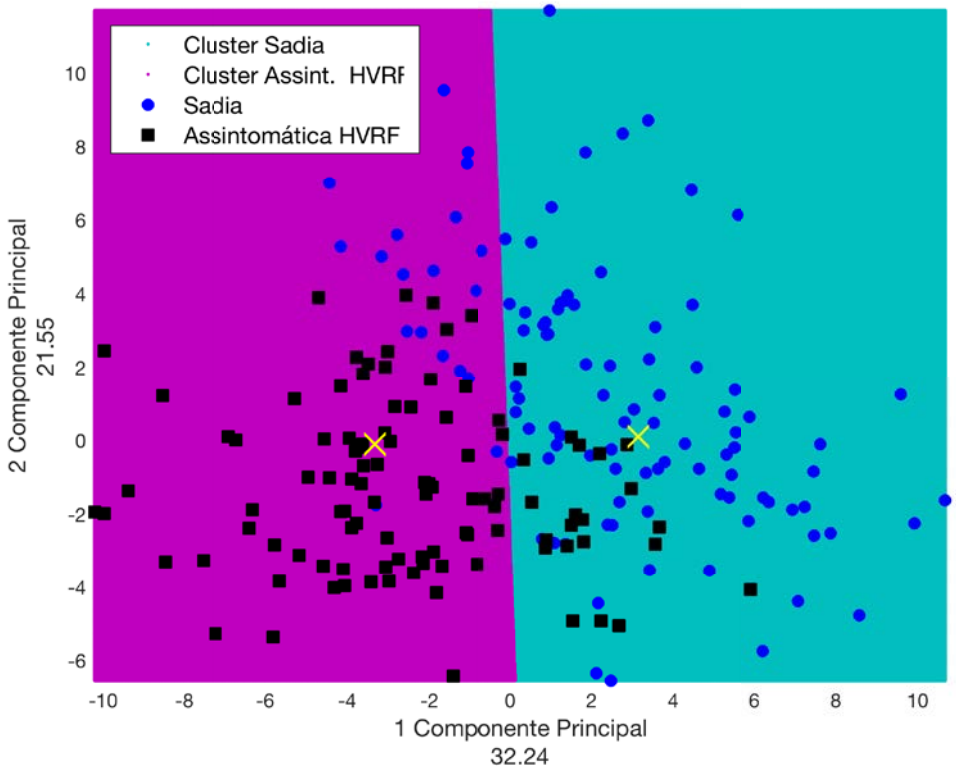

Figura 31 – Projeção bidimensional nas duas componentes principais de maior variância acumulada para as classes Sadia e Assintomática HVRF referente à matriz de dados das imagens LIFI. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes. Além disso, são destacados na figura os centroides de cada *cluster*.
Fonte: Elaborada pelo autor.

A decomposição em duas componentes principais para os dados LIFI não apresentou um agrupamento tão significativo quanto as instrumentações LIFS e EM-365, conclusão essa baseada na quantidade de amostras situadas em regiões diferentes do respectivo agrupamento de classe, e na dispersão relativa de cada grupo projetado. Neste caso, também foi inserida no gráfico a posição do centro de massa dos grupos calculados, a fim de comprovar a qualidade do agrupamento. É possível afirmar então que a 1ª CP é responsável majoritária pela formação dos agrupamentos. Vale a ressalva de que a presente visualização acumulou 53,79% da variância dos dados, enquanto as duas primeiras CPs do LIFS acumularam 95,19% e o EM-365 acumulou 74,3%. A título de curiosidade, a variância acumulada pelas três primeiras componentes principais alcança 77% da variância total, e a projeção tridimensional das três maiores

componentes principais sugere uma capacidade discriminatória maior para os dados, conforme mostrado na Figura 32.

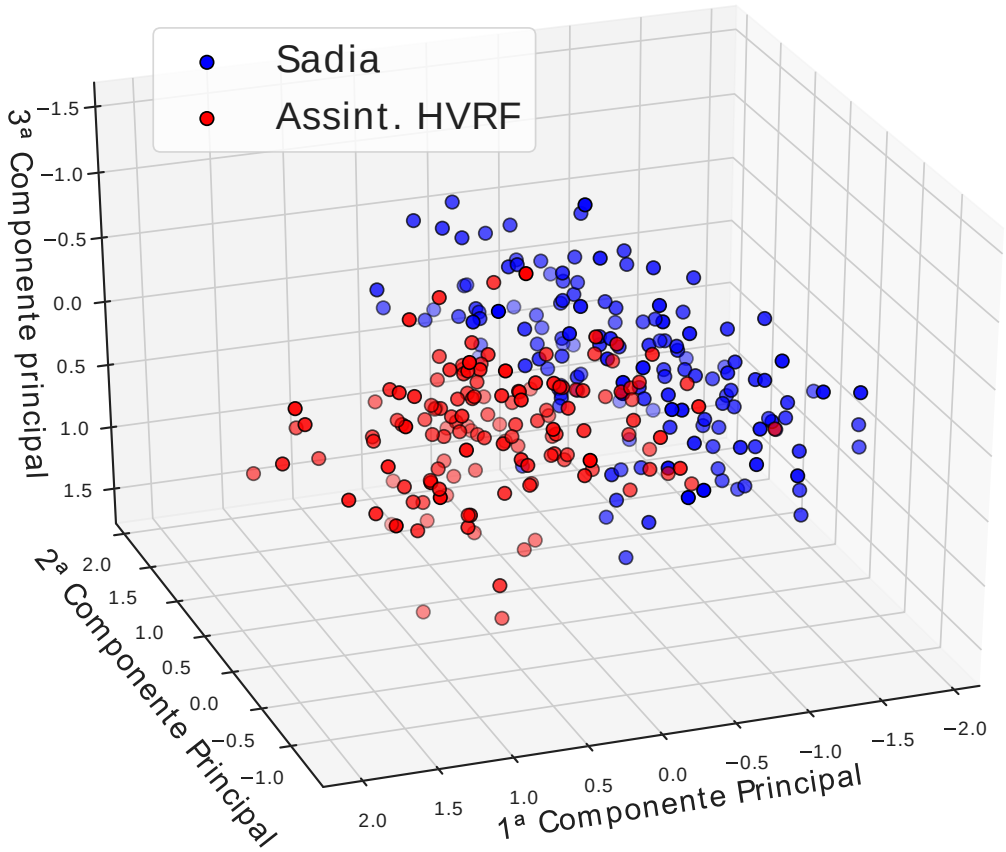

Figura 32 – Projeção tridimensional nas três componentes principais de maior variância acumulada para as classes Sadia e Assintomática HVRF referente à matriz de dados das imagens LIFI.
Fonte: Elaborada pelo autor.

Na sequência, é apresentada na Tabela 5 a contribuição dos 50 principais atributos para a direção da 1ª CP. Novamente, optou-se por avaliar a contribuição dos blocos dos atributos. Dentre os 50 mais relevantes, o espaço de cores HSV apresentou mais atributos, representando 26%. Os atributos de textura LBP mantiveram uma alta porcentagem de aparição, com 24%. Os espaços de cores La*b* e RGB e a cor dominante contabilizaram respectivamente, 20%, 18% e 12% dos atributos relevantes. Para a presente instrumentação, houve relativo equilíbrio na importância dos blocos de atributos, ou seja, nenhum bloco apresentou uma quantidade de atributos excêntrica.

Tabela 5 – Nome e módulo da contribuição dos 50 principais atributos na direção da primeira componente principal.

	Atributos	|Contribuição| (1º CP)		Atributos	|Contribuição| (1º CP)
1	skewness_S	0,20	26	skewness_B	0,14
2	media_S	0,20	27	skewness_V	0,12
3	energia_H	0,20	28	kurtosis_b*	0,11
4	energia_B	0,20	29	media_L	0,1
5	ASM_B	0,20	30	lbp_15	0,1
6	dissimilaridade_H	0,20	31	lbp_11	0,1
7	variancia_H	0,19	32	lbp_10	0,1
8	skewness_H	0,19	33	correlação_H	0,1
9	media_B	0,19	34	lbp_14	0,1
10	desvio_H	0,19	35	lbp_12	0,1
11	dcolor_B2	0,19	36	lbp_9	0,09
12	ASM_H	0,19	37	lbp_16	0,09
13	kurtosis_S	0,18	38	lbp_8	0,09
14	kurtosis_H	0,18	39	lbp_17	0,08
15	dcolor_B3	0,18	40	lbp_7	0,08
16	dcolor_B1	0,18	41	lbp_18	0,07
17	media_H	0,18	42	variancia_a*	0,06
18	media_a*	0,18	43	desvio_a*	0,06
19	media_b*	0,17	44	variancia_L	0,05
20	correlação_a*	0,17	45	desvio_L	0,05
21	media_G	0,16	46	skewness_R	0,05
22	dcolor_G3	0,16	47	lbp_23	0,04
23	dcolor_G2	0,16	48	variancia_G	0,03
24	dcolor_G1	0,15	49	desvio_G	0,03
25	kurtosis_B	0,14	50	desvio_b*	0,03

Fonte: Elaborada pelo autor.

De posse dos atributos relevantes, os algoritmos de classificação foram implementados e validados. O melhor resultado de classificação ocorreu novamente para o classificador SVM, com um núcleo polinomial de ordem 2 e padronização da matriz de dados. A validação do treinamento do modelo apresentou um valor de acurácia de 97,5%, sendo 98% de

especificidade e 97% de sensibilidade. Resultado amplamente satisfatório, tal qual foram os resultados para predição do grupo de teste, que obteve expressivos 98% em todas as métricas, prevenindo assim possibilidades de *overfitting*. A Figura 33 apresentam as matrizes de confusão para a validação cruzada do modelo e da predição do grupo de teste.

A)

	Sadia	***HVRF***
Sadia	95%	5%
HVRF	5%	95%

B)

	Sadia	***HVRF***
Sadia	98%	2%
HVRF	2%	98%

Figura 33 – Matrizes de confusão para a classificação das amostras assintomáticas HVRF com dados LIFI. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste.
Fonte: Elaborada pelo autor.

Esse resultado finaliza as análises para a doença HVRF com considerações importantes. Todas as técnicas de florescência obtiveram acurácia na predição dos grupos de teste superiores a 90%, sendo que o maior valor ocorreu para a técnica LIFI. Esse resultado comprova a capacidade da fluorescência em diagnosticar a doença HVRF em estágio assintomático. É animador para as pretensões de desenvolvimento de uma instrumentação de diagnóstico em campo, com a técnica LIFI obtendo resultados equivalentes às demais técnicas já consagradas.

Na sequência, serão expostos os resultados para a doença FA, utilizando as mesmas ferramentas computacionais utilizadas para a HVRF.

4.4 FERRUGEM ASIÁTICA: LIFS

Conforme apresentado na seção de Metodologia, a doença Ferrugem asiática (FA) é transmitida pelo fungo *phakopsora pachyrhizi* e tem uma ação diferente do nematoide *aphelenchoides besseyi*. Enquanto o nematoide se reproduz mantendo a planta em estado vegetativo perene, reprimindo sua senescência, o fungo age agressivamente durante sua reprodução, ocasionando lesões nas folhas e levando-as a morte. O estágio assintomático da FA é caracterizado pela fixação do esporo na folha, ocasionando micro lesões invisíveis a olho nu. O estágio sintomático ocorre cerca de duas semanas após a fixação, onde essas lesões tornam-se visíveis.

O ensaio da reprodução do fungo gerou três coletas assintomáticas: 4 dai, 7 dai, 11 dai. Cada uma das amostras passou pelos sistemas de fluorescência nos mesmos padrões para a doença HVRF. Cabe destacar a dificuldade na obtenção de amostras assintomáticas para doença FA, pois seu desenvolvimento é rápido e a extração de folhas em períodos de tempo menores induziriam a alterações fisiológicas nas folhas.

Com relação à detecção de *outliers*, as frações de outliers para cada coleta foi mantida em 10% para o LIFS. Foram excluídas 11 das 96 amostras para a classe Sadia, enquanto que para a classe Assintomática FA, foram detectados 10 das 96. A Figura 34 apresenta a evolução dos valores médios de F520 através das coletas 4 dai, 7 dai e 11 dai. A Tabela 6 apresenta os valores médios com seu respectivo desvio padrão para cada uma das coletas, bem como o p-valor resultante do teste de hipótese para diferenciação da média.

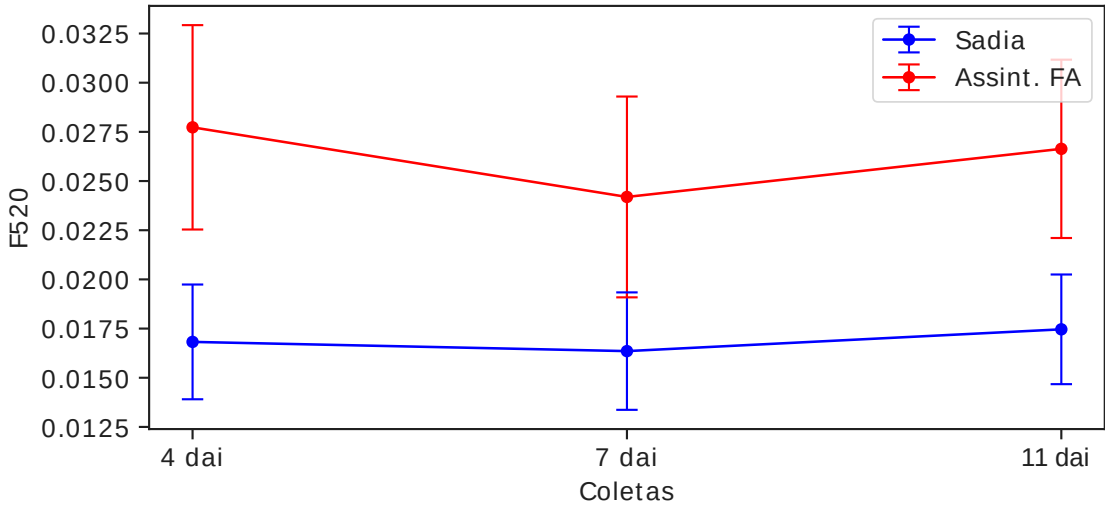

Figura 34 – Evolução dos valores médios da área F520 para as classes Sadia e Assintomática FA através das coletas 4 dai, 7 dai e 11 dai.
Fonte: Elaborada pelo autor.

Tabela 6 – Valores médios da área F520 referente às classes Sadia e Assintomática FA e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias.

F520	4 dai	7 dai	11 dai
Sadia	0,017 ± 0,003	0,016 ± 0,003	0,017 ± 0,003
Assint. FA	0,027 ± 0,005	0,022 ± 0,005	0,025 ± 0,004
p-valor	< 0,05	< 0,05	< 0,05

Fonte: Elaborada pelo autor.

A evolução das médias sugere um comportamento diferenciado entre as classes. A área da banda F520 da classe Assintomática FA apresentou valores, em média, maiores que a mesma

banda para a classe Sadia, conforme o esperado. Esse resultado corrobora com a evolução dos valores para a HVRF. Nota-se também que tais médias são diferenciáveis estatisticamente, baseado nos resultados do p-valor. A seguir, é apresentado na Tabela 7 e na Figura 35 as mesmas disposições referentes à análise da relação F690 / F740.

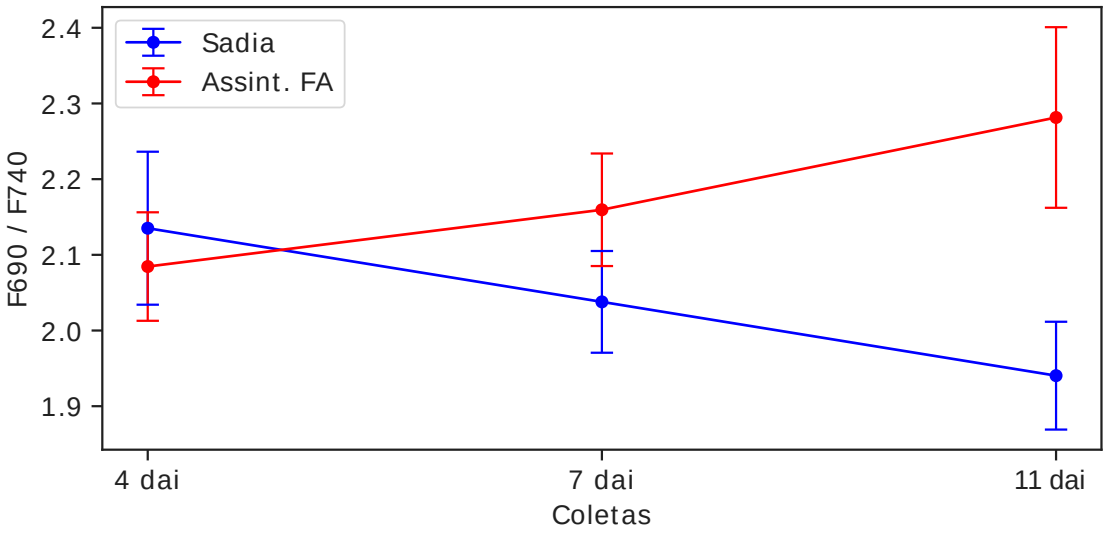

Figura 35 – Evolução dos valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF para o sistema LIFS e o respectivo desvio padrão para cada coleta.
Fonte: Elaborada pelo autor

Tabela 7 – Valores médios da relação das áreas F690 / F740 referente às classes Sadia e Assintomática HVRF e o respectivo p-valor resultante da aplicação do teste de hipótese para diferenciação das médias.

F690 / F740	**4 dai**	**7 dai**	**11 dai**
Sadia	2,1 ± 0,1	2,03 ± 0,07	1,93 ± 0,07
Assint. FA	2,07 ± 0,06	2,16 ± 0,06	2,2 ± 0,1
p-valor	0,177	< 0,05	< 0,05

Fonte: Elaborada pelo autor.

Na coleta 4 dai, não houve diferenciação estatística entre as classes, enquanto as demais coletas mostraram constante aumento na diferenciação entre as classes Sadia e Assintomática FA. Porém, apesar de inicialmente não haver diferenciação, a evolução apresentada está condizente com a teoria proposta, onde o *stress* ocasionado pela infestação fúngica na planta é verificado pelo aumento na relação F690 / F740 no decorrer do tempo.

Comparando os dois gráficos de evolução das bandas F520 (Figura 33) e da relação F690/F740 (Figura 34), nota-se uma complementação de informações. Na coleta 4 dai, houve

a maior diferenciação estatística para a F520 e não houve diferenciação para a relação F690/F740. As coletas 7 dai e 11 dai mantiveram evolução constante para F520 e houve aumento na diferença entre as classes para F690/F740 nas mesmas coletas. Esse resultado sugere, para o presente caso, que ambas as regiões contribuem significativamente para a caracterização do estágio assintomático. Pode-se prever que a etapa de classificação ocorra com sucesso para esses dados, pois os valores espectrais possuem informações com diferenciação estatística em todas as coletas, além de serem complementares.

Na sequência, é apresentada na Figura 36 a disposição dos 95 comprimentos de onda destacados pelos algoritmos de seleção de atributos para a etapa de classificação das classes Sadia e Assintomática FA, bem como a disposição dos espectros médios da coleta 14 dai, referentes ao estágio Sintomático da FA.

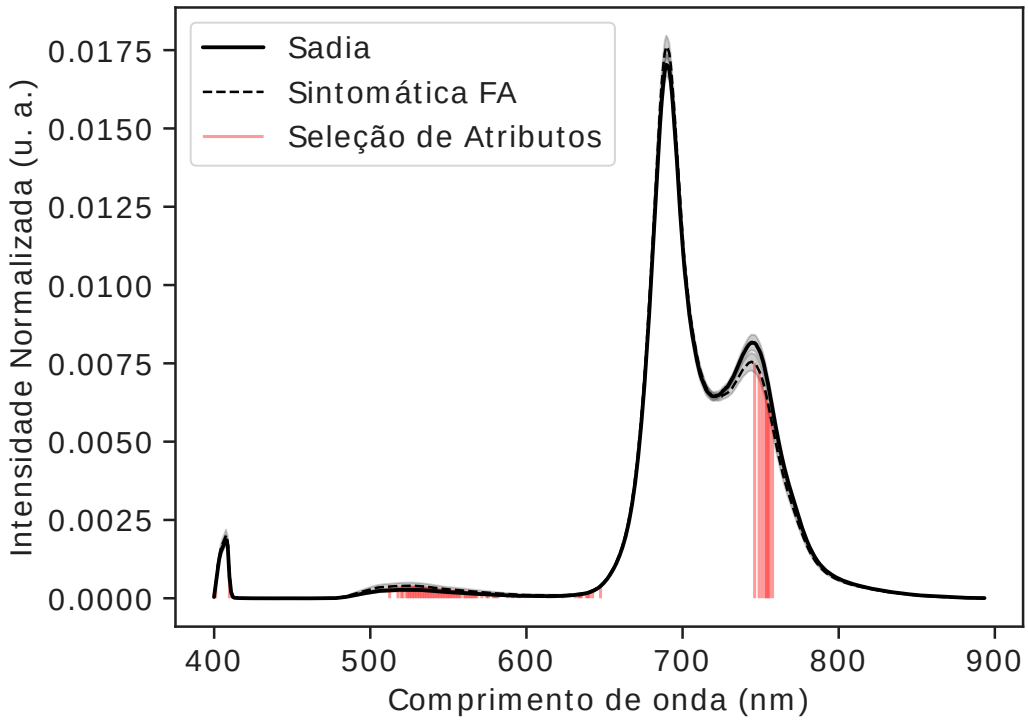

Figura 36 – Distribuição dos 95 comprimentos de onda indicados pelos algoritmos de seleção de atributos para a doença FA.
Fonte: Elaborada pelo autor.

Esse resultado desponta a importância majoritária das bandas de emissão em 520 e 740 nm para a discriminação dos grupos Sadia e Assintomática FA. Alguns poucos comprimentos de onda fora das bandas principais também foram selecionados no presente caso. Na banda em 405 nm há dois atributos selecionados, e há também alguns poucos comprimentos próximos a 640 nm destacados. Cabe lembrar que todas as regiões espectrais podem ser afetadas pela presença do patógeno, pois a alteração nas bandas de emissão pode afetar o comportamento do espectro total. Em comparação com os comprimentos de onda selecionados no LIFS para a

HVRF, coincide a presença de atributos nas bandas de emissão em 520 e 740 nm. Desse resultado, surge uma consideração a respeito de futuras melhorias para o sistema de imagem LIFI, através da utilização de filtros passa-banda nessas regiões. Ao limitar os comprimentos de onda que chegam ao sensor da câmera às bandas essenciais, é possível acessar com maior precisão informações discriminatórias para as doenças. Os mesmos padrões de comportamento, com aumento de intensidade da banda F520 e diminuição da banda F740 foram repetidos tanto para HVRF quanto para a FA. Esse é um indicativo importante sobre a caracterização óptica de plantas assintomáticas e, numa esfera maior, a respeito dos mecanismos biológicos de defesa da planta.

De posse dos atributos selecionados, foi realizada a decomposição dos dados nas componentes principais (PCA) e a Figura 37 apresenta a projeção da matriz dos dados LIFS para duas primeiras componentes, bem como o resultado da aplicação do k-*Means* para agrupamento das classes.

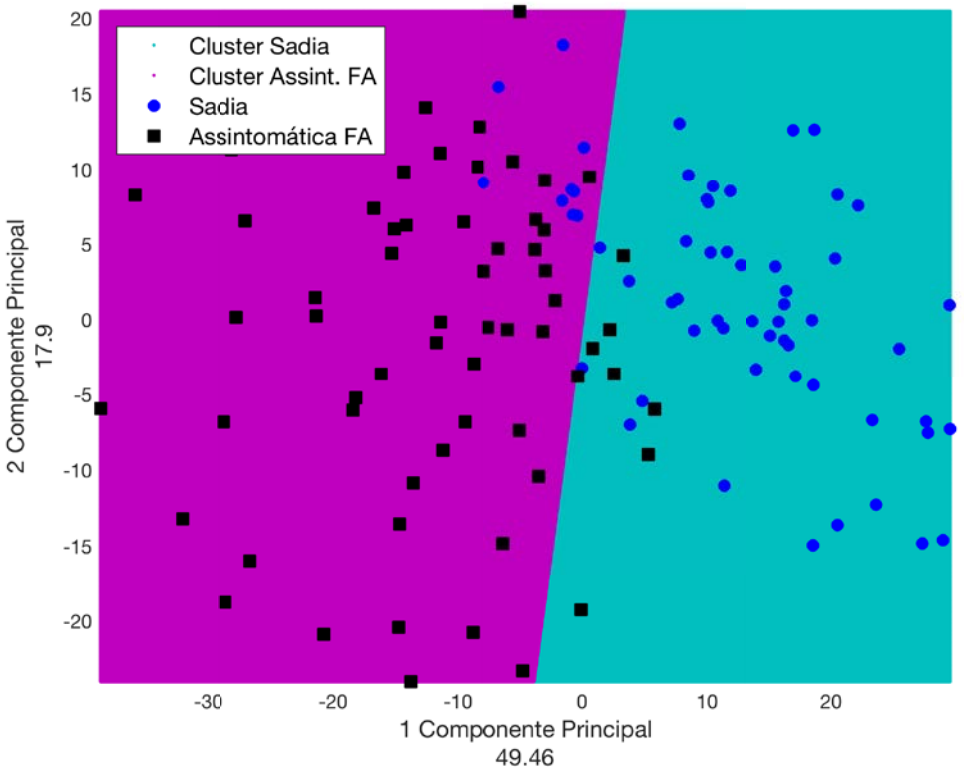

Figura 37 – Projeção bidimensional das duas primeiras componentes principais para as classes Sadia e Assintomática FA referente à matriz de dados dos espectros LIFS. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes.
Fonte: Elaborada pelo autor.

O resultado da projeção dos dados LIFS para a doença FA apresentou agrupamento satisfatório, com poucas amostras fora dos *clusters* calculados. Além do mais, esse agrupamento é similar ao obtido para a doença HVRF (Figura 26), com poucas amostras fora dos *clusters* e maior variabilidade dos dados para a classe Assintomática FA. A primeira componente principal, com variância de 49,46%, é responsável majoritária pela discriminação satisfatória, pois a mesma caracteriza a disposição dos agrupamentos. A segunda componente principal, com 17,9% da variância, visivelmente não contribui para uma discriminação dos grupos. As duas maiores CPs acumulam uma 67,36% da variância total presente nos dados LIFS. Na sequência, a Figura 38 mostra as contribuições de cada comprimento de onda para formação da primeira componente principal, além de um espectro arbitrário de uma planta sadia para referência.

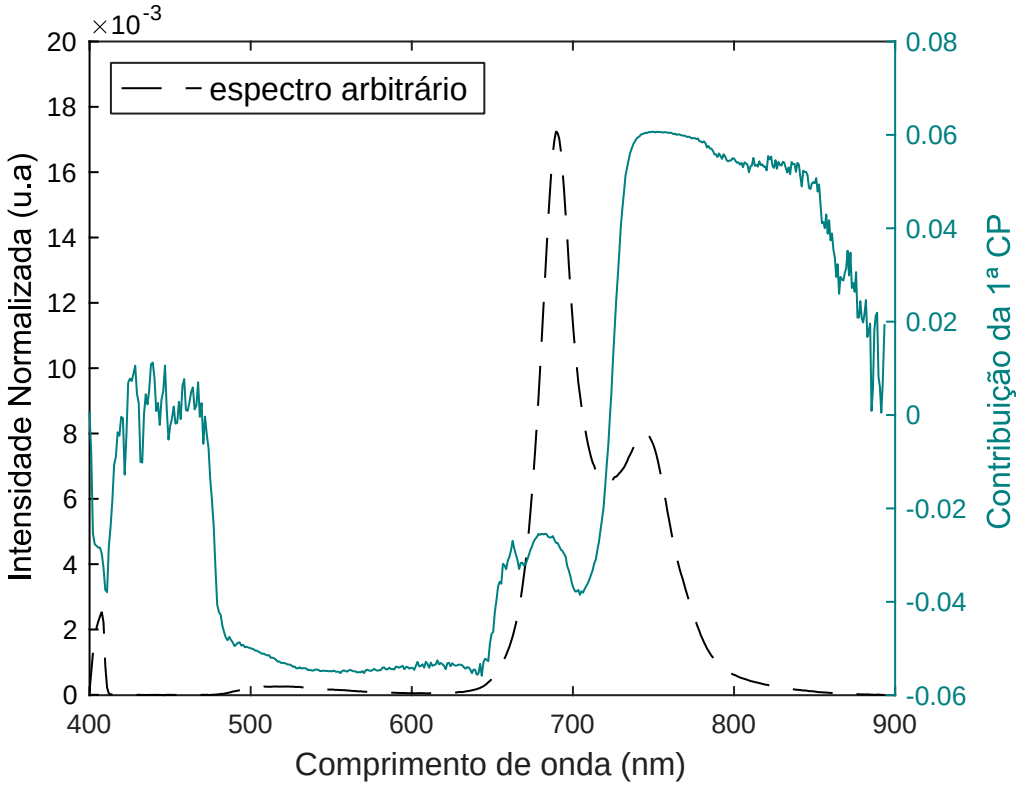

Figura 38 – Contribuição de cada comprimento de onda para a direção da primeira componente principal (1ª CP). A figura ainda traz um espectro arbitrário de uma amostra para referência dos comprimentos de onda.
Fonte: Elaborada pelo autor

Tal qual obtido para a doença HVRF, as bandas de emissão em 520 e 740 nm são as maiores contribuições, em módulo, para a direção da 1ª CP. Este gráfico comprova o comportamento constatado nos espectros médios da Figura 36, onde um aumento na banda 520 nm reflete num comportamento inverso para a banda 740 nm, conferindo uma diminuição nessa banda. Para ambas as doenças, a 1º CP contribuiu para a separação adequada dos grupos. Esse

resultado é significativo e comprova a capacidade discriminatória da técnica de fluorescência LIFS na caracterização óptica das assintomáticas.

Na sequência, os atributos selecionados foram utilizados na elaboração de modelos de classificação. O principal resultado de classificação foi obtido para o algoritmo SVM com núcleo polinomial de ordem 2 com normalização dos dados. O treinamento do modelo de classificação apresentou uma acurácia de 99%, com uma sensibilidade de 98% e especificidade de 100%. A predição do grupo de teste apresentou acurácia total de 94%, com sensibilidade de 93% e especificidade de 95%. A Figura 39 apresenta ambas as matrizes de confusão.

A)

	Sadia	***FA***
Sadia	100%	2%
FA	0%	98%

B)

	Sadia	***FA***
Sadia	95%	7%
FA	5%	93%

Figura 39 – Matrizes de confusão para a classificação das amostras assintomáticas FA com dados LIFS. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste.
Fonte: Elaborada pelo autor.

A capacidade discriminatória da técnica LIFS já havia sido comprovada para a doença HVRF, e agora a mesma consistência é alcançada para a doença FA. O sucesso em ambas as classificações permite concluir que a técnica de fluorescência é uma importante estratégia para avaliar o estado de saúde de plantas de soja. A técnica LIFS permite acessar informações para discriminação das classes assintomáticas para as referidas doenças logo nos estágios iniciais, podendo ser concebida como uma importante ferramenta de diagnóstico e monitoramento de plantas de soja. Essa conclusão é animadora para as pretensões do trabalho, e cabe agora avaliar se esse padrão é mantido para as técnicas de imagem, e se as elas podem também garantir acesso a informações discriminatórias de fluorescência para ambas as doenças.

4.5 FERRUGEM ASIÁTICA: EM-365

Inicia-se a avaliação das imagens do EM-365 para a doença FA a partir da remoção dos *outliers* presentes no banco de dados da referida instrumentação. A taxa de outliers foi a mesma

utilizada para a HVRF, com 20% de amostras anômalas por coleta. A aplicação desta resultou na exclusão de 21 amostras Sadias e de 19 amostras Assintomáticas FA do total de 192 imagens.

Todas as imagens do EM-365 restantes da exclusão de outliers foram submetidas aos algoritmos de seleção de atributos. A estratégia adotada elegeu 55 atributos relevantes para classificação, sendo os mesmos utilizados no cálculo da PCA para visualização dos dados e posteriormente na etapa de classificação e validação. A Figura 40 apresenta a projeção das amostras do EM-365 para duas primeiras componentes principais, bem como o resultado da aplicação do k-NN para agrupamento das classes.

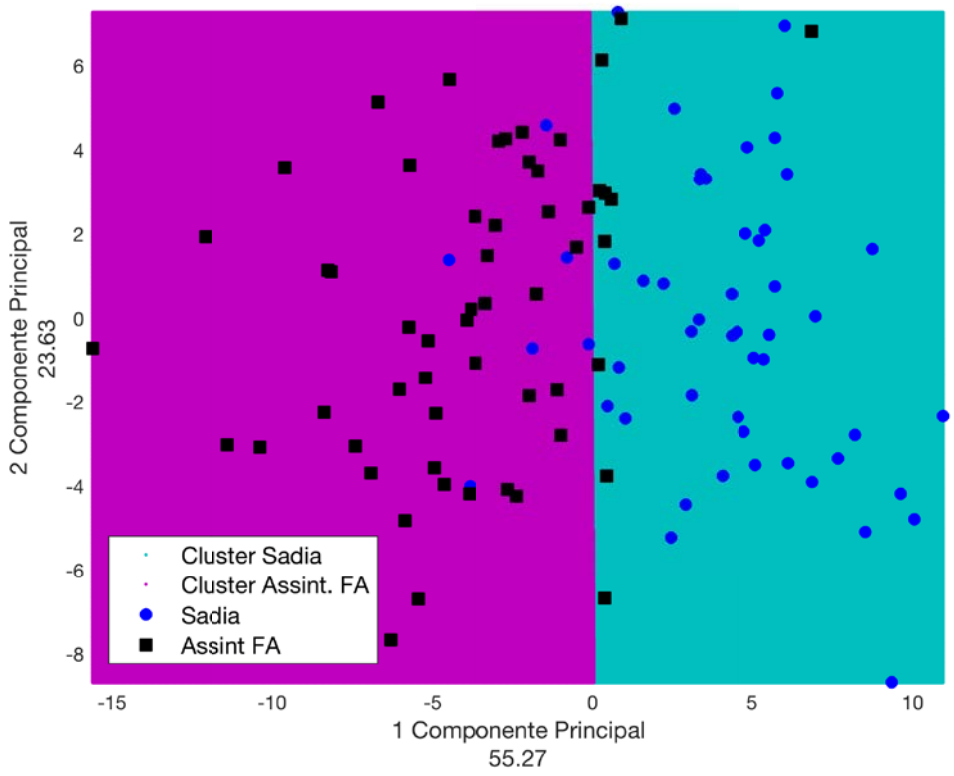

Figura 40 – Projeção bidimensional das duas primeiras componentes principais para as classes Sadia e Assintomática FA para as imagens EM-365. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes.
Fonte: Elaborada pelo autor.

O resultado da projeção dos dados EM-365 para a doença FA, assim como para a HVRF, apresentou agrupamento satisfatório. A primeira componente principal, com variância de 55,27%, é responsável, predominantemente, pela disposição aceitável dos grupos. A segunda componente principal, com 23,63% da variância, visivelmente não contribui efetivamente na disposição dos grupos. As duas maiores CPs totalizam uma variância acumulada de 78,90%. Com relação à segregação, apenas seis amostras sadias estão fora do respectivo grupo, enquanto a classe Assintomática FA apresentou dez. Através da disposição dos pontos no gráfico,

percebe-se maior variabilidade da classe Assintomática FA do que a classe Sadia. Na sequência, a Tabela 8 mostra os 50 principais atributos selecionados e a contribuição relativa para formação da primeira componente principal.

Tabela 8 – Nome e módulo da contribuição dos 50 principais atributos na direção da primeira componente principal.

	Atributos	|Contribuição| (1º CP)		Atributos	|Contribuição| (1º CP)
1	media_S	0,17	26	ASM_G	0,15
2	media_b*	0,17	27	homogeneidade_S	0,14
3	media_B	0,17	28	skewness_B	0,14
4	energia_H	0,17	29	kurtosis_S	0,14
5	media_G	0,16	30	homogeneidade_V	0,13
6	entropia_H	0,16	31	homogeneidade_R	0,13
7	homogeneidade_H	0,16	32	dissimilaridade_V	0,13
8	energia_S	0,16	33	dissimilaridade_R	0,13
9	energia_B	0,16	34	desvio_H	0,13
10	dcolor_G2	0,16	35	homogeneidade_a*	0,12
11	dcolor_G1	0,16	36	dissimilaridade_a*	0,12
12	dcolor_B1	0,16	37	dissimilaridade_L	0,12
13	ASM_H	0,16	38	contraste_a*	0,12
14	ASM_B	0,16	39	contraste_V	0,12
15	entropia_B	0,16	40	contraste_R	0,12
16	energia_G	0,16	41	contraste_L	0,12
17	homogeneidade_B	0,15	42	correlação_L	0,12
18	dcolor_B2	0,15	43	homogeneidade_L	0,11
19	ASM_S	0,15	44	variancia_H	0,11
20	skewness_S	0,15	45	energia_V	0,11
21	media_a*	0,15	46	ASM_V	0,11
22	homogeneidade_b*	0,15	47	skewness_a*	0,1
23	entropia_S	0,15	48	lbp_19	0,09
24	entropia_G	0,15	49	correlação_G	0,09
25	correlação_H	0,15	50	ASM_R	0,09

Fonte: Elaborada pelo autor.

Trazendo a análise dos atributos importantes por blocos, destaca-se nesse caso o conjunto de atributos do espaço de cores HSV, com 38% do total de atributos. Na sequência, aparecem atributos da representação RGB, com 30%, e o espaço de com La*b*, com 22%. Esses atributos totalizam 90% dos atributos e revelam a importância dos mesmos na caracterização do problema. Na comparação com a HVRF, constata-se uma grande diferença na importância dos atributos LBP que, no presente caso, destacou apenas um atributo (apenas 2%) contra os 36% alcançados pela HVRF. Os atributos de cor dominante encerram os blocos com 8% dos atributos, mesma taxa de aparição para o caso anterior.

Por fim, os atributos selecionados foram também utilizados na elaboração de classificadores e predição dos dados. Seguindo as mesmas etapas realizadas para casos anteriores, foi realizada a separação dos grupos de treino e teste coleta a coleta na proporção 70 / 30. O melhor desempenho ocorreu para o algoritmo SVM, com núcleo linear e padronização da matriz de dados. A Figura 41 mostra a matriz de confusão gerada do teste de predição.

A)

	Sadia	***FA***
Sadia	92%	7%
FA	8%	93%

B)

	Sadia	***FA***
Sadia	84%	15%
FA	16%	85%

Figura 41 – Matrizes de confusão para a classificação das amostras assintomáticas FA com dados EM-365. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste.
Fonte: Elaborada pelo autor

A acurácia média alcançada pela validação cruzada do modelo foi de 93,5%, com sensibilidade de 92% e especificidade de 93%. Esse resultado é satisfatório e coerente com o resultado alcançado para o caso das Assintomáticas HVRF. Com relação à predição do grupo de teste, a classe Assintomática HVRF obteve uma acurácia média de 90,5%, enquanto a presente predição das Assintomáticas FA obteve 84,5%, com 85% de sensibilidade e 84 % de especificidade. Apesar dessa diferença, pode-se considerar ambos os resultados satisfatórios para um teste de predição.

4.6 FERRUGEM ASIÁTICA: LIFI

A análise das imagens do LIFI inicia-se pela remoção de *outliers* da matriz de dados originais. A margem para detecção dos pontos anômalos novamente foi de 10% para cada coleta, mesmo critério utilizado para a detecção de *outliers* para o caso HVRF. No presente caso, nove amostras dos 96 referentes à classe Sadia e oito dos 96 referentes à classe Assintomática FA foram consideradas anômalas e excluídas da matriz de dados.

Todos os atributos das imagens do LIFI foram submetidos aos algoritmos de seleção de atributos. Foram eleitos 46 atributos relevantes para classificação, sendo os mesmos utilizados no cálculo da PCA para visualização dos dados e posterior classificação e validação dos dados. A Figura 42 apresenta a projeção das amostras do EM-365 para duas primeiras componentes principais, bem como o resultado da aplicação do k-NN para agrupamento das classes.

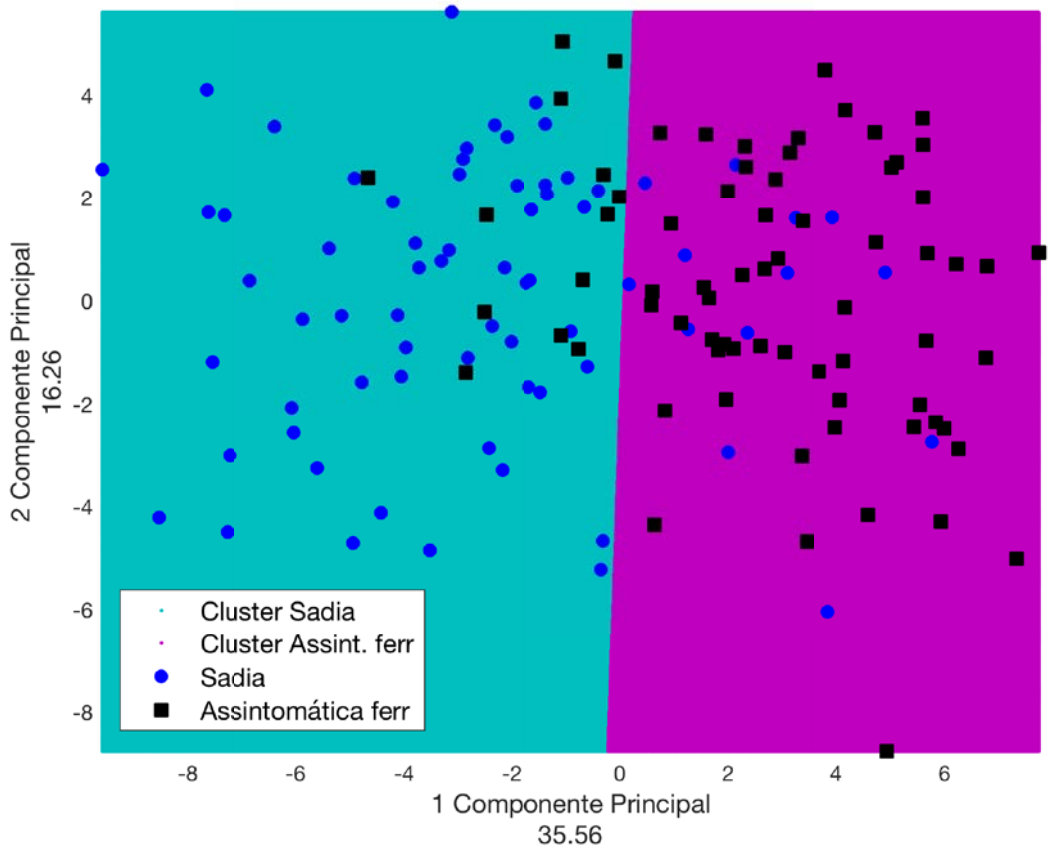

Figura 42 – Projeção bidimensional das duas primeiras componentes principais para as classes Sadia e Assintomática FA para as imagens LIFI. O gráfico também mostra a área resultante da aplicação do algoritmo k-NN para as respectivas classes.
Fonte: Elaborada pelo autor.

O resultado da projeção bidimensional em componentes principais dos dados LIFI não foi satisfatório, quando comparada com as demais instrumentações LIFI e EM-365 para a FA. Constatam-se 13 amostras Sadias e 13 amostras Assintomática FA em regiões diferentes de

seus respectivos grupos, resultados bem diferentes das demais projeções analisadas anteriormente.

A primeira componente principal tem 35,56% da variância, enquanto a segunda componente tem 16,26%, o que resulta numa variância acumulada de apenas 51,82%. Ou seja, praticamente metade da variância dos dados não foi explorada nessa visualização. Dessas informações, pode-se concluir que a variância acumulada por essas componentes não foi capaz de projetar uma discriminação adequada dos grupos. Isso não significa que a técnica LIFI seja ineficiente, pois essa avaliação será realizada com a implementação dos algoritmos de classificação e validação. As técnicas LIFS e EM-365 apresentaram uma característica valiosa: o acumulo significativo de variância logo nas primeiras componentes. Essa característica não foi observada para o sistema LIFI, onde se constatou uma variância acumulada mais bem distribuída entre as componentes.

Outra consideração importante versa sobre a disposição dos grupos. A variância da classe Sadia visualizada nessa projeção é maior do que a da classe Assintomática FA. Esse resultado não seria condizente com a natureza física do problema. Porém, vale a ressalva de que essa projeção acumula aproximadamente 50% da variância total dos dados, isto é, metade da informação da variância está contida em componentes não projetadas.

A título de comparação, também foram avaliados os *loadings* para a primeira componente principal, que visivelmente é a que melhor contribui para a segregação dos dados. A disposição dos 46 atributos relevantes para a formação dessa componente, bem como a contribuição relativa de cada um, é apresentada na Tabela 9.

Os blocos de atributos aqui apresentados estão dispostos, em ordem decrescente, nas seguintes porcentagens: espaço de cor RGB com 32,6%, espaço de cor La*b* com 26,1%, cor dominante com 19,6%, espaço de cor HSV com 15,2% e por fim os atributos LBP com 6,5%. Em comparação com a HVRF, destacam-se a diminuição na quantidade de atributos HSV e um aumento de atributos referentes à cor dominante.

Tabela 9 – Nome e módulo da contribuição dos 46 principais atributos na direção da primeira componente principal para os dados LIFI da doença FA.

	Atributos	|Contribuição| (1º CP)		Atributos	|Contribuição| (1º CP)
1	media_G	0,20	24	ASM_G	0,16
2	media_L	0,20	25	media_R	0,15
3	desvio_b*	0,20	26	skewness_R	0,14
4	homogeneidade_b*	0,20	27	ASM_b*	0,14
5	media_B	0,19	28	variancia_a*	0,13
6	variancia_b*	0,19	29	kurtosis_B	0,13
7	dissimilaridade_b*	0,19	30	desvio_a*	0,13
8	dcolor_G3	0,19	31	kurtosis_b*	0,12
9	dcolor_G2	0,19	32	homogeneidade_V	0,12
10	dcolor_G1	0,19	33	homogeneidade_R	0,12
11	energia_B	0,19	34	dcolor_R3	0,12
12	ASM_B	0,19	35	lbp_1	0,1
13	dcolor_B1	0,18	36	skewness_V	0,09
14	media_V	0,17	37	desvio_R	0,09
15	dcolor_R1	0,17	38	variancia_H	0,08
16	dcolor_B3	0,17	39	kurtosis_a*	0,08
17	dcolor_B2	0,17	40	variancia_R	0,08
18	variancia_B	0,16	41	lbp_3	0,07
19	desvio_B	0,16	42	lbp_24	0,05
20	contraste_b*	0,16	43	f_dcolor_3	0,02
21	variancia_S	0,16	44	kurtosis_R	0,01
22	energia_G	0,16	45	media_S	0,01
23	desvio_S	0,16	46	kurtosis_L	≈ 0

Fonte: Elaborada pelo autor.

Por fim, os atributos selecionados foram também utilizados na elaboração de classificadores e predição dos dados. Nessa etapa, o melhor desempenho ocorreu para o algoritmo SVM, com núcleo polinomial de ordem 2 e padronização da matriz de dados. O resultado final do treinamento do modelo apresentou uma acurácia total de 94% com uma sensibilidade

de 91% e especificidade de 97%. Já a predição do grupo de teste resultou numa acurácia de 92,5%, com sensibilidade de 96% e especificidade de 89%. Destaca-se negativamente a taxa de falsos positivos do teste de predição (11%). Esse índice é superior à mesma taxa alcançada na validação cruzada do modelo, com apenas 3%. Apesar desse resultado inesperado, pode-se concluir que no geral os resultados foram satisfatórios, com valores de acurácia acima de 90% para treino e teste, além de matrizes de confusão balanceadas, configurando assim bom funcionamento do modelo, além da prevenção de *overfitting*. A Figura 43 mostra as matrizes de confusão geradas pelo treinamento dos modelos e do teste de predição.

A)

	Sadia	***FA***
Sadia	97%	9%
FA	3%	91%

B)

	Sadia	***FA***
Sadia	89%	4%
FA	11%	96%

Figura 43 – Matrizes de confusão para a classificação das amostras assintomáticas FA com dados LIFI. A) Matriz de confusão gerada pela validação cruzada e B) Matriz de confusão da predição do grupo de teste.
Fonte: Elaborada pelo autor.

Comparando com os resultados da HVRF, houve uma perda de qualidade nos valores de acurácia total, sensibilidade e especificidade. Porém, o desempenho inferior para a FA não inviabiliza o uso a técnica LIFI para aplicação em campo. Primeiramente, porque na comparação das duas doenças, os resultados do LIFI para a HVRF foram superiores às demais técnicas LIFS e EM-365. Outro ponto é a quantidade de amostras utilizado em ambos os casos. O ensaio da FA gerou menos amostras que a HVRF e, obviamente, menos informação torna a etapa de classificação menos confiável. Vale sempre considerar as dificuldades no desenvolvimento dos ensaios com as plantas de soja e os empecilhos do transporte delas. Levando isso em conta, e somando ao fato de que se trata de amostras assintomáticas, uma acurácia de 92,5% para o teste de predição da técnica LIFI para a FA não pode ser considerado mal resultado.

Essa discussão finaliza as análises para a doença FA com considerações importantes. Todas as técnicas de fluorescência obtiveram acurácia total superiores a 90%. Esse resultado comprova a capacidade da fluorescência em diagnosticar a doença FA em estágio assintomático.

4.7 COMPARAÇÃO DAS TRÊS CLASSES

A última análise apresentada no trabalho são as predições para modelos de classificação com as três classes – Sadia, Assintomática HVRF e Assintomática FA – para as técnicas LIFS, EM-365 e LIFI. Para isso, foram utilizadas apenas as coletas 7, 8 e 9, que fazem referência ao mesmo período de desenvolvimento da planta e possuem amostras das três classes. Neste caso, a classe Assintomática HVRF corresponde aos períodos 21, 24 e 28 dai, enquanto a classe Assintomática FA são as próprias amostras utilizadas anteriormente, referente aos períodos 4, 7 e 11 dai. As etapas de remoção de outliers, seleção de atributos e os critérios para formação dos grupos de treino e teste, seguiram a mesma metodologia para a análise de duas classes. Aqui se optou pela exibição apenas dos resultados finais de predição dos grupos de teste para as instrumentações LIFS, EM-365 e LIFI, que estão exibidos, respectivamente, nas Figura 44, Figura 45 e Figura 46.

A)

	Sadia	***HVRF***	***FA***
Sadia	97%	3%	3%
HVRF	2%	94%	3%
FA	1%	3%	94%

B)

	Sadia	***HVRF***	***FA***
Sadia	96%	4%	0%
HVRF	0%	92%	4%
FA	4%	4%	96%

Figura 44 – Matrizes de confusão A) para validação dos modelos de classificação e B) para predição do conjunto de teste das amostras Sadia, Assintomática HVRF e Assintomáticas FA para os espectros do LIFS.
Fonte: Elaborada pelo autor.

A)

	Sadia	***HVRF***	***FA***
Sadia	91%	6%	4%
HVRF	6%	88%	7%
FA	3%	6%	89%

B)

	Sadia	***HVRF***	***FA***
Sadia	75%	0%	15%
HVRF	4%	83%	15%
FA	21%	17%	70%

Figura 45 – Matrizes de confusão A) para validação dos modelos de classificação e B) para predição do conjunto de teste das amostras Sadia, Assintomática HVRF e Assintomáticas FA para as imagens do EM-365.
Fonte: Elaborada pelo autor.

A)

	Sadia	***HVRF***	***FA***
Sadia	93%	0%	9%
HVRF	0%	98%	2%
FA	7%	2%	89%

B)

	Sadia	***HVRF***	***FA***
Sadia	89%	4%	4%
HVRF	7%	92%	4%
FA	4%	4%	92%

Figura 46 – Matrizes de confusão A) para validação dos modelos de classificação e B) para predição do conjunto de teste das amostras Sadia, Assintomática HVRF e Assintomáticas FA para as imagens do LIFI.
Fonte: Elaborada pelo autor.

Os resultados de predição das três classes, referentes a cada instrumentação, também apresentaram respostas consistentes, assim como obtido para as predições de duas classes. A principal contribuição desse resultado é a comprovação da capacidade diagnóstica da fluorescência, que foi capaz de distinguir diferentes classes assintomáticas. Isso significa que os atributos utilizados conseguem representar coerentemente cada classe analisada, o que amplia significativamente as possibilidades de análise, podendo extrapolar para outras necessidades agrícolas, tais como diferentes doenças, diferentes variedades de soja e déficits nutricionais.

5 CONCLUSÕES

A primeira conclusão versa sobre a hipótese inicial do trabalho, com relação a fluorescência das folhas assintomáticas. Com base em toda explanação dos resultados, foi possível concluir que a fluorescência das folhas de soja é capaz de identificar variações químicas das plantas contaminadas com HVRF e FA com relação a plantas sadias, logo nos primeiros dias após a inoculação. O experimento também mostrou a viabilidade do uso da metodologia descrita no diagnóstico de ambas as doenças em campo. Ambas as instrumentações LIFI e LIFS obtiveram resultados de acurácia para predição do grupo de teste superiores a 90%, com baixa exclusão de *outliers* e sem problemas de *overfitting*.

Cabe destacar uma importante consideração sobre os resultados de predição das imagens do EM-365. A escolha do estereomicroscópio se deu por sua fonte de excitação em 365 nm, faixa próxima ao ideal para fluorescência dos metabólitos secundários, e sua capacidade de ampliação de imagens e detalhamento. Porém, as taxas de acurácia obtidas, em ambos os casos, não foram superiores às instrumentações menos robustas igualmente testadas. A experiência na utilização dessas imagens permite concluir que esse desempenho inferior está associado às dificuldades operacionais ocorridas durante a aquisição das imagens. Dentre as adversidades, pode-se citar a necessidade periódica de adaptação dos parâmetros das imagens, a fragilidade das amostras e seu comportamento variado durante as aquisições. De qualquer maneira, os índices da validação cruzada e predição estão aceitáveis, e também corroboram para comprovar a capacidade discriminatória da fluorescência para o presente caso.

Com relação aos dados LIFS, a acurácia total do teste de predição foi de 94% para a HVRF e para a FA. Em ambos os casos, as bandas em 520 e 740 nm foram os principais atributos para discriminação das classes. A análise da contribuição dos atributos para a formação da 1ª CP revelou uma correlação inversa das bandas, onde o aumento de emissão observado em 520 nm reflete numa diminuição da emissão em 740 nm. Houve também resultados similares entre as doenças na decomposição PCA. O resultado satisfatório do teste de predição com as três classes sugere ainda que esse comportamento possa ser um padrão para contaminações de soja.

Outra conclusão nasce da comparação LIFI com o LIFS, ambos de excitação em 405 nm. A aquisição dos espectros LIFS é realizada a partir do contato direto e pontual com a folha, e isso gera uma informação precisa. Porém, essa análise é feita de maneira pontual, restringindo a informação a uma única região. Já a aquisição de imagens do sistema LIFI é capaz de analisar área foliar total, garantindo assim uma análise global das características da emissão de

fluorescência. Apesar de a câmera ser posicionada a certa distância da amostra, e isso imputar dissipação luminosa e, consequentemente, de informação, as emissões de fluorescência apresentam robustez suficiente para seu uso na caracterização de amostras assintomáticas. Tendo em vista a alta taxa de acerto de ambas as técnicas, pode-se dizer que o problema da discriminação das assintomáticas foi bem caracterizado. Portanto, na comparação entre as duas, concluir-se que ambas são robustas o suficiente para o diagnóstico de assintomáticas, e que a escolha de uma ou outra dependerá das necessidades da lavoura.

O conjunto de atributos das imagens de fluorescência propostos nesse trabalho faz referência à coloração e a textura das mesmas. Conforme apresentado nos resultados, tais atributos extraem satisfatoriamente informações relevantes para a discriminação das classes. Como sugestão de evolução do aparato, propõe-se inicialmente a introdução de outra fonte de excitação no UV, como o 355 nm. Esse comprimento de onda é crucial para a excitação apropriada da região dos metabólitos secundários, centrados em 440 e 520 nm. Sabe-se que o comprimento de onda em 405 nm não é a excitação ideal da faixa de tais metabólitos, mas conforme visto no trabalho, sua fluorescência é importante para a caracterização dos estágios assintomáticos. Com a utilização de duas fontes de excitação, uma dedicada às clorofilas e carotenoides, e outra dedicada aos metabólitos secundários, a capacidade de diagnóstico pode ser ampliada. Além disso, a utilização de filtros em regiões específicas pode melhorar o acesso às informações de fluorescência desejada. Um conjunto de filtros passa-banda e passa-alta relacionados às faixas espectrais relevantes poderia gerar não apenas uma imagem RGB, mas também diversas imagens espectrais, o que facilitaria o acesso à informação de fluorescência de maneira mais precisa, além de ampliar a quantidade de atributos para a elaboração dos classificadores. Por fim, novas fontes de excitação e novos tipos de imagens requerem adequação dos atributos, a fim de maximizar a qualidade da informação. Os algoritmos e atributos propostos nesse trabalho podem servir como base para variações e melhorias futuras do sistema como um todo.

Por fim, conclui-se o trabalho destacando os avanços obtidos com o presente estudo: o principal avanço está no ineditismo da caracterização das assintomáticas para soja por três distintas ferramentas de fluorescência. Outra contribuição está na compilação metodológica de técnicas computacionais para análise dos dados espectrais e de imagens voltados para a classificação, que funcionou adequadamente para todos os casos testados, podendo ser utilizada e adaptada para outras necessidades em trabalhos distintos.

REFERÊNCIAS

1 EMBRAPA SOJA. **Soja em números (safra 2017/2018)**. Disponível em: <https://www.embrapa.br/soja/cultivos/soja1/dados-economicos>. Acesso em: 6 jun. 2019.

2 KREYCI, P. F.; MENTEN, J. O. M. **Limitadores de produtividade.** 2013. Disponível em: <https://www.slideshare.net/AgriculturaSustentavel/ccas-caderno-tcnico-cultivar>. Acesso em: 09 nov. 2019

3 HENNING, A. A. et al. **Manual de identificação de doenças de soja**. Disponível em: <https://www.agencia.cnptia.embrapa.br/Repositorio/Doc256_000g0qwdrfk02wx5ok026zxpgrjzggx0.pdf>. Acesso em: 09 nov. 2019

4 EMBRAPA. **Ferrugem da soja:** manejo e prevenção. Disponível em: < https://www.embrapa.br/soja/ferrugem >. Acesso em: 09 nov. 2019

5 GODOY, C. V. **Ferrugem asiática da soja:** identificação e sintomas. Monitoramento. Manejo. Estratégias antirresistência. Vazio sanitário. Fungicidas para o controle. Resistência do fungo aos fungicidas . Londrina: Embrapa Soja, 2014. Disponível em: < https://ainfo.cnptia.embrapa.br/digital/bitstream/item/112852/1/010001.pdf>. Acesso em: 06 jun. 2019.

6 YORINORI, J. T. et al. **Ferrugem da soja:** identificação e controle. Londrina: Embrapa Soja, 2003. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/59588/1/Documentos-204.pdf>. Acesso em: 23 jan. 2018.

7 EMBRAPA. **Soja Louca II é reconhecida como nova doença da soja pelo Mapa**. Disponível em: <https://www.embrapa.br/busca-de-noticias/-/noticia/5213621/soja-louca-ii-e-reconhecida-como-nova-doenca-da-soja-pelo-mapa>. Acesso em: 6 jun. 2019.

8 EMBRAPA. **Feijão e algodão são hospedeiros do nematoide causador da Soja Louca II**. Disponível em: <https://www.embrapa.br/busca-de-noticias/-/noticia/29877986/feijao-e-algodao-sao-hospedeiros-do-nematoide-causador-da-soja-louca-ii>. Acesso em: 6 jun. 2019.

9MOLIN, J. P.; AMARAL, L. R.; COLAÇO, A. F. **Agricultura de precisão**. São Paulo: Oficina de Textos, 2015.

10 MILORI, D. M. B. P. et al. Aplicações agroambientais das técnicas fotônicas. In: NAIME, J. M. (Ed.). **Conceitos e aplicações da instrumentação para o avanço da agricultura**. Brasilia: EMBRAPA, 2014. p. 47–76.

11 MILORI, D. M. B. P. et al. Identification of citrus varieties using laser-induced fluorescence spectroscopy (LIFS). **Computers and Electronics in Agriculture**, v. 95, p. 11–18, 2013.

12 PEREIRA, F. M. V. et al. Laser-induced fluorescence imaging method to monitor citrus greening disease. **Computers and Electronics in Agriculture**, v. 79, n. 1, p. 90–93, 2011.

13 RANULFI, A. C. et al. Laser-induced fluorescence spectroscopy applied to early diagnosis of citrus Huanglongbing. **Biosystems Engineering**, v. 144, p. 133–144, 2016.

14 REITZ, J. R.; MILFORD, F. J.; ROBERT, W. C. **Fundamentos da teoria eletromagnética**. Rio de Janeiro: Editora Campus, 1982.

15 EISBERG, R. M. **Modern physics**. New York: John Wiley and Sons, 1961.
16 LAKOWICZ, J. R. (ED. . **Principles of fluorescence spectroscopy**. Berlin: Springer Science, 2013.

17 ALEUR, B. **Molecular fluorescence:** digital encyclopedia of applied physics. Weinheim: Wiley, 2003.

18 HEINE, J.; MÜLLER-BUSCHBAUM, K. Engineering metal-based luminescence in coordination polymers and metal-organic frameworks. **Chemical Society Reviews**, v. 42, n. 24, p. 9232–9242, 2013.

19 TAIZ, L., ZEIGER, E., MØLLER, I. M., & MURPHY, A. **Fisiologia vegetal**. Porto Alegre: Artmed, 2017.

20 ACADEMY, K. **Luz e pigmentos fotossintéticos**. Disponível em: <https://pt.khanacademy.org/science/biology/photosynthesis-in-plants/the-light-dependent-reactions-of-photosynthesis/a/light-and-photosynthetic-pigments>. Acesso em: 09 nov. 2019

21 STREIT, N. M. et al. As clorofilas. **Ciência Rural**, v. 35, n. 3, p. 748–755, 2005.

22 KRAUSE, G. H.; WEIS, E. Chlorophyll fluorescence as a tool in plant physiology - II. Interpretation of fluorescence signals. **Photosynthesis Research**, v. 5, n. 2, p. 139–157, 1984.

23 BĄBA, W. et al. Discovering trends in photosynthesis using modern analytical tools: More than 100 reasons to use chlorophyll fluorescence. **Photosynthetica**, v. 57, n. 2, p. 668–679, 2019.

24 BUSCHMANN, C.; LICHTENTHALER, H. K. Principles and characteristics of multi-colour fluorescence imaging of plants. **Journal of Plant Physiology**, v. 152, n. 2–3, p. 297–314, 1998.

25 BUSCHMANN, C. Variability and application of the chlorophyll fluorescence emission ratio red/far-red of leaves. **Photosynthesis Research**, v. 92, n. 2, p. 261–271, 2007.

26 ROLFE, S. A.; SCHOLES, J. D. Chlorophyll fluorescence imaging of plant-pathogen interactions. **Protoplasma**, v. 247, n. 3, p. 163–175, 2010.

27 STOBER, F.; LICHTENTHALER, H. K. Characterization of the laser-induced blue, green and red fluorescence signatures of leaves of wheat and soybean grown under different irradiance. **Physiologia Plantarum**, v. 88, n. 4, p. 696–704, 1993.

28 AGATI, G. et al. The F685/F730 Chlorophyll Fluorescence Ratio as a Tool in Plant Physiology: Response to Physiological and Environmental Factors*. **Journal of Plant Physiology**, v. 145, n. 3, p. 228–238, 1995.

29 HUMPLÍK, J. F. et al. Cell wall bound ferulic acid, the major substance of the blue-green fluorescence emission of plants. **Plant Methods**, v. 11, n. 1, p. 1–10, 2015.

30 LANG, M.; STOBER, F.; LICHTENTHALER, H. K. Fluorescence emission spectra of plant leaves and plant constituents. **Radiation and Environmental Biophysics**, v. 30, n. 4, p. 333–347, 1991.

31 KULBAT, K. Biotechnology and Food Sciences The role of phenolic compounds in plant resistance. **Biotechnol Food Science**, v. 80, n. 2, p. 97–108, 2016.

32 LICHTENTHALER, H. K.; SCHWEIGER, J. Cell wall bound ferulic acid, the major substance of

the blue-green fluorescence emission of plants. **Journal of Plant Physiology**, v. 152, n. 2–3, p. 272–282, 1998.

33 MEYER, S. et al. UV-induced blue-green and far-red fluorescence along wheat leaves: A potential signature of leaf ageing. **Journal of Experimental Botany**, v. 54, n. 383, p. 757–769, 2003.

34 ROHÁČEK, K. et al. **Chlorophyll fluorescence-** a wonderful tool in plant stress physiology. 2008. Disponível em: <https://www.researchgate.net/profile/K_Rohacek/publication/285891590_Chlorophyll_fluorescence_A_wonderful_tool_to_study_plant_physiology_and_plant_stress/links/570bc70808ae8883a1ffd862.pdf>. Acesso em: 09 nov. 2019

35 LICHTENTHALER, H. K.; WENZEL, O.; BUSCHMANN, C.; GITELSON, A. Plant Stress Detection by Reflectance and Fluorescence. **Annals of the New York Academy of Sciences**, v. 851, n.1, p. 271–285, 1998.

36 PITOL, C. et al. Manejo de doenças na cultura da soja. In: **Tecnologia e Produção:** Soja 2014 / 2015. Curitiba: Midiograf, 2015. p. 161.

37 AGROLINK. **Os sintomas causados pela ferrugem asiática**. Disponível em: <https://www.agrolink.com.br/culturas/soja/informacoes/sintomas_361550.html>. Acesso em: 12 set. 2019.

38 EMBRAPA. **SOJA Louca II é reconhecida como nova doença da soja pelo Mapa**. Disponível em: <https://www.embrapa.br/busca-de-noticias/-/noticia/5213621/soja-louca-ii-e-reconhecida-como-nova-doenca-da-soja-pelo-mapa>. Acesso em: 12 set. 2019.

39 MEYER, M. C. et al. Soybean green stem and foliar retention syndrome caused by Aphelenchoides besseyi. **Tropical Plant Pathology**, v. 42, n. 5, p. 403–409, 2017.

40 RANULFI, A. C. et al. Laser-induced fluorescence spectroscopy applied to early diagnosis of citrus Huanglongbing. **Biosystems Engineering**, v. 144, n. April, p. 133–144, 2016.

41 RUSS, J. C. **The image processing handbook**. Boca Raton: CRC press, 2016.

42 OLIPHANT, T. E. **A guide to NumPy**. USA: Trelgol Publishing, 2006.

43 OLIPHANT, T. E. Python for ccientific computing python overview. **Computing in Science and Engineering**, p. 10–20, 2007.

44 PEDREGOSA, F. et al. Scikit-learn: machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

45 BRADSKI, G. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, v.120. p.122-125, 2000.

46 MARQUES FILHO, O.; VIEIRA NETO, H. **Processamento digital de imagens**. New York: Brasport, 1999.

47 GONZALEZ, R. C., WOOD, R. E. **Digital image processing**. Upper Saddle River: Prentice Hall, 2002.

48 EKSTROM, M. P. **Digital image processing techniques**. New York: Academic Press, 2012.

49 VOICU, L. I. Practical considerations on color image enhancement using homomorphic filtering. **Journal of Electronic Imaging**, v. 6, n. 1, p. 108, 1997.

50 NIKHIL, R. P.; SANKAR, K. P. A review on image segmentation techniques. **Pattern recognition**, v. 26, n. 9, p. 1277–1294, 1993.

51 COSTA, L. F.; CESAR JR, R. M. **Shape analysis and classification:** theory and practice. Boca Raton: CRC Press, Inc., 2000.

52 SOILLE, P. **Morphological image analysis:** principles and applications. Berlin: Springer Science & Business Media, 2013.

53 GUYON, I. ET AL. (ED.). **Feature extraction:** foundations and applications. Berlin: Springer, 2008.

54 NIXON, MARK; AGUADO, A. S. **Feature extraction and image processing for computer vision**. New York: Academic Press, 2012.

55 REED, TOOD R.; HANS DU BUF, J. M. A Review of recent texture segmentation and feature extraction techniques. **CVGIP: Image Understanding**, v. 57, p. 359–372, 1992.

56 IMANI, M.; GHASSEMIAN, H. Feature space discriminant analysis for hyperspectral data feature reduction. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 102, p. 1–13, 2015.

57 ILEA, D. E.; WHELAN, P. F. Image segmentation based on the integration of colourtexture descriptors - A review. **Pattern Recognition**, v. 44, n. 10–11, p. 2479–2501, 2011.

58 ZHANG, Z. et al. Learning completed discriminative local features for texture classification. **Pattern Recognition**, v. 67, p. 263–275, 2017.

59 PETROU, MARIA; SEVILLA, P. G. **Image Processing:** Dealing with texture. New York: John Wiley & Sons, 2006.

60 SIEDLECKI, W.; SKLANSKY, J. A note on genetic algorithms for large-scale feature selection In**:** CHEN, C. H.; PAU, L. F.: WANG, P. S. P. (Ed.). **Handbook of pattern recognition & computer vision**. Singapoure: World Scientific, 1993.

61 BACKES, A. R.; CASANOVA, D.; BRUNO, O. M. Color texture analysis based on fractal descriptors. **Pattern Recognition**, v. 45, n. 5, p. 1984–1992, 2012.

62 HE, D. C.; WANG, L. Texture Unit, Texture Spectrum, and Texture Analysis. **IEEE Transactions on Geoscience and Remote Sensing**, v. 28, n. 4, p. 509–512, 1990.

63 OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on feature distributions. **Pattern Recognition**, v. 29, n. 1, p. 51–59, 1996.

64 WIKIPEDIA. **Local binary patterns**. Disponível em: <https://en.wikipedia.org/wiki/Local_binary_patterns>. Acesso em: 30 set. 2019.

65 KOSCHAN, A.; ABIDI, M. **Digital color image processing image**. Hoboken: John Wiley, 2008.

66 ROBERTSON, A. R. The CIE 1976 Color-Difference Formulae. **Color Research & Application**, v. 2, n. 1, p. 7–11, 1977.

67 LUKAC, R.; PLATANIOTIS, K. N. **Color image processing**: methods and applications. Boca

Raton: CRC Press, 2006.

68 DENG, X. et al. Feature selection for text classification: A review. **Multimedia Tools and Applications**, v. 78, n. 3, p. 3797–3816, 2019.

69 STAŃCZYK, U.; JAIN, L. C. **Feature selection for data and pattern recognition**. Berlin: Springer-Verlag, 2015. (Studies in computational intelligence, v. 584).

70 LEE, H. D. **Seleção de atributos importantes para a extração de conhecimento de bases de dados**. 2005. 154p. Tese (Doutorado em Ciências) - Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos, 2005.

71 SHALEV-SHWARTZ, S. et al. **Understanding machine learning**. New York: Cambridge University Press, 2014.

72 BREIMAN, L. Random Forests. **Machine learning**, v. 45, p. 5–32, 2001.

73 KE, G. et al. LightGBM: A highly efficient gradient boosting decision tree. **Advances in Neural Information Processing Systems**, v. 2017, p. 3147–3155, 2017.

74 PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space . **Philosophical Magazine and Journal of Science**, v. 2, n. 11, p. 559–572, 1901.

75 BISHOP, C. M. **Pattern recognition and machine learning**. Berlin: Springer Science & Business Media, [s.d.].

76 ALPAYDIN, E. **An introduction to machine learning**. Cambridge: MIT Press, 2014.

77 FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006.

78 SEBE, N. et al. **Machine learning in computer vision**. Netherlands: Springer, 2005. (Computational imaging and vision, v. 29)

79 SPILIOPOULOU, M.; SCHMIDT-THIEME, L.; JANNING, R. **Data analysis, machine learning and knowledge discovery**. New York: Springer, 2014.

80 ZHAO, Y.; NASRULLAH, Z.; LI, Z. PyOD: a Python toolbox for scalable outlier detection. **Journal of Machine Learning Research**, v. 20, n. 96, p. 1–7, 2019.