UNIVERSIDADE FEDERAL DE MATO GROSSO
DO SUL CAMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM
AGRONOMIA

MARINA VALENTINI ARF

APLICAÇÃO DE ESPECTROSCOPIA VNIR-SWIR E APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DA QUALIDADE DE GRÃOS DE ARROZ EM UNIDADES ARMAZENADORAS E INDÚSTRIAS BENEFICIADORAS

Chapadão do Sul - MS
2025

UNIVERSIDADE FEDERAL DE MATO GROSSO
DO SUL CAMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM
AGRONOMIA

MARINA VALENTINI ARF

APLICAÇÃO DE ESPECTROSCOPIA VNIR-SWIR E APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DA QUALIDADE DE GRÃOS DE ARROZ EM UNIDADES ARMAZENADORAS E INDÚSTRIAS BENEFICIADORAS

Orientador: Prof. Dr. Paulo Carteri Coradi

Dissertação apresentada à Universidade Federal de Mato Grosso do Sul, como requisito para obtenção do título de Mestre em Agronomia, área de concentração: Produção Vegetal.

Chapadão do Sul – MS
2025

Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

CERTIFICADO DE APROVAÇÃO

DISCENTE: Marina Valentini Arf

ORIENTADOR: Prof. Dr. Paulo Carteri Coradi

TÍTULO: APLICAÇÃO DE ESPECTROSCOPIA VNIR-SWIR E APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DA QUALIDADE DE GRÃOS DE ARROZ EM UNIDADES ARMAZENADORAS E INDÚSTRIAS BENEFICIADORAS.

Documento assinado digitalmente
gov.br
PAULO CARTERI CORADI
Data: 08/11/2025 08:25:21-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Paulo Carteri Coradi (UFSM)

Presidente

Documento assinado digitalmente
gov.br
LARISSA PEREIRA RIBEIRO TEODORO
Data: 10/11/2025 10:46:34-0300
Verifique em https://validar.iti.gov.br

_____

Profa. Dra. Larissa Pereira Ribeiro Teodoro (UFMS)

Documento assinado digitalmente
gov.br
JORGE GONZALEZ AGUILERA
Data: 12/11/2025 11:13:10-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Jorge Gonzalez Aguilera (UEMS)

Chapadão do Sul, 13 de outubro de 2025

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, pela vida, pela saúde e por me conceder sabedoria, força e serenidade em todos os momentos desta caminhada. Por iluminar meus passos, me dar fé diante dos desafios e permitir que este sonho se tornasse realidade.

Aos professores que compuseram a banca examinadora, ao meu orientador Prof. Paulo Carteri Coradi, agradeço imensamente pela disponibilidade, pelas valiosas contribuições e por compartilhar seus conhecimentos, que enriqueceram significativamente este trabalho.

Aos colaboradores, em especial ao Ênio Antônio Manfroi Filho, que dedicaram tempo, paciência e empenho para a realização desta pesquisa. Suas orientações, conselhos e incentivos foram fundamentais para o desenvolvimento deste estudo e para o meu crescimento acadêmico e pessoal.

Aos colegas de laboratório e parceiros de pesquisa, obrigada pela troca de experiências, pelas conversas produtivas e pelo apoio mútuo durante todas as etapas do trabalho.

Aos meus familiares, pelo amor incondicional, pela compreensão nos momentos de ausência e por acreditarem em mim mesmo quando eu duvidava. O apoio de vocês foi meu alicerce e motivação para seguir em frente.

Aos amigos, que estiveram presentes nas horas boas e ruins, oferecendo palavras de incentivo, risadas e companhia. A amizade e o carinho de vocês tornaram essa trajetória mais leve e significativa.

Agradecer em especial também a empresa Nutriplant, por me apoiar nessa decisão e me dar total suporte para mais essa conquista em minha vida.

A todos que, de alguma forma, contribuíram para a realização deste trabalho, o meu mais sincero muito obrigada.

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O arroz (*Oryza sativa* L.) é um alimento essencial na dieta mundial e sua qualidade físico-química é um fator determinante na cadeia produtiva. O objetivo deste estudo é avaliar a capacidade de predição e caracterização da qualidade físico-química de diferentes tipos de arroz por meio de sensores hiperespectrais e algoritmos de aprendizado de máquina. Foram utilizadas amostras de arroz dos tipos Branco, Preto, Vermelho e Parboilizado, analisadas por espectroscopia hiperespectral (350–2500 nm) e submetidas a Regressão Linear; modelos tradicionais: Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB); e profundos: Redes Neurais Convolucionais (CNN) e Recorrentes (RNN) de aprendizado de máquina. Os dados espectrais e físico-químicos foram explorados com técnicas de análise multivariada (PCA) e validação cruzada com métricas Coeficiente de Correlação (r); Coeficiente de Correlação ao quadrado (r^2); Mean Absolute Error (MAE); e Root Mean Square Error (RMSE). Os resultados indicaram que os modelos SVM, RF e GB apresentaram desempenho superior, com r e r^2 variando entre 0,95 e 1,0 e MAE e RMSE abaixo de 0,2, enquanto os modelos profundos apresentaram desempenho inferior, especialmente em bases de dados com volume moderado. O arroz preto destacou-se por altos teores de proteínas (~9), lipídios (~2) e cinzas(~1,45); o parboilizado apresentou maior teor de fibras (~2,8), enquanto o arroz branco se destacou pelo teor de amido(~73). A espectroscopia hiperespectral demonstrou ser eficaz na diferenciação entre os tipos de arroz, permitindo a seleção de bandas relevantes para sensores otimizados. Conclui-se que o uso de tecnologias não destrutivas integradas com aprendizado de máquina é promissor para o controle de qualidade do arroz no setor industrial, oferecendo rapidez, precisão e sustentabilidade no processo pós-colheita.

Palavras-chave: Aprendizado de máquina. Espectroscopia NIR. Métodos não destrutivos. Qualidade físico-química.

ABSTRACT

Rice (*Oryza sativa* L.) is an essential food in the global diet, and its physicochemical quality is a determining factor in the production chain. The objective of this study is to evaluate the predictive and characterization capacity of the physicochemical quality of different rice types using hyperspectral sensors and machine learning algorithms. Samples of White, Black, Red, and Parboiled rice were analyzed through hyperspectral spectroscopy (350–2500 nm) and subjected to Linear Regression; traditional models such as Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB); and deep learning models, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Spectral and physicochemical data were explored using multivariate analysis techniques (PCA) and cross-validation with the following metrics: Correlation Coefficient (r); Coefficient of Determination (r^2); Mean Absolute Error (MAE); and Root Mean Square Error (RMSE). The results indicated that the SVM, RF, and GB models showed superior performance, with *r* and r^2 ranging from 0.95 to 1.0 and MAE and RMSE below 0.2, while deep learning models showed lower performance, especially with datasets of moderate size. Black rice stood out for its high levels of protein (~9), lipids (~2), and ash (~1.45); parboiled rice showed a higher fiber content (~2.8), while white rice was characterized by its starch content (~73). Hyperspectral spectroscopy proved to be effective in differentiating rice types, allowing the selection of relevant bands for optimized sensor development. It is concluded that the use of non-destructive technologies integrated with machine learning is promising for quality control in the rice industry, providing speed, accuracy, and sustainability in the post-harvest process.

Keywords: Machine learning. NIR spectroscopy. Non-destructive methods. Physicochemical quality.

SUMÁRIO

1. INTRODUÇÃO

O arroz (*Oryza sativa* L.) é um dos principais cereais cultivados e consumidos no mundo, ocupando posição de destaque tanto na segurança alimentar quanto na economia agrícola. No contexto do agronegócio, o arroz representa uma cadeia produtiva estratégica, que envolve desde o cultivo no campo até a comercialização e exportação de derivados industrializados (DE OLIVEIRA, 2021). Sua relevância é particularmente marcante em países em desenvolvimento, onde constitui a base alimentar da população e gera milhares de empregos diretos e indiretos.

O Brasil é um dos maiores produtores do grão fora da Ásia, com destaque para os estados do Rio Grande do Sul, Santa Catarina, Tocantins, Maranhão e Mato Grosso. A produção brasileira atende majoritariamente ao mercado interno, mas também há exportações significativas para América Latina, África e partes da Europa e Ásia (CONAB, 2025). Além da relevância econômica, o arroz tem grande importância social. É o alimento básico para mais da metade da população mundial, sendo consumido em praticamente todas as regiões do Brasil. Sua produção contribui para o combate à fome e à insegurança alimentar, especialmente em regiões mais pobres e vulneráveis (DOS SANTOS, 2006).

Os tipos de arroz abordados neste trabalho diferenciam-se por cores, beneficiamento e compostos bioquímicos. O arroz branco é obtido a partir do grão integral submetido a processos de beneficiamento que removem a casca, o farelo e o gérmen, resultando em um produto de coloração branca, com maior tempo de conservação e textura mais macia após o cozimento (RATHNAYAKE et al., 2014).

Do ponto de vista nutricional, o arroz branco é composto majoritariamente por carboidratos (aproximadamente 80%), com baixo teor de fibras, lipídios e micronutrientes, uma vez que grande parte desses componentes está concentrada nas partes removidas durante o polimento por esse motivo, embora seja uma importante fonte energética, seu valor nutricional é inferior ao de outras variedades menos processadas, como o arroz integral, vermelho ou preto.

Diferente do branco, o arroz vermelho não é submetido ao processo de polimento, o que permite a preservação de fibras alimentares, vitaminas do complexo B e minerais essenciais como ferro, magnésio e zinco. Os pigmentos fenólicos responsáveis pela coloração avermelhada, como proantocianidinas e antocianinas, ficam concentrados principalmente no pericarpo (FĂRCAȘ et al., 2022), e possuem alta capacidade antioxidante, que pode contribuir

na prevenção de doenças crônicas, como doenças cardiovasculares, diabetes tipo 2 e câncer (MALUMPONG et al., 2021).

O arroz preto é uma variedade que se destaca por sua coloração escura, aroma característico e alto valor nutricional. Seu pigmento roxo-escuro a preto se deve à presença de antocianinas, compostos fenólicos com forte ação antioxidante, especialmente concentrados no pericarpo do grão (KABIR et al., 2024). Além das antocianinas, o arroz preto é rico em fibras alimentares, proteínas, vitaminas do complexo B e minerais como ferro e zinco. Estudos têm demonstrado que sua composição química favorece a saúde cardiovascular, possui potencial anti-inflamatório e pode auxiliar na prevenção de doenças crônicas, como diabetes tipo 2 e certos tipos de câncer (MIN et al., 2012).

Já o arroz parboilizado (ou parcialmente fervido) é obtido por meio de um processo hidrotérmico que envolve três etapas principais: embebição, tratamento térmico (vaporização) e secagem. Esse processo é realizado antes do descascamento e polimento do grão, fazendo com que parte dos nutrientes presentes no farelo e no gérmen migrem para o endosperma, melhorando assim seu valor nutricional mesmo após o beneficiamento. Por este motivo, o arroz parboilizado apresenta maior teor de vitaminas do complexo B, como tiamina, riboflavina e niacina, além de conter mais minerais, como magnésio e fósforo quando comparado ao arroz branco (JULCARIMA et al., 2018). A composição química do arroz abrange macronutrientes e micronutrientes essenciais para a dieta humana, entre eles o amido, proteínas, fibras alimentares, lipídios, minerais (cinzas) e água, e a proporção de tais substâncias reflete na qualidade nutricional que o produto pode ofertar.

O amido é o principal constituinte do arroz, representando aproximadamente 70% a 80% do peso do grão. É composto por duas frações: amilose e amilopectina. O teor de amilose varia entre 10% e 30% (BERNARDO et al., 2010). As proteínas correspondem de 7% a 10% da composição do grão e são consideradas de alta qualidade biológica, pois apresentam boa digestibilidade e aminoácidos equilibrados, com destaque para a lisina em comparação a outros cereais (FAO, 1991).

As fibras alimentares, ainda que presentes em menores quantidades no arroz polido, desempenham papel importante na saúde intestinal (SOUZA et al., 2011). A remoção do farelo durante o processo de beneficiamento reduz significativamente o conteúdo de fibras. As cinzas representam a fração inorgânica do alimento e indicam o conteúdo de minerais, como ferro, zinco, magnésio e fósforo. O teor de cinzas no arroz polido geralmente é inferior a 1%, sendo maior no arroz integral (NACHTIGALL et al., 2004).

Os lipídios do arroz estão concentrados no germe e nas camadas externas do grão, sendo, portanto, mais abundantes no arroz integral. Em geral, o teor lipídico do arroz polido varia de 0,3% a 0,5%, enquanto no arroz integral pode ultrapassar 2% (SCHAFFERT et al., 2011). Esses lipídios são em sua maioria insaturados, destacando-se os ácidos oleico e linoleico. Por fim, a umidade é um parâmetro essencial para a conservação do arroz. Valores acima de 14% podem favorecer o crescimento de microrganismos e reduzir a estabilidade do produto durante o armazenamento (BRASIL, 2011). A umidade ideal para armazenamento do arroz gira em torno de 12% a 14%.

A análise da qualidade do arroz é um fator essencial para garantir a segurança alimentar, a competitividade no mercado e a satisfação do consumidor. O arroz ocupa um lugar destacado na alimentação básica da população brasileira, fornecendo um relevante aporte de calorias e proteínas, especialmente para o estrato de baixa renda (Associação Brasileira das Indústrias de Arroz (ABIA), 2023).

Tradicionalmente, essa avaliação era realizada por métodos destrutivos, como a moagem e a cocção do grão, que inviabilizavam o uso posterior do produto. O teste de cocção em arroz é um dos parâmetros de qualidade muito utilizado por programas de melhoramento genético e indústrias de beneficiamento como forma de avaliar o comportamento culinário das cultivares lançadas e/ou novas linhagens em estudo (EMBRAPA, 2020).

No entanto, os avanços na tecnologia permitiram o desenvolvimento de técnicas de análise não destrutiva, que vêm se consolidando como ferramentas eficazes e sustentáveis na cadeia de produção e comercialização do arroz. A espectroscopia de infravermelho próximo (NIR) tem sido aplicada como método complementar à classificação física, correlacionando defeitos com a composição físico-química nos grãos (CONBEA, 2023).

Essas técnicas não destrutivas representam um avanço significativo na avaliação da qualidade do arroz, oferecendo soluções modernas e sustentáveis que atendem às exigências do mercado e da segurança alimentar. Métodos como a espectroscopia de infravermelho próximo e a análise de imagens computadorizadas têm sido utilizados para classificar o vigor de lotes de sementes de arroz com alta precisão (Locus UFV, 2023).

Essas tecnologias oferecem diversas vantagens, como rapidez nos resultados, redução de desperdício, padronização da análise e maior segurança na tomada de decisões durante o

processamento. A espectroscopia de infravermelho próximo, por exemplo, permite a predição rápida do potencial fisiológico das sementes, com acurácia de 96% (Locus UFV, 2023).

Além disso, contribuem para a rastreabilidade e para o controle de qualidade em tempo real. A implementação de sistemas de monitoramento em tempo real, como os baseados em espectroscopia, tem sido cada vez mais importante na agricultura e na indústria alimentícia (Instituto Nacional de Investigação Agrária e Veterinária (INIAV), 2024).

Entretanto, o padrão adotado de classificação física dos grãos pode gerar dúvidas, controvérsias e erros significativos nos níveis de qualidade, por ser um método subjetivo e interpretativo. A classificação física é a operação realizada para avaliação da qualidade física dos grãos de milho para comercialização, entretanto, devido à subjetividade dos métodos, pode haver variações nos resultados (CONBEA, 2023).

Portanto, o desenvolvimento de novos métodos para melhorar a assertividade da medição da qualidade dos grãos, aumentar a operacionalidade do processo e melhorar os fluxos dos lotes de grãos nas etapas de pré-processamento e armazenamento são requeridos dentro do sistema agroindustrial. A adoção de tecnologias como a espectroscopia de infravermelho próximo tem mostrado ser uma alternativa eficaz para a avaliação da qualidade do arroz integral e polido, complementando a classificação física tradicional (CONBEA, 2023).

As análises não destrutivas baseiam-se em métodos que preservam a integridade do grão, permitindo sua avaliação física, química e até sensorial sem a necessidade de descarte. O ultrassom, por exemplo, é uma tecnologia limpa, rápida, não invasiva, não destrutiva e precisa, sendo ecologicamente correta e utilizada para melhorar diversas características nos alimentos (Universidade Federal de Pelotas (CTI/UFPEL), 2023).

Os métodos ópticos de medição indireta podem ser uma boa alternativa para análises não destrutivas da qualidade de grãos pós-colheita. A espectroscopia de infravermelho próximo tem sido aplicada como método complementar à classificação física, correlacionando defeitos com a composição físico-química nos grãos (CONBEA, 2023).

Esses métodos apresentam algumas vantagens, como baixos custos a médio-longo prazo, simplicidade operacional, não destruição das amostras analisadas, além de oferecer análises rápidas e sem necessidade (ou com um mínimo) de preparo das amostras de caráter

não invasivo. A espectroscopia de infravermelho próximo, por exemplo, permite a predição rápida do potencial fisiológico das sementes, com acurácia de 96% (Locus UFV, 2023).

As técnicas mais utilizadas dentre as análises não destrutivas são: Espectroscopia no infravermelho próximo (NIR); Ressonância magnética nuclear (RMN) e tomografia por impedância elétrica; e Visão computacional e processamento de imagem. Enquanto estas duas últimas técnicas avaliam a estrutura física interna do grão (QU; JIN, 2022), e o formato, cor, integridade e presença de impurezas (SANTOS et al., 2020), respectivamente, a Espectroscopia no infravermelho próximo (NIR) determina teores de umidade, proteína e amido, além de identificar defeitos internos nos grãos, tudo isso por meio de modelos estatísticos calibrados com métodos de referência (QUEVEDO RAMIREZ, 2023).

A Espectroscopia NIR é uma técnica analítica baseada na interação da radiação eletromagnética com a matéria, especificamente na faixa do infravermelho próximo, que compreende comprimentos de onda entre 780 e 2500 nanômetros (METROHM, 2024). Essa técnica é rápida, não destrutiva e requer pouca ou nenhuma preparação da amostra, o que a torna ideal para aplicações em controle de qualidade, agricultura, indústria alimentícia, farmacêutica e ambiental. Além disso, a técnica tem se destacado em pesquisas por possibilitar análises em tempo real e in situ, como na avaliação da qualidade de grãos, frutas, produtos cárneos e lácteos, sem a necessidade de reagentes químicos ou extrações complexas.

Para o tratamento e interpretação de dados obtidos pela Espectroscopia NIR recomenda-se o uso de métodos quimiométricos, que tornam possível a identificação de substâncias em misturas complexas com maior precisão (BRERETON, 2003). A quimiometria é uma área interdisciplinar que aplica métodos estatísticos e matemáticos à química, com o objetivo de extrair informações relevantes de dados experimentais complexos. Essa abordagem é particularmente útil em análises químicas onde se busca maximizar a extração de informações a partir de grandes volumes de dados, muitas vezes ruidosos ou altamente correlacionados.

Segundo Massart et al. (1997), a quimiometria surgiu da necessidade de interpretar dados químicos de maneira mais eficaz, incorporando técnicas como análise de componentes principais (PCA), regressão por mínimos quadrados parciais (PLS), e análise discriminante. No entanto, essas metodologias podem causar multicolinearidade e grande volume de dados. Para

superar esses problemas é necessário utilizar-se de aprendizado de máquina. Essas ferramentas estatísticas permitem a construção de modelos preditivos robustos.

O avanço das tecnologias de aprendizado de máquina tem proporcionado soluções cada vez mais eficientes para problemas complexos, como a predição de propriedades físico-químicas de produtos agrícolas. Entre os modelos mais utilizados estão os algoritmos tradicionais, como Regressão Linear Múltipla (LR), Support Vector Machine (SVM), Random Forest (RF) e Gradient Boosting (GB), e os modelos de aprendizado profundo, como Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN) (SARKER, 2021; MAGNIMIND ACADEMY, 2023; ZHOU et al., 2023)

A Regressão Linear Múltipla é um modelo estatístico clássico que assume uma relação linear entre variáveis independentes e dependentes. Apesar de sua simplicidade e interpretabilidade, ela apresenta limitações em problemas com relações não-lineares ou alta multicolinearidade entre variáveis (DRAPER & SMITH, 1998).

O modelo SVM é eficaz em problemas de classificação e regressão, especialmente em conjuntos de dados de alta dimensionalidade. Ele busca encontrar o hiperplano ótimo que melhor separa as classes ou realiza a regressão com base em margens máximas, sendo robusto mesmo com poucos dados (CORTES & VAPNIK, 1995).

Modelos baseados em árvores, como Random Forest (RF) e Gradient Boosting (GB), têm sido amplamente aplicados por sua capacidade de modelar não linearidades e interações complexas entre variáveis. O RF combina múltiplas árvores de decisão por meio de bagging, o que reduz o risco de overfitting (BREIMAN, 2001), enquanto o GB realiza uma agregação sequencial de árvores ajustadas aos resíduos do modelo anterior, aumentando a acurácia preditiva (FRIEDMAN, 2001).

No campo do aprendizado profundo, as Redes Neurais Convolucionais (CNN) destacam-se pela capacidade de capturar padrões espaciais e são amplamente utilizadas em imagens, espectros e sinais (LECUN et al., 2002). Sua arquitetura permite detectar automaticamente características relevantes sem a necessidade de engenharia manual de atributos. Já as Redes Neurais Recorrentes (RNN) são mais adequadas a dados sequenciais, como séries temporais e espectros hiperespectrais, pois possuem conexões recorrentes que mantêm estados anteriores e modelam dependências temporais (HOCHREITER &

SCHMIDHUBER, 1997). A variante LSTM (Long Short-Term Memory) tem sido eficaz em superar o problema do desaparecimento do gradiente em sequências longas.

Estudos comparativos indicam que, embora modelos profundos tenham grande potencial, algoritmos tradicionais como RF e SVM muitas vezes superam redes profundas em cenários com poucos dados rotulados, alta variabilidade e necessidade de interpretabilidade (XU et al., 2021; GRINSZTAJN, OYALLON, VAROQUAUX, 2022). Isso é particularmente relevante no contexto agrícola, onde a coleta de dados é difícil e sujeita a intempéries. Assim, a escolha do modelo ideal depende do volume e natureza dos dados, da complexidade do problema e do objetivo final da análise, sendo recomendável realizar testes comparativos e validações cruzadas para garantir desempenho e robustez.

2. OBJETIVOS

O objetivo deste estudo é avaliar a capacidade de predição e caracterização da qualidade físico-química de diferentes tipos de arroz por meio de sensores hiperespectrais e algoritmos de aprendizado de máquina.

2.1 Objetivos específicos

1 - Identificar as bandas espectrais mais representativas para a predição de cada componente físico-químico;

2 - Comparar o desempenho de modelos tradicionais de aprendizado de máquina e modelos de aprendizado profundo na predição dos atributos;

3 - Avaliar bandas mais representativas para construção de sensores multiespectrais otimizados;

4 - Demonstrar a aplicabilidade prática dessas tecnologias como ferramentas não destrutivas para uso no setor pós-colheita, visando otimização de processos industriais de triagem e controle de qualidade.

3. MATERIAL E MÉTODOS

3.1. Obtenção e preparação das amostras

As amostras compostas de arroz beneficiadas foram obtidas logo após o beneficiamento dos grãos (sem armazenamento) em uma unidade de beneficiamento localizado no município de Cachoeira do Sul, no centro do Rio Grande do Sul, na região fisiográfica Depressão Central, ao lado do Rio Jacuí, latitude: 30° 0' 45" S, longitude: 52° 55' 11" W e altitude de 73 metros. Na Figura 1 estão representadas as amostras de arroz Branco, Parboilizado, Vermelho e Preto, que pertencem à mesma espécie, *Oryza sativa* L.

Figura 1. Imagens de amostras de arroz beneficiadas: Arroz Branco (A), Arroz Parboilizado (B), Arroz Vermelho (C) e Arroz Preto (D).

As amostras foram separadas conforme a Instrução Normativa nº 6 de 16 de fevereiro de 2009 com base no regulamento técnico do arroz (Ministério da Agricultura, 2009). As amostras foram avaliadas e constituídas pelos limites máximos de defeitos tolerados para os Tipo 1, baseando-se nas Tabelas 1 (Brasil. Instrução Normativa 2/2012. Ministério da Agricultura, 2012). Foram constituídas amostras de 2kg cada que, posteriormente, foram divididas em 100 subamostras de 20g.

Tabela 1. Limites máximos de tolerância de defeitos expressos em % peso-1 e obtidos nas amostras de arroz empregadas no experimento.

Arroz	1	2	3	4	5	6	7	8
Polido	1	1	0,2	0,5	1,50	1,75	0,3	14
Parboilizado	1	1	0,1	0,15	1,5	1,75	0,3	14
Vermelho	1	1	0,2	0,5	1,5	1,75	0,3	14
Preto	1	1	0,2	0,5	1,5	1,75	0,3	14

Legenda: 1. Tipo; 2. Matérias, estranhas e impurezas; 3. Mofados e ardidos; 4. Picados ou manchados; 5. Gessados verdes; 6. Rajados; 7. Amarelos; 8. Total de quebrados e quirera.

O limite máximo de tolerância admitido para grão não parboilizado é de 0,30% (zero vírgula trinta por cento) para todos os tipos. Acima desse limite o produto será considerado como Fora de Tipo.

3.2. Espectroscopia de Infravermelho Próximo (NIR)

A determinação, em percentual, da composição centesimal, constituída proteína bruta (PB), teor de água (UM), extrato etéreo (LP), fibra bruta (FB), cinzas (CZ) e amido (AM) das amostras foi realizada por espectroscopia de infravermelho próximo (Metrohm, espectrômetro DS2500, Herisau, Suíça), denominado NIR com alta precisão óptica, em triplicata. As amostras foram homogeneizadas e colocadas na cápsula de amostragem.

A análise tende a iluminar uma amostra com radiação de um comprimento de onda específico na região do infravermelho próximo e, em seguida, mede a diferença entre as quantidades de energia emitida pelo espectroscópio e refletida pela amostra para o detector. Essa diferença é medida em várias bandas, criando um espectro para cada amostra. O registro dos dados espectrais foi feito no modo de refletância, na faixa espectral de 350 a 2500 nm (Barnes, Dhanoa e Lister, 1989). O resultado obtido foi comparado a uma curva de calibração (Horwitz W, 1970).

3.3. Análises multivariadas, modelagem preditiva e avaliação estatística

As análises deste estudo foram conduzidas predominantemente no ambiente Python 3.10, utilizando bibliotecas como pandas, numpy, scikit-learn, seaborn, matplotlib e tensorflow. Inicialmente, aplicou-se a correlação de Pearson entre variáveis físico-químicas (Umidade, Proteínas, Lipídios, Fibras, Amido e Cinzas), seguida de análises de componentes principais (PCA), realizadas separadamente para dados físico-químicos e espectrais (350–2500 nm), permitindo visualizar padrões de agrupamento entre os tipos de arroz.

Para as predições, foram aplicados dois grupos de modelos: (1) Modelos de aprendizado de máquina tradicional — Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear Multipla (LR); (2) Modelos de aprendizado profundo — Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN).

Os dados foram previamente padronizados (StandardScaler), e os modelos foram avaliados com validação cruzada K-Fold (10 folds) e dez repetições. As métricas de desempenho incluíram: coeficiente de correlação (r), coeficiente de determinação (r^2), erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE). Para os modelos CNN e RNN, foram utilizados tensores 1D (Conv1D, LSTM), e aplicou-se EarlyStopping para evitar overfitting.

Além das métricas preditivas, foi realizada análise de importância das variáveis por RF, identificando as bandas espectrais mais relevantes para cada característica avaliada. Foram também construídos boxplots para comparar a distribuição dos resultados por modelo, bem como análises de resíduos entre valores reais e preditos. A normalidade dos resíduos foi testada com o teste de Shapiro-Wilk, e foram gerados histogramas e boxplots para inspeção visual.

Por fim, análises adicionais de variância (ANOVA) entre modelos e métricas foram conduzidas utilizando o software Sisvar (Ferreira, 2019), com objetivo de validar as diferenças estatísticas identificadas entre os algoritmos.

Tabela 2. Relação dos modelos de aprendizagem de máquinas, utilizados na classificação da qualidade do arroz.

Sigla	Modelo de aprendizagem de máquina	Referência
RF	Floresta aleatória	(Belgiu e Drăguţ, 2016)
SVM	Máquina de vetor suporte	(Cortes e Vapnik, 1995)
GB	Gradient Boosting	(Friedman, 2001)
LR	Regressão Linear Multipla	(Draper e Smith, 1998)
CNN	Redes Neurais Convolucionais	(Lecun *et al.*, 1998)
RNN	Redes Neurais Recorrentes	(Hochreiter e Schmidhuber, 1997)

Tabela 3. Hiperparâmetros Utilizados para Modelos Aprendizado Profundo

Modelo	Tipo de Camadas	Número de Camadas	Neurônios	Dropout	Otimizador	Epochs	Batch Size
CNN	Conv1D + MaxPooling + Dense	4	64 e 128	0.5	Adam	50	32
RNN	LSTM + Dense	4	64 e 128	0.5	Adam	50	32

Os resultados foram representados por meio de boxplots. As médias dos resultados foram comparadas através do teste de Scott-Knott com nível de significância de 5% de probabilidade, no caso de efeitos significativos do teste F ($p \leq 0,05$), utilizando o software estatístico Sisvar versão 5.6 (FERREIRA, 2019).

4. RESULTADOS

4.1. Características Físico-químicas e Assinaturas espectrais por tipo de Arroz.

A Figura 2 apresenta os resultados das principais características físico-químicas de quatro tipos de Arroz: Branco, Parboilizado, Vermelho e Preto. As características físico-químicas analisadas foram Umidade, Amido, Proteína, Fibras, Cinzas e Lipídios. As diferenças entre os tipos foram todas estatisticamente significativas com Pr>Fc = 0.0000 para todas as variáveis analisadas. Os coeficientes de variação (CV%) variaram entre 7,26% e 13,39%, todos dentro de limites aceitáveis, indicando boa homogeneidade dos dados.

O Arroz Preto apresenta maiores teores de Proteínas e Lipídios. No que diz respeito ao Arroz Parboilizado, destaca-se os altos teores de Fibras. O Arroz Branco, por sua vez, apresenta os maiores teores de Amido. Já o Arroz Vermelho apresentou composição intermediária em várias características, porém com destaque negativo para o teor de Fibras, sendo o mais baixo entre os tipos avaliados. Os valores de outliers, de uma forma geral, aparecem discretamente, mas não comprometem a análise e a visualização, permitindo perceber que o Arroz Preto se destaca de forma clara em várias variáveis, como Proteína, Lipídios e Cinzas.

Além disso, há diferenciação química clara entre os tipos de Arroz, destacando novamente o Arroz Preto, que possui valores mais altos em Proteína e Lipídios, e para o Arroz Parboilizado que tem altas quantidades de Fibras em seu teor. Esses resultados reforçam a importância de tipificar o Arroz segundo suas propriedades nutricionais.

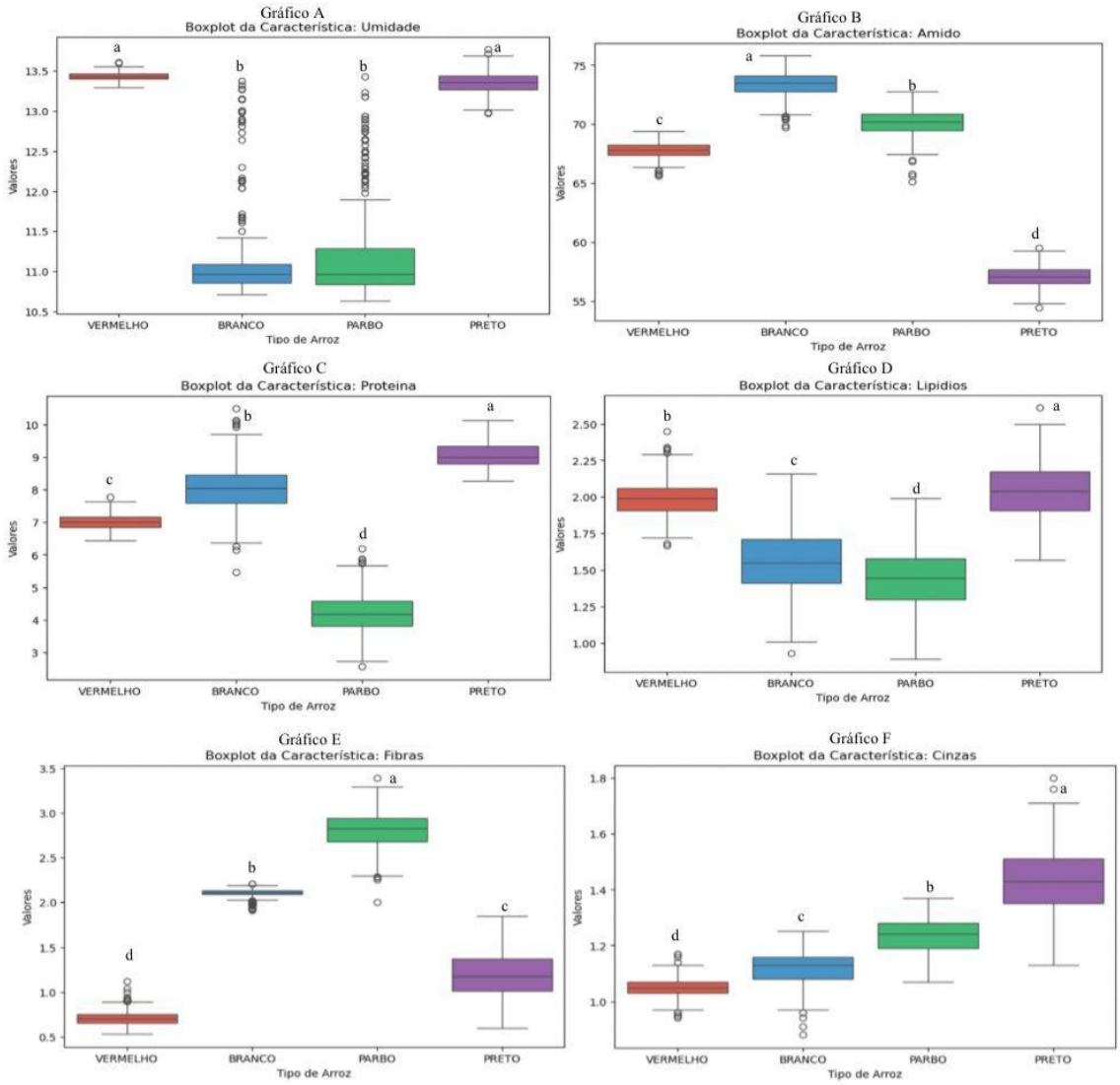

Figura 2. Boxplots para Características Físico-químicas do Tipo de Arroz Vermelho, Branco, Preto e Parboilizado. “a”, “b”, “c”, “d” parâmetros de semelhança estatística.

Já a Figura 3, PCA das Características Físico-Químicas, considera variáveis como Umidade, Proteínas, Lipídios, Fibras, Amido e Cinzas, que são atributos laboratoriais dos grãos. Ao observar o gráfico, como o agrupamento parcial entre os tipos de Arroz, o Arroz Preto tende a se distanciar bem dos demais, possivelmente por apresentar teores mais altos de Lipídios e Proteínas, como mostrado na Figura 2. O Arroz Vermelho e o Branco apresentam maior sobreposição, indicando que, com base apenas nas variáveis físico-químicas, há maior similaridade entre eles (Figura 3A).

Há uma boa dispersão entre as amostras, que pode até ser razoável, mas não suficientemente forte se comparada às espectrais para garantir separação clara entre todos os tipos de Arroz (Figura 3B). Embora útil, o PCA físico-químico mostra limitações na capacidade

de separação entre os tipos de Arroz, especialmente entre os mais similares nos atributos laboratoriais (Gao *et al.*, 2024; Hazrul, Raja Ibrahim; Duralim, 2025).

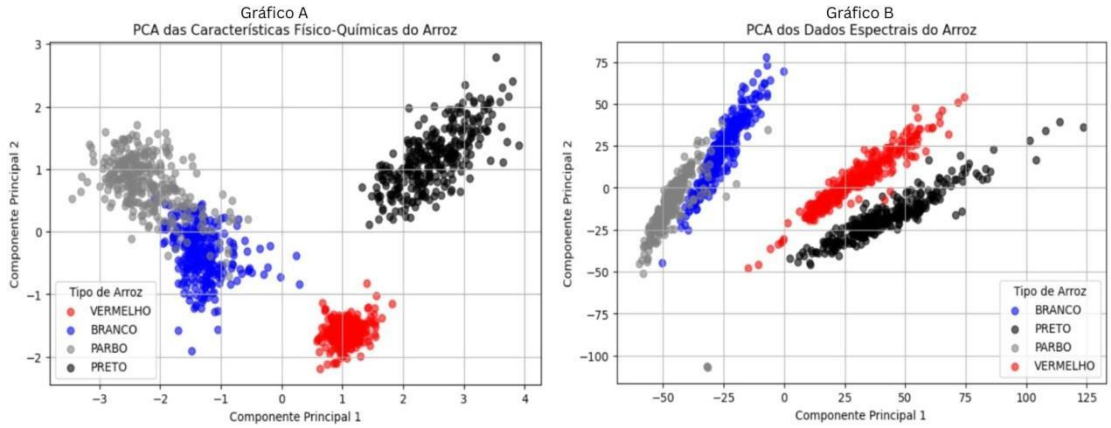

Figura 3. Análise de componentes principais (PCA). Dados de características físico-química e espectrais extraídos dos diferentes tipos de Arroz (Branco, Vermelho, Preto e Parboilizado).

4.2.Assinaturas espectrais dos tipos de Arroz.

A Figura 4 ilustra as curvas médias de refletância espectral para os tipos de Arroz Preto, Vermelho, Branco e Parboilizado, considerando a faixa espectral de 350 a 2500 nm. As curvas apresentam padrões bem definidos entre os tipos de Arroz, especialmente nas faixas do espectro visível (400–700 nm) e do infravermelho próximo (700–1350 nm), o que confirma a capacidade da espectroscopia de diferenciar os grupos.

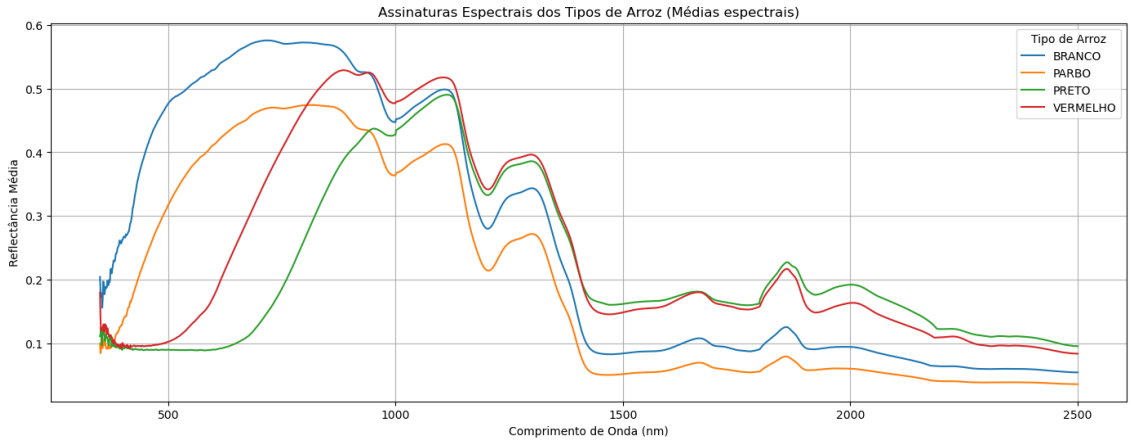

Figura 4. Assinaturas Espectrais médias dos tipos de Arroz Preto, Vermelho, Branco e Parboilizado.

Quanto ao Arroz Preto, nota-se que ele apresenta refletâncias mais baixas ao longo de todo o espectro, especialmente na região do visível 380-700 nm. Já o Arroz Branco exibe maior refletância geral, especialmente nas faixas visível e de transição com o NIR. O Arroz Vermelho e o Parboilizado ocupam posições intermediárias, com assinaturas relativamente próximas, embora apresentem diferenças sutis em regiões específicas, como entre 1100–1400 nm e 1900–2100 nm.

4.3. Modelos de aprendizado de máquina RF, SVM, GB, LR e Modelos de aprendizado profundo CNN e RNN para predição de Proteína.

A Tabela 4 apresenta a ANOVA para os modelos testados na predição do teor de Proteína nos diferentes tipos de Arroz Branco, Vermelho, Preto e Parboilizado. A análise foi realizada com base nas métricas médias de desempenho dos modelos: r, r^2,, MAE e RMSE. A significância estatística Pr>Fc = 0.0000 para todas as métricas demonstra que há diferenças reais entre os desempenhos dos modelos analisados. Entre os modelos, os algoritmos de aprendizado de máquina SVM, RF e GB destacaram-se com os melhores resultados, pertencendo ao mesmo grupo estatístico nas métricas de r, r^2,, MAE e RMSE.

Tabela 4. Análise de variância dos modelos de aprendizado de máquinas para os diferentes tipos de Arroz na predição de Proteína.

Modelo	**R**	R2	**MAE**	**RMSE**
LR	0.826970 a	0.584926 a	0.964798 a	1.212073 a
RNN	0.875092 b	0.760400 b	0.679102 c	0.911554 b
CNN	0.917549 c	0.719705 b	0.756813 b	0.956924 b
GB	0.946986 d	0.894261 c	0.431075 d	0.611450 c
RF	0.949667 d	0.899362 c	0.416296 d	0.597104 c
SVM	0.952401 d	0.903906 c	0.409373 d	0.581563 c
Pr>Fc	0.0000*	0.0000*	0.0000*	0.0000*
CV (%)	2.82	8.58	13.83	14.06
Média Geral	0.9114440	0.7937598	0.6095761	0.8117780

Legenda: r - Coeficiente de Correlação; r^2 - Coeficiente de Correlação ao quadrado; MAE - Mean Absolute Error; RMSE - Root Mean Square Error; * - Significância a 5% de probabilidade. Os valores seguidos por letras diferentes nas colunas indicam diferenças estatisticamente significativas, teste de Scott-Knott a 5% de probabilidade.

A Figura 5 compara as métricas r, r^2, MAE e RMSE entre os modelos. Os modelos de aprendizado profundo CNN e RNN apresentaram desempenho intermediário. A RNN superou a CNN na métrica r^2,, porém com MAE e RMSE superiores, refletindo maior erro absoluto e quadrático. Ambos, no entanto, mostraram desempenho significativamente melhor que o modelo LR, que obteve os piores resultados em todas as métricas, reforçando sua limitação para modelar a complexidade dos dados espectrais.

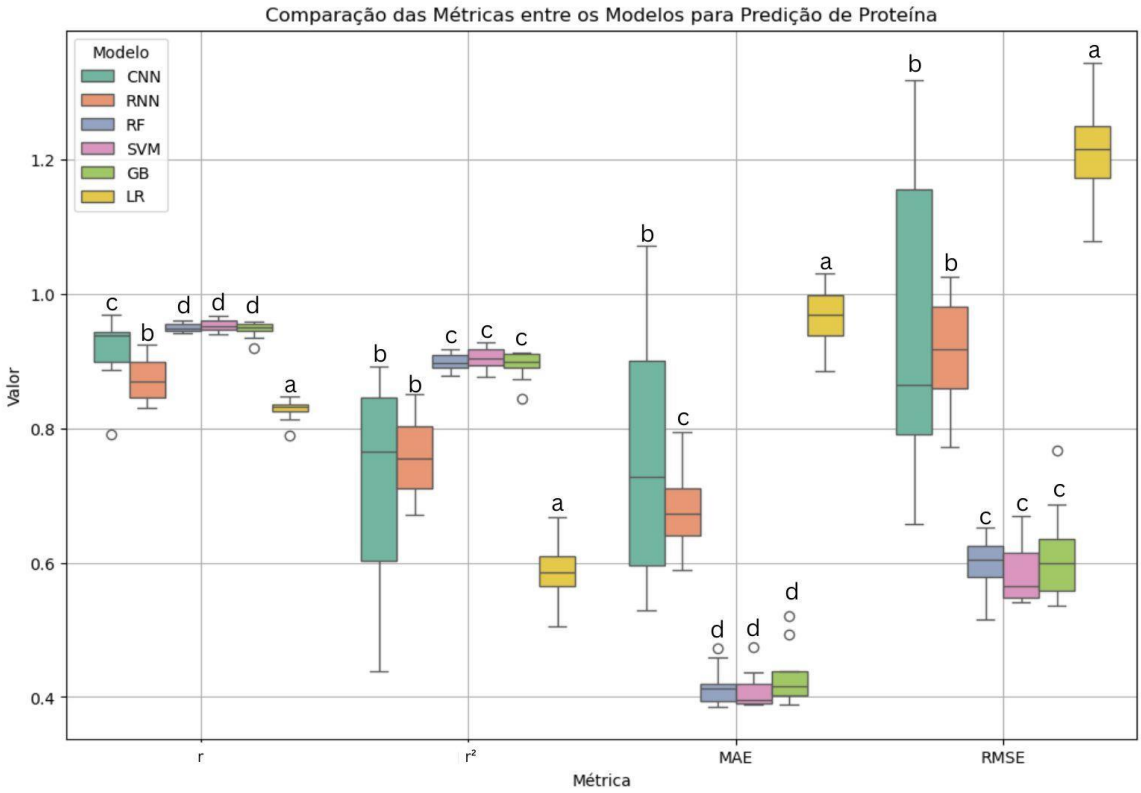

Figura 5. Comparação das métricas r, r², MAE e RMSE entre os modelos Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes para predição de Proteína nos Arroz Preto, Branco, Vermelho e Parboilizado. “a”, “b”, “c”, “d” parâmetros de semelhança estatística.

A Figura 6 apresenta o Boxplot para verificação de distribuição de resíduos entre valores reais e preditos na predição de Proteína nos diferentes tipos de Arroz. Observa-se que os modelos RF, SVM e GB apresentaram histogramas com distribuições simétricas e centradas próximas de zero, sugerindo que os erros foram aleatórios e de variância constante.

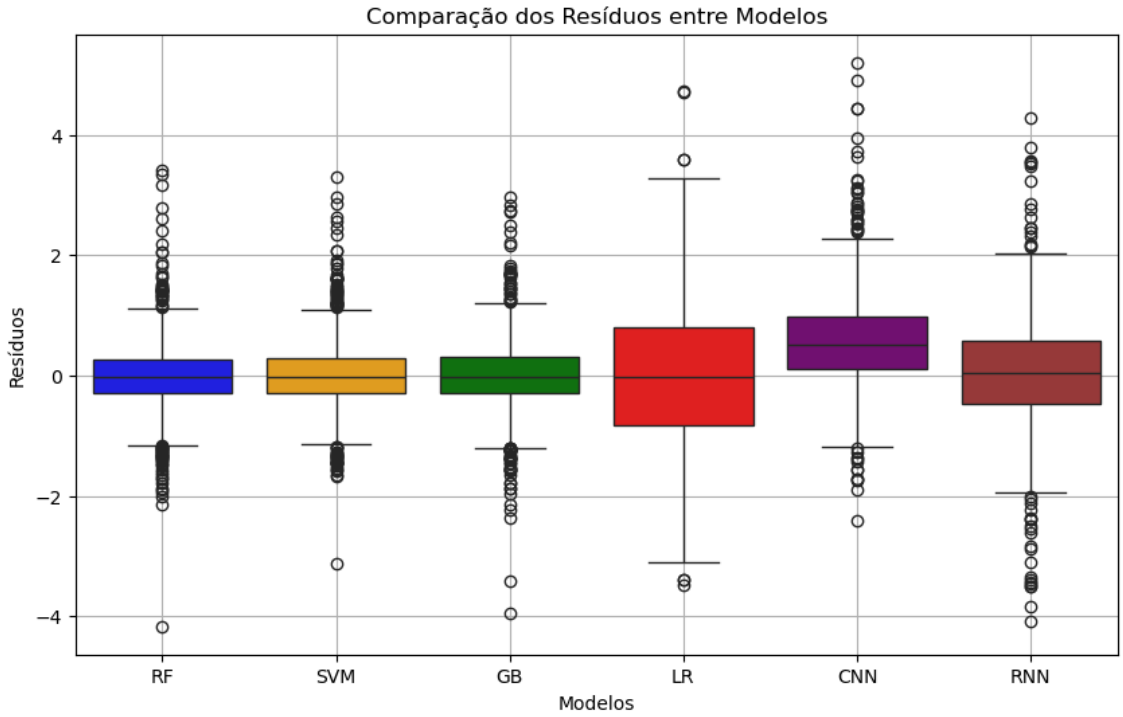

Figura 6. Boxplot para verificação de distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos (Gaussiana) em Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR) para predição de Proteína.

4.4. Assinaturas Espectrais dos teores de Proteína e 50 bandas mais representativas nos tipos de Arroz Branco, Preto, Parboilizado e Vermelho.

A Figura 7 exibe as assinaturas espectrais completas das amostras dos quatro tipos de Arroz: Preto, Vermelho, Branco e Parboilizado, de acordo com os teores de Proteína. A visualização permite identificar padrões distintos de refletância ao longo das bandas espectrais entre os diferentes tipos de Arroz e destacam seus teores de Proteína. Essa representação visual serve como base importante para selecionar regiões espectrais que serão utilizadas na construção dos modelos preditivos.

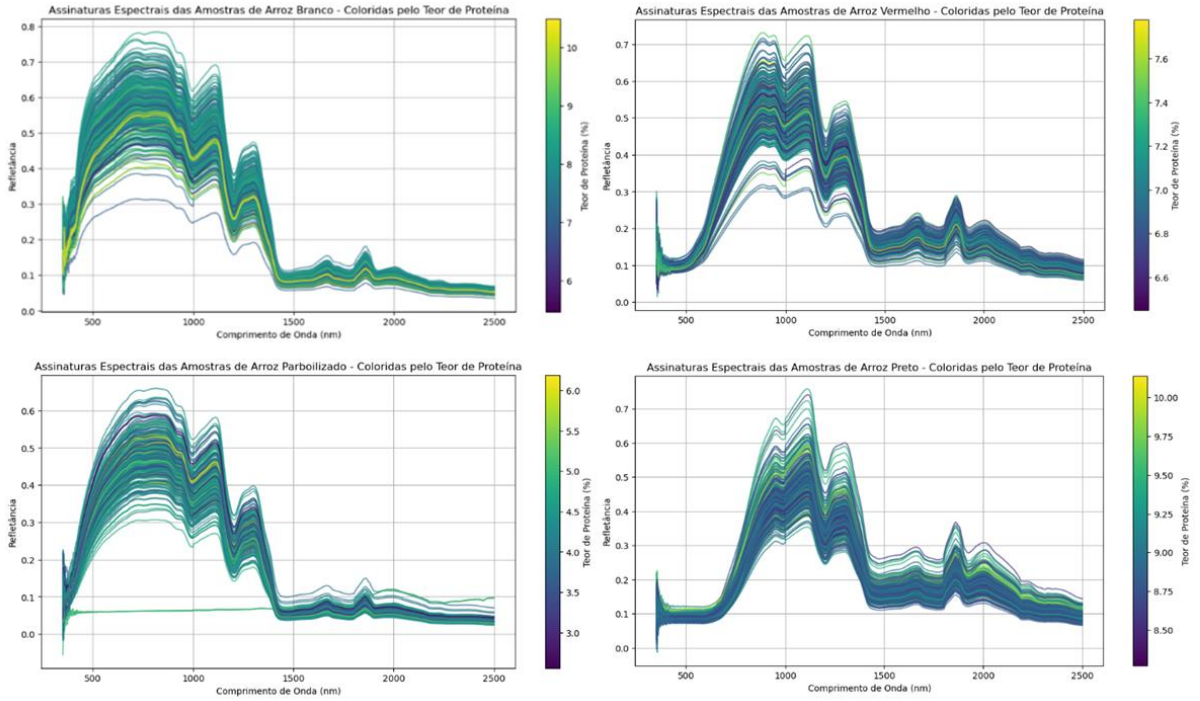

Figura 7. Assinaturas Espectrais dos teores de Proteína presentes do Arroz Preto, Vermelho, Branco e Parboilizado.

É possível observar que as amostras com maiores teores de Proteína apresentam curvas de refletância geralmente mais intensas em determinadas faixas do espectro, indicando potenciais regiões espectrais sensíveis a esse conteúdo.

Já a Figura 8 demonstra as 50 bandas espectrais mais importantes para a predição de Proteína, classificadas de acordo com técnicas de aplicabilidade de modelos em RF: forte concentração de importância entre (1200–1450 nm) e também uma contribuição secundária próxima de (1500–1550 nm), regiões tradicionalmente associadas a absorções relacionadas a grupos N–H e O–H, característicos de proteínas e seus componentes (Kaur *et al.*, 2020; Yang *et al.*, 2024b).

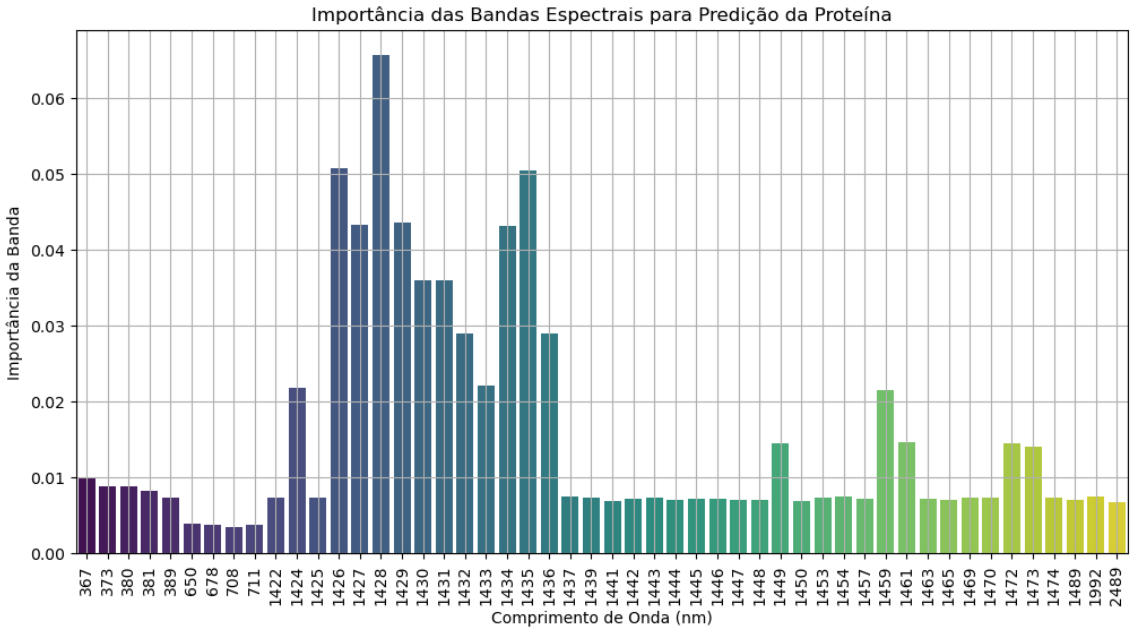

Figura 8. Classificação de 50 bandas espectrais mais representativas para predição de Proteína nos tipos de arroz.

4.5. Modelos de aprendizado de máquina RF, SVM, GB, LR e Modelos de aprendizado profundo CNN e RNN para predição de Amido.

A ANOVA da Tabela 5, que trata da predição de Amido para os diferentes modelos analisados, indicou diferenças estatísticas com valores de Pr > Fc iguais a 0,0000 em todas as métricas. Confirmando que houve real variação no desempenho entre os modelos. Os modelos RF, SVM e GB obtiveram os melhores resultados e com menores valores de MAE e RMSE e altos R2 e r agrupando-se todos no grupo “d”. Já os modelos RNN e LR, demonstraram os piores desempenhos, ficando agrupados e sem diferenças estatísticas entre si no grupo “a”. A CNN, de forma distinta, apresentou resultados intermediários, com métricas melhores que RNN e LR, entretanto abaixo dos modelos de Aprendizado de Máquina. Sua maior variabilidade entre os folds contribuiu para essa classificação.

Tabela 5. Análise de variância dos modelos de aprendizado de máquinas para os diferentes tipos de Arroz na predição de Amido.

Modelo	**r**	**r²**	**MAE**	**RMSE**
RNN	0.833538 a	0.677346 a	2.585056 d	3.468162 d
LR	0.913455 b	0.808589 a	2.151020 c	2.671538 c
CNN	0.977529 c	0.947339 c	1.088146 b	1.394123 b
GB	0.986908 c	0.973679 d	0.766993 a	0.984199 a
RF	0.987449 c	0.974739 d	0.748048 a	0.966918 a
SVM	0.987520 c	0.974626 d	0.761022 a	0.972568 a
Pr>Fc	0.0000*	0.0000*	0.0000*	0.0000*
CV (%) =	1.35	2.75	8.97	7.96
Média Geral:	0.9477333	0.8927196	1.3500474	1.7429178

Na figura 9, que trata do Boxplot de comparação das métricas r, R2, MAE e RMSE entre os modelos na predição para Amido, os resultados obtidos demostram desempenhos muitos similares nas Predições realizadas para Proteína.

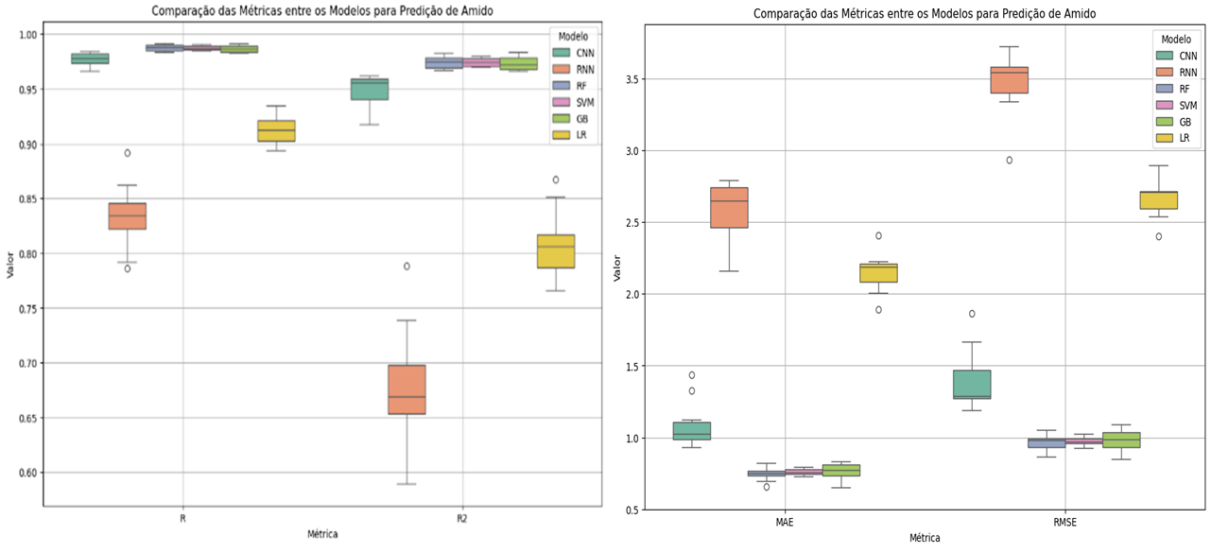

Figura 9. Comparação das métricas r e R2 MAE E RMSE, entre os modelos Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes para predição de Amido no Arroz Preto, Branco, Vermelho e Parboilizado.

Os modelos de aprendizado de máquina RF, SVM e GB mostram-se superiores quando comparados aos modelos de aprendizado profundo, com melhores valores de r e R2, bem como menores valores para MAE e RMSE. Entre estes modelos de aprendizado de máquina RF e SVM apresentaram as melhores médias de desempenho, com valores médios de r e R2 superiores e menores valores de erro. Já GB também se destacou, com desempenho muito próximo ao RF e SVM. A regressão linear, repete seus resultados apresentados em Proteína, ficando significativamente atrás em todos as métricas analisas, indicando baixa capacidade preditiva para essa variável em comparação aos demais modelos.

Entre os modelos de Aprendizado Profundo, na predição de Amido, a CNN obteve resultados intermediários. Embora tenha apresentado um bom r, os valores de erro ainda foram maiores que os dos modelos RF, SVM e GB. Já a RNN, por sua vez, foi o modelo com pior desempenho, apresentando valores muito superiores de erro, o que pode indicar dificuldade significativa em modelar os dados espectrais do Amido, assim como em Proteína.

A Figuras 10 apresenta a distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos (Gaussiana) em Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR) para predição de Amido.

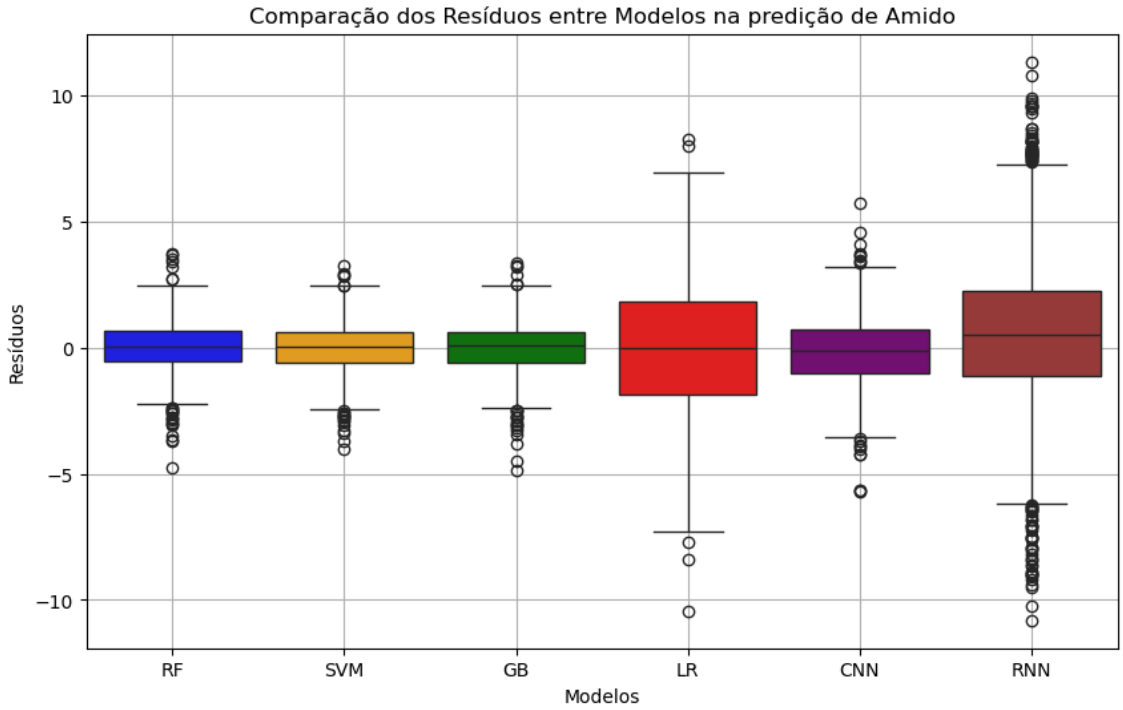

Figura 10. Boxplot para verificação de distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos (Gaussiana) em Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR) para predição de Amido.

4.6. Assinaturas Espectrais dos teores de Amido e 50 bandas mais representativas presentes no Arroz Preto, Vermelho, Branco e Parboilizado.

Na figura 11, que trata das Assinaturas Espectrais dos teores de Amido no Arroz Preto, Vermelho, Branco e Parboilizado, percebe-se diferenciação clara entre os padrões de Amido entre os diferentes tipos de Arroz. Na região do visível percebemos as maiores variações (400-700 nm) e no início o Infravermelho próximo de forma (700 – 1100 nm), o que indica a sensibilidade mais representativa destas faixas ao predizer a variável Amido. Essas correlações de refletância são mais aparentes no Arroz Parboilizado, que mostra mais dispersão entre as amostras com diferentes teores de Amido.

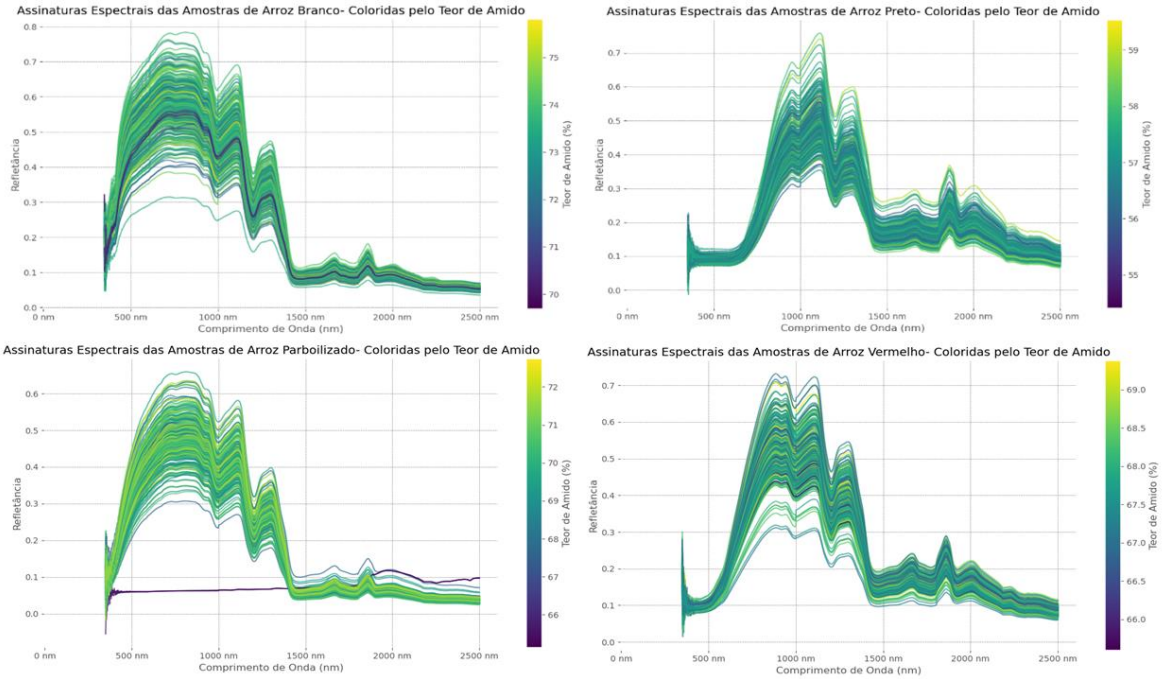

Figura 11. Assinaturas Espectrais dos teores de Amido no Arroz Preto, Vermelho, Branco e Parboilizado.

A identificação destas Assinaturas Espectrais de Amido demonstra a correlação espectral com essa variável. Evidência mais uma vez da capacidade funcional de ser modelada por algoritmos de aprendizado de máquina. Ou seja, mesmo com a variação entre os diversos tipos de Arroz analisados, há pradronização comum espectral ligadas aos teores de Amido (John *et al.*, 2022; Wei *et al.*, 2023; Xie *et al.*, 2022), o que justifica o bom desempenho dos modelos de Aprendizado de Máquina como RF, SVM e GB para predição dessa variável.

Na Figura 12, que trata da classificação das 50 bandas espectrais mais representativas para predição de Amido nos tipos de Arroz Preto, Vermelho, Branco e Parboilizado, essa análise é complementada revelando regiões do espectro mais concentradas entre 530 e 730 nm. Os picos de importância, como demonstrado na Figura 11, estão concentrados nos 381 nm e 642 nm, ambos localizados na região do visível, e foram identificados a partir da análise de importância de variáveis do algoritmo RF.

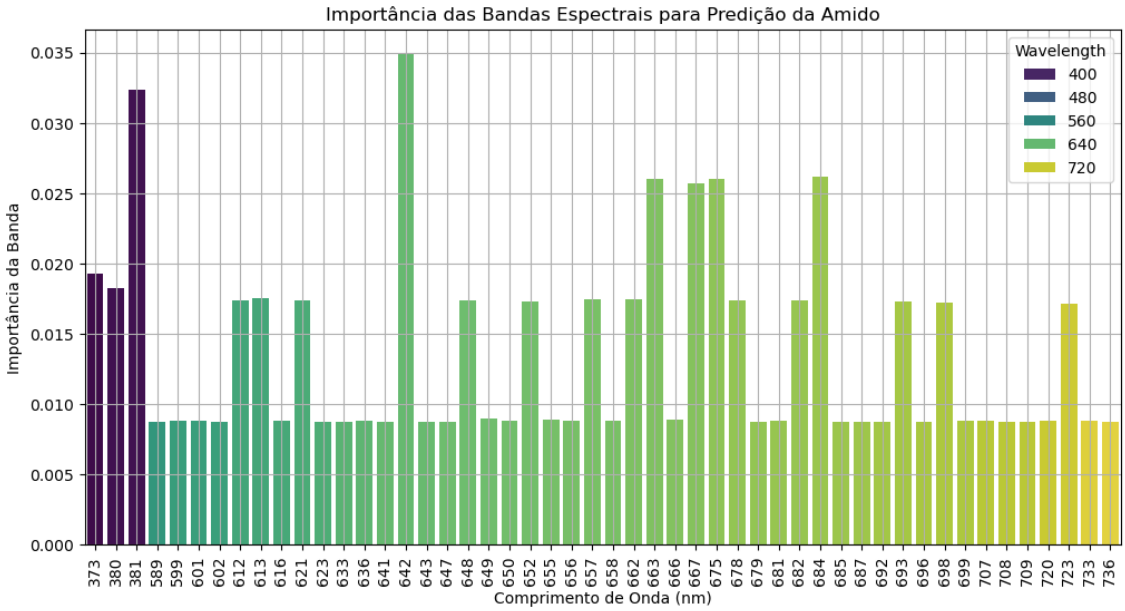

Figura 12. Classificação de 50 bandas espectrais mais representativas para predição de Amido nos tipos de Arroz Preto, Vermelho, Branco e Parboilizado

4.7. Modelos de aprendizado de máquina RF, SVM, GB, LR e Modelos de Aprendizado Profundo CNN e RNN para predição de Umidade.

A análise de variância da Tabela 6 confirma as estatísticas significativas entre os modelos, agrupando SVM, RF, GB e RNN em um único grupo “c”. Com Destaque para SVM, RF e GB que obtiveram melhores r, R2 e menores MAE e RMSE. Esses modelos também apresentaram estabilidade entre os folds e baixa variabilidade dos resíduos.

Tabela 6. Análise de variância dos modelos de aprendizado de máquinas para os diferentes tipos de Arroz na predição de Umidade.

Modelo	**R**	**R2**	**MAE**	**RMSE**
LR	0.761479 a	0.376972 b	0.740377 b	0.933110 b
CNN	0.850031 b	-0.232026 a	0.912414 b	1.122891 b
RNN	0.927148 c	0.852417 c	0.280969 a	0.452444 a
GB	0.935030 c	0.872243 c	0.229157 a	0.418233 a
RF	0.941759 c	0.884839 c	0.223402 a	0.397731 a
SVM	0.945649 c	0.889280 c	0.199913 a	0.388168 a
Pr>Fc	0.0000*	0.0000*	0.0000*	0.0000*
CV (%) =	2.59	97.80	60.45	45.55
Média Geral:	0.8935161	0.6072874	0.4310389	0.6187627

Na Figura 13, observamos os boxplots de comparação das métricas entre os modelos para predição de Umidade nos diversos tipos de arroz. Neste contexto, ao serem utilizados modelos preditivos para análise da variável Umidade, repetem-se os desempenhos de forma muito similar aos apresentados nas predições de Proteína e Amido para os diferentes modelos testados. RF, SVM e GB apresentaram desempenhos superiores, e da mesma forma, esses modelos mostraram altos valores de r e r^2,, além de baixos valores de MAE e RMSE, com consistência entre os folds, demonstrando precisão e robustez assim como nas outras características preditas anteriormente.

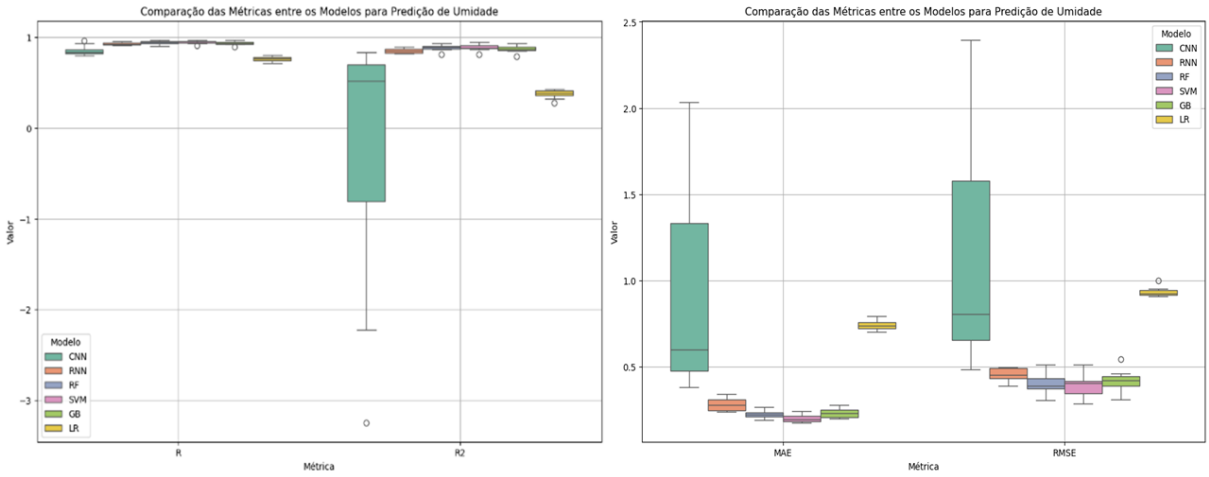

Figura 13. Comparação das métricas r e R2, entre os modelos Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes para predição de Umidade no Arroz Preto, Branco, Vermelho e Parboilizado.

Neste caso, sobre os resultados da CNN, observa-se um elevado intervalo interquartil nas métricas r^2,, MAE e RMSE. Isso evidencia uma alta variabilidade no desempenho do modelo entre os folds, sugerindo instabilidade da arquitetura frente à predição de Umidade. Embora em alguns casos os erros tenham permanecido dentro de uma faixa aceitável, em outros o modelo apresentou forte queda de desempenho com evidência de valores negativos de r^2,.

Essa oscilação compromete a confiabilidade do modelo, e pode estar associada a limitações na arquitetura utilizada, ou mesmo à dificuldade da CNN em generalizar padrões relacionados à Umidade dentro da base de dados utilizada.

Já a RNN, diferentemente do que foi visto nas predições de Proteína e Amido, aqui saiu-se melhor com a variável Umidade. No entanto, ainda demonstrou maior instabilidade quando comparada aos modelos mais robustos, indicando que continua enfrentando dificuldades para se ajustar a este conjunto de dados. Por fim, a LR voltou a apresentar desempenho insatisfatório, com baixa correlação r e valores negativos de r^2 em alguns folds.

Como evidenciado na Figuras 14, a CNN apresentou comportamento instável nesta variável, com distribuição de resíduos assimétrica e presença de outliers, o que reflete em sua baixa capacidade preditiva observada na métrica r^2 negativa. Já a LR, como nas variáveis anteriores, foi o modelo com pior desempenho, com erros amplos, baixa correlação e resíduos altamente dispersos. Portanto, RF, SVM e GB apresentam resíduos concentrados e estáveis, enquanto a LR exibe maior dispersão e presença de outliers, confirmando sua baixa precisão.

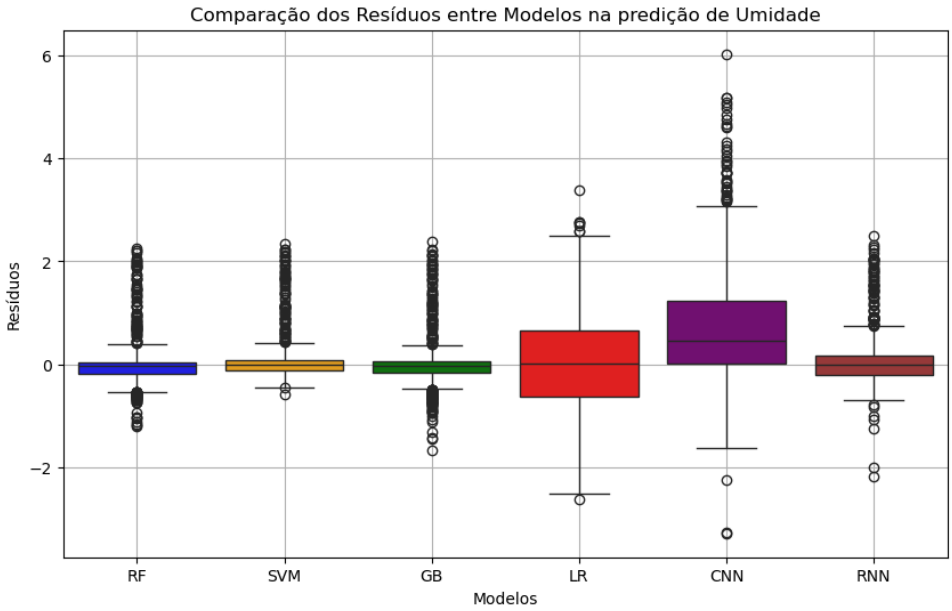

Figura 14. Boxplot para verificação de distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos (Gaussiana) em Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR) para predição de Umidade.

4.8. Assinaturas Espectrais dos teores de Umidade e 50 bandas mais representativas presentes no Arroz.

A figura 15, que trata as assinaturas espectrais para os quatro tipos de arroz Branco, Preto, Parboilizado e Vermelho, coloridas de acordo com o teor de Umidade, demonstra que há padrões distintos e bem definidos entre as amostras. Em todos os tipos, a região do visível (400–700 nm) apresenta variação perceptível na refletância conforme o teor de Umidade. Apesar das variações entre os tipos, observa-se uma tendência comum: quanto maior o teor de Umidade, menor a refletância em determinadas faixas, principalmente entre 420 e 450 nm.

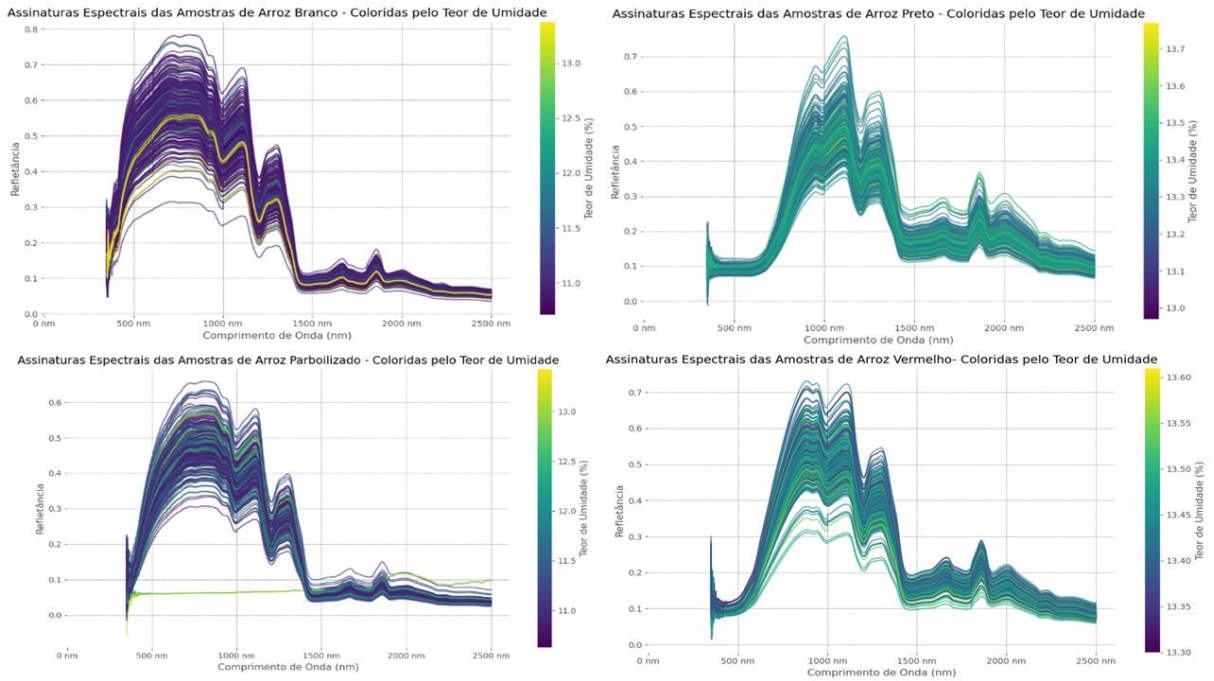

Figura 15. Assinaturas Espectrais dos teores de Umidade no Arroz Preto, Vermelho, Branco e Parboilizado.

A Figura 16, que apresenta a classificação das 50 bandas espectrais mais representativas para a predição de Umidade nos diferentes tipos de arroz, revela um padrão de destaque claro nas faixas iniciais do espectro visível, especialmente entre (420 e 430 nm). Indicativo de que essas bandas são as que carregam maior poder preditivo nos modelos utilizados.

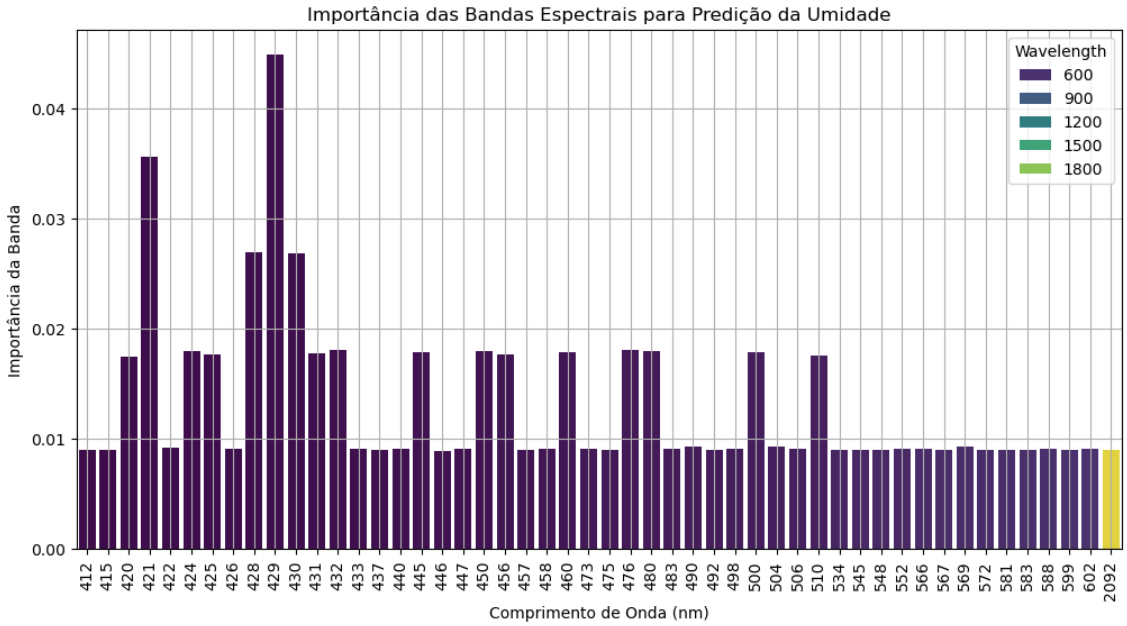

Figura 16. Classificação de 50 bandas espectrais mais representativas para predição de Umidade nos tipos de Arroz Preto, Vermelho, Branco e Parboilizado.

4.9. Modelos a aprendizado de máquina RF, SVM, GB, LR e Modelos de aprendizado profundo CNN e RNN para predição de Fibras, Cinzas e Lipídios.

A análise de variância aplicada às características Fibras - Tabela 7, Cinzas - Tabela 8 e Lipídios - Tabela 9, demonstrou um padrão de desempenho dos modelos testados, reforçando as tendências já observadas nos boxplots anteriores. As métricas r e r^2 mostram que os algoritmos de aprendizado de máquina, RF, GB e SVM foram os mais consistente. Por último, mas não mesmo importante, foi realizada a predição de Lipídios entre os modelos. Lipídios e ácidos graxos afetam significativamente a qualidade do arroz, incluindo textura e valor nutricional.

Tabela 7. Análise de variância dos modelos de aprendizado de máquinas para os diferentes tipos de Arroz na predição de Fibras.

Modelo	**R**	R2	**MAE**	**RMSE**
LR	0.912111 a	0.806076 a	0.284225 b	0.362201 b
RNN	0.944472 b	0.890028 b	0.206893 c	0.271541 c
CNN	0.954765 c	0.704363 c	0.379771 a	0.442052 a
GB	0.974320 d	0.948111 d	0.121865 d	0.187390 d
RF	0.976409 d	0.952487 d	0.120353 d	0.179621 d
SVM	0.979788 d	0.958903 d	0.125506 d	0.165917 d
Pr>Fc	0.0000*	0.0000*	0.0000*	0.0000*
CV (%) =	0.87	5.24	13.35	12.34
Média Geral:	0.9569776	0.8766614	0.2064355	0.2681203

Tabela 8. Análise de variância dos modelos de aprendizado de máquinas para os diferentes tipos de Arroz na predição de Cinzas.

Modelo	**R**	**R2**	**MAE**	**RMSE**
RNN	0.479685 a	0.224316 a	0.108010 b	0.141589 b
LR	0.680372 b	0.111755 b	0.118993 b	0.151002 b
SVM	0.831289 c	0.685680 d	0.072255 a	0.090049 a
CNN	0.845071 c	0.511615 c	0.115809 b	0.137690 b
GB	0.891771 d	0.791237 e	0.054818 a	0.073521 a
RF	0.894952 d	0.797437 e	0.054389 a	0.072392 a
Pr>Fc	0.0000*	0.0000*	0.0000*	0.0000*
CV (%) =	4.44	12.69	39.03	30.43
Média Geral:	0.7705233.	0.5203401	0.0873789	0.1110405

Tabela 9. Análise de variância dos modelos de aprendizado de máquinas para os diferentes tipos de Arroz na predição de Lipídios.

Modelo	**R**	**R2**	**MAE**	**RMSE**
LR	0.498860 a	-0.651551 a	0.325620 c	0.410776 c
RNN	0.794359 b	0.628090 c	0.150674 a	0.195300 a
GB	0.799374 b	0.632075 c	0.153244 a	0.194911 a
RF	0.801374 b	0.636410 c	0.150537 a	0.193632 a
CNN	0.809633 b	0.430318 b	0.226324 b	0.272005 b
SVM	0.810526 b	0.650851	0.150325 a	0.189891 a
Pr>Fc	0.0000*	0.0000*	0.0000*	0.0000*
CV (%) =	4.01	27.83	20.49	16.21
Média Geral:	0.7523545	0.3876988	0.1927873	0.2427526

A análise das variáveis Fibras – Figura 17, Cinzas – Figura 18 e Lipídios - Figura 19, em grãos de Arroz revelou um consistente desempenho entre os modelos de aprendizado de máquina, novamente aqui, os modelos generalizaram bem, com SVM, RF, GB apresentando resultados similares aos encontrados nas variáveis Proteína, Amido e Umidade. Mais do que resultados estatisticamente robustos, esses achados reforçam o potencial prático dessas abordagens para o setor pós-colheita.

Os algoritmos de aprendizado de máquina, apresentaram os melhores resultados em todas as três variáveis, com valores elevados de r e r^2 e menores erros MAE e RMSE. Esses modelos também demonstraram baixa variabilidade entre os folds, o que é evidenciado, pelos boxplots compactos em seus intervalos interquartis, indicando estabilidade, acurácia e capacidade de generalização mesmo em variáveis com dados de menor variância, como o teor de Cinzas.

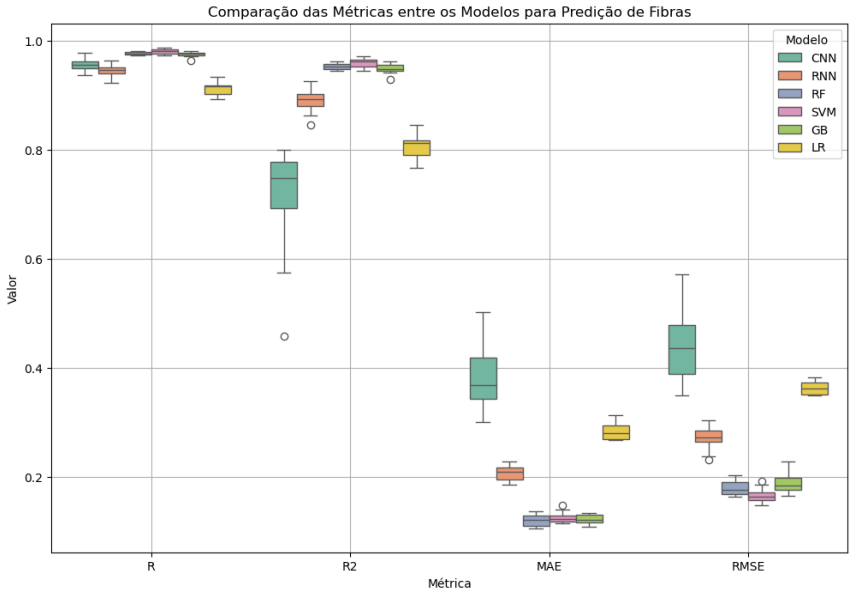

Figura 17. Comparação das métricas r, r², MAE e RMSE entre os modelos Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes para predição de Fibras nos Arroz Preto, Branco, Vermelho e Parboilizado.

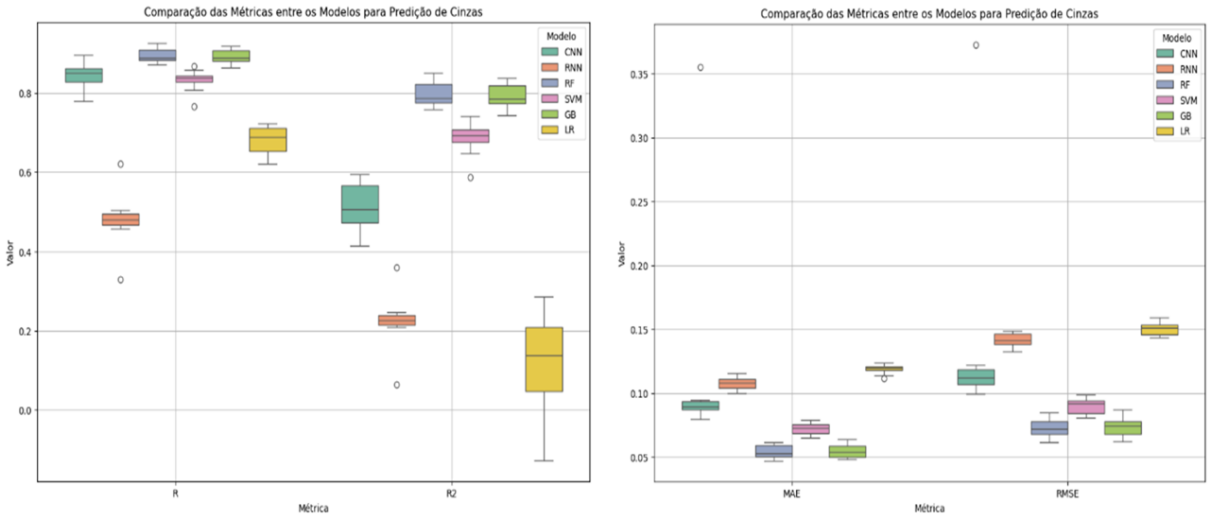

Figura 18. Comparação das métricas MAE e RMSE entre os modelos Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes para predição de Cinzas no Arroz Preto, Branco, Vermelho e Parboilizado.

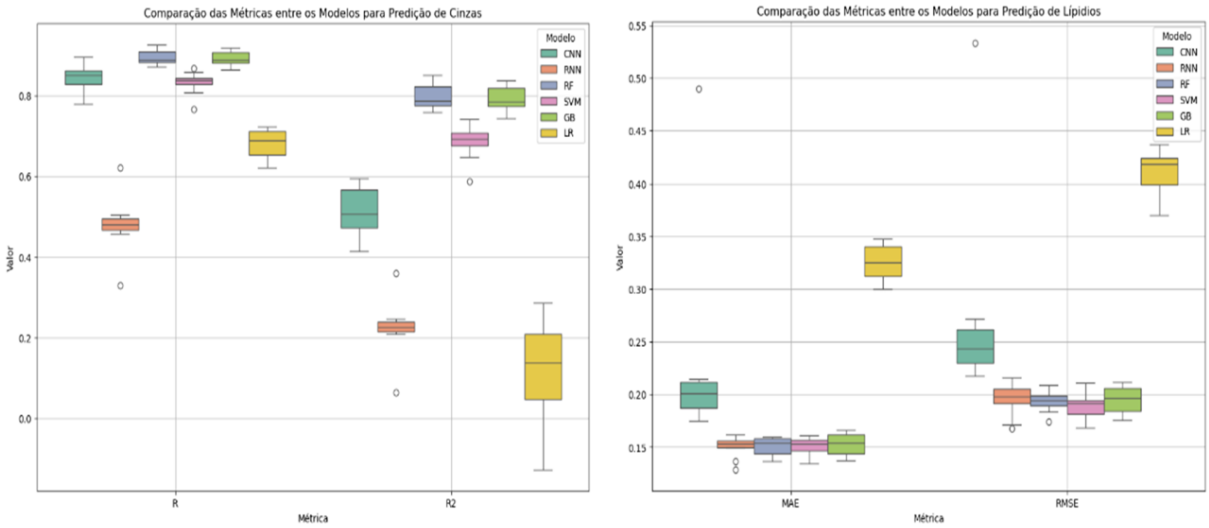

Figura 19. Comparação das métricas r, r entre os modelos Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes para predição de Lipídios no Arroz Preto, Branco, Vermelho e Parboilizado.

No caso das redes neurais, a CNN apresentou desempenho intermediário, com resultados promissores na predição de Fibras, mas variações maiores em Cinzas e Lipídios. Já a RNN manteve desempenho inferior aos modelos tradicionais em quase todas as variáveis, com grande dispersão dos erros em Cinzas. A LR foi, mais uma vez, o modelo com pior desempenho em todas as variáveis analisadas. Os baixos valores de r^2 e a alta variabilidade dos erros reforçam que os atributos físico-químicos aqui estudados não apresentam relação linear com os dados espectrais.

Tendo em vista essas qualidades, durante a tarefa de predição, os valores médios de r e r^2 em todos os modelos ficaram abaixo dos obtidos em outras variáveis, refletindo a maior dificuldade dos modelos. Ainda assim, SVM, RF, GB e RNN apresentaram resultados equivalentes estatisticamente, com destaque para o SVM, que obteve o menor RMSE. A LR foi novamente a pior, com r^2 negativo e agrupamento no grupo "a", mais uma vez evidenciando ineficácia em lidar com relações não lineares envolvidas na predição deste atributo.

Em todas as características, os valores de Pr>Fc foram inferiores a 0,0000 e confirmam a existência de diferença estatística significativa entre os modelos para as métricas avaliadas. Ocorreu baixa variabilidade e os bons CVs reforçam a confiabilidade.

Desta forma, os modelos de aprendizado de máquina foram mais estáveis, com menor dispersão e desempenho superior frente às redes neurais CNN e RNN. Essa consistência entre os métodos estatísticos e gráficos fortalece a conclusão de que RF, SVM e GB são ferramentas eficientes, mesmo em variáveis de baixa magnitude como Fibras, Cinzas e Lipídios — um ponto fundamental para aplicações em triagem pós-colheita e controle de qualidade com espectroscopia não destrutiva.

As análises gráficas dos resíduos nas Figuras 20, 21 e 22, que tratam dos boxplots para verificação de distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos, reforçam os resultados estatísticos obtidos nas análises de variância, demonstrando melhor desempenho dos modelos baseados em árvores RF, GB e SVM.

Nestes modelos os resíduos mostram-se bem distribuídos e com poucos outliers e os histogramas estão próximos à distribuição normal, ou seja, próximo de zero. Em contrapartida, os modelos de LR e, em menor grau, as redes neurais CNN e RNN estão com maior dispersão e resíduos assimétricos, com concentração negativa para a predição de Cinzas e Lipídios. Esses resultados indicam que os modelos de árvore e SVM são mais adequados para predição das características físico-químicas do arroz, apresentando maior robustez e estabilidade preditiva.

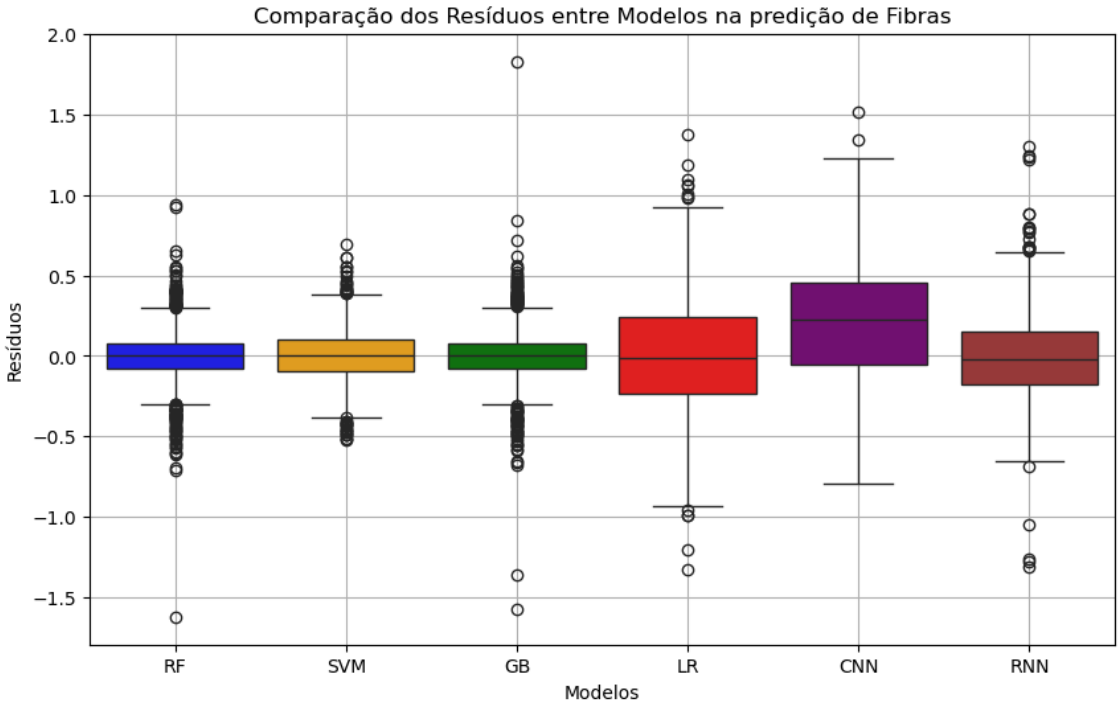

Figura 20. Boxplot para verificação de distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos (Gaussiana) em Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR) para predição de Fibras.

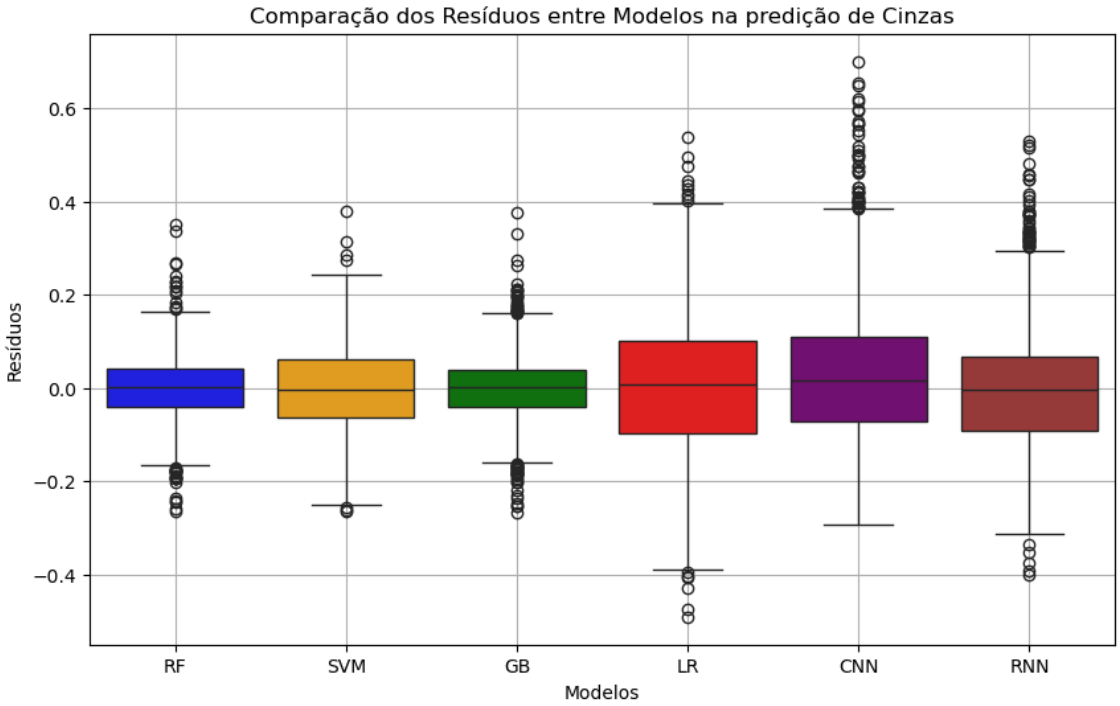

Figura 21. Boxplot para verificação de distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos (Gaussiana) em Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR) para predição de Cinzas no Arroz.

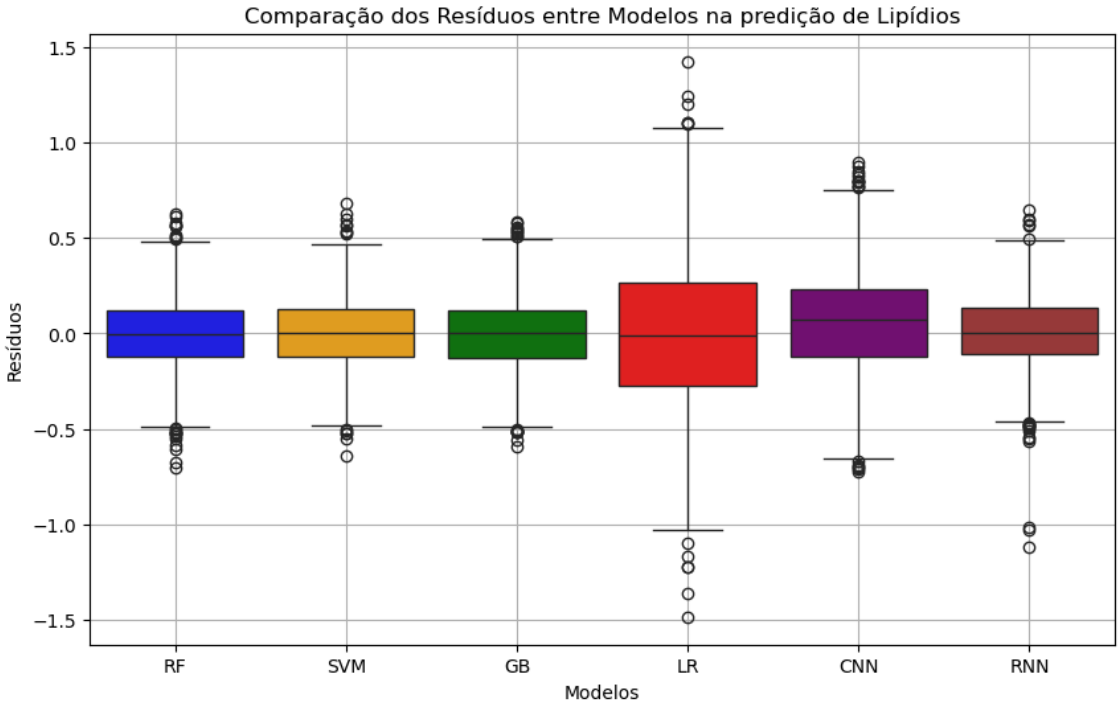

Figura 22. Boxplot para verificação de distribuição de resíduos entre valores reais e preditos e verificação da normalidade dos modelos (Gaussiana) em Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) e Regressão Linear (LR) para predição de Lipídios.

4.10. Assinaturas Espectrais dos teores de Fibras, Cinzas e Lipídios e 50 bandas mais representativas presentes no Arroz.

A Figura 23 mostra as assinaturas espectrais médias associadas aos teores de Fibras nos diferentes tipos de arroz Preto, Vermelho, Branco e Parboilizado, em função do comprimento de onda. Observa-se que as assinaturas espectrais variam entre os tipos de arroz, indicando que cada um apresenta padrões de absorção específicos. Essas diferenças espectrais justificam a capacidade dos modelos preditivos em diferenciar os tipos de arroz com base no teor de Fibras, destacando a utilidade da espectroscopia como ferramenta não destrutiva e precisa.

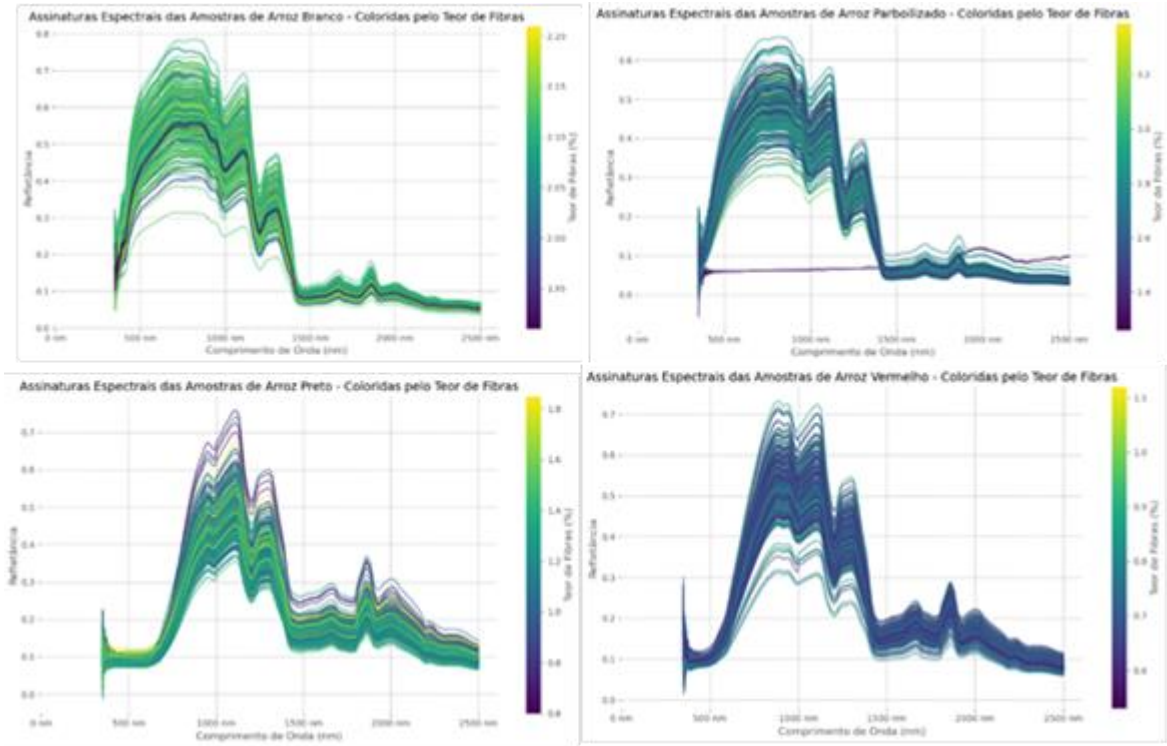

Figura 23. Assinaturas Espectrais dos teores de Fibras presentes do Arroz Preto, Vermelho, Branco e Parboilizado.

A Figura 24 apresenta a importância relativa das 50 bandas espectrais mais representativas para predição do teor de Fibras nos diversos tipos de arroz analisados e considerando sua contribuição em cada comprimento de onda para os modelos de aprendizado de máquina.

É importante observar que a distribuição não se concentrou em uma única região do espectro, mas sim em uma combinação multiespectral, o que demonstra a necessidade de modelos que considerem múltiplas faixas em conjunto. Esse comportamento reforça, ainda, o potencial para redução de custos nas aplicações práticas, visto que sensores multiespectrais — mais acessíveis que os hiperespectrais — poderiam ser empregados com eficiência na predição destas características para fins industriais.

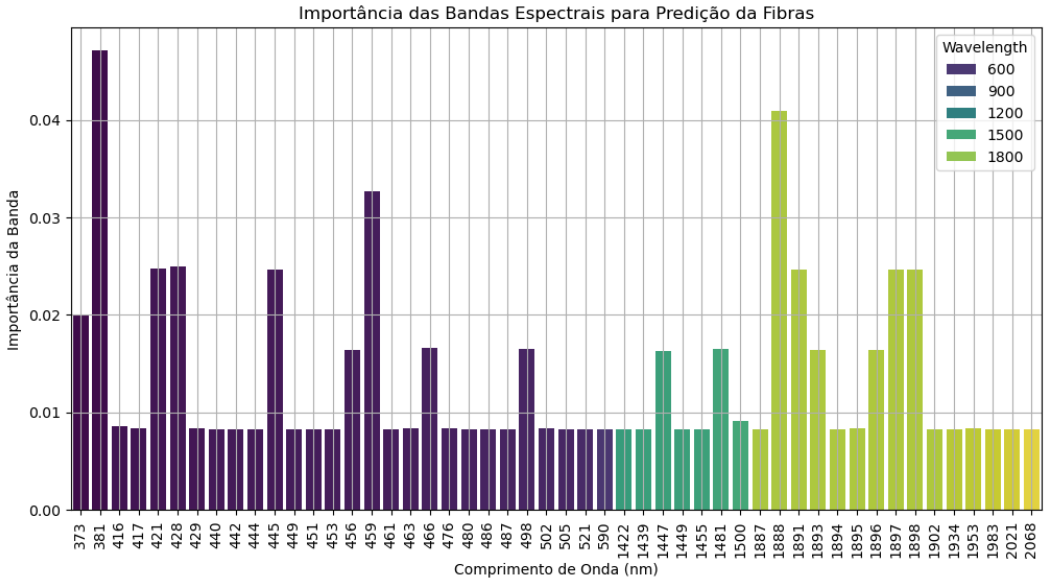

Figura 24. Classificação de 50 bandas espectrais mais representativas para predição de Fibras nos tipos de Arroz Preto, Vermelho, Branco e Parboilizado.

A Figura 25 apresenta as assinaturas espectrais das amostras de arroz Preto, Branco, Parboilizado e Vermelho, coloridas de acordo com o teor de Cinzas. Em todos eles é possível observar variações em suas curvas espectrais ao longo do comprimento de onda, com destaque para picos nas regiões entre 500–700 nm e entre 950–1350 nm. No arroz preto e vermelho há dispersão nas curvas na região do NIR. Já no arroz Branco, as curvas apresentam maior uniformidade, coerente com o processamento de polimento que remove parte significativa da casca e das camadas externas. O arroz parboilizado também apresenta curvas mais compactas.

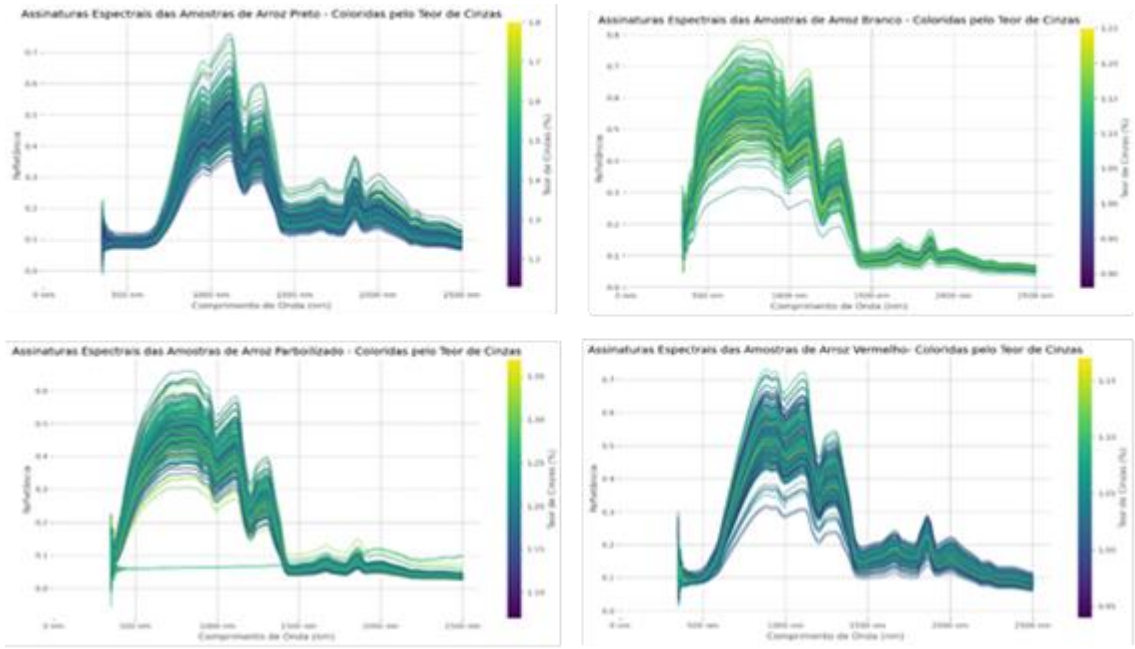

Figura 25. Assinaturas Espectrais dos teores de Cinzas presentes do Arroz Preto, Vermelho, Branco e Parboilizado.

Ao analisarmos a Figura 26, que trata das assinaturas espectrais dos teores de Lipídios entres do Arroz Preto, Vermelho, Branco e Parboilizado, observa-se uma distribuição espectral com destaque para alguns picos nas regiões visível (350–700 nm) e de forma mais determinante no infravermelho próximo (1100–1900 nm). O gradiente de cores aplicado aos espectros apresenta aparentemente que, embora o teor de Lipídios seja baixo de forma geral, ele exerce um padrão detectável no comportamento espectral para esta variável, especialmente nas faixas NIR.

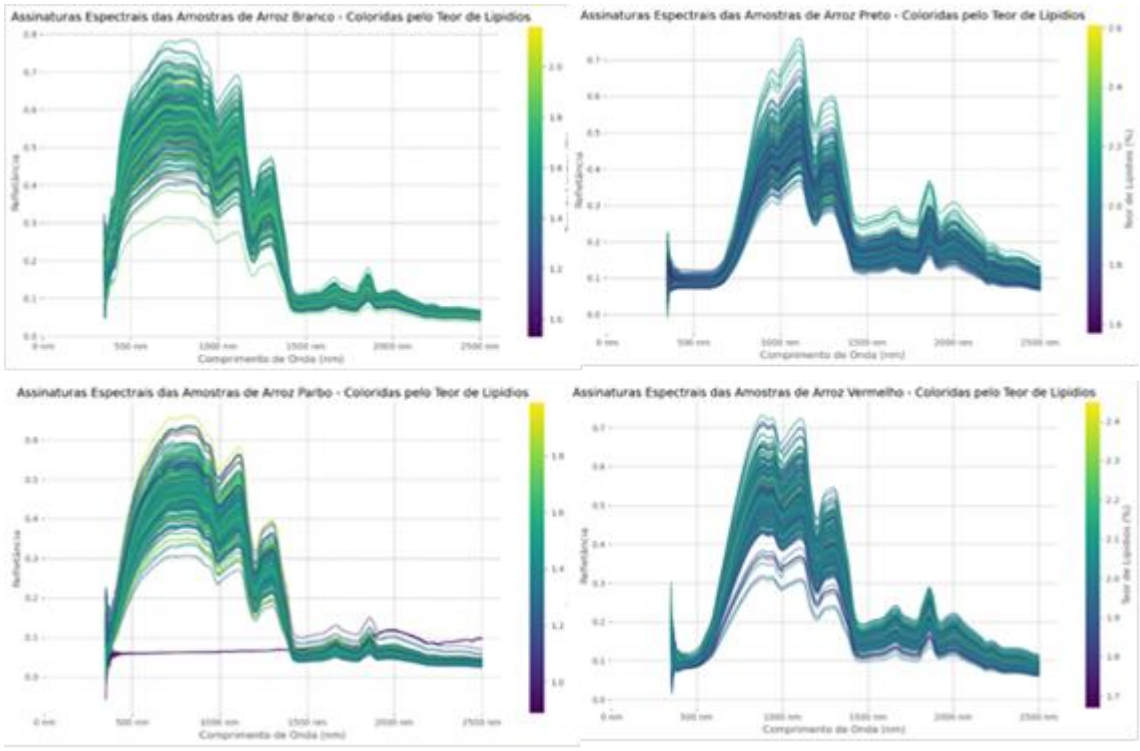

Figura 26. Assinaturas Espectrais dos teores de Lipídios entes do Arroz Preto, Vermelho, Branco e Parboilizado.

Ao analisarmos a figura 27, em conjunto com a interpretação da figura 28 que trata classificação de 50 bandas espectrais mais representativas, diferentemente de Cinzas e Fibras, observa-se uma distribuição bem-marcada nas faixas do NIR, especialmente entre 1400–2000 nm. Os picos mais representativos concentram-se em torno de 1442 nm, 1538 nm, 1834 nm e 1981 nm. As bandas do visível, neste caso, apresentam relevância menor, reforçando que uma abordagem multiespectral seria a mais recomendada, embora a região NIR carregue a maior carga preditiva para os Lipídios.

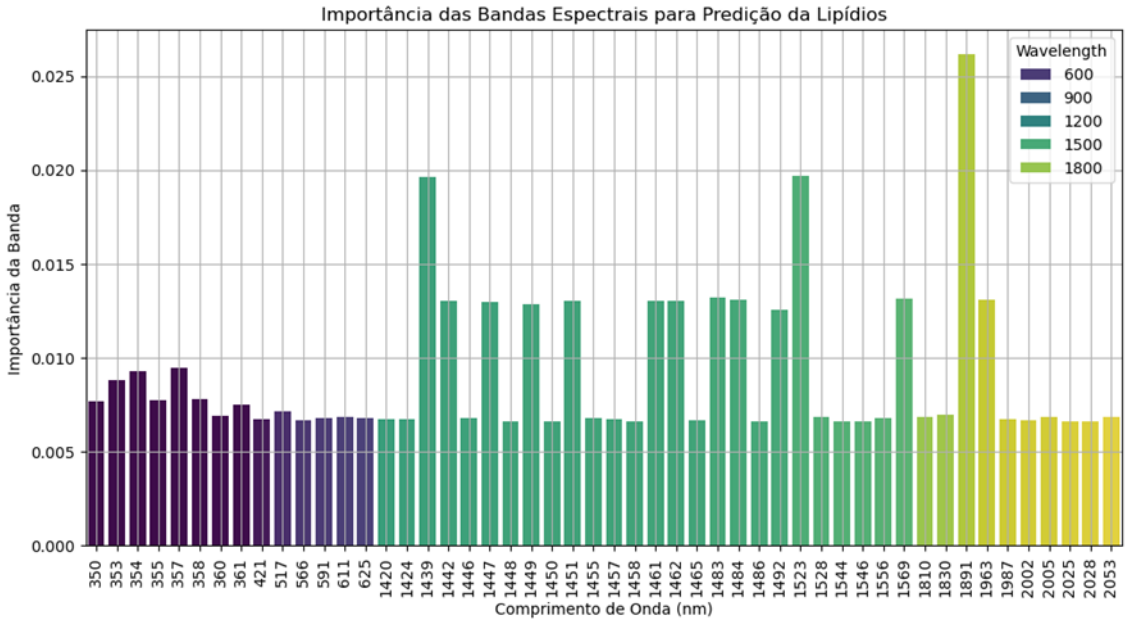

Figura 27. Classificação de 50 bandas espectrais mais representativas para predição de Lipídios nos tipos de Arroz Preto, Vermelho, Branco e Parboilizado.

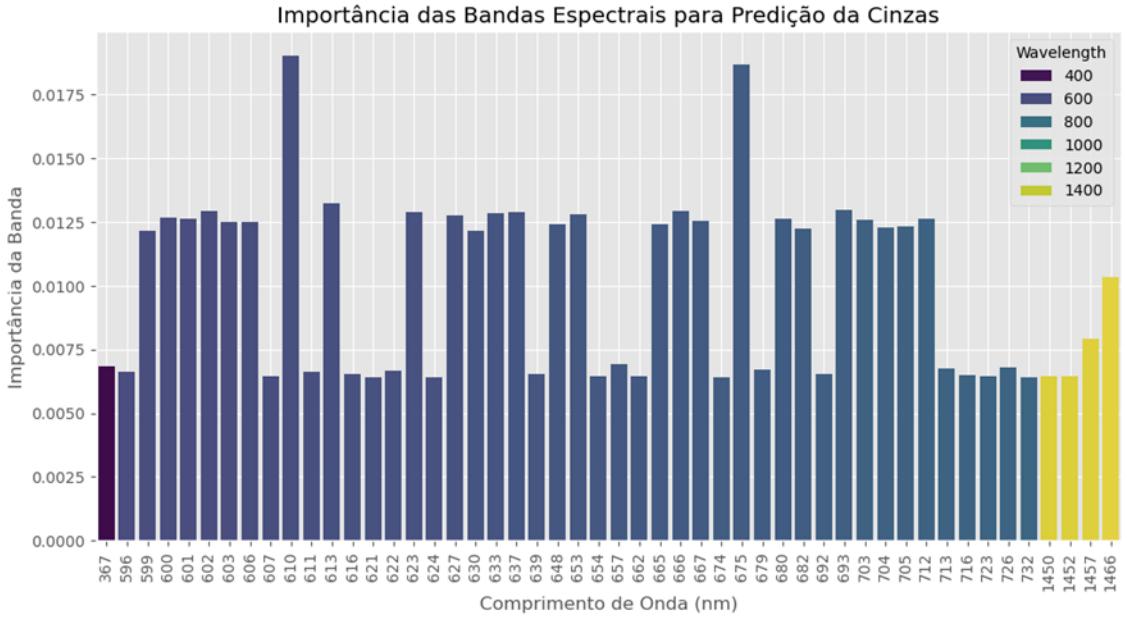

Figura 28. Classificação de 50 bandas espectrais mais representativas para predição de Cinzas nos tipos de Arroz Preto, Vermelho, Branco e Parboilizado.

5. DISCUSSÃO

Nestas análises os resultados corroboram com achados de estudos anteriores que destacam o Arroz Preto como o tipo com maior valor nutricional de lipídeos e proteínas entre as variedades analisadas (BANERJEE et al., 2023; KHATUN; MOLLAH, 2024). O Arroz Preto apresenta maiores teores de Proteínas, Lipídios e compostos fenólicos, sendo frequentemente classificado como um alimento funcional (DAS et al., 2023; RAHIMZADEH ARASHLOO; KITTLER, 2022; ZHAO et al., 2021). Essa composição diferenciada justifica os altos teores nutricionais observados neste estudo, refletindo a presença de componentes bioativos e minerais concentrados na casca e pericarpo do grão (BANERJEE et al., 2023).

No que diz respeito ao Arroz Parboilizado, os altos teores de Fibras identificados neste estudo estão em consonância com os resultados da literatura (AKHTER et al., 2023). Além disso, o processo de parboilização promove a migração de nutrientes da casca para o endosperma (KALITA; GOHAIN; HAZARIKA, 2021), aumentando a concentração de Fibras insolúveis. Entretanto, esse mesmo processo pode levar à perda de parte das Proteínas (NIRMAGUSTINA; HANDAYANI, 2023) e de Lipídios (MUDGAL; SINGH, 2024), o que pode explicar os menores teores encontrados para essas variáveis no Arroz Parboilizado aqui avaliado.

O Arroz Branco, por sua vez, apresentou maiores teores de Amido (YOVIONO; SANDRA; ARIFANDI, 2022). Este tipo é composto principalmente de Amido, armazenado no endosperma, o tecido que compõe a maior parte do grão. A concentração, a estrutura e a distribuição do Amido no endosperma são influenciadas por fatores genéticos, metabólicos e ambientais (YANG et al., 2020), que afetam a qualidade do Arroz, a digestibilidade e o desenvolvimento do grão (CAO et al., 2022). Essa característica torna o Arroz Branco menos nutritivo em comparação com os demais tipos (HASHIMOTO et al., 2022; ZHANG et al., 2023), entretanto mais palatável e de rápida cocção, o que explica sua popularidade no consumo humano (GONDAL et al., 2021; ZHAO et al., 2021).

Já o Arroz Vermelho apresentou composição intermediária em várias características, porém com destaque negativo para o teor de Fibras, sendo o mais baixo entre os tipos avaliados. Estes resultados divergem dos resultados encontrados por ABEYSIRIWARDENA; GUNASEKARA (2020); GOGOI et al. (2025), que destacam seu alto teor de Fibras

alimentares. Embora o Arroz Vermelho seja reconhecido por sua coloração rica em antocianinas, seus valores de macronutrientes podem variar amplamente conforme a variedade genética e o manejo agrícola (GOGOI et al., 2025).

Essas diferenças evidenciam o potencial funcional e nutricional diferenciado entre os tipos de Arroz, e justificam o interesse crescente na utilização de espectroscopia e modelagem preditiva como ferramentas para diferenciação rápida e não destrutiva de atributos de qualidade nutricional (JOHNSON et al., 2021; LIU et al., 2020; XU et al., 2025).

A Figura 02 apresenta os Boxplots que ilustram visualmente a distribuição das grandezas numéricas das características físico-químicas para os quatro tipos de Arroz. Nele há diferenças entre os grupos, confirmando graficamente os resultados da ANOVA. As distribuições são observadas para algumas características, como em Fibras no Arroz Parboilizado, indicando possível variação do material. Os valores de outliers, de uma forma geral, aparecem discretamente, mas não comprometem a análise e a visualização, permitindo perceber que o Arroz Preto se destaca de forma clara em várias variáveis, como Proteína, Lipídios e Cinzas.

Além disso, esta figura evidencia que há diferenciação química clara entre os tipos de Arroz, destacando novamente o Arroz Preto, que possui valores mais altos em Proteína e Lipídios, e para o Arroz Parboilizado que tem altas quantidades de Fibras em seu teor. Esses resultados reforçam a importância de tipificar o arroz segundo suas propriedades nutricionais.

A Figura 03, PCA das Características Físico-Químicas, considera variáveis como Umidade, Proteínas, Lipídios, Fibras, Amido e Cinzas, que são atributos laboratoriais dos grãos. Ao observar o gráfico, como o agrupamento parcial entre os tipos de Arroz, o Arroz Preto tende a se distanciar bem dos demais, possivelmente por apresentar teores mais altos de Lipídios e Proteínas, como mostrado na Figura 02.

O Arroz Vermelho e o Branco apresentam maior sobreposição, indicando que, com base apenas nas variáveis físico-químicas, há maior similaridade entre eles. Isso está de acordo com os achados de PEREIRA et al. (2023), que relataram que as características físico-químicas tradicionais, embora relevantes, nem sempre são suficientes para distinguir com clareza variedades de arroz com morfologia e composição intermediária.

Há uma boa dispersão entre as amostras, que pode até ser razoável, mas não suficientemente forte se comparada às espectrais para garantir separação clara entre todos os tipos de Arroz. Embora útil, o PCA físico-químico mostra limitações na capacidade de separação entre os tipos de Arroz, especialmente entre os mais similares nos atributos laboratoriais (GAO et al., 2024; HAZRUL; RAJA IBRAHIM; DURALIM, 2025).

Já no PCA das Características Espectrais, que considera as bandas espectrais de 350 a 2500 nm, observa-se uma discriminação expressivamente mais acentuada entre os grupos. Neste PCA, diferentemente do que ocorre no físico-químico, há maior separação entre os tipos de Arroz. Cada grupo, ou seja, Preto, Vermelho, Branco e Parboilizado tende a concentrar-se em regiões distintas do gráfico. Isso indica que as assinaturas espectrais capturam detalhes estruturais e de composição específicas de cada tipo, aparentemente imperceptíveis em análises laboratoriais.

O PCA espectral apresenta melhor discriminação entre os tipos de Arroz, demonstrando que os dados espectrais são mais sensíveis à composição do grão do que os dados físico-químicos tradicionais utilizados. Essa sensibilidade dos dados espectrais à composição do grão também foi observada por KANG et al. (2024) na discriminação de arroz, entretanto, com outras variedades de Arroz.

Essa análise pode ser crucial no contexto pós-colheita, onde decisões impactam na preservação da qualidade, o valor de mercado e a eficiência operacional (MAHMOOD et al., 2024; MÜLLER et al., 2022). Desta forma, o PCA não contribui somente com modelos preditivos que virão, mas também fortalece o papel das tecnologias não destrutivas como ferramentas de triagem e controle em tempo real nos processos pós-colheita.

A Figura 4 ilustra as curvas médias de refletância espectral para os tipos de Arroz Preto, Vermelho, Branco e Parboilizado, considerando a faixa espectral de 350 a 2500 nm. As curvas apresentam padrões bem definidos entre os tipos de Arroz, especialmente nas faixas do espectro visível (400–700 nm) e do infravermelho próximo (700–1350 nm), o que confirma a capacidade da espectroscopia de diferenciar os grupos.

Quanto ao Arroz Preto, ao observarmos esta figura, notamos que ele apresenta refletâncias mais baixas ao longo de todo o espectro, especialmente na região do visível. A cor

visível do arroz preto é resultado da luz refletida na camada de farelo rica em antocianinas, ou seja, quanto mais antocianinas presentes, mais escura e intensa a cor parece.

Grãos de arroz preto mais escuros apresentam concentrações mais elevadas desses compostos benéficos em comparação com grãos de arroz mais claros ou brancos (BORAH et al., 2025; BRUNET-LOREDO et al., 2023; THEIVENTHIRAN et al., 2020). Já o Arroz Branco exibe maior refletância geral, especialmente nas faixas visível e de transição com o NIR, resultado esperado por sua coloração clara (AEKRAM et al., 2025; GARCÍA-SALCEDO et al., 2023).

O Arroz Vermelho e o Parboilizado ocupam posições intermediárias, com assinaturas relativamente próximas, embora apresentem diferenças sutis em regiões específicas, como entre 1100–1400 nm e 1900–2100 nm. A visualização reforça que as assinaturas espectrais são características únicas de cada tipo de Arroz, refletindo suas diferenças espectrais e estruturais.

Esses padrões justificam o uso de modelos de aprendizado de máquina e análises espectrais para predição de atributos nutricionais. Esses achados são particularmente relevantes no contexto pós-colheita, pois permitem a triagem rápida e não destrutiva de lotes, otimizando processos de classificação e direcionamento comercial dos grãos. A figura também contribui como evidência visual complementar à Figura 2, reforçando que os tipos de Arroz apresentam assinaturas distintas, o que possibilita sua identificação.

Ao analisarmos os resultados das tabelas e Boxplots apresentados para cada uma das características físico-químicas é evidente que os modelos baseados em aprendizado de máquina tradicional, como Support Vector Machine (SVM), Random Forest (RF) e Gradient Boosting (GB), foram superiores em relação aos modelos de aprendizado profundo (CNN e RNN). Essa superioridade foi observada de forma consistente nas métricas de desempenho: apresentaram maiores valores de r e r^2 e menores erros médios absolutos (MAE) e raízes do erro quadrático médio (RMSE), além de demonstrarem maior estabilidade entre os folds, com menor variabilidade estatística.

Esse comportamento reafirma evidências recentes na literatura científica que destacam a robustez de modelos como SVM e RF em tarefas de regressão com dados hiperespectrais (FEYISA et al., 2020; LIU et al., 2022; TUNCA et al., 2023), especialmente em contextos com

conjuntos de dados de tamanho moderado e sem grandes variações temporais (LI et al., 2023; NAGY et al., 2024).

Isso se deve, em parte, à capacidade desses algoritmos de explorar relações não lineares nos dados sem depender de ajustes complexos, como camadas ocultas profundas ou sequências temporais de treinamento. Já o modelo GB mostrou-se uma ferramenta relevante na predição de propriedades físico-químicas em milho (ZHENG et al., 2024; ZOU et al., 2025) e soja (HUBER et al., 2022; LI et al., 2023), utilizando dados espectroscópicos complexos, achados que também foram replicados neste estudo.

Já os modelos de aprendizado profundo, embora promissores, demonstraram maior variabilidade entre os folds, sendo mais suscetíveis ao overfitting ou à necessidade de ajustes finos nos hiperparâmetros (ERCANLI, 2024). A CNN, por exemplo, apresentou desempenho inferior em quase todas as variáveis analisadas, provavelmente por não capturar adequadamente as dependências sequenciais dos espectros.

A RNN teve desempenho ligeiramente melhor, por ser mais adaptada a sequências, mas ainda ficou abaixo dos modelos tradicionais, o que pode estar associado ao volume de dados relativamente pequeno, uma limitação comum em estudos experimentais com amostras físicas (LIU, Y. et al., 2024).

Os resultados evidenciam que, no contexto analisado neste estudo, os modelos clássicos de aprendizado de máquina continuam figurando entre as abordagens mais eficazes (OLIVEIRA CARNEIRO et al., 2023), tanto pela precisão quanto pela consistência estatística, demonstrando serem ferramentas altamente eficazes em diversos cenários agrícolas.

A determinação do teor de proteína no arroz é um fator crítico na pós-colheita, já que pode influenciar seu valor nutricional, o posicionamento de mercado e o direcionamento do grão para diferentes finalidades comerciais, como consumo humano, formulações especiais ou produtos industriais (YAN et al., 2021; YANG et al., 2024a). Métodos tradicionais de quantificação exigem análises laboratoriais demoradas, destrutivas e muitas vezes custosas, criando atrasos nos processos de triagem e controle de qualidade (AZEEZ; QASIM; MOHIE, 2024; BHUPENDRA et al., 2022).

Para esta sensível tarefa de predição do teor de proteína, a Figura 5 apresenta boxplots comparativos entre os modelos para cada métrica avaliada r, r^2,, MAE e RMSE, permitindo visualizar a variabilidade dos resultados nos diferentes folds da validação cruzada. Entre os modelos, SVM destacou-se com o maior r médio e menor RMSE, mantendo também a menor variabilidade entre os folds. Esse comportamento pode ser atribuído à sua capacidade de maximizar margens e generalizar bem (YOOSEFZADEH-NAJAFABADI et al., 2021), mesmo em contextos com dados multivariados de alta dimensionalidade, como os hiperespectrais.

Já RF obteve desempenho muito próximo ao SVM e mostrou-se igualmente robusto, porém com ligeira elevação nos erros médios em alguns folds, o que pode estar relacionado à sua maior sensibilidade à heterogeneidade amostral em algumas repetições (XIAO et al., 2023; ŽÍŽALA et al., 2024). Mesmo assim, seu desempenho confirma observações de outros estudos que destacam a eficiência de modelos baseados em árvores em tarefas de predição agrícola utilizando espectroscopia (ELAVARASAN; VINCENT, 2021; GAROFALO et al., 2024). Estes modelos são robustos, lidam bem com dados de alta dimensão e exigem menos ajustes de parâmetros (MARQUES RAMOS et al., 2020; MIRANDA RAMOS SOARES et al., 2020; ZHOU et al., 2023).

O GB apresentou desempenho intermediário, com valores de r muito similares aos outros dois modelos. Embora apresente maior variabilidade em alguns folds, sua média geral ainda o posiciona entre os modelos de bom desempenho, sendo adequado para tarefas em que se deseja balancear precisão e interpretabilidade (RAZAVI et al., 2024; SHAWON et al., 2025). A LR, como esperado, foi o modelo de pior desempenho, com r baixos e MAE e RMSE elevados, demonstrando incapacidade de modelar relações não lineares complexas (RAHIMZADEH ARASHLOO; KITTLER, 2022; ZAKERI et al., 2020), especialmente comuns em dados espectrais de alimentos.

No grupo de aprendizado profundo, a CNN apresentou resultados inconsistentes. Apesar de um r médio elevado, observou-se uma grande variação entre os folds e valores altos de MAE e RMSE, o que indica que a arquitetura da rede pode não estar totalmente ajustada para lidar com este tipo de dado espectral. Como destacado por WANG, Y. et al., (2022), as CNNs foram originalmente projetadas para dados com estrutura espacial, como imagens, e sua aplicação direta em espectros demanda cuidados específicos, tanto na definição da arquitetura quanto no pré-processamento.

Esses fatores podem ter influenciado os resultados obtidos, especialmente considerando que o objetivo deste estudo foi manter uma metodologia padronizada entre os modelos de aprendizado de máquina e os de aprendizado profundo, buscando sempre equilíbrio entre simplicidade e robustez.

Já o modelo RNN, mesmo com desempenho inferior ao SVM e RF, mostrou mais estabilidade que a CNN, com bons resultados de r médio e menor variabilidade nos folds. Isso se deve ao fato de que dados espectrais apresentam forte estrutura sequencial (PAYNE; KUROUSKI, 2021) e RNNs, por sua natureza temporal, capturam dependências longitudinais mais eficientemente (SINGH; KASANA, 2022). No entanto, é importante destacar que redes profundas, como RNNs, requerem grandes volumes de dados para generalização adequada (BALA KRISHNAN; GOKILA, 2024; SHROTRIYA et al., 2024). O desempenho inferior observado neste estudo pode estar relacionado à limitação amostral ou à necessidade de maior ajuste de hiperparâmetros.

Assim, os resultados reforçam que, na ausência de grandes volumes de dados, modelos tradicionais de aprendizado de máquina como SVM e RF seguem sendo altamente eficazes (HUANG; CHEN; LIU, 2023; SHARMA et al., 2023), inclusive superando redes profundas em tarefas de regressão espectral agrícola, devido à escassez de bases rotuladas (ATTRI et al., 2023; VICTOR; NIBALI; HE, 2025).

Desta forma, a possibilidade de predizer o teor de proteína por meio de modelos baseados em dados espectrais e aprendizado de máquina mostra-se como uma alternativa estratégica, permitindo operações mais ágeis, padronizadas e sustentáveis no setor pós-colheita (SHI et al., 2023; TIAN et al., 2024; XUAN et al., 2024).

A Tabela 2 apresenta a ANOVA para os modelos testados na predição do teor de proteína nos diferentes tipos de Arroz Branco, Vermelho, Preto e Parboilizado. A análise foi realizada com base nas métricas médias de desempenho dos modelos: r, r^2,, MAE e RMSE. A significância estatística Pr>Fc = 0,0000 para todas as métricas demonstra que há diferenças reais entre os desempenhos dos modelos analisados. Entre os modelos, os algoritmos de aprendizado de máquina SVM, RF e GB destacaram-se com os melhores resultados, pertencendo ao mesmo grupo estatístico nas métricas de r, r^2,, MAE e RMSE.

Os modelos de aprendizado profundo CNN e RNN apresentaram desempenho intermediário. A RNN superou a CNN na métrica r^2,, porém com MAE e RMSE superiores, refletindo maior erro absoluto e quadrático. Ambos, no entanto, mostraram desempenho significativamente melhor que o modelo LR, que obteve os piores resultados em todas as métricas, reforçando sua limitação para modelar a complexidade dos dados espectrais.

A Figura 6 apresenta o Boxplot para verificação de distribuição de resíduos entre valores reais e preditos na predição de proteína nos diferentes tipos de Arroz. A distribuição de resíduos é uma ferramenta de diagnóstico fundamental para avaliar a qualidade dos ajustes de modelos em pesquisas agrícolas (CHAPUIS, 2023; RAN et al., 2022), bem como a presença de viés sistemático e a adequação ao pressuposto de normalidade dos erros.

Observa-se que os modelos RF, SVM e GB apresentaram histogramas com distribuições simétricas e centradas próximas de zero, sugerindo que os erros foram aleatórios e de variância constante. Essa característica é desejável em modelos preditivos, pois indica que as previsões não estão sistematicamente subestimando ou superestimando os valores reais (SCHIELZETH et al., 2020; XIAO et al., 2023).

Além disso, a forma de sino (gaussiana) dos histogramas desses três modelos sugere normalidade nos resíduos. Essa condição é essencial para validação de muitos testes estatísticos e indica maior robustez dos modelos (OBERPRILLER; SOUZA LEITE; DE; PICHLER, 2022).

A Figura 7 exibe as assinaturas espectrais completas das amostras dos quatro tipos de Arroz Preto, Vermelho, Branco e Parboilizado, coloridas de acordo com os teores de proteína. A visualização permite identificar padrões distintos de refletância ao longo das bandas espectrais entre os diferentes tipos de Arroz e destaca seus teores de proteína. É possível observar que as amostras com maiores teores de proteína apresentam curvas de refletância geralmente mais intensas em determinadas faixas do espectro, indicando potenciais regiões espectrais sensíveis a esse conteúdo.

Os tipos Preto e Vermelho exibem maior variação na assinatura espectral associada à proteína, especialmente nas faixas intermediárias e próximas ao infravermelho (RIBEIRO et al., 2020; UIVARASAN et al., 2024). Já o Arroz Branco e o Parboilizado apresentam padrões

mais homogêneos, sugerindo menor variação proteica detectável por reflectância, ou ainda menor sensibilidade espectral nestes tipos específicos (AKHTER et al., 2023; GARCÍA-SALCEDO et al., 2023). Essa representação visual serve como base importante para selecionar regiões espectrais que serão utilizadas na construção dos modelos preditivos.

A Figura 8 demonstra as 50 bandas espectrais mais importantes para a predição de proteína, classificadas de acordo com técnicas de explicabilidade de modelos em RF: forte concentração de importância entre 1200–1450 nm e também uma contribuição secundária próxima de 1500–1550 nm, regiões tradicionalmente associadas a absorções relacionadas a grupos N–H e O–H, característicos de proteínas e seus componentes (KAUR et al., 2020; YANG et al., 2024b).

Isso confirma a sensibilidade do infravermelho próximo (NIR) para detectar variações nos teores de Proteína por meio da reflectância. A seleção de apenas 50 bandas, entre mais de 2000, pode permitir redução de dimensionalidade e aumentar o entendimento dos modelos sem perda substancial de desempenho.

Na figura 9 que trata do Boxplot de comparação das métricas r, R2, MAE e RMSE entre os modelos na predição para Amido, os resultados obtidos demostram desempenhos muitos similares nas Predições realizadas para Proteína. Os modelos de aprendizado de máquina RF, SVM e GB mostram-se superiores quando comparados aos modelos de aprendizado profundo, com melhores valores de r e R2, bem como menores valores para MAE e RMSE. Entre estes modelos de aprendizado de máquina RF e SVM apresentaram as melhores médias de desempenho, com valores médios de r e R2 superiores e menores valores de erro.

Já GB também se destacou, com desempenho muito próximo ao RF e SVM. A regressão linear, repete seus resultados apresentados em Proteína, ficando significativamente atrás em todos as métricas analisas, indicando baixa capacidade preditiva para essa variável em comparação aos demais modelos.

Entre os modelos de Aprendizado Profundo, na predição de Amido, a CNN obteve resultados intermediários. Embora tenha apresentado um bom r, os valores de erro ainda foram maiores que os dos modelos RF, SVM e GB. Já a RNN, por sua vez, foi o modelo com pior

desempenho, apresentando valores muito superiores de erro, o que pode indicar dificuldade significativa em modelar os dados espectrais do Amido, assim como em Proteína.

O teor de amido no arroz, entre outras variáveis, destaca-se entre as mais importantes estudadas, pois pode determinar a qualidade final dos grãos e impactar em atributos como textura, viscosidade e propriedades culinárias após o cozimento (BOWEN et al., 2024; LI et al., 2020; TAO et al., 2020). Além disso, a composição amilosa-amilpectina está relacionada ao índice glicêmico do arroz, sendo um fator decisivo para mercados especializados, como produtos destinados a consumidores diabéticos ou dietas funcionais (KUMAR et al., 2022; RITHESH et al., 2024; SAHOO et al., 2025).

Para a indústria, conhecer o teor de amido de forma rápida e não destrutiva poderá permitir que lotes sejam direcionados para os processos mais adequados, otimizando etapas como polimento, secagem e empacotamento. Assim, a predição eficiente dessa variável é fundamental não apenas para a padronização e controle de qualidade, mas também para maximizar o valor comercial dos diferentes tipos de arroz analisados, seja branco, preto, vermelho ou parboilizado.

Desta forma, a espectroscopia no infravermelho próximo (NIR) e imagens hiperespectrais estão sendo amplamente utilizadas para prever o teor de amido em frutas, feijões e arroz. WANG et al., (2024), por exemplo, aplicou imagens hiperespectrais na predição de qualidade interna de frutas, incluindo teor de amido. No feijão-comum foi possível demonstrar a eficácia do NIR para a quantificação precisa de amido e proteínas nestes grãos (SINGH et al., 2024). Já em arroz, e diferentemente do presente estudo que não utiliza meios destrutivos, estudos comprovaram a viabilidade da espectroscopia hiperespectral combinada a métodos quimiométricos para estimar com precisão o conteúdo de amilose no arroz (GUO et al., 2021; YANG et al., 2024a).

Entretanto, ao associarmos esses dados a modelos de redes convolucionais, são necessários ajustes e adaptações específicas para lidar com espectros (GOMES et al., 2021); isso pode ter influenciado os resultados observados. No presente estudo, como o ocorrido em proteína, os resultados mantiveram uma estrutura metodológica padronizada entre os algoritmos e com resultados similares, o que reforça a superioridade dos métodos tradicionais de aprendizado de máquina com essa estrutura de dados utilizada.

A ANOVA da Tabela 3, que trata da predição de amido para os diferentes modelos analisados, indicou diferenças estatísticas com valores de Pr > Fc iguais a 0,0000 em todas as métricas, confirmando que houve real variação no desempenho entre os modelos. Os modelos RF, SVM e GB obtiveram os melhores resultados, com menores valores de MAE e RMSE e altos r^2 e r, agrupando-se todos no grupo "d". Já os modelos RNN e LR demonstraram os piores desempenhos, ficando agrupados e sem diferenças estatísticas entre si no grupo "a". A CNN, de forma distinta, apresentou resultados intermediários, com métricas melhores que RNN e LR, entretanto abaixo dos modelos de aprendizado de máquina.

Em se tratando da análise de resíduos, a Figura 10 demonstra que os modelos de aprendizado de máquina RF, SVM e GB apresentam dispersão simétrica e centrada próxima de zero, em formato aproximadamente gaussiano (sino), confirmado também por meio dos boxplots (Figura 11), por serem compactos e sem presença de outliers, demonstrando que os erros foram aleatórios e constantes.

Essas condições descritas demonstram a robustez dos modelos e sua capacidade de generalizar. Já RNN e LR, como evidente também na Figura 9, repetem esses resultados, demonstrando dispersão, caudas alongadas e outliers, comprometendo sua confiabilidade. Já CNN manteve-se como modelo mediano entre os outros, com resíduos normais e bons resultados, mas ainda muito inferiores ao aprendizado de máquina.

Esses resultados reforçam como RF, SVM e GB apresentam maior robustez na modelagem de dados espectrais (ADUGNA; XU; FAN, 2022; AGJEE et al., 2018; SHAO et al., 2023), especialmente quando o conjunto de dados possui ruídos ou variações naturais entre amostras, como ocorre com alimentos in natura (GUO et al., 2023). Por outro lado, os modelos baseados em redes neurais, especialmente a RNN, parecem ter sido sensíveis à arquitetura e aos dados analisados, o que reforça a necessidade de ajustes finos.

Portanto, os métodos tradicionais de aprendizado de máquina demonstraram maior estabilidade e menor variabilidade nos folds, o que é um indicativo importante de confiabilidade quando se pensa na aplicação prática desses modelos para substituição de métodos destrutivos. Além disso, os resultados apresentados entre os modelos para a predição de amido repetem-se de forma similar aos observados na predição de proteína, reforçando ainda mais a

aplicabilidade, confiabilidade e adequação desses modelos para este tipo de conjunto de dados, implicando diretamente na demonstração de sua capacidade de generalização.

Na Figura 11, que trata das assinaturas espectrais dos teores de amido no arroz preto, vermelho, branco e parboilizado, percebe-se diferenciação clara entre os padrões de amido entre os diferentes tipos de arroz. Na região do visível percebe-se as maiores variações (400–700 nm) e no início do infravermelho próximo (700–1100 nm), indicando a sensibilidade mais representativa destas faixas ao predizer a variável amido. Essas correlações de refletância são mais aparentes no arroz parboilizado, que mostra mais dispersão entre as amostras com diferentes teores de amido.

A identificação destas assinaturas espectrais de amido demonstra a correlação espectral com essa variável, evidenciando mais uma vez a capacidade funcional de ser modelada por algoritmos de aprendizado de máquina. Ou seja, mesmo com a variação entre os diversos tipos de arroz analisados, há padronização comum espectral ligada aos teores de amido (JOHN et al., 2022; WEI et al., 2023; XIE et al., 2022), o que justifica o bom desempenho dos modelos de aprendizado de máquina como RF, SVM e GB para predição dessa variável.

Na Figura 12, que trata da classificação das 50 bandas espectrais mais representativas para predição de amido nos tipos de arroz preto, vermelho, branco e parboilizado, essa análise é complementada revelando regiões do espectro mais concentradas entre 530 e 730 nm. Os picos de importância, como demonstrado na Figura 13, estão concentrados nos 381 nm e 642 nm, ambos localizados na região do visível, e foram identificados a partir da análise de importância de variáveis do algoritmo RF. Tais regiões têm sido frequentemente destacadas em estudos que utilizam espectroscopia VIS-NIR para caracterização de alimentos (AMDANI et al., 2023; PELLACANI et al., 2023).

Esses achados validam a utilização da espectroscopia para predição de amido e demonstram que, mesmo em culturas com processamentos distintos, como no caso do arroz parboilizado, e com diferenças de coloração, é possível encontrar bandas robustas para suas distinções. Com isso, abre-se um caminho para o desenvolvimento de sensores personalizados e mais baratos, otimizados apenas nas bandas mais relevantes, sem a necessidade de capturar todo o espectro, favorecendo aplicações comerciais.

Na Figura 14, observam-se os boxplots de comparação das métricas entre os modelos para predição de umidade nos diversos tipos de arroz. O aprendizado de máquina pode ser amplamente utilizado para prever e monitorar a umidade em diversos tipos de grãos (OLIVEIRA CARNEIRO et al., 2023; RIZWANA; HAZARIKA, 2024), um fator crítico para armazenamento, processamento e controle de qualidade.

A predição do teor de umidade no arroz desempenha um importante papel no contexto pós-colheita, já que influencia diretamente aspectos como armazenamento, estabilidade microbiológica e qualidade final do grão (QU et al., 2023; SAHA et al., 2025; ZHU et al., 2024). Altos teores de umidade favorecem o crescimento de fungos, bactérias e insetos, comprometendo a segurança alimentar e gerando perdas econômicas significativas (HE et al., 2022; QI et al., 2022).

Por outro lado, um teor de umidade muito baixo pode afetar negativamente o rendimento industrial, reduzir a textura desejada no cozimento e impactar a qualidade sensorial do arroz (FEDERICI et al., 2021; WANG, K. et al., 2024). Com isso, a aplicação de modelos preditivos que possam ser precisos e não destrutivos para monitorar a umidade nos grãos oferece uma ferramenta poderosa para otimizar a secagem, controlar o armazenamento e definir a classificação comercial dos lotes, promovendo maior eficiência e sustentabilidade nos processos pós-colheita.

Neste contexto, ao serem utilizados modelos preditivos para análise da variável umidade, repetem-se os desempenhos de forma muito similar aos apresentados nas predições de proteína e amido para os diferentes modelos testados. RF, SVM e GB apresentaram desempenhos superiores e, da mesma forma, esses modelos mostraram altos valores de r e r^2,, além de baixos valores de MAE e RMSE, com consistência entre os folds, demonstrando precisão e robustez, assim como nas outras características preditas anteriormente.

A análise de variância da Tabela 4 confirma as estatísticas significativas entre os modelos, agrupando SVM, RF, GB e RNN em um único grupo “c”, com destaque para SVM, RF e GB, que obtiveram melhores r, r^2 e menores MAE e RMSE. Esses modelos também apresentaram estabilidade entre os folds e baixa variabilidade dos resíduos, como evidenciado na Figura 14, reforçando sua robustez e confiabilidade.

A CNN, por outro lado, apresentou comportamento instável nesta variável, com distribuição de resíduos assimétrica e presença de outliers, refletindo em sua baixa capacidade preditiva observada na métrica r^2 negativa. Já a LR, como nas variáveis anteriores, foi o modelo com pior desempenho, com erros amplos, baixa correlação e resíduos altamente dispersos.

A Figura 15, que trata das assinaturas espectrais para os quatro tipos de arroz branco, preto, parboilizado e vermelho, coloridas de acordo com o teor de umidade, demonstra que há padrões distintos e bem definidos entre as amostras. Em todos os tipos, a região do visível (400–700 nm) apresenta variação perceptível na refletância conforme o teor de umidade.

Apesar das variações entre os tipos, observa-se uma tendência comum: quanto maior o teor de umidade, menor a refletância em determinadas faixas, principalmente entre 420 e 450 nm. Essa relação entre umidade e absorção óptica é esperada, já que a presença de água interfere diretamente na forma como a luz é absorvida e espalhada, especialmente nas bandas sensíveis à vibração da molécula de água (MALLET et al., 2021).

A Figura 16, que apresenta a classificação das 50 bandas espectrais mais representativas para a predição de umidade nos diferentes tipos de arroz, revela um padrão de destaque claro nas faixas iniciais do espectro visível, especialmente entre 420 e 430 nm, indicativo de que essas bandas carregam maior poder preditivo dos modelos utilizados. Isso ocorre porque a interação entre luz e água, principalmente nas regiões onde há absorção característica associada às vibrações das ligações O–H, altera a assinatura espectral capturada (LIU et al., 2021).

É importante destacar que, normalmente, a água apresenta absorção máxima no NIR em torno de 1450 nm e 1950 nm; desta forma, essas faixas não apareceram entre as bandas mais representativas neste estudo. Isso ocorre porque o FieldSpec 4 mede refletância, e em regiões de forte absorção pela água, quase não há luz refletida disponível para capturar variações úteis. Assim, os algoritmos priorizaram bandas no VIS, onde diferenças associadas à cor, espalhamento e características estruturais superficiais do grão forneceram informações mais discriminantes para a predição da umidade (MALEGORI et al., 2022; YADAV et al., 2025).

Entender quais bandas são mais representativas é essencial para aplicações práticas, pois permite otimizar os sensores utilizados. Em vez de depender de equipamentos hiperespectrais completos, seria possível trabalhar com sensores multiespectrais especificamente para essas

faixas, reduzindo custos operacionais sem comprometer a acurácia. Esses achados abrem caminhos promissores para avanços tecnológicos no setor pós-colheita, permitindo a implementação de soluções mais rápidas, automatizadas e economicamente viáveis para o monitoramento da qualidade de grãos em escala industrial.

A análise das variáveis fibras – Figura 17, cinzas – Figura 18 e lipídios – Figura 19, em grãos de arroz, revelou consistente desempenho entre os modelos de aprendizado de máquina; novamente, os modelos generalizaram bem, com SVM, RF e GB apresentando resultados similares aos encontrados nas variáveis proteína, amido e umidade. Mais do que resultados estatisticamente robustos, esses achados reforçam o potencial prático dessas abordagens para o setor pós-colheita.

A possibilidade de prever, com acurácia, características como teor de cinzas (RODRIGUES et al., 2024), fibras (JOHN et al., 2022) e lipídios (LU; JIANG; CHEN, 2021) utilizando técnicas não destrutivas representa um ganho expressivo em agilidade, padronização e automação de processos (DÍAZ et al., 2023; OLIVEIRA CARNEIRO, DE et al., 2023). Os algoritmos de aprendizado de máquina apresentaram os melhores resultados em todas as três variáveis, com valores elevados de r e r^2 e menores erros MAE e RMSE.

Esses modelos também demonstraram baixa variabilidade entre os folds, evidenciada pelos boxplots compactos em seus intervalos interquartis, indicando estabilidade, acurácia e capacidade de generalização mesmo em variáveis com dados de menor variância, como o teor de cinzas. Essas características são importantes e podem definir a qualidade nutricional, valor de mercado e destino destes grãos, influenciando decisões de secagem, armazenamento e até classificação industrial (MÜLLER et al., 2022; SCIAROT et al., 2020).

Estudos como os de (MANAFIFARD, 2024; RODRIGUES et al., 2024) mostram que modelos baseados em árvore de decisão e máquinas de vetores de suporte apresentam desempenho superior à LR e às redes neurais na predição de compostos como fibra bruta e lipídios em cereais, quando aplicados a espectros no intervalo do visível ao infravermelho próximo (VIS-NIR). Segundo PEREIRA et al. (2023), essa superioridade decorre da robustez desses algoritmos diante da multicolinearidade das bandas espectrais e da capacidade de modelar relações não lineares entre os dados espectrais e os atributos químicos (DERRAZ et al., 2023; SU et al., 2024).

No caso das redes neurais, a CNN apresentou desempenho intermediário, com resultados promissores na predição de fibras, mas variações maiores em cinzas e lipídios. Essa instabilidade pode ser atribuída à sensibilidade do modelo à arquitetura e à quantidade de dados disponíveis (WANG, Y. et al., 2022). Já a RNN manteve desempenho inferior aos modelos tradicionais em quase todas as variáveis, com grande dispersão dos erros em cinzas — o que pode ser explicado pela ausência de estrutura sequencial nos dados espectrais utilizados, tornando sua aplicação menos adequada (PERICH et al., 2023).

A LR foi, mais uma vez, o modelo com pior desempenho em todas as variáveis analisadas. Os baixos valores de r^2 e a alta variabilidade dos erros reforçam que os atributos físico-químicos aqui estudados não apresentam relação linear com os dados espectrais (KADAM; JADHAV, P. A., et al., 2024; SAMPAIO; ALMEIDA; BRITES, 2021; YU et al., 2023).

Esses resultados reforçam o papel da espectroscopia aliada ao aprendizado de máquina como ferramenta poderosa para predição não destrutiva de compostos químicos na pós-colheita, com destaque para os modelos baseados em árvores e SVM. Além disso, com avanços tecnológicos promovidos nas ciências agrárias e do sensoriamento remoto, esses modelos podem ser incorporados a sistemas automatizados de triagem e controle de qualidade em tempo real, otimizando recursos e reduzindo perdas pós-colheita, o que pode ser promissor para aplicação em escala industrial.

A análise de variância aplicada às características fibras – Tabela 05, cinzas – Tabela 06 e lipídios – Tabela 07, demonstrou um padrão de desempenho dos modelos testados, reforçando as tendências já observadas nos boxplots anteriores. As métricas r e r^2 mostram que os algoritmos de aprendizado de máquina RF, GB e SVM foram os mais consistentes.

Prever o teor de fibra de arroz na pós-colheita é importante para avaliar a qualidade e o valor nutricional do arroz (WATTANAVANITCHAKORN et al., 2021, 2023). Neste estudo, na tarefa de predição do teor de fibras, os modelos RF, GB e SVM alcançaram os maiores valores de r e r^2,, com os menores MAE e RMSE, todos agrupados estatisticamente no mesmo grupo “d”. Isso demonstra não apenas precisão, mas também robustez dos modelos, por fornecerem uma abordagem rápida, precisa e não destrutiva para prever o teor de fibra do arroz pós-colheita (JOHN et al., 2022; KADAM; JADHAV, P., et al., 2024).

Diferentemente do que ocorreu em proteína, amido e umidade, foi o RF que se destacou, possivelmente pela pequena variância apresentada nesta variável e por ser eficaz com variáveis contínuas e categóricas, tornando-o adequado para diversos conjuntos de dados agrícolas (AKBARI et al., 2020).

Para cinzas, uma variável que naturalmente possui menor variação — fator quantitativo ainda mais expressivo do que em fibras — podendo gerar desafios preditivos pelas pequenas concentrações apresentadas, os resultados apontam superioridade dos modelos de árvore e SVM. As cinzas da casca de arroz são importantes variáveis a serem observadas, pois podem diminuir o acúmulo de elementos tóxicos, como arsênio (As) e cádmio (Cd) nos grãos de arroz, tornando-os mais seguros para consumo (GUPTA et al., 2024; JIANG et al., 2024; WANG; WANG; PENG, 2022).

Portanto, um fator não menos importante a ser analisado do que outras variáveis já testadas pelos modelos. Nessa tarefa, RF e GB, respectivamente, foram os modelos que apresentaram os melhores resultados em todas as métricas, com MAE baixos e r^2 superiores, agrupando-se no grupo "e". Já a CNN, apesar de seus bons resultados de r, apresentou desempenho mediano por possuir maiores MAE e RMSE se comparado a SVM. Mais uma vez, RNN e LR ficaram nos grupos inferiores, com altos erros e r^2 insatisfatórios, apresentando inclusive resultados negativos no caso da LR.

Por último, mas não menos importante, foi realizada a predição de lipídios entre os modelos. Lipídios e ácidos graxos afetam significativamente a qualidade do arroz, incluindo textura e valor nutricional (CHANG et al., 2024). Além disso, isso torna as variedades de arroz com alto teor de lipídios benéficas para o desenvolvimento de alimentos funcionais com menor impacto na glicemia e maior disponibilidade do composto bioativo (KHATUN; WATERS; LIU, 2022; SHEN et al., 2021). Estes fatores podem impactar diretamente na tomada de decisão pós-colheita, servindo como parâmetro para classificação comercial ou exportação (LAN et al., 2025; SHEN et al., 2021).

Tendo em vista essas qualidades, durante a tarefa de predição, os valores médios de r e r^2 em todos os modelos ficaram abaixo dos obtidos em outras variáveis, refletindo a maior dificuldade dos modelos. Ainda assim, SVM, RF, GB e RNN apresentaram resultados equivalentes estatisticamente, com destaque para o SVM, que obteve o menor RMSE. A LR foi

novamente a pior, com r^2 negativo e agrupamento no grupo “a”, evidenciando a ineficácia em lidar com relações não lineares envolvidas na predição deste atributo.

Em todas as características, os valores de Pr>Fc foram inferiores a 0,0000 e confirmam a existência de diferença estatística significativa entre os modelos para as métricas avaliadas. Ocorreu baixa variabilidade e os bons CVs reforçam a confiabilidade.

Esses dados contribuem com os boxplots apresentados, pois os modelos de aprendizado de máquina foram mais estáveis, com menor dispersão e desempenho superior frente às redes neurais CNN e RNN. Essa consistência entre os métodos estatísticos e gráficos fortalece a conclusão de que RF, SVM e GB são ferramentas eficientes, mesmo em variáveis de baixa magnitude como fibras, cinzas e lipídios — um ponto fundamental para aplicações em triagem pós-colheita e controle de qualidade com espectroscopia não destrutiva.

As Figuras 20, 21 e 22 apresentam boxplots que ilustram a distribuição dos resíduos para os diferentes modelos de predição aplicados às variáveis fibras, cinzas e lipídios no arroz e reforçam visualmente as tendências observadas nos histogramas, confirmando que os modelos baseados em árvores RF, GB e SVM são os mais consistentes na predição das variáveis. A distribuição compacta e simétrica dos resíduos desses modelos indica melhor capacidade preditiva e menor viés.

A Figura 23 mostra as assinaturas espectrais médias associadas aos teores de fibras nos diferentes tipos de arroz preto, vermelho, branco e parboilizado, em função do comprimento de onda. Observa-se que as assinaturas espectrais variam entre os tipos de arroz, indicando que cada um apresenta padrões de absorção específicos, possivelmente associados à composição físico-química de suas cascas e camadas externas (LI et al., 2024; ONMANKHONG et al., 2022; WANG; TAN, 2021).

O arroz preto e o vermelho são mais ricos em compostos bioativos e fibras insolúveis, podendo apresentar maior variação na região do visível (350–700 nm), e também nas faixas do infravermelho próximo (AMDANI et al., 2023; WANG et al., 2023). O arroz parboilizado e o branco, por sua vez, apresentam espectros mais homogêneos, o que pode estar relacionado à redução dos constituintes fibrosos no processo industrial de parboilização ou no polimento do arroz branco (GARCÍA-SALCEDO et al., 2023; PAL et al., 2018).

Essas diferenças espectrais justificam a capacidade dos modelos preditivos em diferenciar os tipos de arroz com base no teor de Fibras, destacando a utilidade da espectroscopia como ferramenta não destrutiva e precisa.

A Figura 24 apresenta a importância relativa das 50 bandas espectrais mais representativas para predição do teor de Fibras nos diversos tipos de arroz analisados e considerando sua contribuição em cada comprimento de onda para os modelos de aprendizado de máquina.

As bandas que mais representam fibras para os modelos estão na faixa dos 390–400 nm, ou seja, no início do espectro visível, indicando forte relação com componentes que absorvem luz, como fenólicos e ligninas, relacionados à presença de fibras (BALASUBRAMANIAN; VENKATACHALAM, 2023; KAUR; GOYAL, 2024). Já o segundo grupo de bandas mais representativas está presente nas faixas de 1420 a 1460 nm e 1880 a 1940 nm, que correspondem a regiões do infravermelho próximo (NIR) associadas a ligações moleculares do tipo C–H e O–H, comuns em celulose e hemicelulose (FAN et al., 2023; WANG, N. et al., 2022).

É importante observar que a distribuição não se concentrou em uma única região do espectro, mas sim em uma combinação multiespectral, o que demonstra a necessidade de modelos que considerem múltiplas faixas em conjunto. Esse comportamento reforça, ainda, o potencial para redução de custos nas aplicações práticas, visto que sensores multiespectrais — mais acessíveis que os hiperespectrais — poderiam ser empregados com eficiência na predição desta característica para fins industriais.

A Figura 25 apresenta as assinaturas espectrais das amostras de arroz preto, branco, parboilizado e vermelho, coloridas de acordo com o teor de cinzas. Em todos eles é possível observar variações em suas curvas espectrais ao longo do comprimento de onda, com destaque para picos nas regiões entre 500–700 nm e entre 950–1350 nm, sugerindo relação com componentes minerais e matéria orgânica associada às cinzas (SINGH; SINGH, 2021). As regiões entre 500–700 nm normalmente são mais representativas para compostos orgânicos, já os componentes minerais são mais detectáveis fora da faixa de 500–700 nm (GARCÍA-SALCEDO et al., 2023).

No arroz preto e vermelho, observa-se dispersão nas curvas na região do NIR, refletindo uma diversidade química e de compostos fenólicos que podem influenciar no teor de cinzas

(GUPTA et al., 2024; NATH et al., 2022). Já no arroz branco, as curvas apresentam maior uniformidade, coerente com o processamento de polimento que remove parte significativa da casca e das camadas externas — regiões onde os minerais se concentram (AKHTER et al., 2023; HENSAWANG et al., 2020; YAO; CHEN; SUN, 2020). O arroz parboilizado também apresenta curvas mais compactas, possivelmente resultantes do rearranjo físico-químico ocorrido pelo tratamento térmico (BALBINOTI et al., 2022; ZHU et al., 2021).

Ao analisarmos a Figura 26, que trata das assinaturas espectrais dos teores de lipídios entre o arroz preto, vermelho, branco e parboilizado, observa-se uma distribuição espectral com destaque para alguns picos nas regiões visível (350–700 nm) e, de forma mais determinante, no infravermelho próximo (1100–1900 nm). O gradiente de cores aplicado aos espectros sugere que, embora o teor de lipídios seja baixo de forma geral, ele exerce um padrão detectável no comportamento espectral para esta variável, especialmente nas faixas NIR.

Em se tratando do processo de pós-colheita, a predição e quantificação de lipídios é altamente importante. Os lipídios são altamente reativos, apresentando rápida oxidação em condições de calor, umidade e exposição ao ar (BHUNIA et al., 2023; LIU, X. et al., 2024). Sob esses efeitos podem gerar produtos degradáveis que reduzem o valor nutricional e causam efeitos negativos à saúde (PRASAD, C. T. et al., 2022; YAN et al., 2020). Tendo estes fatores em vista, a predição rápida e não destrutiva de lipídios no pós-colheita ajudaria a definir padrões ideais e, com isso, priorizar o escoamento rápido de lotes mais ricos em lipídios, reduzindo perdas.

Ao analisarmos a Figura 27, em conjunto com a interpretação da Figura 28, observa-se, diferentemente de cinzas e fibras, uma distribuição bem marcada nas faixas do NIR, especialmente entre 1400–2000 nm. Os picos mais representativos parecem concentrados, por exemplo, em torno de 1442 nm, 1538 nm, 1834 nm e 1981 nm, que são regiões associadas à absorção por grupos metila e metileno (C–H), típicos de compostos lipídicos (FAN et al., 2022; LIU et al., 2021; XU et al., 2022).

As bandas do visível, neste caso, apresentam relevância menor, reforçando que uma abordagem multiespectral seria a mais recomendada, embora a região NIR carregue a maior carga preditiva para os lipídios (XU et al., 2022). Além disso, o predomínio das bandas NIR demonstra que, para aplicações pós-colheita, industriais e comerciais, a utilização de sensores específicos nessa faixa pode minimizar custos e simplificar o processo, sem comprometer a capacidade de predição.

Este estudo atingiu os objetivos propostos ao demonstrar que é possível prever com alta acurácia os atributos físico-químicos de diferentes tipos de arroz — incluindo proteína, amido, umidade, fibras, cinzas e lipídios — utilizando espectroscopia hiperespectral (350–2500 nm), cobrindo as regiões VIS, NIR e SWIR, e algoritmos de aprendizado de máquina, com destaque para o SVM, que foi superior em todas as variáveis testadas, demonstrando forte estabilidade, acurácia e generalização entre variáveis.

O estudo identificou bandas espectrais representativas para cada componente, revelando que sensores multiespectrais também garantiriam eficiência preditiva. A seleção de faixas específicas pode permitir a utilização de sensores multiespectrais otimizados, capazes de reduzir custos, simplificar processos e acelerar operações industriais.

Além disso, os resultados reforçam a relevância prática das tecnologias não destrutivas para o setor pós-colheita, com triagem automatizada, controle de qualidade em tempo real e tomada de decisão orientada por dados. Ao integrar espectroscopia e aprendizado de máquina, este trabalho avança a fronteira científica no monitoramento da qualidade de grãos e contribui para a inovação tecnológica e sustentabilidade na cadeia produtiva agrícola.

6. CONCLUSÕES

Este estudo comprovou a eficácia da integração entre espectroscopia hiperespectral (350–2500 nm) e algoritmos de aprendizado de máquina na predição de atributos físico-químicos de diferentes tipos de arroz seus atributos como: proteína, amido, umidade, fibras, cinzas e lipídios. Entre os modelos avaliados, SVM destacou-se pelo desempenho superior, evidenciando robustez estatística e capacidade de generalização entre as variáveis. A análise detalhada das bandas espectrais permitiu identificar regiões representativas para a predição, reforçando o potencial do desenvolvimento de sensores multiespectrais otimizados. De forma geral, os resultados obtidos consolidam a aplicação de modelos não destrutivos no contexto pós-colheita, oferecendo avanços para a otimização de operações industriais, bem como para o aprimoramento da precisão nos processos de classificação e controle de qualidade.

7. REFERÊNCIAS

ABDOLLAHPOUR, S.; KOSARI-MOGHADDAM, A.; BANNAYAN, M. Prediction of wheat moisture content at harvest time through ANN and SVR modeling techniques. **Information Processing in Agriculture**, v. 7, n. 4, p. 500–510, dez. 2020.

ABEYSIRIWARDENA, D. S. DE Z.; GUNASEKARA, D. C. S. Development of a red rice variety with excellent health properties and attractive grain qualities. **Indian Journal of Genetics and Plant Breeding (The)**, v. 80, n. 01, 29 abr. 2020.

AEKRAM, S. *et al.* Differentiation of white rice Khao Dawk Mali 105 (KDML105) from other Thai white rice using FTIR microscope-attenuated total reflectance-focal plane array detector. **Food Research**, v. 9, n. 1, p. 282–287, 28 fev. 2025.

AKBARI, E. *et al.* Crop Mapping Using Random Forest and Particle Swarm Optimization based on Multi-Temporal Sentinel-2. **Remote Sensing**, v. 12, n. 9, p. 1449, 3 maio 2020.

AKHTER, K. T. *et al.* Variations in the Major Nutrient Composition of Dominant High-Yield Varieties (HYVs) in Parboiled and Polished Rice of Bangladesh. **Foods**, v. 12, n. 21, p. 3997, 1 nov. 2023.

AMDANI, R. Z. *et al.* The Potency of Visible and Near-Infrared Reflectance Spectroscopy to Profiling and Classify the Common Rice Flour. **IOP Conference Series: Earth and Environmental Science**, v. 1168, n. 1, p. 012003, 1 abr. 2023.

ASSOCIAÇÃO BRASILEIRA DAS INDÚSTRIAS DE ARROZ (ABIA). **Qualidade do arroz no Brasil: evolução e padronização.** São Paulo, 2023.

ATTRI, I. *et al.* A review of deep learning techniques used in agriculture. **Ecological Informatics**, v. 77, p. 102217, nov. 2023.

AZEEZ, Z. N.; QASIM, ASEEL. A.; MOHIE, N. M. Classification of Rice Grains by Image Processing. **Academia Open**, v. 9, n. 2, 15 out. 2024.

BALA KRISHNAN, V.; GOKILA, S. Exploring the Impact of Different Learning Layers within Recurrent Neural Networks on Crop Yield Prediction, 2024, Second International Conference on Data Science and Information System (ICDSIS). **Anais**...IEEE, 17 maio 2024.

BALASUBRAMANIAN, S.; VENKATACHALAM, P. Valorization of rice husk agricultural waste through lignin extraction using acidic deep eutectic solvent. **Biomass and Bioenergy**, v. 173, p. 106776, jun. 2023.

BALBINOTI, T. C. V. *et al.* Multiphysics simulation and characterisation of parboiling of long grain rice during hydration. **Journal of Cereal Science**, v. 103, p. 103391, jan. 2022.

BANERJEE, R. *et al.* Quality evaluation of different black rice varieties of northeastern region of India. **Phytochemical Analysis**, v. 34, n. 5, p. 507–517, 16 jul. 2023.

BARNES, R. J.; DHANOA, M. S.; LISTER, S. J. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. **Applied Spectroscopy**, v. 43, n. 5, p. 772–777, 1 jul. 1989.

BELGIU, M.; DRĂGUŢ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, abr. 2016.

BERNARDO, A. C. et al. Qualidade tecnológica de cultivares de arroz irrigado em dois sistemas de cultivo. *Ciência e Agrotecnologia*, Lavras, v. 34, n. 2, p. 394-400, 2010.

BHUNIA, R. K. *et al.* A Holistic View of the Genetic Factors Involved in Triggering Hydrolytic and Oxidative Rancidity of Rice Bran Lipids. **Food Reviews International**, v. 39, n. 1, p. 441–466, 2 jan. 2023.

BHUPENDRA *et al.* Deep CNN-based damage classification of milled rice grains using a high-magnification image dataset. **Computers and Electronics in Agriculture**, v. 195, p. 106811, abr. 2022.

BORAH, J. L. *et al.* A Validated Method for Identification and Quantification of Anthocyanins in Different Black Rice, Oryza sativa. Varieties Using High-Performance Thin-Layer Chromatography (HPTLC). **Phytochemical Analysis**, 21 jan. 2025.

BOWEN, D. *et al.* Genome-Wide Association Study of Cooked Rice Textural Attributes and Starch Physicochemical Properties in indica Rice. **Rice Science**, v. 31, n. 3, p. 300–316, maio 2024.

BRASIL. INSTRUÇÃO NORMATIVA 2/2012. MINISTÉRIO DA AGRICULTURA, P. E ABASTECIMENTO. 2012. **Brasil. 2/2012.** Brasil: [s.n.].

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa nº 6, de 16 de fevereiro de 2011. *Padrões oficiais de classificação do arroz*. Diário Oficial da União: seção 1, Brasília, DF, 17 fev. 2011.

BREIMAN, L. Random forests. Machine learning, v. 45, p. 5-32, 2001.

BRERETON, R. G. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. Chichester: **Wiley,** 2003.

BRINKHOFF, J.; DUNN, B. W.; DUNN, T. The influence of nitrogen and variety on rice grain moisture content dry-down. **Field Crops Research**, v. 302, p. 109044, out. 2023.

BRUNET-LOREDO, A. *et al.* Assessing Grain Quality Changes in White and Black Rice under Water Deficit. **Plants**, v. 12, n. 24, p. 4091, 7 dez. 2023.

CAO, R. *et al.* OPAQUE3, encoding a transmembrane bZIP transcription factor, regulates endosperm storage protein and starch biosynthesis in rice. **Plant Communications**, v. 3, n. 6, p. 100463, nov. 2022.

CHANG, L. *et al.* Molecular Basis of Lipid Metabolism in Oryza sativa L. **Plants**, v. 13, n. 23, p. 3263, 21 nov. 2024.

CHAPUIS, R. P. Fitting models for a grain size distribution: a review. **Bulletin of Engineering Geology and the Environment**, v. 82, n. 11, p. 427, 25 nov. 2023.

CONAB. Acompanhamento da safra brasileira: grãos - safra 2024/25 – oitavo levantamento. Brasília: **Companhia Nacional de Abastecimento**, v. 12, n. 8, 2025.

CONBEA. **Anais do CONBEA 2023: Ciência e Tecnologia Pós-Colheita - CTP 5.** São Paulo, 2023. Disponível em: https://conbea.org.br/anais/publicacoes/conbea-2023/anais-2023/ciencia-e-tecnologia-pos-colheita-ctp-5/3624-espectroscopia-de-infravermelho-proximo-para-determinacao-da-composicao-centesimal-do-arroz-polido-e-arroz-integral-como-alternativa-ao-metodo-da-classificacao-fisica/file. Acesso em: 2 out. 2025.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

CORTES, C.; VAPNIK, V. Support-vector networks. Machine learning, v. 20, p. 273-297, 1995.

DAS, B. *et al.* Evaluation of different water absorption bands, indices and multivariate models for water-deficit stress monitoring in rice using visible-near infrared spectroscopy. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 247, p. 119104, fev. 2021.

DAS, M. *et al.* Black rice: A comprehensive review on its bioactive compounds, potential health benefits and food applications. **Food Chemistry Advances**, v. 3, p. 100462, dez. 2023.

DE OLIVEIRA, M.; AMATO, G. W. Arroz: tecnologia, processos e usos. **Editora Blucher**, 2021.

DERRAZ, R. *et al.* Ensemble and single algorithm models to handle multicollinearity of UAV vegetation indices for predicting rice biomass. **Computers and Electronics in Agriculture**, v. 205, p. 107621, fev. 2023.

DÍAZ, E. O. *et al.* Non-destructive quality classification of rice taste properties based on near-infrared spectroscopy and machine learning algorithms. **Food Chemistry**, v. 429, p. 136907, dez. 2023.

DOS SANTOS, A. B. **A cultura do arroz no Brasil**. 2006.

DRAPER, N. R.; SMITH, H. Applied regression analysis. John Wiley & Sons, 1998.

ELAVARASAN, D.; VINCENT, P. M. D. R. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. **Journal of Ambient Intelligence and Humanized Computing**, v. 12, n. 11, p. 10009–10022, 1 nov. 2021.

EMBRAPA. **Comunicado técnico 84: avaliação da qualidade do arroz.** Brasília, 2020.

ERCANLI, İ. Deep learning algorithms for addressing overfitting and biological realism in tree taper and volume predictions. **Canadian Journal of Forest Research**, v. 54, n. 12, p. 1500–1518, 1 dez. 2024.

FAN, S. *et al.* Establishment of Non-Destructive Methods for the Detection of Amylose and Fat Content in Single Rice Kernels Using Near-Infrared Spectroscopy. **Agriculture**, v. 12, n. 8, p. 1258, 19 ago. 2022.

FĂRCAȘ, A. C., et al. An update regarding the bioactive compound of cereal by-products: Health benefits and potential applications. **Nutrients**, n. 14, v. 17, p. 3470, 2022.

FEDERICI, E. *et al.* Ready to eat shelf-stable brown rice in pouches: effect of moisture content on product's quality and stability. **European Food Research and Technology**, v. 247, n. 11, p. 2677–2685, 15 nov. 2021.

FERREIRA, D. F. SISVAR: A COMPUTER ANALYSIS SYSTEM TO FIXED EFFECTS SPLIT PLOT TYPE DESIGNS. **REVISTA BRASILEIRA DE BIOMETRIA**, v. 37, n. 4, p. 529–535, 20 dez. 2019.

FEYISA, G. L. *et al.* Characterizing and mapping cropping patterns in a complex agro-ecosystem: An iterative participatory mapping procedure using machine learning algorithms and MODIS vegetation indices. **Computers and Electronics in Agriculture**, v. 175, p. 105595, ago. 2020.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. Annals of statistics, p. 1189-1232, 2001.

GAO, Y. *et al.* Cooked Rice Textural Properties and Starch Physicochemical Properties from New Hybrid Rice and Their Parents. **Foods**, v. 13, n. 7, p. 1035, 28 mar. 2024.

GAONA, J. *et al.* Predictive value of soil moisture and concurrent variables in the multivariate modelling of cereal yields in water-limited environments. **Agricultural Water Management**, v. 282, p. 108280, maio 2023.

GARCÍA-SALCEDO, Á. J. *et al.* Analysis of compositional differences between commercial rice grains by the study of the photoluminescence response. **Journal of Cereal Science**, v. 111, p. 103681, maio 2023.

GAROFALO, S. PIETRO *et al.* Predicting carob tree physiological parameters under different irrigation systems using Random Forest and Planet satellite images. **Frontiers in Plant Science**, v. 15, 19 mar. 2024.

GOGOI, P. *et al.* Nutritional profile and mineral bioaccessibility of pigmented rice landraces. **Journal of Food Measurement and Characterization**, v. 19, n. 2, p. 1513–1530, 16 fev. 2025.

GOMES, V. *et al.* Prediction of Sugar Content in Port Wine Vintage Grapes Using Machine Learning and Hyperspectral Imaging. **Processes**, v. 9, n. 7, p. 1241, 19 jul. 2021.

GONDAL, T. A. *et al.* Consumer Acceptance of Brown and White Rice Varieties. **Foods**, v. 10, n. 8, p. 1950, 22 ago. 2021.

GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. Why do tree-based models still outperform deep learning on typical tabular data?. Advances in neural information processing systems, v. 35, p. 507-520, 2022.

GUO, T. *et al.* FAPD: An Astringency Threshold and Astringency Type Prediction Database for Flavonoid Compounds Based on Machine Learning. **Journal of Agricultural and Food Chemistry**, v. 71, n. 9, p. 4172–4183, 8 mar. 2023.

GUO, Y. *et al.* Integrated phenology and climate in rice yields prediction using machine learning methods. **Ecological Indicators**, v. 120, p. 106935, jan. 2021.

GUPTA, Y. *et al.* Recycled Household Ash in Rice Paddies of Bangladesh for Sustainable Production of Rice Without Altering Grain Arsenic and Cadmium. **Exposure and Health**, v. 16, n. 1, p. 87–99, 9 fev. 2024.

HASHIMOTO, M. *et al.* The journey from white rice to ultra-high hydrostatic pressurized brown rice: an excellent endeavor for ideal nutrition from staple food. **Critical Reviews in Food Science and Nutrition**, v. 62, n. 6, p. 1502–1520, 16 fev. 2022.

HAZRUL, A. F. H.; RAJA IBRAHIM, R. K.; DURALIM, M. Discrimination of Rice Varieties using Laser-Induced Breakdown Spectroscopy assisted with Principal Component Analysis (PCA). **Journal of Physics: Conference Series**, v. 2974, n. 1, p. 012009, 1 mar. 2025.

HE, X. *et al.* Analysis of rice microbial communities under different storage conditions using culture-dependent and -independent techniques. **Quality Assurance and Safety of Crops & Foods**, v. 14, n. 1, p. 1–11, 20 jan. 2022.

HENSAWANG, S. *et al.* Probabilistic assessment of the daily intake of microelements and toxic elements via the consumption of rice with different degrees of polishing. **Journal of the Science of Food and Agriculture**, v. 100, n. 10, p. 4029–4039, 17 ago. 2020.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1 nov. 1997.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. Neural computation, v. 9, n. 8, p. 1735-1780, 1997.

HORWITZ W, C. P. R. H. **Official methods of analysis of the Association of Official Analytical Chemists**. Washington, DC,USA: Association of Official Analytical Chemists, 1970.

HUANG, Y.; CHEN, Z.; LIU, J. Limited agricultural spectral dataset expansion based on generative adversarial networks. **Computers and Electronics in Agriculture**, v. 215, p. 108385, dez. 2023.

HUBER, F. *et al.* Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches. **Computers and Electronics in Agriculture**, v. 202, p. 107346, nov. 2022.

INSTITUTO NACIONAL DE INVESTIGAÇÃO AGRÁRIA E VETERINÁRIA (INIAV). **Tecnologia rápida e não destrutiva**. Lisboa, 2024.

INTERNATIONAL RICE COMMISSION, AND INTERNATIONAL YEAR OF RICE (PROGRAM). Rice is Life: International Year of Rice 2004 and Its Implementation. **Food & Agriculture Org**., 2005.

JIANG, Y. *et al.* Regulation of rhizosphere microenvironment by rice husk ash for reducing the accumulation of cadmium and arsenic in rice. **Journal of Environmental Sciences**, v. 136, p. 1–10, fev. 2024.

JOHN, R. *et al.* Germplasm variability-assisted near infrared reflectance spectroscopy chemometrics to develop multi-trait robust prediction models in rice. **Frontiers in Nutrition**, v. 2022, 4 ago. 2022.

JOHNSON, J.-M. *et al.* Application of infrared spectroscopy for estimation of concentrations of macro- and micronutrients in rice in sub-Saharan Africa. **Field Crops Research**, v. 270, p. 108222, ago. 2021.

JULCARIMA, R.; BEATRIZ C.; LLANOS, E. M. R. **Efecto del parbolizado en las propiedades sensoriales, contenido de vitaminas y minerales en dos variedades de arroz (oryza sativa) producido en el país**. 2018.

KABIR, M. S., et al. "Free and Insoluble-Bound Phenolics in Parboiled and Non-Parboiled Rice after Hydrothermal Treatments." **Maruf, Free and Insoluble-Bound Phenolics in Parboiled and Non-Parboiled Rice after Hydrothermal Treatments**. 2024.

KADAM, S.; JADHAV, P. A.; *et al.* Characterization of rice cultivars using Raman spectroscopy and multivariate analysis. **Biocatalysis and Agricultural Biotechnology**, v. 60, p. 103280, set. 2024.

KADAM, S.; JADHAV, P.; *et al.* Raman Spectroscopic Characterization of Local Rice Germplasm from Konkan Region of Maharashtra. **Food Analytical Methods**, v. 17, n. 3, p. 426–435, 23 mar. 2024.

KALITA, T.; GOHAIN, U. P.; HAZARIKA, J. Effect of Different Processing Methods on the Nutritional Value of Rice. **Current Research in Nutrition and Food Science Journal**, v. 9, n. 2, p. 683–691, 31 ago. 2021.

KANG, Z. *et al.* The Rapid Non-Destructive Differentiation of Different Varieties of Rice by Fluorescence Hyperspectral Technology Combined with Machine Learning. **Molecules**, v. 29, n. 3, p. 682, 1 fev. 2024.

KAUR, H.; GOYAL, D. Alkaline treatment for targeted lignin breakdown in rice straw: Maximizing phenolic content and product characterization. **Food and Bioproducts Processing**, v. 148, p. 478–490, dez. 2024.

KAUR, N. *et al.* Structural insights into rice SalTol QTL located SALT protein. **Scientific Reports**, v. 10, n. 1, p. 16589, 6 out. 2020.

KHATUN, A.; WATERS, D. L. E.; LIU, L. The Impact of Rice Lipid on In Vitro Rice Starch Digestibility. **Foods**, v. 11, n. 10, p. 1528, 23 maio 2022.

KHATUN, S.; MOLLAH, MD. M. I. Analysis of black rice and some other cereal grains for protein, sugar, polyphenols, antioxidant and anti-inflammatory properties. **Journal of Agriculture and Food Research**, v. 16, p. 101121, jun. 2024.

KUMAR, A. *et al.* Biochemical markers for low glycemic index and approaches to alter starch digestibility in rice. **Journal of Cereal Science**, v. 106, p. 103501, jul. 2022.

LAN, L. *et al.* A Comprehensive Investigation of Lipid Profile During the Solid-State Fermentation of Rice by Monascus purpureus. **Foods**, v. 14, n. 3, p. 537, 6 fev. 2025.

LECUN, Y. *et al.* Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LECUN, Yann et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, v. 86, n. 11, p. 2278-2324, 2002.

LI, C. *et al.* Causal relations among starch chain-length distributions, short-term retrogradation and cooked rice texture. **Food Hydrocolloids**, v. 108, p. 106064, nov. 2020.

LI, C. *et al.* Rice Origin Tracing Technology Based on Fluorescence Spectroscopy and Stoichiometry. **Sensors**, v. 24, n. 10, p. 2994, 9 maio 2024.

LI, Y. *et al.* A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. **International Journal of Applied Earth Observation and Geoinformation**, v. 118, p. 103269, abr. 2023.

LIU, X. *et al.* Lipidomics analysis unveils the dynamic alterations of lipid degradation in rice bran during storage. **Food Research International**, v. 184, p. 114243, maio 2024.

LIU, Y. *et al.* Detection of fraud in high-quality rice by near-infrared spectroscopy. **Journal of Food Science**, v. 85, n. 9, p. 2773–2782, 26 set. 2020.

LIU, Y. *et al.* Estimation of Potato Above-Ground Biomass Using UAV-Based Hyperspectral images and Machine-Learning Regression. **Remote Sensing**, v. 14, n. 21, p. 5449, 29 out. 2022.

LIU, Y. *et al.* Spectral Data-Driven Prediction of Soil Properties Using LSTM-CNN-Attention Model. **Applied Sciences**, v. 14, n. 24, p. 11687, 14 dez. 2024.

LOXUS UFV. **Análise de imagens computadorizadas para classificação de vigor de sementes de arroz.** Viçosa, 2023.

LU, H.; JIANG, H.; CHEN, Q. Determination of Fatty Acid Content of Rice during Storage Based on Feature Fusion of Olfactory Visualization Sensor Data and Near-Infrared Spectra. **Sensors**, v. 21, n. 9, p. 3266, 9 maio 2021.

MAGNIMIND ACADEMY. **All Machine Learning Algorithms You Should Know In 2023.** 2023.

MAHMOOD, N. *et al.* Influences of emerging drying technologies on rice quality. **Food Research International**, v. 184, p. 114264, maio 2024.

MALEGORI, C. *et al.* Analysing the water spectral pattern by near-infrared spectroscopy and chemometrics as a dynamic multidimensional biomarker in preservation: rice germ storage

monitoring. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 265, p. 120396, jan. 2022.

MALLET, A. *et al.* Relating Near-Infrared Light Path-Length Modifications to the Water Content of Scattering Media in Near-Infrared Spectroscopy: Toward a New Bouguer–Beer–Lambert Law. **Analytical Chemistry**, v. 93, n. 17, p. 6817–6823, 4 maio 2021.

MALUMPONG, C., et al. Rice (*Oryza sativa* L.) breeding for novel black short grain associated with yield, cooking quality, high nutrition and antioxidant potential derived from indica black rice x japonica white rice. **Research Square**, 2021.

MANAFIFARD, M. A new hyperparameter to random forest: application of remote sensing in yield prediction. **Earth Science Informatics**, v. 17, n. 1, p. 63–73, 22 fev. 2024.

MARQUES RAMOS, A. P. *et al.* A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. **Computers and Electronics in Agriculture**, v. 178, p. 105791, nov. 2020.

MASSART, D. L. et al. Chemometrics: a textbook. Amsterdam: **Elsevier**, 1997.

METROHM. **What is NIR spectroscopy***?* Metrohm Blog, 2024.

MIN, B., GU, et al. Free and bound total phenolic concentrations, antioxidant capacities, and profiling of proanthocyanidins and anthocyanins in whole grain rice. **Journal of Cereal Science**, n. 56, v. 3, p. 603–610, 2012.

MINISTÉRIO DA AGRICULTURA, P. E A. **Aprova o regulamento técnico do arroz. Diário Oficial da União, Brasília. Instrução Normativa nº 6**. BRASIL: [s.n.].

MIRANDA RAMOS SOARES, A. P. DE *et al.* Random Forest as a promising application to predict basic-dye biosorption process using orange waste. **Journal of Environmental Chemical Engineering**, v. 8, n. 4, p. 103952, ago. 2020.

MUDGAL, S.; SINGH, N. Effect of parboiling treatment times on the physicochemical, cooking, textural, and pasting properties and amino acid, phenolic, and sugar profiles of germinated paddy rice from different rice varieties. **Journal of Food Science**, v. 89, n. 6, p. 3208–3229, 18 jun. 2024.

MÜLLER, A. *et al.* Rice Drying, Storage and Processing: Effects of Post-Harvest Operations on Grain Quality. **Rice Science**, v. 29, n. 1, p. 16–30, jan. 2022.

NACHTIGALL, G. R. et al. Teores de minerais em arroz integral e polido cultivado no Rio Grande do Sul. *Revista Brasileira de Agrociência*, Pelotas, v. 10, n. 2, p. 205-208, 2004.

NAGY, A. *et al.* Hyperspectral indices data fusion-based machine learning enhanced by MRMR algorithm for estimating maize chlorophyll content. **Frontiers in Plant Science**, v. 15, 16 out. 2024.

NATH, S. *et al.* Grain characteristics, proximate composition, phytochemical capacity, and mineral content of selected aromatic and non-aromatic rice accessions commonly cultivated in the North-East Indian plain belt. **Applied Food Research**, v. 2, n. 1, p. 100067, jun. 2022.

NIRMAGUSTINA, D. E.; HANDAYANI, S. Nutritional Content of Brown Rice and White Rice from Organic Rice of Mentik Susu Varieties with Parboiled Method. **International Journal of Life Science and Agriculture Research**, v. 02, n. 11, 1 nov. 2023.

OBERPRILLER, J.; SOUZA LEITE, M. DE; PICHLER, M. Fixed or random? On the reliability of mixed-effects models for a small number of levels in grouping variables. **Ecology and Evolution**, v. 12, n. 7, 24 jul. 2022.

OLIVEIRA CARNEIRO, L. DE *et al.* Characterizing and Predicting the Quality of Milled Rice Grains Using Machine Learning Models. **AgriEngineering**, v. 5, n. 3, p. 1196–1215, 4 jul. 2023.

ONMANKHONG, J. *et al.* Cognitive spectroscopy for the classification of rice varieties: A comparison of machine learning and deep learning approaches in analysing long-wave near-infrared hyperspectral images of brown and milled samples. **Infrared Physics & Technology**, v. 123, p. 104100, jun. 2022.

PAL, P. *et al.* Effect of Parboiling on Phenolic, Protein, and Pasting Properties of Rice from Different Paddy Varieties. **Journal of Food Science**, v. 83, n. 11, p. 2761–2771, 29 nov. 2018.

PAYNE, W. Z.; KUROUSKI, D. Raman spectroscopy enables phenotyping and assessment of nutrition values of plants: a review. **Plant Methods**, v. 17, n. 1, p. 78, 15 dez. 2021.

PELLACANI, S. *et al.* Near Infrared and UV-Visible Spectroscopy Coupled with Chemometrics for the Characterization of Flours from Different Starch Origins. **Chemosensors**, v. 12, n. 1, p. 1, 22 dez. 2023.

PEREIRA, C. L. *et al.* Relationship between Physicochemical and Cooking Quality Parameters with Estimated Glycaemic Index of Rice Varieties. **Foods**, v. 13, n. 1, p. 135, 30 dez. 2023.

PERICH, G. *et al.* Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. **Field Crops Research**, v. 292, p. 108824, mar. 2023.

PRASAD C. T., M. *et al.* Experimental rice seed aging under elevated oxygen pressure: Methodology and mechanism. **Frontiers in Plant Science**, v. 13, 1 dez. 2022.

QI, Z. *et al.* Distribution of mycotoxin-producing fungi across major rice production areas of China. **Food Control**, v. 134, p. 108572, abr. 2022.

QU, L. *et al.* Untargeted Lipidomics Reveal Quality Changes in High-Moisture Japonica Brown Rice at Different Storage Temperatures. **Foods**, v. 12, n. 23, p. 4218, 22 nov. 2023.

QU, Qian; JIN, Lan. Application of nuclear magnetic resonance in food analysis. **Food Science and Technology**, v. 42, p. e43622, 2022.

QUEVEDO RAMIREZ, L., F. Espectroscopia no infravermelho próximo e análise de imagens na avaliação da qualidade fisiológica de sementes de arroz. **Dissertação de Mestrado da Universidade Federal de Viçosa**. Viçosa/MG, 2023.

RAHIMZADEH ARASHLOO, S.; KITTLER, J. Multi-target regression via non-linear output structure learning. **Neurocomputing**, v. 492, p. 572–580, jul. 2022.

RAN, H. *et al.* A framework to quantify uncertainty of crop model parameters and its application in arid Northwest China. **Agricultural and Forest Meteorology**, v. 316, p. 108844, abr. 2022.

RATHNAYAKE, H. A.; NAVARATNE, S. B.; NAVARATNE, C. M. Effect of Size Reduction Processes on Rice Flour Properties and the Quality of their Bakery Products. **Ceylon Journal of Science,** v. 53, 2024.

RAZAVI, M. A. *et al.* Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data. **Artificial Intelligence in Agriculture**, v. 14, p. 99–114, dez. 2024.

RITHESH, B. N. *et al.* Evaluation of glycemic index, glycemic load and biochemical traits of rice associated with anti-diabetic properties. **Plant Science Today**, 21 dez. 2024.

RIZWANA, S.; HAZARIKA, M. K. Study of the Soaking Process of a ready-to-eat rice of Assam (Komal Chaul): A Mechanistic and a Machine Learning Based Approach for spectra-based Estimation of Endpoint. **Food Biophysics**, v. 19, n. 3, p. 771–783, 5 set. 2024.

RODRIGUES, D. M. *et al.* Monitoring and predicting corn grain quality on the transport and post-harvest operations in storage units using sensors and machine learning models. **Scientific Reports**, v. 14, n. 1, p. 6232, 14 mar. 2024.

SAHA, S. *et al.* Exploring Moisture Content in 50 Rough Rice through Micro Oven for Deeper Dietary Insights. **Current Functional Foods**, v. 03, 27 jan. 2025.

SAHOO, U. *et al.* Rice with lower amylose content could have reduced starch digestibility due to crystallized resistant starch synthesized by linearized amylopectin. **Journal of the Science of Food and Agriculture**, v. 105, n. 5, p. 3064–3072, 30 mar. 2025.

SAMPAIO, P. S.; ALMEIDA, A. S.; BRITES, C. M. Use of Artificial Neural Network Model for Rice Quality Prediction Based on Grain Physical Parameters. **Foods**, v. 10, n. 12, p. 3016, 5 dez. 2021.

SANTOS, T. T., et al. Visão computacional aplicada na agricultura. **Agricultura digital: pesquisa, desenvolvimento e inovação nas cadeias produtivas**, 2020.

SARKER, I. H. **Machine Learning: Algorithms, Real-World Applications and Research Directions**. 2021.

SCARIOT, M. A. *et al.* Effect of drying air temperature and storage on industrial and chemical quality of rice grains. **Journal of Stored Products Research**, v. 89, p. 101717, dez. 2020.

SCHAFFERT, R. E. et al. Composição lipídica do arroz integral e polido de cultivares cultivadas no sul do Brasil. *Boletim do Centro de Pesquisa de Arroz Irrigado*, Pelotas, v. 21, p. 73-81, 2011.

SCHIELZETH, H. *et al.* Robustness of linear mixed-effects models to violations of distributional assumptions. **Methods in Ecology and Evolution**, v. 11, n. 9, p. 1141–1152, 16 set. 2020.

SHARMA, P. *et al.* Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning. **IEEE Access**, v. 11, p. 111255–111264, 2023.

SHAWON, S. M. *et al.* Crop yield prediction using machine learning: An extensive and systematic literature review. **Smart Agricultural Technology**, v. 10, p. 100718, mar. 2025.

SHEN, Y. *et al.* Rice varieties with a high endosperm lipid content have reduced starch digestibility and increased γ-oryzanol bioaccessibility. **Food & Function**, v. 12, n. 22, p. 11547–11556, 2021.

SHI, S. *et al.* Combination of near-infrared spectroscopy and key wavelength-based screening algorithm for rapid determination of rice protein content. **Journal of Food Composition and Analysis**, v. 118, p. 105216, maio 2023.

SHROTRIYA, A. *et al.* Hybrid Ensemble Learning With CNN and RNN for Multimodal Cotton Plant Disease Detection. **IEEE Access**, v. 12, p. 198028–198045, 2024.

SINGH, A.; SINGH, B. Characterization of rice husk ash obtained from an industrial source. **Journal of Sustainable Cement-Based Materials**, v. 10, n. 4, p. 193–212, 6 ago. 2021.

SINGH, N. *et al.* Integrating NIR spectroscopy with machine learning and heuristic algorithm-assisted wavelength selection algorithms for protein content prediction in rice bean (Vigna umbellata L.). **Food and Humanity**, v. 3, p. 100399, dez. 2024.

SINGH, S.; KASANA, S. S. Quantitative estimation of soil properties using hybrid features and RNN variants. **Chemosphere**, v. 287, p. 131889, jan. 2022.

SOUZA, J. R. et al. Caracterização físico-química e sensorial de arroz integral orgânico. *Alimentos e Nutrição*, Araraquara, v. 22, n. 3, p. 445-451, 2011.

SU, X. *et al.* Multispectral Inversion of Starch Content in Rice Grains from Yingjiang County Based on Feature Band Selection Algorithms. **Agronomy**, v. 15, n. 1, p. 86, 31 dez. 2024.

TAO, K. *et al.* Investigating cooked rice textural properties by instrumental measurements. **Food Science and Human Wellness**, v. 9, n. 2, p. 130–135, jun. 2020.

THEIVENTHIRAN, T. V. *et al.* Screening Out the Phytochemicals and Colour Values in White and Black Rice Lines and Its Interrelationship. **International Journal of Biochemistry Research & Review**, p. 34–41, 1 maio 2020.

TIAN, Y. *et al.* Quantitative detection of crude protein in brown rice by near-infrared spectroscopy based on hybrid feature selection. **Chemometrics and Intelligent Laboratory Systems**, v. 247, p. 105093, abr. 2024.

TUNCA, E. *et al.* Accurate estimation of sorghum crop water content under different water stress levels using machine learning and hyperspectral data. **Environmental Monitoring and Assessment**, v. 195, n. 7, p. 877, 23 jul. 2023.

UIVARASAN, A. *et al.* Characterization of Polyphenol Composition and Starch and Protein Structure in Brown Rice Flour, Black Rice Flour and Their Mixtures. **Foods**, v. 13, n. 11, p. 1592, 21 maio 2024.

UNIVERSIDADE FEDERAL DE PELOTAS (CTI/UFPEL). **Tecnologia de ultrassom na avaliação de qualidade de alimentos.** Pelotas, 2023.

VICTOR, B.; NIBALI, A.; HE, Z. A Systematic Review of the Use of Deep Learning in Satellite Imagery for Agriculture. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 18, p. 2297–2316, 2025.

WANG, H.; WANG, X.; PENG, B. Using an improved Si-rich husk ash to decrease inorganic arsenic in rice grain. **Science of The Total Environment**, v. 803, p. 150102, jan. 2022.

WANG, K. *et al.* Insight into the relationship between the starch crystalline structure and textural quality and physicochemical properties of reconstituted rice: Influence of feed moisture content. **International Journal of Biological Macromolecules**, v. 280, p. 135758, nov. 2024.

WANG, N. *et al.* Rapid Determination of Cellulose and Hemicellulose Contents in Corn Stover Using Near-Infrared Spectroscopy Combined with Wavelength Selection. **Molecules**, v. 27, n. 11, p. 3373, 24 maio 2022.

WANG, W. *et al.* Deciphering the Genetic Architecture of Color Variation in Whole Grain Rice by Genome-Wide Association. **Plants**, v. 12, n. 4, p. 927, 17 fev. 2023.

WANG, X. *et al.* Non-destructive assessment of apple internal quality using rotational hyperspectral imaging. **Frontiers in Plant Science**, v. 15, 6 nov. 2024.

WANG, Y. *et al.* Mark-Spectra: A convolutional neural network for quantitative spectral analysis overcoming spatial relationships. **Computers and Electronics in Agriculture**, v. 192, p. 106624, jan. 2022.

WANG, Y.; TAN, F. Extraction and classification of origin characteristic peaks from rice Raman spectra by principal component analysis. **Vibrational Spectroscopy**, v. 114, p. 103249, maio 2021.

WATTANAVANITCHAKORN, S. *et al.* **Biochemical and Molecular Assessment of Cooking Quality and Nutritional Value of Pigmented and Non-pigmented Whole Grain Rice**, 1 set. 2021.

WEI, X. *et al.* Confocal Raman microspectroscopy combined with spectral screening algorithms for quantitative analysis of starch in rice. **Food Hydrocolloids**, v. 141, p. 108737, ago. 2023.

XIAO, C. *et al.* Prediction of soil salinity parameters using machine learning models in an arid region of northwest China. **Computers and Electronics in Agriculture**, v. 204, p. 107512, jan. 2023.

XIE, L.-H. *et al.* Simultaneous determination of apparent amylose, amylose and amylopectin content and classification of waxy rice using near-infrared spectroscopy (NIRS). **Food Chemistry**, v. 388, p. 132944, set. 2022.

XU, H. et al. When are deep networks really better than decision forests at small sample sizes, and how?. arXiv preprint arXiv:2108.13637, 2021.

XU, Y. *et al.* Combination of near infrared spectroscopy with characteristic interval selection for rapid detection of rice protein content. **Journal of Food Composition and Analysis**, v. 137, p. 106995, jan. 2025.

XU, Z. *et al.* Data fusion of near-infrared diffuse reflectance spectra and transmittance spectra for the accurate determination of rice flour constituents. **Analytica Chimica Acta**, v. 1193, p. 339384, fev. 2022.

XUAN, G. *et al.* Protein content prediction of rice grains based on hyperspectral imaging. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 320, p. 124589, nov. 2024.

YADAV, S. *et al.* Near-infrared reflectance spectroscopy (NIRS): An innovative, rapid, economical, easy and non-destructive whole grain analysis method for nutritional profiling of

pearl millet genotypes. **Journal of Food Composition and Analysis**, v. 142, p. 107373, jun. 2025.

YAN, W. *et al.* Inhibition of Lipid and Aroma Deterioration in Rice Bran by Infrared Heating. **Food and Bioprocess Technology**, v. 13, n. 10, p. 1677–1687, 31 out. 2020.

YAN, Y. *et al.* Laboratory shortwave infrared reflectance spectroscopy for estimating grain protein content in rice and wheat. **International Journal of Remote Sensing**, v. 42, n. 12, p. 4467–4492, 18 jun. 2021.

YANG, H.-E. *et al.* Prediction of protein content in paddy rice (Oryza sativa L.) combining near-infrared spectroscopy and deep-learning algorithm. **Frontiers in Plant Science**, v. 15, 31 jul. 2024a.

YANG, Y. *et al.* Rice starch accumulation at different endosperm regions and physical properties under nitrogen treatment at panicle initiation stage. **International Journal of Biological Macromolecules**, v. 160, p. 328–339, out. 2020.

YAO, B.; CHEN, P.; SUN, G. Distribution of elements and their correlation in bran, polished rice, and whole grain. **Food Science & Nutrition**, v. 8, n. 2, p. 982–992, 9 fev. 2020.

YOOSEFZADEH-NAJAFABADI, M. *et al.* Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. **Frontiers in Plant Science**, v. 11, 12 jan. 2021.

YOVIONO, F.; SANDRA, Y.; ARIFANDI, F. Perbandingan Kadar Pati Pada Beras Hitam Dibandingkan Dengan Beras Putih Menggunakan Uji Iodida. **Cerdika: Jurnal Ilmiah Indonesia**, v. 2, n. 11, p. 976–981, 21 nov. 2022.

YU, Y. *et al.* Prediction of Potassium Content in Rice Leaves Based on Spectral Features and Random Forests. **Agronomy**, v. 13, n. 9, p. 2337, 7 set. 2023.

ZAKERI, Z. *et al.* Cross-validating models of continuous data from simulation and experiment by using linear regression and artificial neural networks. **Informatics in Medicine Unlocked**, v. 21, p. 100457, 2020.

ZHANG, Y. *et al.* Structural changes of starch under different milling degrees affect the cooking and textural properties of rice. **Food Chemistry: X**, v. 17, p. 100627, mar. 2023.

ZHAO, M. *et al.* How anthocyanin biosynthesis affects nutritional value and anti-inflammatory effect of black rice. **Journal of Cereal Science**, v. 101, p. 103295, set. 2021.

ZHENG, R. *et al.* Optimizing feature selection with gradient boosting machines in PLS regression for predicting moisture and protein in multi-country corn kernels via NIR spectroscopy. **Food Chemistry**, v. 456, p. 140062, out. 2024.

ZHOU, X. *et al.* Fault diagnosis of silage harvester based on a modified random forest. **Information Processing in Agriculture**, v. 10, n. 3, p. 301–311, set. 2023.

ZHU, D. *et al.* Quality changes in Chinese high-quality indica rice under different storage temperatures with varying initial moisture contents. **Frontiers in Nutrition**, v. 11, 11 mar. 2024.

ZHU, Y. *et al.* The effect of dry heat parboiling processing on the short-range molecular order structure of highland barley. **LWT**, v. 140, p. 110797, abr. 2021.

ŽÍŽALA, D. *et al.* Soil sampling design matters - Enhancing the efficiency of digital soil mapping at the field scale. **Geoderma Regional**, v. 39, p. e00874, dez. 2024.

ZOU, X. *et al.* Fusion of convolutional neural network with XGBoost feature extraction for predicting multi-constituents in corn using near infrared spectroscopy. **Food Chemistry**, v. 463, p. 141053, jan. 2025.

ZHOU, Z. et al. A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models. **Scientific Reports**, v. 13, p. 22420, 2023.