Predição da Ocorrência de Sinistros Agrícolas por meio de Imagens de Satélite e Técnicas de Machine Learning

1st Anne I. R. Carvalho
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil

***Abstract*—No dinâmico panorama agrícola do Brasil, os seguros rurais desempenham um papel fundamental, abordando preocupações econômicas, de sustentabilidade ambiental, de inclusão social e de desenvolvimento rural. As atividades agrícolas, susceptíveis a diversos riscos, necessitam de estratégias sofisticadas de mitigação e gestão de riscos.**

Este estudo tem como objetivo coletar e caracterizar dados de referência, desenvolvendo um método para prever potenciais sinistros em áreas agrícolas ao longo do período contratual. Combinando análise de imagens de satélite, dados espectrais, informações descritivas das áreas seguradas e dados climáticos, o objetivo é classificar cada apólice de seguro. Os resultados previstos são a implementação de um modelo de redes neurais, robusto e capaz de fazer distinção entre propriedades que apresentaram ou não a ocorrência de sinistros.

***Index Terms*—redes neurais, geoprocessamento, seguros, agronomia**

I. Introdução

Atualmente, a subvenção e o gerenciamento de sinistros rurais desempenham um papel crucial no contexto agrícola do Brasil, devido a diversas razões, que incluem aspectos econômicos, sustentabilidade ambiental, inclusão social e desenvolvimento rural. No entanto, a atividade agrícola está sujeita a uma série de riscos, desde a incerteza da produção decorrente da instabilidade climática até ameaças sanitárias, tornando-a um setor de elevada complexidade e risco.

O seguro rural emerge como um instrumento essencial para os agricultores mitigarem esses riscos na atividade agropecuária e assegurarem sua renda. O governo federal e os estados, por padrão, fornecem subvenções que oferecem ao agricultor a oportunidade de segurar sua produção com custo reduzido, impulsionando o desenvolvimento do setor. Esta prática, além de incentivar a modernização da agricultura e garantir competitividade, possui também um papel regulatório, uma vez que a concessão do seguro implica na análise das restrições às quais o agricultor está sujeito.

Em caso de sinistro, ou seja, o acionamento do seguro em decorrência de um incidente coberto, a seguradora assume a obrigação de indenizar o prejuízo eventualmente sofrido pelo segurado. As especificidades de cada contrato são formalizadas nas apólices de seguro, e as taxas e prêmios são estabelecidos de acordo com o nível de risco, seguindo uma lógica similar a outros tipos de seguro, como os automotivos.

O Seguro Agrícola, por definição, cobre perdas relacionadas à vida da cultura. Geralmente, os seguros são contratados para proteger contra perdas decorrentes de fenômenos meteorológicos, incêndios, secas e variações extremas de temperatura.

Conforme os dados mais recentes divulgados pelo Governo no Relatório Geral do Programa do Seguro Rural [18], a atividade apresentou resultados recordes, com um aumento nominal no valor da produção garantida e o maior número de culturas atendidas, totalizando 62 diferentes culturas. Além disso, mais do que o dobro da área foi assegurada em relação ao ano anterior, demonstrando um cenário progressista, porém também suscitando preocupações, devido à alta de 47,1% das indenizações [10].

Neste contexto, o gerenciamento de risco assume extrema importância para a seguradora, visto que, na obrigação de pagar o prêmio em caso de sinistro, não é viável que a seguradora se exponha a riscos que possam resultar em grandes prejuízos.

Diante disso, a geotecnologia tornou-se fundamental para o controle e monitoramento das culturas asseguradas, tanto na manutenção das lavouras e na avaliação mercadológica, quanto na avaliação das propostas de seguro. As imagens de satélite obtidas ao longo do ciclo de vida das culturas oferecem indicativos sobre o histórico de saúde das plantações, identificam danos e avaliam as condições específicas das culturas, como período de cultivo e forma de plantio, conforme indicado pelo zoneamento agrícola formulado pelo IPEA [8].

II. Objetivo

A previsão da ocorrência de sinistros pode revelar-se como um fator substancial para mitigar efeitos adversos na atividade agrícola. Além disso, a vigilância por meio de imagens de satélite pode se tornar uma ferramenta essencial para identificar eventos impactantes, permitindo respostas rápidas e efetivas.

Este trabalho tem como objetivo primário coletar e caracterizar os dados de referência e projetar um método capaz de predizer se uma área de plantio agrícola, em algum momento do período contratual, indicará a ocorrência de sinistros. A intenção final é de combinar a análise de imagens de satélite com dados de índices de vegetação, informações descritivas

da área assegurada e dados climáticos, a fim de classificar cada apólice de seguro. Em relação ao Projeto Orientado a Computação II, é esperado que a metodologia apresentada pelo presente trabalho, dsenvolva com satisfação a implementação, avaliação de modelos e a análise dos resultados.

III. Trabalhos Correlatos

Bauer [2], publicou a primeira revisão científica dos usos do sensoriamento remoto aplicado à agronomia. O capítulo, além de delinear a base física do sensoriamento remoto, descrevendo as propriedades espectrais relevantes e sua relação com as características de uma cultura, delineia os esforços iniciais para a aplicação do sensoriamento remoto para várias estimativas agrícolas, dentre elas, a predição de produção. De fato, os estudos deixam claro o grande potencial do sensoriamento remoto de detectar e caracterizar vários fenômenos agrícolas.

O sensoriamento remoto, desde então, tem se destacado como uma ferramenta essencial para auxiliar a agricultura diante dos novos desafios climáticos e ambientais, demonstrando também a capacidade de minimizar os danos causados pelos impactos ambientais da atividade [1]. Em um curto espaço de tempo, grandes desenvolvimentos computacionais como o desenvolvimento de scanners multiespectrais capazes de realizar o escaneamento da superfície da Terra, o desenvolvimento de técnicas de reconhecimento de padrões, o avanço da capacidade computacional e a ampla disponibilidade de imagens de satélite têm incentivado a pesquisa voltada para as tarefas de Observação da Terra [17]. Isso culminou em um considerável volume de estudos que giram em torno do tema central deste trabalho: o monitoramento de áreas de cultivo por meio de imagens de satélite e a constante busca por predizer a produtividade das culturas.

Várias pesquisas subsequentes obtiveram êxito ao empregar imagens de satélite para estimar a produtividade de culturas. O artigo "Global Crop Forecasting" [17], que descreve o Large Area Crop Inventory Experiment (LACIE), delineou seu principal objetivo: antecipar o conhecimento da colheita global de grãos para tomar decisões informadas no âmbito do comércio internacional e da cadeia alimentar global. O sistema de inventário de colheita foi desenvolvido utilizando dados de proporção de área (extraídos de imagens do LANDSAT), dados climáticos e de produção. Após anos de experimentos, os cientistas conseguiram atestar uma previsão de produção de alta qualidade nos Estados Unidos da América, com uma acurácia de cerca de 90% em anos sem condições meteorológicas extremas. Entretanto, os resultados de outros países, incluindo o Brasil, que apresentaram previsões pouco relevantes, indicam a necessidade de modelos menos dependentes de dados históricos. O sistema representou um avanço significativo na capacidade do sensoriamento remoto para monitoramento da produção.

A introdução do Google Earth Engine democratizou o acesso a imagens globais e simplificou a realização de análises geoespaciais robustas, proporcionando acesso completo a dados meteorológicos, altimétricos, atmosféricos, sociais, entre outros. [6] apresenta o amplo impacto da disponibilização do Google Earth Engine e elaboram um resumo de todos os satélites e funcionalidades disponíveis na plataforma.

Sakamoto et al. [24], na tentativa de estimar a produção de milho, utilizam imagens do MODIS para o cálculo do Índice de Vegetação de Faixa Dinâmica Ampla (WDRVI) a fim de linearizar o relacionamento do Índice de Área Foliar (LAI), desenvolvendo um modelo capaz de revelar uma distribuição espaço-temporal detalhada dos estágios de crescimento do milho nos EUA com um modelo de Vizinhos Próximos. Entretanto, na avaliação final do modelo de predição, foram indicadas tendências de erro em algumas localidades; no entanto, confirmou-se que o modelo de MODIS WDRVI pode prever com alta acurácia os estados de maior produção.

Huang et al. [9], ao indicar a necessidade de prever o crescimento do trigo para a segurança alimentar e o desenvolvimento sustentável, desenvolve um método de previsão com base em dados do LAI, NDVI (Índice de Vegetação por Diferença Normalizada), SAVI (Índice de Vegetação Ajustado ao Solo) derivados do MODIS e do LANDSAT TM, posteriormente inserindo os dados no modelo estatísticos WOFOST de crescimento de grãos para prever a produção.

Guan et al. [7] realiza um estudo inédito sobre a contribuição de dados de satélite em diversas faixas espectrais, como EVI (Índice de Vegetação Melhorado), SIF (Fluorescência Induzida pelo Sol), Ku-band (banda K) e VOD (profundidade óptica de Vegetação), para o processo de predição. O estudo ainda utiliza o Partial Least Square Regression (PLSR) para distinguir informações individuais comumente compartilhadas e únicas entre os vários dados de satélite e outras informações climáticas auxiliares para a estimativa do rendimento das culturas. A conclusão do estudo é de extrema importância e deixa claro o relacionamento dos estágios da agricultura para cada espectro, além das informações conjuntas e diferenciais de cada índice em relação aos componentes de biomassa e estresse ambiental, explicando a variabilidade no modelo estatístico de produção.

Os métodos de estimativa, por muito tempo, basearam-se fortemente em técnicas convencionais, principalmente em modelos agrometeorológicos e modelos estatísticos de regressão, que limitavam os resultados e estimativas [9], [24]. Esses métodos foram gradualmente substituídos por modelos de Inteligência Artificial, inseridos em um contexto de maior capacidade computacional, demonstrando desempenho robusto em comparação às técnicas tradicionais.

Khaki et al. [13] desenvolveram, para o Syngenta Crop Challenge de 2018, métodos utilizando deep neural network para prever, checar e verificar a diferença de produção de milhos híbridos por genótipo e dados ambientais. O modelo proposto não utiliza imagens de satélite, mas sim dados de solo e climáticos, como temperaturas, precipitação, duração do dia, pressão de vapor e radiação solar. No ano seguinte, [14] desenvolvem novamente um modelo híbrido CNN-RNN para a predição de produção de culturas. Desta vez, o modelo captura os fatores ambientais e o avanço das sementes ao longo do tempo, mas não tem acesso à informação genotípica. O

estudo é capaz de demonstrar também a extensão da influência de diversos fatores ambientais mencionados para a produção da cultura.

Liu et al. [15] cria um conjunto de dados múltiplos que contém várias gerações de imagens de sensoriamento remoto do MODIS, incluindo temperatura e índice de vegetação, com informações espaciais e temporais. O modelo desenvolvido utiliza CNN como estrutura básica, com a introdução do RNN, tornando-se um modelo híbrido, e com a incorporação do LSTM para tornar o elemento mais sensível aos dados temporais. O modelo foi considerado um bom preditor da produção de grãos na China.

Cao et al. [3] utiliza diversas fontes de dados, como índices de vegetação ESI (combinado do EVI e SIF), índices meteorológicos, propriedades do solo em modelos de aprendizado de máquina, incluindo regressão e deep learning, concluindo pela eficácia do modelo LSTM. Khaki et al. ainda [12] busca desenvolver um modelo de estimação para múltiplas culturas por meio de uma nova abordagem de Redes Neurais Convolucionais, obtendo resultados promissores em relação a outras técnicas.

Luo et al. [16] apresentam um mapeamento abrangente das variáveis de relevância na predição da produção robusta, desenvolvendo modelos de predição baseados em RF, LightGBM e algoritmos LSTM. Os autores adotam a estratégia "Leave One Year Out" para evitar potenciais autocorrelações temporais nas amostras.

Qiao et al. [23] propõem um modelo 3D-CNN para extrair características tanto espaciais quanto temporais, inerentes aos dados volumétricos de imagens, sendo pioneiros na extração conjunta de características de imagens multiespectrais, resultando em um desempenho superior até mesmo a outros modelos de deep learning. Chen et al. [4] utilizam imagens de alta resolução e CNN em conjunto com diversas estratégias de pooling para a tarefa de mapeamento de cobertura do solo. Em outra abordagem, Xie et al. [26] emprega algoritmos de LSTM, 1D CNN e RF para estimar a produção agrícola, utilizando dados de LAI e NDVI.

Os trabalhos internacionais que abrangem o tema são abundantes e extensos, apresentando diversas metodologias de previsão e integrando dezenas de diferentes fontes de dados. Nacionalmente, foram identificados poucos trabalhos relacionados ao tema de pesquisa, especialmente no âmbito da previsão de sinistros rurais. Embora alguns estudos nacionais tenham sido desenvolvidos, nota-se que o tema ainda carece de uma exploração mais aprofundada, dada sua relevância.

Barros et al. [1] propõem uma metodologia de aprendizado de máquina utilizando imagens ópticas e de radar de satélite para prever a taxa de prêmio a ser cobrada em eventualidades de sinistros. O estudo emprega modelos como KNN, ANN, SVM, DT e RF. A robustez do modelo é validada por meio da implementação de técnicas como validação cruzada k-fold e eliminação recursiva de características. Diferentemente de outros estudos, características municipais, como escolaridade e renda per capita, também foram consideradas nas estimativas.

Mota et al. [21] utilizam o Programa de Subvenção ao Prêmio do Seguro Rural como base de dados para prever a ocorrência de sinistros agrícolas. Para essa previsão, empregam algoritmos tradicionais como Random Forest, Support Vector Machine e k-Nearest Neighbour, com foco principalmente em dados meteorológicos para a predição dos eventos.

IV. Desenvolvimento

No âmbito deste projeto, o presente capítulo destaca e analisa de maneira detalhada a condução das atividades. Essa seção não apenas delineia as atividades realizadas, mas também destaca como essas atividades podem impactar diretamente os resultados obtidos.

A. Coleta de Dados

Foram coletados dados de diversas fontes. Podemos classificá-los como Base de Dados Principal, Dados Complementares e Imagens de Satélite.

1) Base de Dados Principal: A fonte primária dos dados é o Sistema de Subvenção Econômica ao Prêmio Rural disponibilizado pelo Ministério da Agricultura, Pecuária e Abastecimento [19]. Este sistema abrange milhares de apólices asseguradas pelo governo no período de 2016 a 2021. As informações contidas nas apólices incluem valores de subvenção, tipo e localização das culturas, modalidade de seguro, montante da eventual indenização e a ocorrência (ou não) de sinistros, juntamente com a indicação do motivo do sinistro. Esses dados são armazenados e apresentados em formato csv. Devido a natureza das ocorrências, esse conjunto de dados são inerentemente desequilibrados em relação à nossa variável de classificação.

Em adição, foram extraídos os polígonos de imóveis rurais de todo o Território nacional do Sistema de Cadastro Ambiental Rural (SICAR). Todos os imóveis rurais devem ser registrados no CAR para fins de integrar e gerenciar informações ambientais das propriedades rurais em todo o país e são abertos ao público. Esses dados, com formato .shp, foram armazenados em dísco rígido externo.

2) Dados Complementares: Foram coletados dados que de alguma forma auxiliariam no processo de Tratamento, Análise e Desempenho do modelo proposto. Esses dados incluem:

Catálogo de Municípios - Utilizado principalmente para identificação da localização político-geográfica das apólices, de forma a aprimorar a contextualização geográfica das análises, com ênfase no tratamento necessário para garantir a consistência e a precisão.

Indicadores Socioeconômicos e de Desenvolvimento - Dados municipais do Censo 2010, dentre eles, podemos citar expectativa de vida, taxa de envelhecimento, taxa de analfabetismo, índice de Gini, renda per capita, índice de theil, população rural, população urbana, subindice de escolaridade, subindice de frequencia escolar, IDHM, IDHM Educação, IDHM Longevidade, IDHM Renda [11].

Ocorrência de El Niño e La Niña - Dados baseados no Oceanic Niño Index (ONI), indicam a ocorrência e intensidade dos fenômenos por conjunto trimestral [22]. Para a adição de um canal nos dados de input do modelo, foram usados

os índices disponibilizados pela Ferramenta para o monitoramento dos padrões de teleconexão na América do Sul [25].

3) Imagens de Satélite: Todas as imagens de Satélite foram coletadas a partir da API do Google Earth Engine. O GEE nos permite acessar uma coleção de imagem correspondente às propriedades, bem como calcular os diversos indicadores referentes a cada produto e banda. Os dados coletados foram os seguintes:

Land Surface Temperature (LST) - Derivada pro produto MOD11A1, descreve a temperatura na superfície da terra durante o dia.

Leaf Area Index (LAI) - Comumente utilizada para medir a produtividade em uma escala espacial, quantifica a quantidade de material foliar em uma copa [20]. Derivada do produto MCD15A3H.

Fraction of Photosynthetically Active Radiation (FPAR) - Adequada para monitorar o ciclo de sazonalidade e a variabilidade interanual da atividade vegetativa relacionada à fotossíntese da superfície terrestre [5]. Derivada do produto MCD15A3H.

Normalized Difference Vegetation Index (NDVI) - índice usado para medir a saúde e a densidade da vegetação. calculado a partir de dados espectrométricos em duas bandas específicas: vermelho e infravermelho próximo. Derivada do produto MOD13Q1. Resolução de 250 metros.

Enhanced Vegetation Index (EVI) - Assim, como o NDVI, é usado para medir a saúde e a densidade da vegetação. Porém incorpora uma melhor sensibilidade em regiões de alta biomassa e desacoplamento do sinal de fundo do dossel e uma redução nas influências atmosféricas com a adição da banda Azul. Derivada do produto MOD13Q1. Resolução de 250 metros.

Precipitação - Fornece uma taxa de chuva global por hora com resolução de 0,1 x 0,1 graus.

É essencial notar que foram inicialmente selecionados aleatóriamente 50 municípios brasileiros para constituir os dados para a coleta de imagens de satélite. Isso se dá pelo fato da extensão do número de propriedades e dos requisitos de processamento e análise de uma grande quantidade de dados.

B. Tratamento de Dados

Vale ressaltar que, dentre as milhares de apólices disponíveis, serão selecionadas inicialmente as apólices referentes ao seguro agrícola de culturas de soja, em virtude de ser a cultura mais amplamente cultivada no território nacional. Além disso, foram desqualificadas apólices que não tem a cobertura de sinistro associada à Produtividade.

1) Mapeamento da Apólice para Propriedade: A partir das coordenadas fornecidas pelo Sistema de Subvenção Econômica ao Prêmio Rural (SSEPR), foi possível identificar a área agrícola correspondente à apólice indicada. Como sabemos apenas um ponto de localidade informada por grau, minuto e segundo, precisamos tratar os dados para as coordenadas de latitude e longitude e operar a operação de junção entre os dados geográficos para identificar os pontos residente dentro de um dado polígono. Na Fígura 1 podemos ver uma conglomeração dos polígonos de imóveis do Estado de são Paulo e, em seguida, na Figura 2, três pontos quaisquer dentro de imóveis (polígonos). Dessa forma, nos é indicado que devemos tratar o polígono específico.

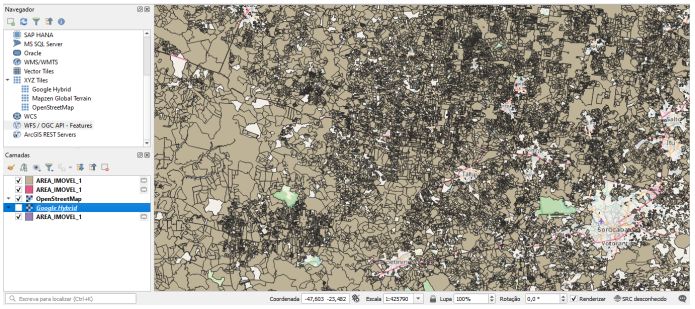

Fig. 1. Recorte de polígonos de imóveis cadastrados no CAR do Estado de São Paulo. Executado no QGIS.

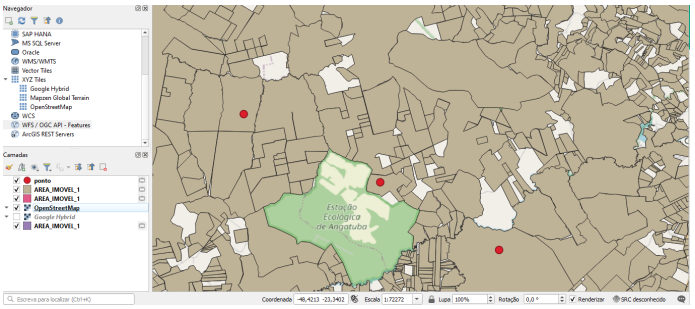

Fig. 2. Recorte de polígonos com um ponto interno, representando a localização da cultura informada na apólice. Executado no QGIS.

Pode acontecer, ainda, de várias apólices se referirem à mesma propriedade, com localizações relativas diferentes (para várias plantações). Dessa forma, o polígono encontrado, é retornado à todos os registros de apólice. Esse caso é retrato na figura 3, em que um mesmo imóveis foi referenciado para 4 apólices diferentes. Todas as apólices, nesse caso, indicaram a ocorrência de sinistro, mostrando uma homogeneidade nos eventos em relação ao espaço geográfico e temporal.

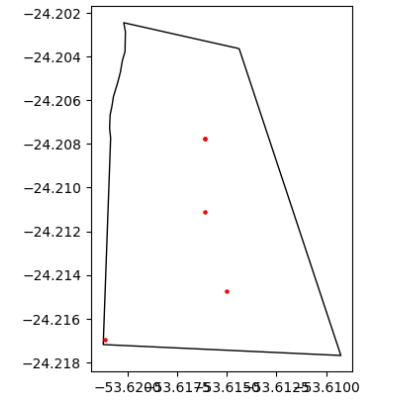

Fig. 3. Exemplo de Propriedade com múltiplas apólices associadas

2) Tratamentos Gerais: Para além disso, foram performados tratamentos envolvendo estruturas geográficas do Google Earth Engine e tratamentos de tipagem, filtragem, merge e seleção de dados.

C. Análise de Dados

Na presente seção, abordamos a análise dos dados coletados, delineando a contextualização dos conjuntos de dados e a metodologia adotada para extrair insights significativos. A análise é um componente crítico deste estudo, sendo conduzida com o intuito de identificar padrões, correlações e tendências que podem impactar diretamente a predição.

1) Análise Descritiva e Exploratória: A cultura de Soja é a mais predominante assegurada nos dados de 2016-2021, com 313.8075 apólices registradas. Isso equivale a cerca de 45% das apólices diferentes a 65 culturas totais. Em relação ao Produto assegurado, a modalidade de Produtividade é a mais predominante, seguida de Custeio. Outras Classificações presentes no relatório são Pecuário, Florestas e Receita.

Considerando apenas as apólices de Soja e Produtividade, temos 174.625 apólices, caracterizando esta, como uma amostra claramente desequilibrada, com muitas apólices não sinistralizadas e poucas apólices sinistralizadas. É possível ver um aumento bruto de Apólices a partir do ano de 2019, porém, o número de apólices sinistralizadas se manteve relativamente constante. Os valores de Subvenção Federal, também tiveram um aumento considerável a partir de 2017. É possível observar também, um aumento gradual no valor de Indenizações a partir de 2019.

Em relação às seguradoras, 14 foram citadas, e dentre elas, Mapfre Seguros Gerais S.A. foi a que mais indenizou na amostra citada e Essor Seguros S.A. a que mais recebeu valor de subvenção federal. Além disso, FairFax foi a que mais apresentou contratos de apólice.

Em relação à justificativa da ocorrência de sinistro, a seca aparece como o maior motivador, seguido de chuva excessiva, granizo e inundação.

Em relação à ocorrência do El Niño e La Niña, a presença de sinistros parece estar mais relacionada à ocorrência do El Niño Moderado, tanto no ano anterior ao contrato, quanto no ano do contrato. Operamos uma análise de variância sobre os índices trimestrais, um método estatístico usado para testar as diferenças entre duas ou mais médias. A sua finalidade é entender se existe uma diferença significativa entre os grupos de apólice. No caso da Análise dos fenômenos, foram encontrados valores F elevados, podendo indicar que a variação entre os grupos (sinistro ou não) é maior do que seria esperado ao acaso, o que sugere que há uma diferença significativa entre os grupos. Além disso, o baixo valor p demonstra que diferença entre os grupos é estatisticamente significativa.

Em relação aos indicadores socioeconômico e desenvolvimento de municípios, comparamos o número de apólices, número de sinistro, a taxa de sinistro por apólice e valores de subvenção por município com seu indicadores do Censo 2010. Foram encontrados correlações baixas, de no máximo 0.25 entre as variáveis de município.

	sum_sq	df	F	PR(>F)
Intercept	12552.799942	1.0	40532.242470	0.000000e+00
C(sinistro_t)	106.869686	1.0	345.075842	5.394571e-77
C(variable)	6125.553414	3.0	6593.015566	0.000000e+00
C(sinistro_t):C(variable)	541.938667	3.0	583.295880	0.000000e+00
Residual	123877.171567	399992.0	NaN	NaN

TABLE I
RESULTADO DA ANÁLISE DE VARIÂNCIA DOS ÍNDICES DE OCORRÊNCIA DO EL NIÑO E LA NIÑA, COM 'VARIABLE' REPRESENTANDO OS TRIMESTRES DAS ESTAÇÕES DO ANO DA APÓLICE

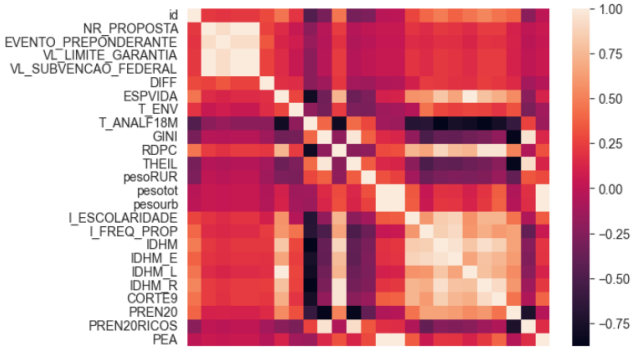

Fig. 4. Heatmap da Correlação entre as variáveis estudadas

D. Método de Predição

1) Definição dos dados de Entrada: Os dados usados para a entrada do modelo de predição foram capturados da seguinte forma: as imagens convertidas em numpy arrays são combinadas de acordo com a banda espectral selecionada e o aspecto temporal mensal de cada uma. Dessa forma, cada propriedade terá um objeto de shape (24, 256, 256, 3, spectral bands), que corresponde aos 24 meses da característica temporal das imagens, 256 pixels de altura e largura da imagem, 3 dimensões de cores referente ao RGB e o adicional das bandas variáveis de feature, correspondentes às bandas citadas pela subseção A. Essas bandas foram definidas de acordo com uma Forward Feature Selection, em que a variável com a maior acurácia de validação apresentada no modelo, era adicionada ao experimento seguinte de forma iterativa. Esses dados, agrupados por propriedade e labelados se apresentaram ou não um sinistro, são os dados de entrada do modelo de predição.

2) Arquitetura do Modelo de Predição: Para a predição, foram utilizados 3 modelos diferentes, para a comparação de resultados e performance. Todos eles foram definidos considerando a eficiência em termos de parâmetros e operações de computação, dado ao tamanho extremo dos dados de entrada.

Foi desenvolvido um modelo personalizado (vamos chamá-lo de custom) e foram usados também os modelos de redes neurais convolucionais Xception (Extreme Inception) e EfficientNetB0. Estes são modelos muito usados para tarefas de visão computacional, como classificação de imagens e Detecção de objetos, e são disponibilizados pré-treinados pela biblioteca Keras.

A arquitetura do modelo de rede neural custom é definida com uma camada de entrada que recebe os dados de acordo com a forma especificada. Em seguida, há um bloco inicial de

camadas convolucionais, onde três camadas convolucionais de 128 filtros são aplicadas sequencialmente, cada uma seguida por uma normalização de lote para estabilizar o treinamento, e finalizando com uma camada de pooling máximo que reduz a dimensão dos dados.

O segundo bloco de convolução repete essa estrutura, mas com um número menor de filtros, utilizando 64 filtros em cada uma das três camadas convolucionais, seguidas novamente por normalizações de lote e uma camada de pooling máximo.

Nesses blocos, éutilizada a camada relu, com um kernel size de 3.

Após esses blocos convolucionais, o modelo possui três camadas LSTM bidirecionais de 64 filtros, que processam as sequências de entrada em ambas as direções para capturar dependências de longo prazo. Entre essas camadas LSTM, são aplicadas camadas de dropout para evitar overfitting e normalizações de lote para estabilizar o treinamento. Em seguida, é aplicada uma camada de atenção local, com um window size de 50.

Finalmente, o modelo tem três camadas densas. Cada uma dessas camadas aplica 64 unidades com ativação ReLU, e um regularizador l2, seguidas por camadas de dropout para regularização. A saída final do modelo é produzida por uma camada densa com uma unidade e ativação sigmoid, adequada para problemas de classificação binária.

Este modelo é projetado para capturar características espaciais e temporais complexas nos dados de entrada, utilizando uma combinação de camadas convolucionais e LSTM, e refinando a saída através de camadas densas com regularização para evitar overfitting.

V. Experimentos

Inicialmente, 15% do conjunto de dados foi separado como conjunto de teste. Para a validação cruzada, foi utilizado um k-fold estratificado com 8 folds. Os experimentos foram conduzidos com os três modelos descritos na seção anterior e utilizaram as bandas de Precipitação, LSTDay, NDVI, FPAR, EVI e LAI.

Os experimentos foram realizados em etapas, começando com cada banda espectral de forma isolada. Em seguida, a banda com a melhor acurácia de validação foi adicionada à execução seguinte, formando uma matriz numpy com uma stack de bandas espectrais. Como experimento final, foi adicionada uma camada extra com informações sobre a ocorrência do El Niño às bandas variáveis com melhor desempenho de validação.

Esse procedimento permitiu verificar a mudança nos resultados conforme novas camadas foram adicionadas, proporcionando uma análise detalhada do impacto de cada variável no desempenho dos modelos.

A. Desempenho de Treino e Validação

Aqui, temos a intenção de comparar o desempenho do treino dos três modelos preditivos: Custom, EfficientNetB0 e Xception, utilizando diferentes combinações de variáveis.

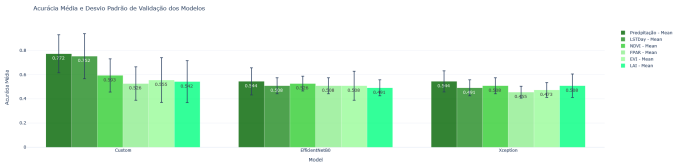

Fig. 5. Validação do Treino com uma variável

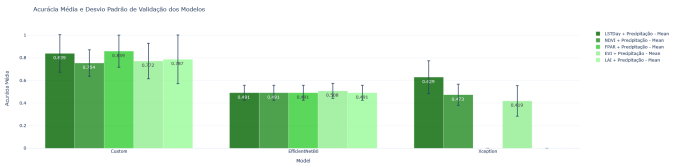

Fig. 6. Validação do Treino com duas variáveis

As métricas avaliadas foram acurácia e o desvio padrão da acurácia apresentada na validação das folds.

O modelo Custom apresentou consistentemente os melhores resultados em todas as combinações de variáveis e condições testadas, com destaque para as condições acrescida com os dados do El Niño, onde obteve uma acurácia máxima de 0.843 e desvio padrão de 0.147 para a combinação com as bandas de LAI + Precipitação + FPAR. O modelo EfficientNetB0 teve um desempenho inferior em todas as combinações, enquanto o modelo Xception apresentou resultados variados e geralmente inferiores ao modelo Custom.

Algumas variações não foram calculadas para os modelos EfficientNetB0 e Xception, e isso se deveu ao limite de recursos computacionais disponíveis para a execução desse trabalho.

Isolando as variáveis que alcançaram as melhores acurácias durante os experimentos, apresentamos uma análise detalhada do desempenho de treinamento e validação do modelo custom. O foco é fornecer uma visão abrangente do desempenho do modelo que parece ter se saído melhor nos treinos por meio de gráficos de acurácia e perda (loss).

Para as variáveis com as bandas FPAR + Precip, a acurácia de treinamento aumenta de forma geral, mas com algumas variações e um padrão mais irregular, especialmente após cerca de 60 epochs. A acurácia de validação mostra maior variabilidade e flutuações, com quedas abruptas em vários pontos, indicando instabilidade no desempenho do modelo nos dados de validação e sugerindo que o modelo está se ajustando de forma inconsistente aos dados de validação.

A perda no treinamento diminui de forma consistente ao longo do tempo, indicando que o modelo está aprendendo a minimizar o erro nos dados de treinamento. A perda de validação segue um padrão semelhante à perda de treinamento, mas com picos e flutuações significativas, especialmente após cerca de 60 epochs, reafirmando que o modelo está se ajustando de forma inconsistente aos dados de validação.

Para as variáveis LSTDAy + Precipitação + Fpar e dados do El Nino, a acurácia de treinamento mostra um aumento

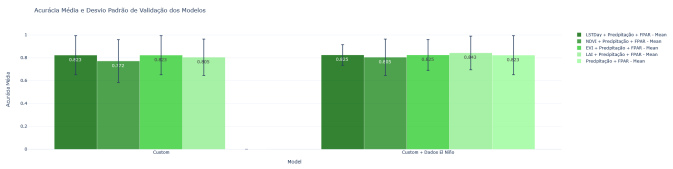

Fig. 7. Validação do Treino três variáveis e com dados do El Niño

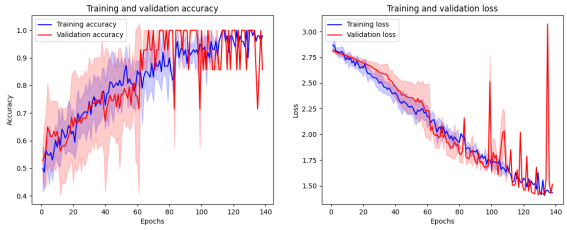

Fig. 8. Gráfico de Acurácia e Loss de Treinamento e Validação - FPAR + Precipitação

consistente ao longo do tempo, atingindo valores próximos de 1 à medida que o modelo aprende. A acurácia de validação também melhora ao longo do tempo, mas apresenta uma variabilidade maior e algumas quedas abruptas, especialmente nas últimas epochs. Essa variabilidade pode indicar que o modelo está se ajustando muito aos dados de treinamento e não generaliza bem nos dados de validação, constituindo um overfitting.

A perda no treinamento diminui de forma consistente ao longo das epochs, indicando que o modelo está aprendendo a minimizar o erro nos dados de treinamento. A perda de validação também diminui, mas com uma variabilidade maior e algumas picos de aumento, similar às quedas abruptas vistas na acurácia de validação. Apesar das variações, a acurácia de validação em muitos momentos é bastante alta (próxima de 1), indicando que o modelo tem um bom potencial de desempenho, mas precisa ser ajustado para melhorar sua generalização.

Em relação as variáveis com as bandas LAI + Precipitação + Fpar e dados do El Niño, a acurácia de treinamento mostra um aumento gradual, com algumas variações, e parece estabilizar perto de 1 (100%) após cerca de 80 epochs. A acurácia de validação também aumenta e parece seguir de perto a acurácia de treinamento, mas com mais flutuações e picos. A variabilidade na acurácia de validação sugere que o modelo pode estar se ajustando aos dados de validação de forma inconsistente, o que pode ser um sinal de overfitting.

A perda no treinamento diminui de forma constante, indicando que o modelo está aprendendo a minimizar o erro nos dados de treinamento. A perda de validação também diminui, seguindo um padrão similar ao da perda de treinamento, mas com flutuações maiores, especialmente nas últimas epochs. Apesar das flutuações, a perda de validação está em uma tendência geral de diminuição, o que é um sinal positivo. A performance geral parece boa, com acurácia de validação alta e

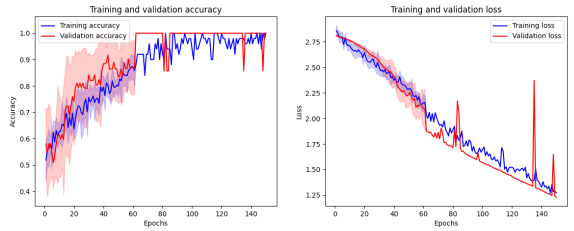

Fig. 9. Gráfico de Acurácia e Loss de Treinamento e Validação - LSTDAy + Precipitação+ Fpar + El Nino

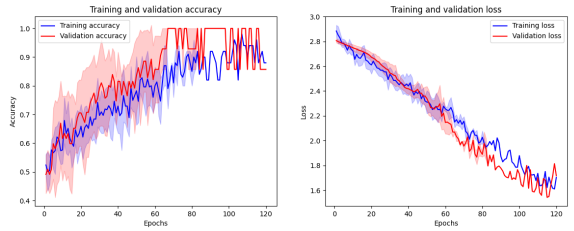

Fig. 10. Gráfico de Acurácia e Loss de Treinamento e Validação - LAI + Precip + Fpar + El Nino

perda de validação baixa na maior parte do tempo, indicando que o modelo está, em geral, funcionando bem, apesar das flutuações.

VI. Resultados

Foram feitas análises dos resultados de predição dos três modelos preditivos: Custom, EfficientNetB0 e Xception, utilizando as várias combinações de variáveis das bandas espectrais. As métricas avaliadas foram precisão, f1-score e acurácia de predição.

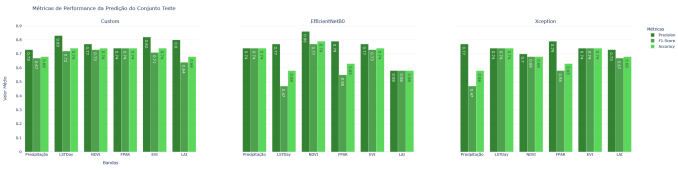

Fig. 11. Gráfico de Precisão, F1-Score e Acurácia de Predição - Com uma variável

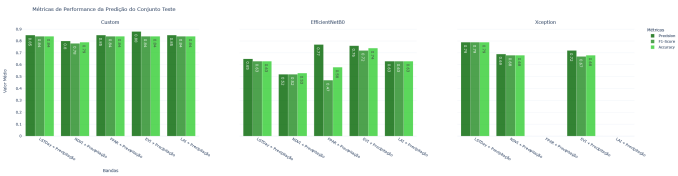

Fig. 12. Gráfico de Precisão, F1-Score e Acurácia de Predição - Com duas variáveis

O modelo Custom apresentou consistentemente os melhores resultados em todas as combinações de variáveis. Em especial, quando combinadas três variáveis, o modelo Custom mostrou

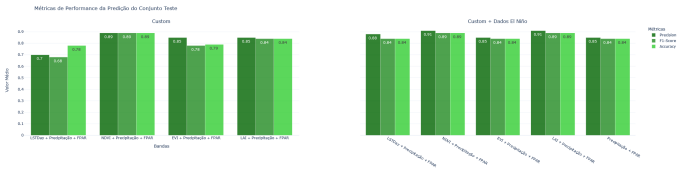

Fig. 13. Gráfico de Precisão, F1-Score e Acurácia de Predição - Com três variáveis e com dados do El Niño

valores muito altos de precisão, f1-score e acurácia, particularmente com as combinações NDVI + Precipitação + FPAR e LAI + Precipitação + FPAR, ambas com resultados de 0.91, 0.89, 0.89 para, respectivamente, índices de precisão, f1-score e acurácia. O modelo EfficientNetB0 teve um bom desempenho em algumas variáveis individuais, como Precipitação e NDVI, mas não foi competitivo quando mais variáveis foram adicionadas. O modelo Xception apresentou resultados variáveis, com alguns bons desempenhos em combinações de duas variáveis, mas não competiu em combinações de três variáveis.

Ao combinar variáveis como Precipitação, LSTDay e FPAR, o modelo Custom conseguiu melhorar significativamente a acurácia. Isso ocorre porque cada uma dessas variáveis pode influenciar os resultados de maneiras diferentes e complementares. A inclusão de condições climáticas, como o El Niño, também demonstrou ser valiosa, de forma a proporcionar um contexto adicional que pode afetar as variáveis, aumentando os valores de acurácia de alguns modelos. É esperado que aumentar o número de propriedades usadas nos dados de entrada, melhore significamente o os valores de acurácia e seja também, um potencial redutor do desvio padrão apresentado na validação do modelo.

Em relação ao modelo Custom, apresentamos os resultados obtidos pelos testes que demonstraram o melhor desempenho durante a validação e avaliação relatados. A avaliação dos resultados foi realizada com base em métricas de desempenho como a curva ROC (Receiver Operating Characteristic) e a matriz de confusão. Essas métricas fornecem uma visão abrangente da capacidade dos modelos em distinguir entre as diferentes classes e a eficácia geral na previsão dos dados de teste. Dessa forma, podemos definir a capacidade de generalização e precisão no melhor modelo e variáveis definidas.

Em relação às variáveis com bandas FPAR e Precipitação, a curva ROC e o valor de AUC indicam que o modelo tem uma boa capacidade de distinguir entre as classes positivas e negativas e que modelo é eficiente e equilibrado na classificação das amostras, com um médio desempenho geral.

Nos estudos com as bandas LSTDay + Precipitação + FPAR e dados do El Niño, acurva ROC sobe rapidamente, o que significa que o modelo tem uma alta taxa de verdadeiros positivos em comparação com a taxa de falsos positivos para a maioria dos limiares. O valor do AUC é 0.86, o que é um bom indicador de desempenho, sugerindo que o modelo tem uma boa capacidade discriminativa, sendo eficaz em separar as

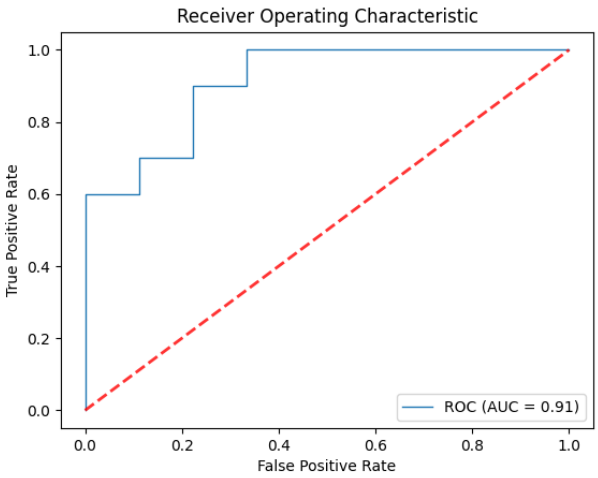

Fig. 14. Curva ROC - FPAR + Precip

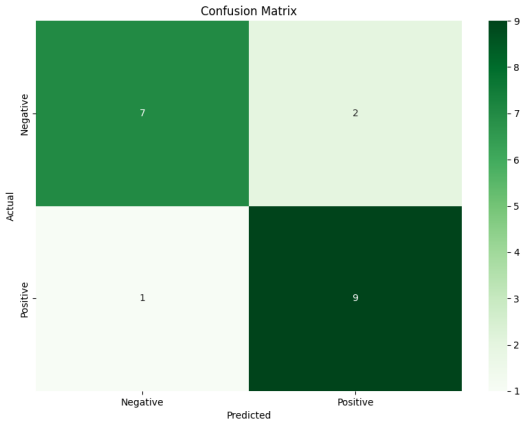

Fig. 15. Matriz de Confusão - FPAR + Precip

classes positivas das negativas. É possível definir que existe alguma dificuldade na identificação de instâncias negativas, ainda que tenha apresentado um bom equilíbrio entre precisão e recall.

A matriz de confusão mostra que o modelo com variáveis com as bandas LAI + Precipitação + FPAR e dados do El Niño tem uma alta acurácia e precisão, com uma boa capacidade de identificar corretamente as instâncias positivas e negativas. A ausência de falsos positivos é um ponto muito positivo e mostra que o modelo é confiável ao classificar instâncias negativas. A curva ROC sobe rapidamente, o que significa que o modelo tem uma alta taxa de verdadeiros positivos em comparação com a taxa de falsos positivos para a maioria dos limiares. O valor do AUC é 0.92, o que é um excelente indicador de desempenho. É possível definir que o modelo tem um desempenho excelente frente aos testes realizados.

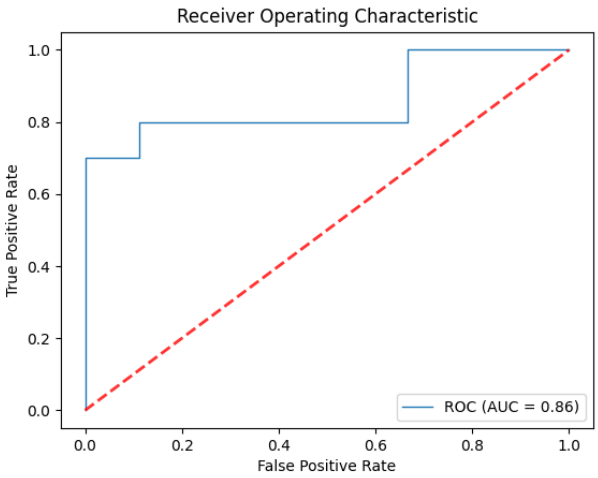

Fig. 16. Curva ROC - LSTDay + Precipitação + FPAR + El Niño

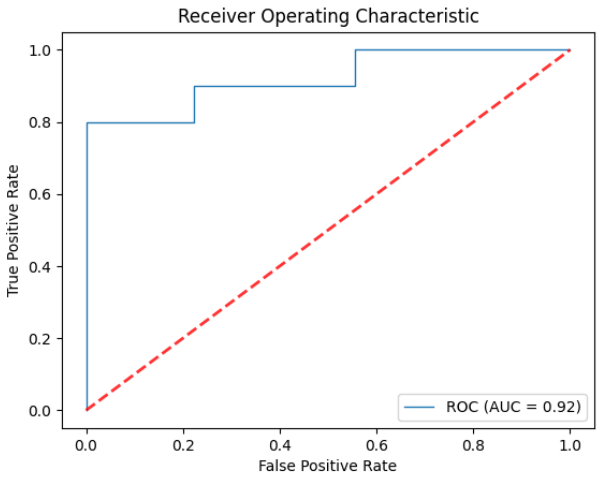

Fig. 18. Curva ROC - LAI + Precipitação + FPAR + El Niño

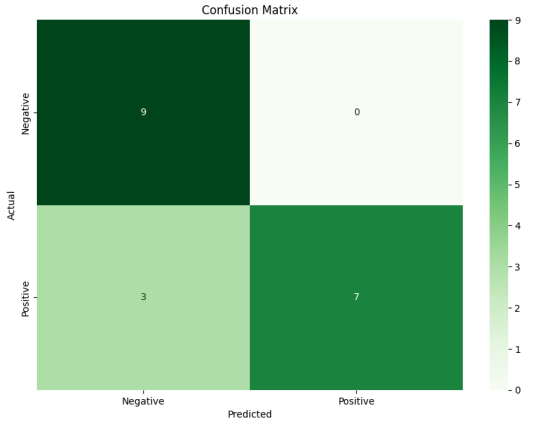

Fig. 17. Matriz de Confusão - LSTDay + Precipitação + FPAR + El Niño

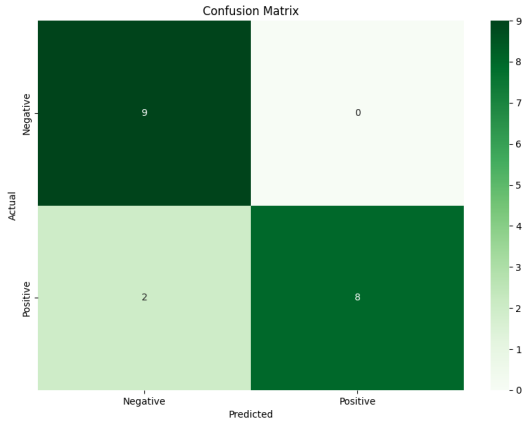

Fig. 19. Matriz de Confusão - LAI + Precipitação + FPAR + El Niño

VII. Conclusão

Com base na análise abrangente dos dados coletados, tratados e analisados, desenvolvemos um modelo preditivo de redes neurais convolucionais aliada ao LSTM e testamos outros dois modelos CNN utilizando diversas combinações de variáveis.

No geral, o modelo Custom apresentou consistentemente os melhores resultados em todas as combinações de variáveis e condições testadas. O modelo EfficientNetB0 teve um desempenho inferior em todas as combinações, enquanto o modelo Xception apresentou resultados variados e geralmente inferiores ao modelo Custom.

Durante o treinamento e validação, é possível perceber que ao combinar variáveis como Precipitação, LSTDay, FPAR, LAI e NDVI, o modelo Custom conseguiu melhorar significativamente a acurácia. Isso demonstra a capacidade das diversas bandas espectrais de descreverem a saúde e produção das culturas e a forma como esses dados influenciam os resultados de maneiras diferentes e complementares. A inclusão de condições climáticas, como o El Niño, demonstrou ser uma aliada à classificação, proporcionando um contexto adicional, aumentando os valores de acurácia de alguns modelos. Posteriormente, é considerado significativo, aumentar o número de dados de entrada (ou propriedades analisadas) de forma a auxiliar na melhora dos resultados de acurácia e também como um potencial redutor do desvio padrão apresentado na validação do modelo. Houve uma distinta dificuldade do presente trabalho em aumentar os dados de entrada devido à limitação de recursos computacionais.

Houve a validação da importância de índices vegetativos como LAI, NDVI, FPAR, dados climáticos e de temperatura na predição da produção e da saúde das culturas. Conseguimos desenvolver um modelo que apresenta um bom desempenho na classificação das instâncias positivas e negativas e que é muito eficaz em distinguir entre as classes.

Estas análises representam um passo inicial em direção ao desenvolvimento de um modelo de previsão e gerenciamento

de riscos robusto. No entanto, é evidente que o estudo contínuo nessa área é crucial, dada a constante evolução tecnológica e literária. Dessa forma, os resultados deste estudo buscam fornecer insights valiosos para projetos de pesquisa futuros. O foco subsequente em desenvolver modelos de deep learning capazes de prever com resiliência a ocorrência de sinistros agrários representa uma direção promissora para avançar na compreensão e na gestão eficaz dos riscos no setor de seguro rural.

REFERENCES

[1] Pedro Henrique Batista de Barros and Adirson Maciel de Freitas Junior. Combinando inteligência artificial e imagens de satélite para a previsão de sinistros agrícolas: Uma nota. *Revista Brasileira de Economia*, 77:e012023, 2023.

[2] Marvin E. Bauer. The role of remote sensing in determining the distribution and yield of crops. volume 27 of *Advances in Agronomy*, pages 271–304. Academic Press, 1975.

[3] Juan Cao, Zhao Zhang, Fulu Tao, Liangliang Zhang, Yuchuan Luo, Jing Zhang, Jichong Han, and Jun Xie. Integrating multi-source data for rice yield prediction across china using machine learning and deep learning approaches. *Agricultural and Forest Meteorology*, 297:108275, 02 2021.

[4] Xie Chen, Zhang, and Xue. Deep convolutional neural network for mapping smallholder agriculture using high spatial resolution satellite image. *Sensors*, 19:2398, 05 2019.

[5] Climate, Energy and Tenure Division. Terrestrial essential climate variables for climate change assessment, mitigation and adaptation. Technical Report GTOS 52 - Biennial Report Supplement, Global Terrestrial Observing System, 2008.

[6] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 07 2017.

[7] Kaiyu Guan, Jin Wu, J. Kimball, Martha Anderson, Steve Frolking, Bo Li, Christopher Hain, and David Lobell. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sensing of Environment*, 199:333–349, 09 2017.

[8] Daniel Pereira Guimarães. Zoneamento agrícola de risco climático (zarc) para o sorgo granífero no brasil, 2020.

[9] Jianxi Huang, Liyan Tian, Shunlin Liang, Hongyuan Ma, Inbal Becker-Reshef, Yanbo Huang, Wei Su, Xiaodong Zhang, Dehai Zhu, and Wu Wenbin. Improving winter wheat yield estimation by assimilation of the leaf area index from landsat tm and modis data into the wofost model. *Agricultural and Forest Meteorology*, 204:106–121, 05 2015.

[10] InfoMoney. Seguro rural cresce 39% no país em 2022, mas indenizações sobem ainda mais, 2022.

[11] Instituto Brasileiro de Geografia e Estatística (IBGE). Censo 2010, 2010.

[12] Saeed Khaki, Hieu Pham, and Lizhi Wang. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Scientific Reports*, 11, 05 2021.

[13] Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, 2019.

[14] Saeed Khaki, Lizhi Wang, and Sotirios Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10, 01 2020.

[15] Fan Liu, Xiangtao Jiang, and Zhenyu Wu. Attention mechanism-combined lstm for grain yield prediction in china using multi-source satellite imagery. *Sustainability*, 15:9210, 06 2023.

[16] Yuchuan Luo, Zhao Zhang, Juan Cao, Liangliang Zhang, Jing Zhang, Jichong Han, Huimin Zhuang, Fei Cheng, and Fulu Tao. Accurately mapping global wheat production system using deep learning algorithms. *International Journal of Applied Earth Observation and Geoinformation*, 110:102823, 06 2022.

[17] R. B. MacDonald and F. G. Hall. Global crop forecasting. *Science*, 208(4445):670–679, 1980.

[18] MAPA. Relatório geral do programa do seguro rural 2020, 2020.

[19] MAPA. Sisser 3.0 - sistema de seguro rural, 2021.

[20] METER Group. Researcher's complete guide to leaf area index (lai), 2023.

[21] Arthur Lula Mota, Daniel Lima Miquelluti, and Vitor Augusto Ozaki. Predição de sinistros agrícolas: uma abordagem comparativa utilizando aprendizagem de máquina. *Economia Aplicada*, 24(4):533–554, dez. 2020.

[22] National Centers for Environmental Prediction (NCEP). Oceanic niño index (oni), 2023.

[23] Mengjia Qiao, Xiaohui He, Xijie Cheng, Panle Li, Haotian Luo, Zhihui Tian, and Hengliang Guo. Exploiting hierarchical features for crop yield prediction based on 3-d convolutional neural networks and multikernel gaussian process. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–1, 04 2021.

[24] Toshihiro Sakamoto, Anatoly Gitelson, and Timothy Arkebauer. Modis-based corn grain yield estimation model incorporating crop phenology information. *Remote Sensing of Environment*, 131:215–231, 04 2013.

[25] Christie Andre de Souza and Michelle Simões Reboita. Ferramenta para o monitoramento dos padrões de teleconexão na américa do sul. *Terrae Didatica*, 17(00):e021009, fev. 2021.

[26] Yi Xie and Jianxi Huang. Integration of a crop growth model and deep learning methods to improve satellite-based yield estimation of winter wheat in henan province, china. *Remote Sensing*, 13(21), 2021.

Todo o código desenvolvido referente aos resultados obtidos neste trabalho, podem ser encontrados no repósitorio[1] do projeto.

[1] https://github.com/AnneIsabelleRodrigues/Predicao-Sinistro-Rural