v7.n11.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

INTEGRAÇÃO ENTRE INTELIGÊNCIA ARTIFICIAL E ARQUIVOLOGIA EM SETORES PÚBLICOS

INTEGRATION BETWEEN ARTIFICIAL INTELLIGENCE AND ARCHIVAL SCIENCE IN PUBLIC SECTORS

INTEGRACIÓN ENTRE INTELIGENCIA ARTIFICIAL Y ARCHIVOLOGÍA EN SECTORES PÚBLICOS

Alexandre Braga Gonçalves[1], Felipe Diniz Dallilo[2], João Henrique Gião Borges[3]

https://doi.org/10.47820/recima21.v7i11.9627

PUBLICADO: 11/2026

RESUMO

O presente trabalho analisa o comportamento e a aplicabilidade de três motores de Reconhecimento Óptico de Caracteres (OCR) — Tesseract, EasyOCR e PaddleOCR — no contexto da gestão documental de arquivos na administração pública. Caracterizada como um estudo de caso exploratório de amostragem intencional, a investigação observou o funcionamento dos motores perante nove documentos agrupados em três tipologias representativas: nato-digitais, híbridos (formulários com letra de forma) e manuscritos cursivos. Os dados observados indicaram a adequação específica de cada ferramenta conforme o suporte: o Tesseract destacou-se na precisão de documentos nato-digitais padronizados, enquanto o EasyOCR e o PaddleOCR demonstraram maior resiliência em *layouts* variáveis e textos de menor contraste, apresentando todas as ferramentas desafios acentuados na leitura de escrita cursiva. Conclui-se que não há um motor universalmente superior, mas sim ferramentas adequadas a cenários específicos, recomendando-se a escolha orientada pela tipologia do acervo e pela infraestrutura computacional disponível. Como perspectiva para trabalhos futuros, aponta-se a relevância do processamento documental em larga escala para mensurar parâmetros estatísticos abrangentes de desempenho.

PALAVRAS-CHAVE: Reconhecimento Óptico de Caracteres. Tesseract. EasyOCR. PaddleOCR. Administração Pública.

ABSTRACT

This study analyzes the behavior and applicability of three Optical Character Recognition (OCR) engines — Tesseract, EasyOCR, and PaddleOCR — within the context of document management in public administration archives. Characterized as an exploratory case study with purposive sampling, the investigation observed the performance of the engines across nine documents grouped into three representative typologies: born-digital, hybrid (forms filled in block letters), and handwritten cursive. The observed data indicated specific suitability for each tool depending on the medium: Tesseract stood out for its accuracy in standardized born-digital documents, whereas EasyOCR and PaddleOCR demonstrated greater resilience in variable layouts and lower-contrast texts, while all tools faced significant challenges in reading cursive handwriting. It is concluded that there is no universally superior engine, but rather tools suited to specific scenarios, with the recommendation that selection be guided by collection typology and available computational infrastructure. As a perspective for future work, the relevance of large-

[1] Graduando em Sistemas de Informação pela Universidade de Araraquara - UNIARA.

[2] Doutor em Engenharia de Software pela Universidade de São Paulo. Docente do curso de Sistemas de Informação da UNIARA.

[3] Doutor em Química Orgânica pela Universidade de São Paulo. Docente do curso de Sistemas de Informação da UNIARA.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

scale document processing is highlighted to measure comprehensive statistical performance parameters.

KEYWORDS: *Optical Character Recognition. Tesseract. EasyOCR. PaddleOCR. Public Administration.*

RESUMEN
El presente trabajo analiza el comportamiento y la aplicabilidad de tres motores de Reconocimiento Óptico de Caracteres (OCR) — Tesseract, EasyOCR y PaddleOCR — en el contexto de la gestión documental de archivos en la administración pública. Caracterizada como un estudio de caso exploratorio de muestreo intencional, la investigación observó el funcionamiento de los motores ante nueve documentos agrupados en tres tipologías representativas: nativos digitales, híbridos (formularios en letra de imprenta) y manuscritos cursivos. Los datos observados indicaron la adecuación específica de cada herramienta según el soporte: Tesseract se destacó en la precisión de documentos nativos digitales estandarizados, mientras que EasyOCR y PaddleOCR demostraron mayor resiliencia en diagramaciones variables y textos de menor contraste, presentando todas las herramientas desafíos acentuados en la lectura de escritura cursiva. Se concluye que no existe un motor universalmente superior, sino herramientas adecuadas a escenarios específicos, recomendándose una selección orientada por la tipología del acervo y la infraestructura computacional disponible. Como perspectiva para trabajos futuros, se señala la relevancia de realizar un procesamiento de documentos a gran escala para medir parámetros estadísticos exhaustivos de desempeño.

PALABRAS CLAVE*: Reconocimiento Óptico de Caracteres. Tesseract. EasyOCR. PaddleOCR. Administración Pública.*

1. INTRODUÇÃO

Historicamente, a estrutura administrativa dos setores públicos esteve atrelada ao acúmulo massivo de documentos impressos, consolidando vastos arquivos físicos de manuseio lento e complexo. Embora a transição desse acervo para os meios digitais represente um avanço significativo, a simples conversão de suporte não garante a recuperação eficiente da informação. A arquivologia moderna deve transcender a custódia do papel e adaptar-se às tecnologias contemporâneas. Desse modo, não basta apenas digitalizar essas informações; elas precisam ser acessíveis ao arquivista ou a quem necessite dos dados.

Frequentemente, essa digitalização resulta em arquivos de imagem ou *Portable Document Format* – Formato de Documento Portátil (PDF) que atuam como verdadeiras “fotos cegas”. Do ponto de vista da gestão informacional, Schäfer e Lima (2015) alertam que sistemas eletrônicos tradicionais falham ao atuar de forma compartimentada, ignorando atributos orgânicos essenciais. Na prática tecnológica, essa falta de estruturação impede que os arquivos sejam plenamente pesquisáveis, um desafio que se agrava drasticamente ao lidarmos com textos manuscritos ou degradados. Para transpor esse hiato entre a digitalização primária e a

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

interpretação estruturada, será utilizada uma avaliação de inteligências artificiais baseadas em *Optical Character Recognition* – Reconhecimento Óptico de Caracteres (OCR).

O objetivo deste trabalho é observar e comparar a aplicabilidade prática de três motores de reconhecimento óptico de código aberto (Tesseract, EasyOCR e PaddleOCR) em um ambiente experimental que simula as demandas documentais de uma prefeitura municipal. Por meio de um estudo de caso exploratório, busca-se identificar qual ferramenta apresenta maior adequação operacional a cada tipologia documental específica, fornecendo subsídios técnicos para apoiar a tomada de decisão no setor público sem generalizar os resultados para além do recorte analisado.

Avanços da visão computacional, que evoluíram de algoritmos básicos de reconhecimento de padrões para complexas arquiteturas de redes neurais, tornam possível à inteligência artificial possuir uma capacidade analítica aprimorada na leitura e interpretação de documentos. É por meio desse aparato tecnológico que se viabiliza ao Estado o cumprimento rigoroso do que preconiza a Lei nº 14.129/2021 (Lei do Governo Digital) (BRASIL, 2021) e o Decreto nº 10.278/2020 (BRASIL, 2020). Ao regulamentar o disposto na Lei nº 13.874/2019 (Lei da Liberdade Econômica) (BRASIL, 2019) e na Lei nº 12.682/2012 (Lei de Digitalização de Documentos) (BRASIL, 2012), o referido decreto estabelece a técnica e os requisitos para a digitalização de documentos públicos e privados, garantindo que os representantes digitalizados produzam os mesmos efeitos legais dos documentos originais. Para que tal equivalência seja alcançada, ao fixar os padrões técnicos mínimos, a norma exige que o processo assegure a integridade e a indexação do conteúdo. Desse modo, a informação rompe com a limitação do arquivo estático, convertendo-se em um ativo de dados plenamente pesquisável e acessível, em consonância com o princípio constitucional da eficiência administrativa. A gestão documental otimizada transcende a mera obrigação burocrática, consolidando-se como um pilar indispensável para a transparência pública e para a efetivação da Lei nº 12.527/2011 (Lei de Acesso à Informação – LAI) (BRASIL, 2011) no âmbito dos municípios (SILVA, 2020).

Por mais atualizado que seja um determinado departamento público, raramente se leva em consideração qual seria o próximo passo após a digitalização de um documento. Se o arquivo estiver degradado ou for manuscrito, há grande dificuldade em extrair determinadas informações, tornando-o uma simples representação visual, com dados não pesquisáveis internamente ou de difícil leitura. Por esse motivo, a proposta busca identificar uma ferramenta de inteligência artificial com acurácia satisfatória e baixo custo de implementação para o cenário nacional. Diante desse contexto, o presente trabalho é norteado pela seguinte questão de pesquisa: de que maneira diferentes motores gratuitos de Inteligência Artificial baseados em OCR (Tesseract,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

EasyOCR e PaddleOCR) se comportam na extração e conversão de dados em um acervo municipal, e qual o grau de adequação de cada ferramenta perante as diferentes tipologias documentais analisadas?

Para responder a essa questão, adotou-se uma abordagem de natureza exploratória, aplicada e comparativa, fundamentada na observação empírica dos modelos de Inteligência Artificial e na revisão de suas documentações oficiais. O desenvolvimento estruturou-se a partir da configuração de um ambiente de programação na linguagem Python, utilizando a biblioteca CustomTkinter para a criação de uma interface gráfica de apoio aos experimentos, operada pelo próprio pesquisador. Em seguida, procedeu-se à análise detalhada das ferramentas de OCR delineadas para o escopo deste estudo, avaliando suas características operacionais e limitações perante a língua portuguesa. Para garantir a integridade dos dados extraídos, desenvolveu-se um algoritmo de normalização capaz de processar os textos lidos como *strings* únicas, mitigando potenciais distorções decorrentes de falhas de formatação. Por fim, a observação dos resultados foi amparada por critérios matemáticos consolidados, empregando-se a Taxa de Erro de Caracteres (CER — *Character Error Rate*), o índice de acurácia e a quantificação exata das operações de edição — substituições, deleções e inserções — efetuadas pelas ferramentas de OCR, utilizados como indicadores práticos para compreender o grau de adequação de cada motor às diferentes tipologias documentais analisadas.

2. REFERENCIAL TEÓRICO

Esta seção articula os preceitos clássicos da gestão documental com as inovações tecnológicas da inteligência artificial. Para tanto, a discussão inicia-se com uma abordagem sobre os conceitos e princípios essenciais da Arquivologia e o funcionamento do ciclo vital dos documentos no setor público. Em seguida, examina-se o panorama das tecnologias de OCR, detalhando o funcionamento dos motores computacionais selecionados para os testes. Por fim, delineiam-se as ferramentas de desenvolvimento e os recursos de programação que viabilizaram a construção do laboratório computacional deste estudo.

2.1. Fundamentos e Princípios da Arquivologia

A Arquivologia constitui a disciplina científica dedicada ao estudo da origem, organização, preservação e difusão da informação registrada, independentemente de seu suporte físico. Seu objetivo central não se limita à mera guarda ou custódia de documentos, mas sim à garantia do acesso rápido e seguro aos dados, resguardando o valor probatório e histórico dos registros. Conforme argumenta Fonseca (2005), o foco da ciência arquivística deve ser a

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

preservação das características intrínsecas do documento, as quais garantem a sua fidedignidade e autenticidade como reflexo fiel das atividades institucionais. Dentre os pilares dessa disciplina, destaca-se o princípio da organicidade, que compreende o documento não como um item isolado, mas como parte de um conjunto natural e inter-relacionado, gerado organicamente em decorrência das funções de uma entidade. Na administração pública, a aplicação rigorosa desses preceitos arquivísticos — incluindo o princípio da proveniência, que impede a mistura de documentos de origens diversas — é o que assegura a integridade do patrimônio documental, estabelecendo as bases primárias para a continuidade administrativa, a transparência e a preservação da memória institucional.

2.2. A Integração entre Arquivologia e os Sistemas da Informação

A gestão documental no setor público exige uma visão interdisciplinar entre os arquivos físicos e os sistemas informatizados. Fonseca (2005) argumenta que a Arquivologia e a Ciência da Informação compartilham o estudo do fenômeno da informação, em que o foco deve ser a preservação das características intrínsecas do documento — como organicidade e contexto — independentemente de o suporte ser papel ou eletrônico. O desafio contemporâneo reside em aplicar esses princípios de arquivamento aos complexos ambientes digitais das administrações municipais.

Complementando essa visão, Schäfer e Lima (2015) advertem que a simples digitalização sem instrumentos técnicos de classificação torna a gestão documental deficiente. Para os autores, os sistemas eletrônicos costumam falhar quando realizam a mera transposição do documento físico para o meio digital, ignorando sua validade como prova. A negligência com a recuperação dessas informações impacta diretamente na governança. Silva (2020) destaca que, se as políticas de arquivo são fracas e as informações não podem ser pesquisadas, a tomada de decisão é prejudicada, configurando uma falha na transparência e no cumprimento da Lei de Acesso à Informação (LAI) (BRASIL, 2011).

2.3. Parâmetros de Avaliação: Texto-Ouro e Métricas de Erro

A coleta de dados para a aferição da acurácia das inteligências artificiais no escopo deste trabalho baseia-se em dois pilares metodológicos fundamentais: o estabelecimento de um referencial de controle e a aplicação de uma métrica matemática padronizada de distanciamento.

O primeiro pilar consiste no conceito de Texto-Ouro (do inglês, *Ground Truth*). No contexto do processamento de imagens e da transcrição automatizada, o Texto-Ouro representa a versão exata, verídica e incontestável do conteúdo contido no documento original, transcrita ou rigorosamente revisada de forma humana pelo pesquisador. Este fragmento textual atua

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

como o gabarito absoluto do experimento, tendo sido digitado manualmente pelo pesquisador a partir da observação visual direta dos originais e submetido a dupla revisão para eliminar vieses humanos. Previamente normalizado e livre de ruídos de formatação, é contra este Texto-Ouro que todas as saídas de dados (*outputs*) geradas pelas ferramentas de OCR serão confrontadas. Adicionalmente, elementos puramente gráficos presentes nos documentos, como assinaturas e trechos manuscritos, foram inseridos manualmente no gabarito para garantir a integridade da avaliação das métricas.

O segundo pilar metodológico compreende a Taxa de Erro de Caractere (CER), consolidada na literatura científica como a métrica mais confiável para a avaliação de sistemas de reconhecimento de texto. Diferentemente de avaliações superficiais baseadas apenas na contagem de palavras, o CER mensura a diferença em nível de caractere valendo-se da Distância de Levenshtein. Esse conceito matemático calcula o número mínimo de edições operacionais necessárias para transformar o resultado gerado pela IA de volta no Texto-Ouro exato. Na prática, o algoritmo varre as cadeias de texto e contabiliza as falhas agrupando-as em três tipologias de edição: as substituições (quando um caractere é lido de forma equivocada), as inserções (quando a máquina adiciona símbolos irreais devido a ruídos na imagem) e as deleções (quando a ferramenta omite ou ignora um caractere existente).

O cálculo final, conforme detalhado na Equação (1) a seguir, consiste em somar todas essas operações de distanciamento e dividi-las pelo número total de caracteres de referência. Complementarmente, a acurácia global da ferramenta é extraída como o inverso exato dessa taxa de erro, expressando em percentual o nível de exatidão do motor testado.

$$CER = \frac{S + D + I}{N} \quad (1)$$

Onde:

S - Substituições, número de caracteres que foram reconhecidos de forma errada no lugar correto.

D - Deleções, número de caracteres que estavam no texto original, mas foram omitidos.

I - Inserções, número de caracteres extras/inválidos adicionados por engano.

N - Número total de caracteres no texto de referência/original (*Ground Truth*).

2.4. Tecnologias de Reconhecimento Óptico de Caracteres (OCR) e Inteligência Artificial (IA)

As tecnologias de Reconhecimento Óptico de Caracteres (OCR) atuam como a ponte computacional necessária para transformar imagens estáticas e documentos digitalizados em dados de texto plenamente pesquisáveis. No contexto da gestão documental e arquivística,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

essas ferramentas são indispensáveis para extrair informações contidas em arquivos PDF ou imagens escaneadas que, sob a ótica computacional, constituem apenas matrizes bidimensionais de pixels sem valor semântico direto. A evolução dessas *engines* reflete a própria evolução da Ciência da Computação: a transição de métodos determinísticos de Visão Computacional para modelos complexos baseados em Inteligência Artificial (IA) e Aprendizado Profundo (*Deep Learning*).

Historicamente, os sistemas de OCR baseavam-se na correspondência de matrizes (*template matching*) e na extração de características geométricas rígidas. Nesses modelos tradicionais, o sistema dependia da binarização estrita da imagem (conversão em preto e branco) e da segmentação do texto em caracteres isolados para compará-los com uma biblioteca pré-definida de fontes. Contudo, essa abordagem apresentava severas limitações ao enfrentar documentos com degradação física, ruídos de digitalização, variação de fontes ou inclinação de página.

A integração da Inteligência Artificial superou essas barreiras ao substituir regras manuais por Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs) e Redes Neurais Recorrentes (*Recurrent Neural Networks* - RNNs). Por meio dessas arquiteturas, as IAs modernas não analisam apenas a forma isolada da letra, mas aprendem representações visuais abstratas e dependências sequenciais do texto, reconhecendo palavras completas mesmo sob condições adversas. Essa transformação converte massas de dados não estruturados em informações indexáveis, otimizando a recuperação da informação e garantindo o processamento automatizado de acervos em escala.

2.4.1. Tesseract

O Tesseract é um dos motores de OCR de código aberto (*open-source*) mais maduros e amplamente difundidos no mundo (TESSERACT-OCR, 2026), mantido pelo Google após a sua disponibilização pública pela HP no final de 2005. Conforme documentado por Smith (2007), a arquitetura tradicional do sistema estrutura-se em um *pipeline* sequencial que envolve a análise detalhada do *layout* da página, a localização de linhas de texto, a segmentação de palavras e o reconhecimento de caracteres por meio de classificadores adaptativos.

A partir da versão 4, o Tesseract incorporou um motor baseado em redes neurais do tipo LSTM (*Long Short-Term Memory*), um tipo especial de Rede Neural Recorrente projetada para processar sequências de dados. Apesar dessa atualização arquitetural, o Tesseract preserva forte dependência das etapas clássicas de pré-processamento de imagem. Para alcançar taxas elevadas de acurácia, a ferramenta exige que o documento passe por ajustes rigorosos de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

binarização, remoção de ruídos e correção de orientação (*deskew*). Quando submetido a textos tipográficos padronizados e imagens com alto contraste, o Tesseract exibe excelente velocidade de processamento e baixo custo computacional, consolidando-se como uma "linha de base" (*baseline*) para comparação em pesquisas acadêmicas.

2.4.2. EasyOCR

O EasyOCR é um *framework* moderno de OCR desenvolvido pela Jaided AI (2020), projetado nativamente para a linguagem Python com foco em flexibilidade e alta precisão. Diferenciando-se das ferramentas tradicionais, o EasyOCR opera fundamentado integralmente em arquiteturas de *Deep Learning*, dividindo a tarefa de extração de texto em dois estágios neurais distintos: detecção de texto e reconhecimento de caracteres.

Para a etapa de detecção, o EasyOCR utiliza o algoritmo CRAFT (*Character Region Awareness for Text Detection*), capaz de identificar regiões de texto em nível de caractere e correlacioná-las, permitindo a detecção precisa de palavras mesmo em geometrias irregulares, textos curvos ou rotacionados. Na etapa de reconhecimento, o *framework* combina uma Rede Neural Convolucional (ResNet) para a extração de recursos visuais com uma rede BiLSTM (*Bidirectional LSTM*) e uma camada CTC (*Connectionist Temporal Classification*) para a transcrição da sequência de texto sem a necessidade de segmentar caractere por caractere. Essa estrutura confere ao EasyOCR elevada robustez contra fundos complexos, sombras e ruídos visuais, dispensando etapas exaustivas de pré-processamento manual.

2.4.3. PaddleOCR

O PaddleOCR é um ecossistema de Reconhecimento Óptico de Caracteres de código aberto (*open-source*) desenvolvido pela Baidu (PADDLEPADDLE, 2026), reconhecido por sua alta eficiência e precisão em cenários industriais e acadêmicos. Conforme explicitado por Du *et al*. (2020), a arquitetura do PaddleOCR opera por meio de uma abordagem de duas etapas principais: a Detecção de Texto e o Reconhecimento de Texto. Na primeira fase, utiliza-se o algoritmo DBNet (*Differentiable Binarization*) para localizar e delimitar as regiões contendo caracteres. Em seguida, a etapa de reconhecimento decodifica a sequência visual por meio de modelos neurais como a arquitetura CRNN (*Convolutional Recurrent Neural Network* ou Rede Neural Recorrente Convolucional), podendo também empregar modelos mais recentes como o SVTR (*Single Visual Model for Text Recognition* ou Modelo Visual Único para Reconhecimento de Texto) nas versões atualizadas do ecossistema.

Ao contrário de modelos com vocabulário restrito, o PaddleOCR possui suporte nativo e abrangente ao idioma português, mapeando corretamente diacríticos e pontuações da língua

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

nacional. Essa característica elimina o fenômeno de "alucinação" semântica comum em motores focados apenas no inglês. O modelo demonstra elevada robustez no processamento de textos tipográficos (*Printed*) e documentos caligráficos (*Handwritten*). Em sua aplicação prática neste estudo, optou-se por sua operação integral e autônoma. Diferente de abordagens que exigem segmentação prévia (recortes em linhas) via ferramentas clássicas, o PaddleOCR demonstrou eficácia superior ao receber a imagem da página completa, utilizando sua própria inteligência (algoritmo DBNet) para localizar as regiões de texto, o que evitou perdas de contexto e erros de detecção causados por recortes irregulares.

2.5. Linguagem Python

O Python é uma linguagem de programação de alto nível, interpretada e de propósito geral, amplamente reconhecida por sua sintaxe clara e expressividade. Para o desenvolvimento e execução do ambiente experimental deste trabalho, adotou-se a versão Python 3.10. A escolha dessa versão específica justifica-se pelo equilíbrio entre as otimizações de desempenho do interpretador CPython e a estabilidade na compatibilidade de pacotes (*wheels*) com os ecossistemas de Inteligência Artificial e Visão Computacional, garantindo, sobretudo, a plena operação das dependências do motor PaddleOCR.

A utilização do Python 3.10 viabilizou a integração ágil de motores heterogêneos de extração em um único fluxo automatizado de testes, permitindo que o Tesseract, o EasyOCR e o PaddleOCR fossem orquestrados em um ambiente controlado, porém respeitando as especificidades de entrada e os *pipelines* de pré-processamento otimizados para cada arquitetura. Ademais, a linguagem provê ferramentas robustas para a manipulação estruturada dos textos extraídos e a automação do cálculo de métricas numéricas, como o CER.

2.6. Ambiente de Desenvolvimento: Visual Studio Code

Para estruturar o laboratório computacional desta pesquisa e executar os testes empíricos, adotou-se o Visual Studio Code (comumente chamado de VS Code) como ambiente principal. Desenvolvido pela Microsoft, este editor de código-fonte destaca-se por ser leve, gratuito e multiplataforma, diferenciando-se de Ambientes de Desenvolvimento Integrado (IDEs) tradicionais por oferecer alta performance e robustez sem o consumo excessivo de recursos computacionais. Na prática, o ambiente de desenvolvimento funciona como a infraestrutura central do projeto: é o espaço onde o código é efetivamente redigido, organizado e testado. A escolha do VS Code justifica-se pela sua vasta capacidade de customização via extensões e pelo suporte avançado à linguagem Python. A ferramenta unifica de forma eficiente o gerenciamento de diretórios, a

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

criação de ambientes virtuais e os recursos de depuração de código (debugging). Toda essa integração técnica tornou a criação e a execução do *script* de avaliação das ferramentas de OCR muito mais ágeis, dando origem à base computacional que estrutura o presente trabalho.

3. METODOLOGIA

O percurso metodológico deste trabalho caracteriza-se como um estudo de caso exploratório com amostragem intencional, estruturado para observar o comportamento de ferramentas de visão computacional no contexto das rotinas documentais municipais. Diante da diversidade dos acervos governamentais, a pesquisa delimita-se a analisar como cada arquitetura de OCR responde a cenários específicos de complexidade, buscando identificar os cenários de maior adequação para cada ferramenta sem emitir juízos universais sobre a eficácia dos motores.

3.1. Tratamento de Imagens e Normalização

No que tange à ingestão e ao pré-processamento dos dados, o carregamento e a conversão das páginas de arquivos PDF em imagens rasterizadas (sob uma resolução de 300 DPI) foram realizados diretamente pelas funções nativas do pacote PyMuPDF. Durante a concepção do laboratório, testou-se a utilização da biblioteca OpenCV (cv2) para aplicar técnicas de segmentação e recorte de linhas; contudo, testes empíricos revelaram que a extração de recortes muito restritos ("fatias" de texto) prejudicava a capacidade dos detectores internos das IAs, resultando em elevadas taxas de erro.

Diante disso, o *pipeline* foi otimizado: o OpenCV e a biblioteca NumPy atuaram no pré-processamento direcionado, garantindo os espaços de cor adequados a cada arquitetura, mantendo a matriz em RGB para o EasyOCR e convertendo para BGR no PaddleOCR. Para o Tesseract, aplicou-se a conversão para escala de cinza e o filtro de desfoque mediano, preparando a imagem de forma ideal para cada motor. Essa abordagem garante que os motores de OCR recebam o documento de forma integral, delegando a identificação do *layout* e a detecção dos blocos de texto exclusivamente às redes neurais, o que mitigou as perdas de leitura. Para o tratamento final das *strings* geradas pelos motores, utilizou-se o módulo de Expressões Regulares (re), focado em normalizar quebras de linha (\n), remover caracteres nulos e padronizar espaços múltiplos, formatando os textos extraídos de maneira contínua, viabilizando a comparação com o padrão-ouro.

3.2. Motores de Leitura (OCR) e Avaliação de Métricas

O núcleo de reconhecimento óptico foi orquestrado mediante a integração de três dependências principais: o PyTesseract (versão 0.3.13), atuando como ponte de comunicação

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

para o motor clássico do Tesseract (versão 5.5.3.20260724), equipado com o modelo linguístico por.traineddata sob arquitetura de rede neural LSTM; o pacote nativo do EasyOCR (versão 1.7.2), executando o modelo CRAFT para detecção de texto e o modelo pré-treinado de reconhecimento latin_g2 (baseado em arquitetura CRNN) para os idiomas português e inglês; e a biblioteca oficial do PaddleOCR (versão 2.7.3), empregando o conjunto de modelos ultraleves latin_PP-OCRv3 (detecção DB e reconhecimento CRNN). Para garantir a padronização experimental, todos os motores receberam as mesmas imagens rasterizadas a 300 DPI, sendo os referidos *arrays* submetidos apenas às adequações exigidas pela arquitetura de cada ferramenta (como ajustes de cor e filtros de desfoque), conforme descrito no escopo de pré-processamento. A aferição observacional foi apoiada pela integração do pacote JiWER, responsável por processar o alinhamento das *strings* e contabilizar as edições (substituições, inserções e deleções), gerando a Taxa de Erro de Caractere (CER) e a correspondente taxa de acerto como indicadores quantitativos de suporte à análise do cenário.

3.3. Arquitetura do Sistema

Sob a ótica da Engenharia de Software, o presente estudo demandou o desenvolvimento de uma aplicação centralizadora para orquestrar os múltiplos motores de OCR e as bibliotecas de processamento. A arquitetura do sistema foi projetada num modelo de fluxo de dados (*pipeline*), onde a imagem transita desde o pré-processamento até a avaliação de métricas, garantindo isolamento entre a lógica de extração (*Back-end*) e a camada de apresentação (*Front-end*).

Para materializar essa arquitetura e desenvolver a interface gráfica moderna e responsiva (*Front-end*), empregou-se a biblioteca CustomTkinter, enquanto o código-fonte lógico foi desenvolvido integralmente em linguagem Python. A estrutura do *software* foi concebida de forma modular. O módulo de entrada é responsável por receber os caminhos dos arquivos gerados pelo usuário; o módulo de processamento aciona as dependências (OpenCV, PyTesseract, EasyOCR e PaddleOCR); e o módulo de avaliação submete os retornos ao pacote JiWER para o cálculo do CER e da acurácia, comparando a transcrição gerada com o padrão-ouro.

3.4. Diagrama de Atividades (Fluxo de Execução)

O Diagrama de Atividades é uma representação comportamental da UML (*Unified Modeling Language*) empregada para mapear o fluxo dinâmico de controle e a sequência lógica de passos de um sistema. Sua principal função é ilustrar o caminho de execução, desde o

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

acionamento inicial até a conclusão do processo, destacando as ações sequenciais e os pontos de decisão (bifurcações) da aplicação.

No escopo deste laboratório (Figura 1), o diagrama modela o ciclo de vida completo de um teste. O fluxo estrutura-se da seguinte maneira:

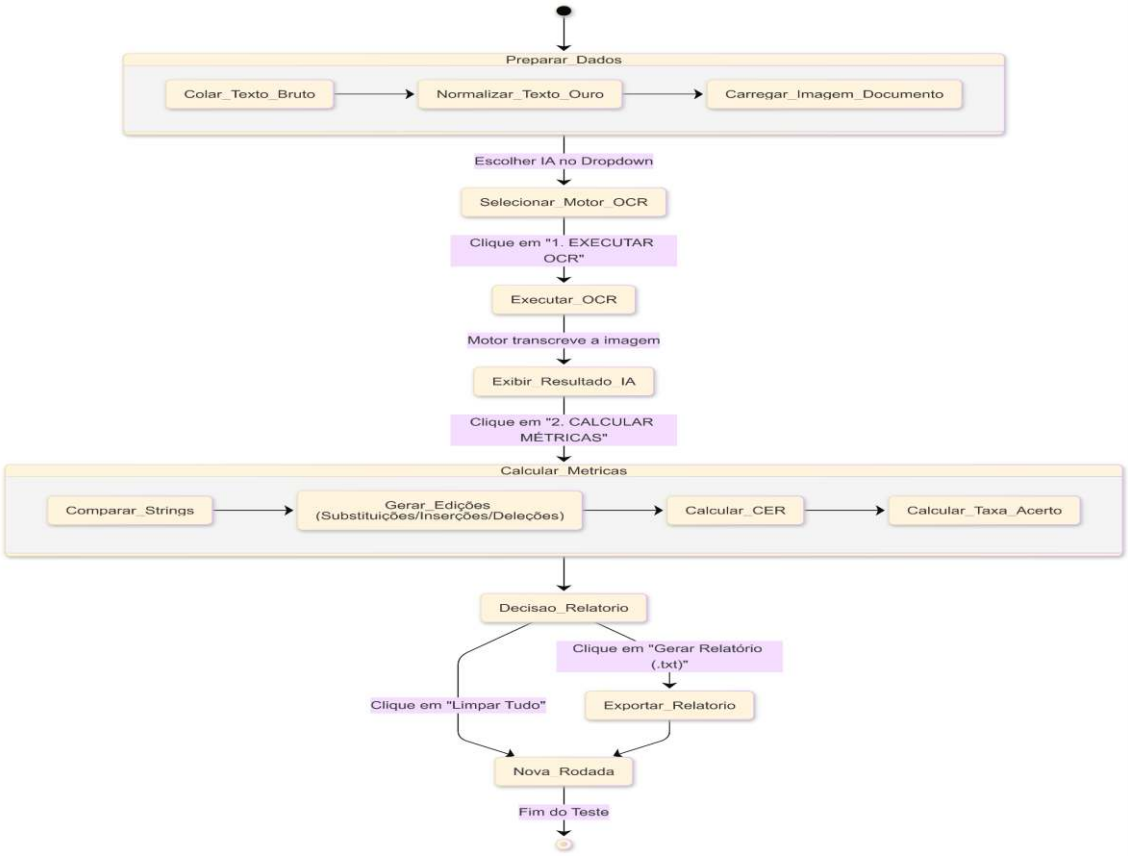
a) Preparação: Inicia-se na fase de Preparar Dados, onde o operador insere o texto bruto, consolida o padrão-ouro e carrega a imagem do documento selecionado.

b) Processamento: O fluxo avança para a seleção da IA e o acionamento do motor óptico, resultando na transcrição primária da imagem.

c) Validação: Ao acionar o cálculo, o sistema adentra o bloco de Calcular Métricas, comparando as *strings* para contabilizar as operações de edição (inserções, deleções e substituições), o que viabiliza a geração do CER.

d) Decisão: O fluxo encerra-se no nó lógico Decisão_Relatório, onde o pesquisador opta por exportar os resultados consolidados (em formato *.txt*) ou acionar o comando de limpeza para iniciar uma nova rodada.

Figura 1. Diagrama de Atividades



Fonte: Elaborado pelo Autor (2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

3.5. Modelagem de Dados: Diagrama Entidade-Relacionamento

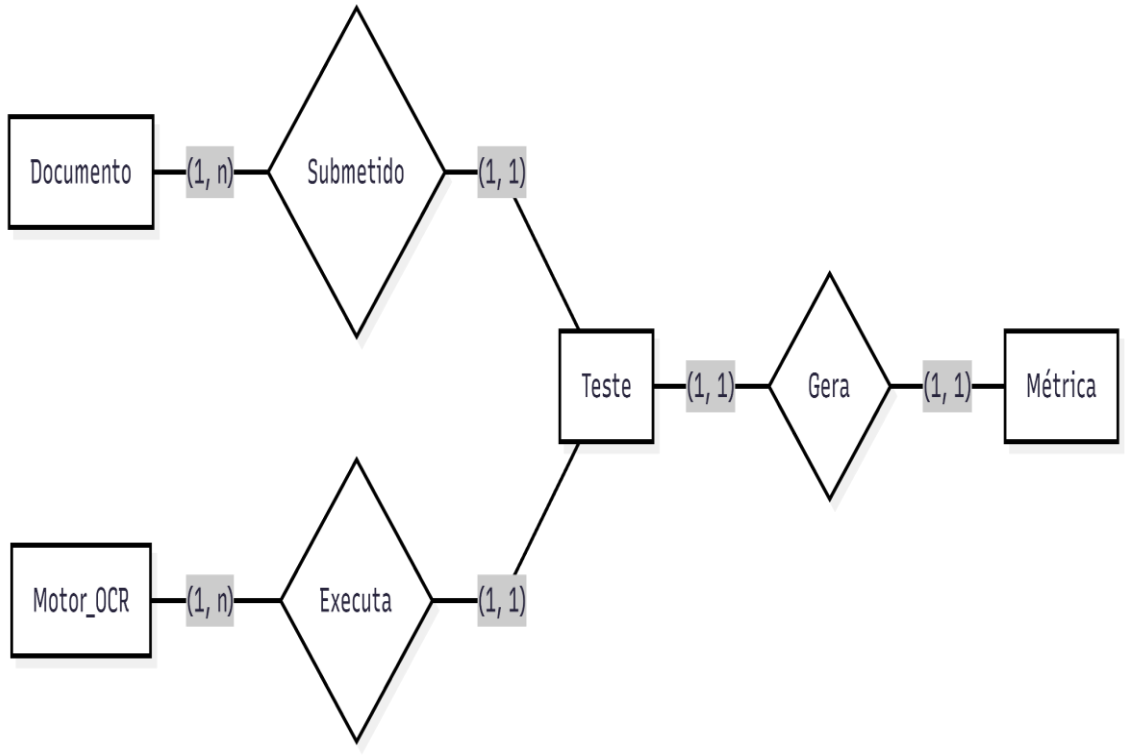
O Diagrama Entidade-Relacionamento (DER) é um instrumento de modelagem conceitual que descreve a estrutura lógica das informações que trafegam em um *software*. Utilizando a notação clássica de Peter Chen, este modelo formaliza as entidades do domínio do problema e define rigorosamente como elas interagem através de cardinalidades (regras de multiplicidade).

Conforme ilustrado na Figura 2, a arquitetura de dados do experimento fundamenta-se em quatro entidades centrais, tendo a entidade Teste como o núcleo relacional de intersecção:

a) Relação de Entrada (1, n): As entidades Documento e Motor_OCR ligam-se ao processo central demonstrando que um mesmo arquivo físico ou um mesmo modelo de Inteligência Artificial podem participar de múltiplos testes distintos no sistema.

b) Relação de Saída (1, 1): O fluxo de exportação obedece a uma relação estrita e exclusiva. Cada teste isolado gera, obrigatoriamente, um único quadro de métrica, atrelando os resultados analíticos apenas àquela rodada específica de execução.

Figura 2. Diagrama de Entidade-Relacionamento



Fonte: Elaborado pelo Autor (2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

3.6. Arquitetura de Software: Diagrama em Camadas

O Diagrama Arquitetural tem como objetivo representar a organização estrutural do sistema em alto nível, evidenciando a adoção do princípio de Separação de Responsabilidades (*Separation of Concerns*). Em modelos de fluxo de dados (*pipeline*), a esquematização mapeia a hierarquia dos componentes tecnológicos, demonstrando como as bibliotecas se integram sem sobreposição de funções. Na aplicação desenvolvida (Figura 3), o *software* foi segregado em quatro camadas lógicas isoladas:

a) Camada de apresentação (GUI): gerenciada pela biblioteca CustomTkinter, atua como o ponto de controle do pesquisador, responsável por injetar os documentos no sistema e receber a exibição dos resultados analíticos;

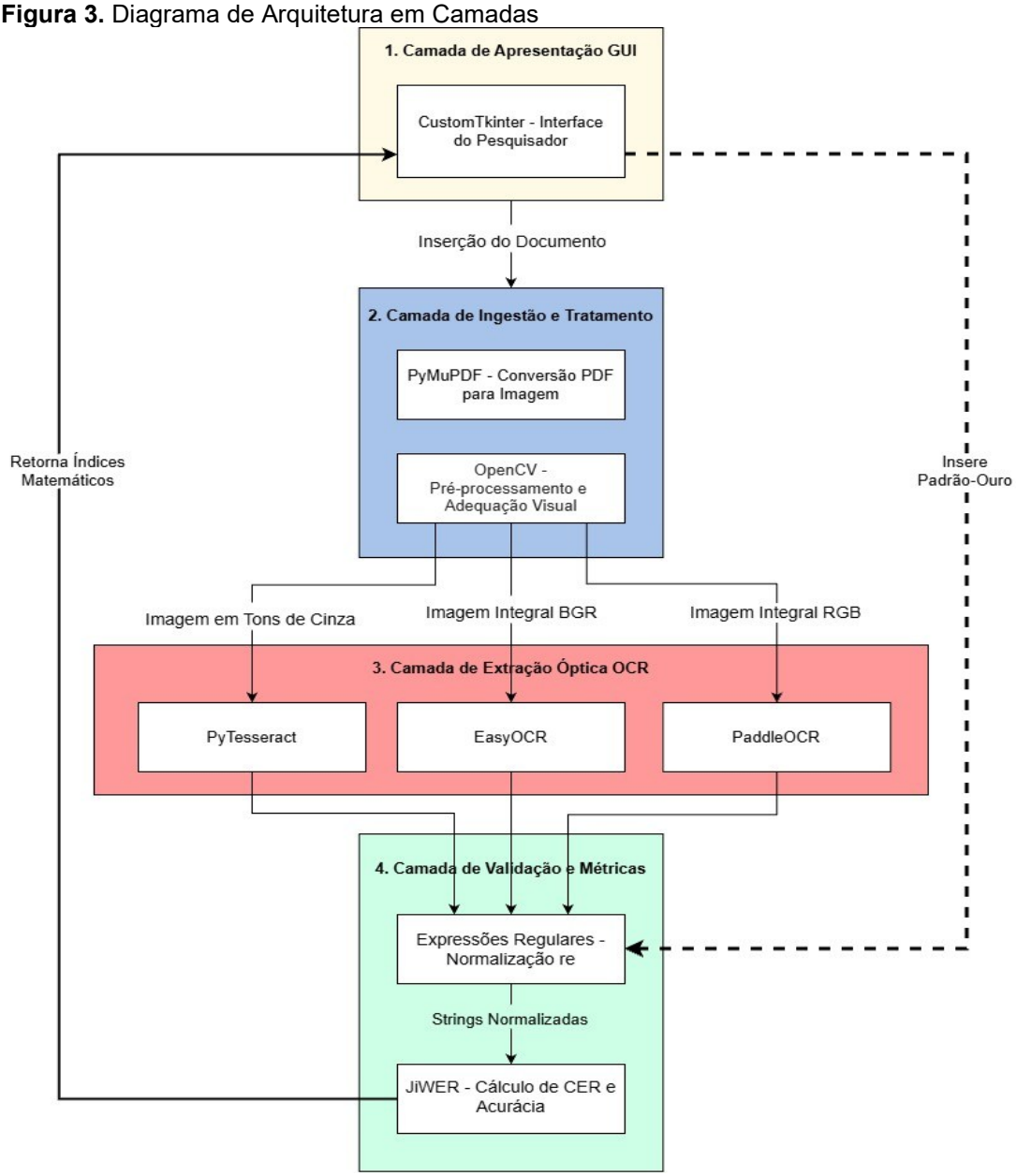
b) Camada de ingestão e tratamento: módulo de preparação visual que utiliza o PyMuPDF para rasterizar as páginas de documentos em formato PDF. Conforme otimizado no ambiente experimental, a biblioteca OpenCV atua nesta fase de forma pontual na conversão de espaços de cor (como de RGB para BGR para o PaddleOCR) e na aplicação de filtro mediano e escala de cinza para o Tesseract. Para o EasyOCR, as matrizes visuais são repassadas mantendo o padrão RGB nativo da captura via NumPy, garantindo a adequação das imagens para cada arquitetura antes da leitura;

c) Camada de extração óptica (OCR): bloco de processamento central, desenhado para acionar, conforme a seleção do pesquisador, o PyTesseract, o EasyOCR ou o ecossistema do PaddleOCR. É nesta camada que os motores utilizam seus respectivos algoritmos clássicos ou redes neurais para transformar a imagem tratada em texto pesquisável (predição) de forma autônoma;

d) Camada de validação e métricas: etapa em que ocorre o rigor analítico. Utiliza a biblioteca de Expressões Regulares (re) para a sanitização transversal e remoção de caracteres invisíveis/nulos do padrão-ouro e dos *outputs*, submetendo as *strings* tratadas ao pacote JiWER, que calcula o CER e retorna os dados de desempenho validados para o painel de apresentação.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 3. Diagrama de Arquitetura em Camadas



Fonte: Elaborado pelo Autor (2026).

4. RESULTADOS E DISCUSSÃO

Conforme a arquitetura de software estabelecida na metodologia, o desenvolvimento do ambiente experimental resultou em uma aplicação funcional. Nas subseções a seguir, são apresentadas a interface do sistema, a discussão dos resultados e as métricas de acurácia observadas.

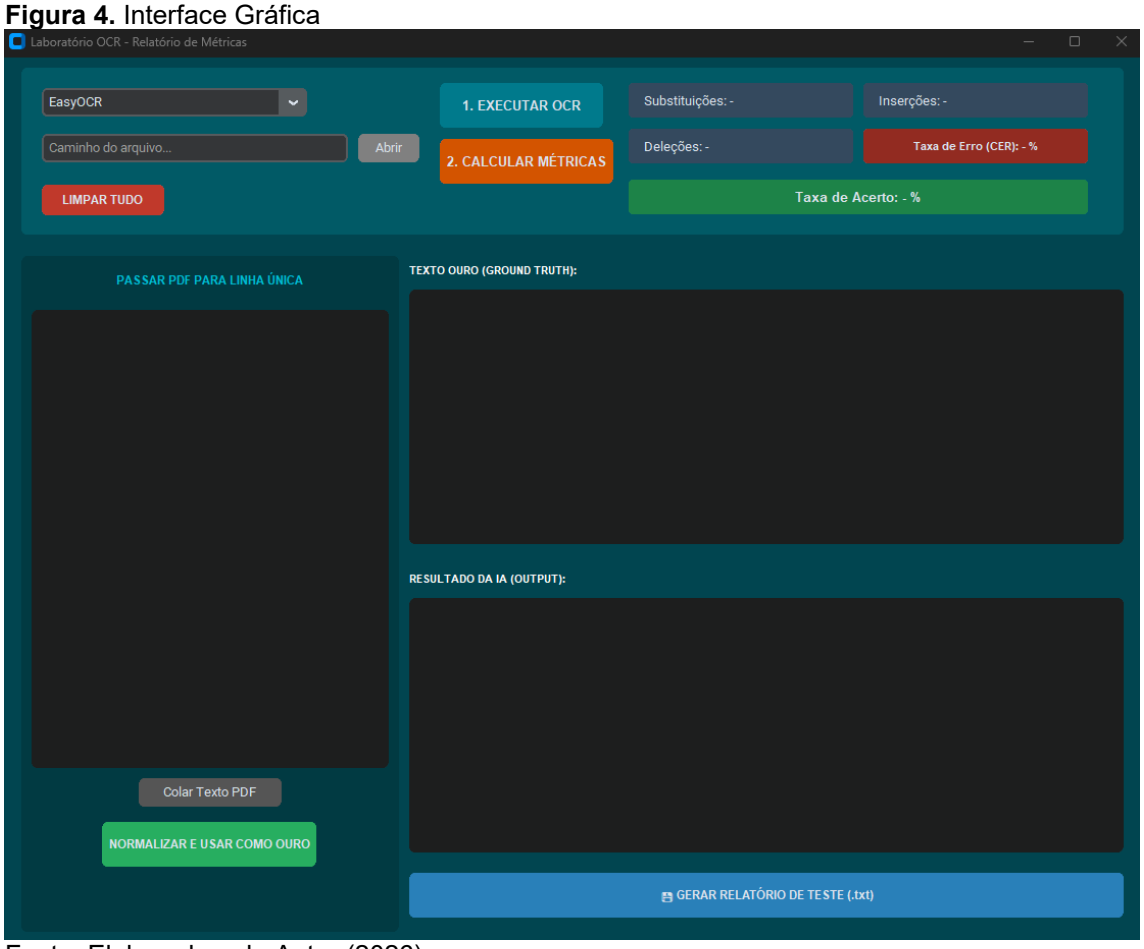
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

4.1. Interface Gráfica de Usuário (Pesquisador)

A camada visual, integralmente concebida por meio da biblioteca CustomTkinter, atua como o painel de controle prático do experimento. A Figura 4 apresenta a interface gráfica finalizada, na qual todos os testes empíricos foram conduzidos.

Observa-se no ambiente a consolidação de recursos essenciais para a agilidade da pesquisa, viabilizando a seleção dinâmica dos modelos de inteligência artificial e a inserção automatizada dos caminhos de arquivos. O destaque operacional da ferramenta consiste em sua capacidade analítica: o sistema dispõe, de forma comparativa e em tempo real, o texto de referência (Padrão-Ouro) e o resultado transcrito pelas ferramentas de reconhecimento óptico, fornecendo simultaneamente o cálculo exato das métricas de acurácia e CER. Foi por meio deste ambiente estruturado que os documentos foram submetidos aos motores de OCR, gerando os resultados quantitativos.

Figura 4. Interface Gráfica



Fonte: Elaborado pelo Autor (2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

4.2. Cenários de Teste

Para viabilizar a observação empírica dentro do recorte proposto, delineou-se um *corpus* de teste composto por uma amostragem intencional de nove documentos, subdivididos em três categorias de complexidade (três arquivos por classe, com amostras ilustradas na Figura 5). A seleção deste acervo fundamentou-se na representatividade tipológica de formatos recorrentes nos protocolos municipais, visando simular especificidades de legibilidade e estruturação arquivística. Para assegurar o cumprimento da Lei Geral de Proteção de Dados (LGPD) — dispensando a submissão ao Comitê de Ética —, optou-se pela utilização de documentos administrativos simulados e arquivos manuscritos redigidos especificamente para o experimento. A amostra foi delimitada conforme a seguinte estrutura:

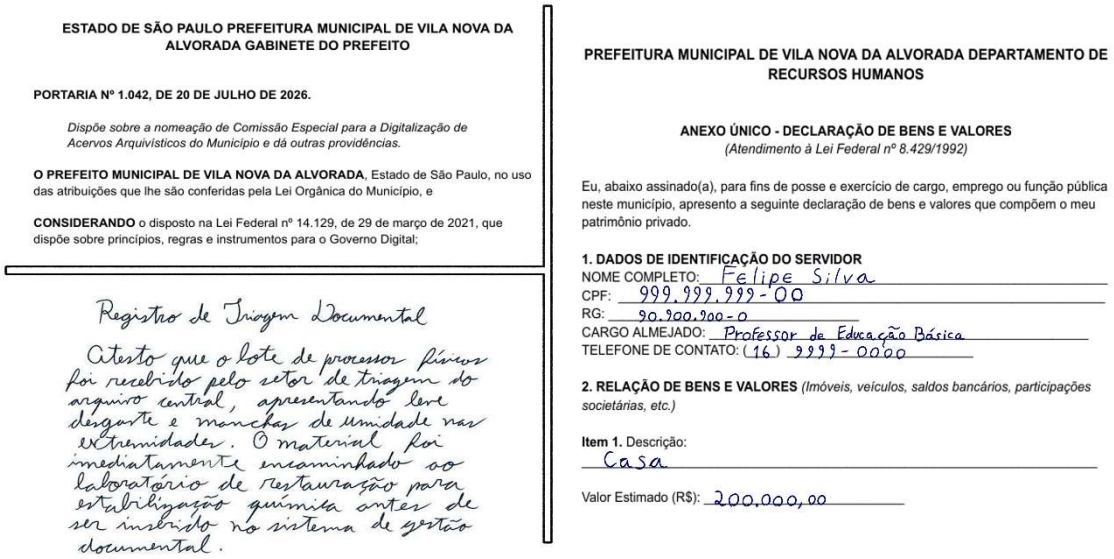
a) Categoria 1 (Nato-Digital): composta por três documentos administrativos tipográficos criados e exportados diretamente no formato PDF sendo duas Portarias e um Memorando Circular. Esta categoria atua como linha de base (*baseline*) ideal, testando a capacidade dos motores em reconhecer caracteres impressos com alta nitidez, diferenciar numerações de matrículas e leis, sem interferências de ruídos visuais;

b) Categoria 2 (Híbrido / Formulários Administrativos): composta por três formulários padronizados de rotinas do setor público e Recursos Humanos (Declaração de Bens e Valores, Requerimento Padrão de Férias e Termo de Recebimento de Cartão Vale-Alimentação). Caracteriza-se por apresentar *layouts* com caixas de seleção (*checkboxes*), lacunas destinadas ao preenchimento manual em letra de forma e assinaturas, exigindo que as inteligências artificiais processem a transição entre o texto tipográfico impresso e os elementos de marcação;

c) Categoria 3 (Manuscrito): composta por três documentos estritamente manuscritos com letra cursiva, redigidos para este trabalho e constituindo o nível máximo de complexidade do experimento, testando a inteligência artificial diante de caligrafias cursivas, variação de caracteres e densidade de escrita.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 5. Amostras das tipologias documentais submetidas aos motores de OCR



ESTADO DE SÃO PAULO PREFEITURA MUNICIPAL DE VILA NOVA DA ALVORADA GABINETE DO PREFEITO

PORTARIA Nº 1.042, DE 20 DE JULHO DE 2026.

Dispõe sobre a nomeação de Comissão Especial para a Digitalização de Acervos Arquivísticos do Município e dá outras providências.

O PREFEITO MUNICIPAL DE VILA NOVA DA ALVORADA, Estado de São Paulo, no uso das atribuições que lhe são conferidas pela Lei Orgânica do Município, e

CONSIDERANDO o disposto na Lei Federal nº 14.129, de 29 de março de 2021, que dispõe sobre princípios, regras e instrumentos para o Governo Digital;

Registro de Triagem Documental

Atesto que o lote de processos físicos foi recebido pelo setor de triagem do arquivo central, apresentando leve desgaste e manchas de umidade nas extremidades. O material foi imediatamente encaminhado ao laboratório de restauração para estabilização química antes de ser inserido no sistema de gestão documental.

PREFEITURA MUNICIPAL DE VILA NOVA DA ALVORADA DEPARTAMENTO DE RECURSOS HUMANOS

ANEXO ÚNICO - DECLARAÇÃO DE BENS E VALORES
(Atendimento à Lei Federal nº 8.429/1992)

Eu, abaixo assinado(a), para fins de posse e exercício de cargo, emprego ou função pública neste município, apresento a seguinte declaração de bens e valores que compõem o meu patrimônio privado.

1. DADOS DE IDENTIFICAÇÃO DO SERVIDOR
NOME COMPLETO: Felipe Silva
CPF: 999.999.999-00
RG: 90.900.900-0
CARGO ALMEJADO: Professor de Educação Básica
TELEFONE DE CONTATO: (16) 9999-0000

2. RELAÇÃO DE BENS E VALORES *(Imóveis, veículos, saldos bancários, participações societárias, etc.)*

Item 1. Descrição:
Casa

Valor Estimado (R$): 200.000,00

Fonte: Elaborado pelo Autor (2026).

4.3. Análise e Observação

Esta seção apresenta os dados observados nas baterias de teste, utilizando as taxas de acerto e a contagem de edições operacionais como indicadores de suporte para diagnosticar as características de cada motor no cenário avaliado. Sob a perspectiva da gestão pública, a taxa de acerto (representada pela acurácia) evidencia o nível de recuperação textual alcançado em cada tipologia específica, permitindo identificar as condições em que cada ferramenta apresenta maior utilidade.

4.3.1. Comportamento nos Documentos Nato-Digitais

A submissão do primeiro cenário de testes (Categoria Nato-Digital) aos motores avaliados revelou um contraste significativo entre as arquiteturas. Conforme a Tabela 1, o Tesseract destacou-se nos textos estruturados e de tipografia padrão, impulsionado por sua varredura linear por contraste e dicionários, ideal para acervos nato-digitais. Em contrapartida, as soluções de *Deep Learning* focadas em *Scene Text Recognition* apresentaram ligeira degradação: o EasyOCR demonstrou dificuldade com microcaracteres e pontuações (confundindo 'º' por 'ª' e vírgulas por ponto e vírgula), enquanto o PaddleOCR evidenciou limitações no suporte ao português, falhando em diacríticos e cedilhas ao converter sufixos como 'ção' em 'gäo'.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Tabela 1. Dados de Acurácia na Categoria Nato-Digital

Testes	Métricas	Tesseract	EasyOCR	PaddleOCR
Documento 1	Acurácia (%)	99,61%	96,95%	95,72%
	Substituições	6	26	63
	Inserções	0	6	0
	Deleções	0	15	3
Documento 2	Acurácia (%)	100%	98,96%	98,03%
	Substituições	0	6	15
	Inserções	0	0	0
	Deleções	0	3	2
Documento 3	Acurácia (%)	99,43%	95,37%	96,60%
	Substituições	5	17	31
	Inserções	1	11	0
	Deleções	0	21	5

Fonte: Elaborado pelo Autor (2026).

4.3.2. Comportamento nos Documentos Híbridos

O segundo cenário, demonstrado na Tabela 2, evidenciou as vulnerabilidades arquiteturais dos motores diante de formulários, mesmo com o preenchimento manual padronizado em letra de forma. O Tesseract demonstrou ineficácia severa ao tentar enquadrar a escrita humana em seu dicionário visual, gerando alucinações de caracteres e convertendo nomes em sequências alfanuméricas caóticas. Já o PaddleOCR falhou primariamente no seu algoritmo de segmentação espacial, aglutinando palavras impressas, suprimindo espaços em branco e ignorando blocos inteiros de preenchimento manual. O EasyOCR, por sua vez, conseguiu decifrar a semântica da letra de forma com precisão, mas falhou na interpretação do *layout* ao processar as linhas das lacunas como caracteres válidos e perder a ordenação estrutural da página. Fica evidente, portanto, que o desafio tecnológico em acervos híbridos transcende a leitura caligráfica, exigindo motores capazes de isolar espacialmente o texto tipográfico, a inserção humana e os elementos gráficos estruturais do formulário.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Tabela 2. Dados de Acurácia na Categoria Texto Híbrido

Testes	Métricas	Tesseract	EasyOCR	PaddleOCR
Documento 1	Acurácia (%)	93,15%	92,90%	82,08%
	Substituições	47	40	62
	Inserções	22	16	10
	Deleções	14	30	145
Documento 2	Acurácia (%)	90,77%	91,32%	83,88%
	Substituições	23	26	47
	Inserções	12	7	5
	Deleções	32	30	65
Documento 3	Acurácia (%)	91,47%	88,53%	87,65%
	Substituições	31	43	37
	Inserções	21	13	6
	Deleções	6	22	41

Fonte: Elaborado pelo Autor (2026).

4.3.3. Comportamento nos Documentos Manuscritos

O terceiro cenário (documentos manuscritos em letra cursiva) simulou a realidade dos órgãos públicos brasileiros, nos quais requerimentos feitos a próprio punho são recorrentes. Diante dessa elevada complexidade, os resultados evidenciaram um colapso generalizado nos motores avaliados. O Tesseract apresentou a falha mais crítica ao suprimir quase a totalidade das sentenças e retornar apenas fragmentos residuais. O PaddleOCR demonstrou incapacidade estrutural de segmentar o fluxo contínuo dos traços manuais, omitindo blocos inteiros de texto. Já o EasyOCR, embora tenha processado a morfologia da escrita, gerou alucinações severas, convertendo o conteúdo em sequências ilegíveis de caracteres e símbolos aleatórios. Atesta-se que o OCR tradicional é inviável para manuscritos cursivos, evidenciando que a realidade documental brasileira exige tecnologias focadas no reconhecimento de escrita manual (HTR — *Handwritten Text Recognition*).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Tabela 3. Dados de Acurácia na Categoria Texto Manuscrito

Testes	Métricas	Tesseract	EasyOCR	PaddleOCR
Documento 1	Acurácia (%)	32,83%	39,14%	33,33%
	Substituições	38	131	38
	Inserções	1	30	2
	Deleções	227	80	224
Documento 2	Acurácia (%)	9,03%	36,34%	39,90%
	Substituições	7	161	46
	Inserções	0	8	2
	Deleções	376	99	205
Documento 3	Acurácia (%)	20,47%	33,65%	27,06%
	Substituições	55	105	38
	Inserções	0	18	0
	Deleções	283	159	272

Fonte: Elaborado pelo Autor (2026).

4.3.4. Consolidação Geral das Métricas

Para fins de consolidação geral, as médias desta seção foram calculadas de forma aritmética simples entre os resultados percentuais, visando avaliar o comportamento médio da ferramenta por tipologia observada, atribuindo peso igualitário aos diferentes desafios de *layout* e caligrafia. Sob essa ótica, o Tesseract atingiu uma média de acurácia de 99,68% na categoria Nato-Digital, apresentando um ligeiro declínio para 91,80% nos documentos híbridos e caindo drasticamente para 20,77% nos manuscritos. O EasyOCR e o PaddleOCR demonstraram perfis de resiliência semelhantes na estabilidade média cruzada: no cenário nato-digital alcançaram 97,09% e 96,78%, respectivamente; nos formulários híbridos registraram 90,92% e 84,54%; e, nos manuscritos, atingiram 36,38% e 33,43%. Este contraste evidencia que, embora as soluções baseadas em *Deep Learning* apresentem maior versatilidade estrutural, nenhuma das ferramentas avaliadas superou a barreira dos 40% na leitura de escrita cursiva livre.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

5. CONSIDERAÇÕES FINAIS

A partir da observação do comportamento dos algoritmos nos cenários analisados, conclui-se que a adoção de motores de OCR no ambiente administrativo exige o mapeamento prévio das características e vocações de cada ferramenta. Soluções como EasyOCR, PaddleOCR e Tesseract demonstraram utilidade prática para a estruturação de buscas e extração de informações, variando sua adequação conforme a natureza e a qualidade do documento.

Neste estudo de caso específico, o Tesseract evidenciou elevada estabilidade e precisão em documentos nato-digitais padronizados, ao passo que o EasyOCR e o PaddleOCR demonstraram maior adaptabilidade em *layouts* variáveis e formulários híbridos preenchidos em letra de forma. Na leitura de manuscritos em escrita cursiva, todas as ferramentas analisadas apresentaram limitações significativas, apontando para a conveniência de soluções voltadas especificamente ao reconhecimento de escrita manual (HTR) para esse tipo de suporte.

Sob a ótica da implementação, a escolha da ferramenta deve ponderar a relação entre a taxa de acerto requerida, o tempo de processamento e a infraestrutura computacional disponível. O Tesseract destaca-se pela extrema leveza operacional em ambientes restritos (processamento exclusivo em CPU), enquanto os ecossistemas baseados em *Deep Learning* (EasyOCR e PaddleOCR) oferecem recursos avançados de análise visual que demandam, contudo, suporte de *hardware* com aceleração gráfica (GPU) para fluxos de trabalho contínuos.

Salienta-se que as evidências reunidas neste trabalho derivam de um estudo de caso exploratório focado em uma amostragem delimitada de nove documentos. Como perspectiva de continuidade e desdobramento desta pesquisa, recomenda-se a realização de trabalhos futuros contemplando a submissão massiva e automatizada de grandes volumes documentais, viabilizando a mensuração de parâmetros estatísticos globais de desempenho e a consolidação de modelos preditivos em larga escala para a gestão pública.

REFERÊNCIAS

BRASIL. Decreto nº 10.278, de 18 de março de 2020. Regulamenta o disposto no inciso X do *caput* do art. 3º da Lei nº 13.874, de 2019, e no art. 2º-A da Lei nº 12.682, de 2012, para estabelecer a técnica e os requisitos para a digitalização de documentos públicos ou privados, a fim de que os documentos digitalizados produzam os mesmos efeitos legais dos documentos originais. **Diário Oficial da União**, Brasília, DF, 19 mar. 2020. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/decreto/d10278.htm. Acesso em: 20 jun. 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal. **Diário Oficial da União**, Brasília, DF, 18 nov. 2011. Edição extra. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 20 jun. 2026.

BRASIL. Lei nº 12.682, de 9 de julho de 2012. Dispõe sobre a elaboração e o arquivamento de documentos em meios eletromagnéticos. **Diário Oficial da União**, Brasília, DF, 10 jul. 2012. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12682.htm. Acesso em: 20 jun. 2026.

BRASIL. Lei nº 13.874, de 20 de setembro de 2019. Institui a Declaração de Direitos de Liberdade Econômica; estabelece garantias de livre mercado e dá outras providências. **Diário Oficial da União**, Brasília, DF, 20 set. 2019. Edição extra. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/lei/l13874.htm. Acesso em: 20 jun. 2026.

BRASIL. Lei nº 14.129, de 29 de março de 2021. Dispõe sobre princípios, regras e instrumentos para o Governo Digital e para o aumento da eficiência pública. **Diário Oficial da União**, Brasília, DF, 30 mar. 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14129.htm. Acesso em: 20 jun. 2026.

DU, Yuning *et al*. **PP-OCR: A Practical Ultra Lightweight OCR System**. arXiv preprint, arXiv:2009.09941, 2020. Disponível em: https://arxiv.org/abs/2009.09941. Acesso em: 28 ago. 2026.

FONSECA, Maria Odila Kahl. **Arquivologia e Ciência da Informação**. Rio de Janeiro: FGV Editora, 2005.

JAIDED AI. **EasyOCR**: Ready-to-use OCR with 80+ supported languages. 2020. Disponível em: https://github.com/JaidedAI/EasyOCR. Acesso em: 11 abr. 2026.

PADDLEPADDLE. **PaddleOCR**: Awesome Multilingual OCR Toolkits. [S. l.], 2026. Disponível em: https://github.com/PaddlePaddle/PaddleOCR. Acesso em: 28 ago. 2026.

SCHÄFER, Murilo Billig; LIMA, Eliseu dos Santos. A classificação e a avaliação de documentos: análise de sua aplicação em um sistema de gestão de documentos arquivísticos digitais. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 20, n. 4, p. 111-131, out./dez. 2015.

SILVA, Luiz Carlos da. **Desafios da implementação da política pública de arquivos:** as funções arquivísticas e o acesso à informação na Prefeitura Municipal de Vitória, ES. 2020. 134 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2020.

SMITH, Ray. An Overview of the Tesseract OCR Engine. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION (ICDAR), 9., 2007, Curitiba. **Proceedings...** Curitiba: IEEE, 2007. p. 629-633.

TESSERACT-OCR. **Tesseract OCR User Manual**. [S. l.], 2026. Disponível em: https://tesseract-ocr.github.io/. Acesso em: 11 abr. 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.