# Detecção de Documentos Acadêmicos Falsificados: Uma Solução Baseada em Aprendizado de Máquina

**Samuel M. Ransolin**[1], **Giovana N. Inocêncio**[1], **Jean E. Martina**[1]

[1]Departamento de Informática e Estatística – Centro Tecnológico
Universidade Federal do Santa Catarina (UFSC) – Florianópolis, SC – Brasil

samuransolin@gmail.com, gioinocencio017@gmail.com, jean.martina@ufsc.br

***Abstract.*** *In recent years in Brazil, the growth in entrants, graduates, and higher education institutions has intensified challenges in validating academic credentials, since verification remains largely manual, error-prone, and vulnerable to fraud. This article proposes a hybrid prototype that combines multimodal analysis, clustering, anomaly detection, and graded classification to assign a legitimacy score to academic documents. By integrating the prototype into Rede de Rastreabilidade de Dados da Educação Superior – the official initiative for tracking academic records and certifications –, documents can be automatically validated before being recorded in its distributed network, increasing the security and reliability of accreditation.*

***Resumo.*** *Nos últimos anos, no Brasil, o crescimento de ingressantes, de formandos e de instituições de ensino superior intensificou os desafios relacionados à validação de certificados acadêmicos, já que a verificação é majoritariamente manual, sujeita a erros e a aceitação de fraudes. Este trabalho propõe um protótipo híbrido que combina análise multimodal, clustering, detecção de anomalias e classificação por grau de legitimidade para detectar documentos acadêmicos falsificados. Ao integrar o protótipo à Rede de Rastreabilidade de Dados da Educação Superior – iniciativa oficial para o acompanhamento de registros e comprovações acadêmicas –, documentos podem ser validados automaticamente antes do registro em sua rede distribuída, aumentando a segurança e a confiabilidade do credenciamento.*

## 1. Introdução

Ao longo da última década, observa-se no Brasil um crescimento contínuo na emissão de diplomas de ensino superior, com um aumento superior a 31% de formandos desde 2013 [INEP 2024]. Isso traz à tona uma série de desafios a serem superados, entre eles a temática explorada neste estudo: a melhoria nos processos de regulação, supervisão e avaliação dessas emissões por parte do Ministério da Educação do Brasil (MEC).

Atualmente, a gerência, armazenamento e emissão de documentos acadêmicos, como diplomas e históricos escolares, é responsabilidade da instituição de ensino que os emite. O processo, burocrático e não computadorizado, é suscetível a erros humanos em verificações manuais, à ação de funcionários desonestos que podem forjar ou vender certificados, e à perda definitiva de registros por desastres ou encerramento de instituições – como ocorrido com a Universidade Gama Filho [Palma et al. 2019]. Essa falta de transparência e redundância cria brechas conhecidas e também utilizadas por

agentes mal-intencionados, que criam falsas instituições especializadas na venda de certificados contrafeitos [Dias and Leal 2022].

É neste cenário que a Rede de Rastreabilidade de Dados da Educação Superior (RRDES), iniciativa do MEC em parceria com o Ministério da Economia e diversas universidades federais, disponibiliza um sistema que permite que discentes acompanhem suas trajetórias estudantis junto ao acesso a seus documentos acadêmicos pertinentes. Além disso, o projeto também tem o potencial de tornar-se uma plataforma conjunta para a emissão e registro destes certificados e até mesmo dados regulatórios das instituições de ensino superior [RNP 2023]. Em consonância a essa iniciativa, o presente estudo aproveita a temática de inteligência artificial aplicada à educação e trata da implementação e validação de um protótipo de software que combina diferentes técnicas de aprendizado de máquina, capaz de identificar certificados falsos antes de sua inserção nesse ambiente.

## 2. Trabalhos Relacionados

A pesquisa sobre detecção de fraudes em documentos é ampla e o uso de inteligência artificial neste campo tem crescido significativamente, no entanto, a detecção de documentos acadêmicos falsificados permanece pouco explorada na literatura. Enquanto a detecção de fraudes foca em adulterações de arquivos originais, a de documentos falsificados busca identificar aqueles completamente forjados desde sua criação. Embora distintos, os métodos se sobrepõem, permitindo que este estudo aproveite referências de ambas as áreas.

No domínio, predominam estratégias de visão computacional, por exemplo: [Jaiswal et al. 2022] utilizam autoencoders convolucionais sobre imagens hiperespectrais para identificar incompatibilidades entre tintas; [Alameri et al. 2023] propõem abordagem não supervisionada usando correlações espectrais para gerar redes ponderadas e aplicar clustering; e [James et al. 2020] reformulam o problema como comparação de grafos via OCR para detectar manipulação de pixels. Alternativamente, outros métodos também são propostos, como o trabalho de [Boonkrong 2024], que utiliza funções hash para verificação preventiva de documentos previamente validados.

As abordagens mais robustas combinam múltiplas tecnologias. [Jain and Wigington 2019] demonstram eficácia da análise multimodal no dataset RVL-CDIP, combinando OCR, representações textuais (ULMFiT, FastText, n-grams) e visuais (VGG-16), alcançando alta acurácia. [Mohammed et al. 2024] aplicam K-means sobre features visuais de milhares de certificados para detecção de anomalias, identificando documentos distantes dos centroides com bons resultados.

## 3. Proposta

O objetivo do estudo é rotular documentos com base em um nível de probabilidade de falsificação. Para isso, é realizada a análise, extração e fusão multimodal de características visuais e textuais dos documentos, o que resulta em uma representação unificada e concisa de cada um. Emprega-se aprendizado não-supervisionado para agrupar essas representações de acordo com suas similaridades, assim, detectores de anomalias são utilizados para a classificação de novos documentos submetidos, que se dá através da avaliação do grau de desvio em relação aos grupos identificados. Finalmente, essa pontuação determina se o documento é classificado como ”Normal”ou ”Suspeito”.

A escolha dessa abordagem tem por base a premissa de que documentos falsificados apresentam inconsistências sutis, tornando-os atípicos em relação aos padrões estabelecidos por documentos legítimos, e assim são detectáveis através da análise multimodal das características extraídas de diversos contextos. Dessa forma, o processo completo consiste em duas etapas: treinamento dos modelos de referência e classificação de novos documentos.

## 3.1. Treinamento dos Modelos de Referência

A fase de treinamento utiliza um dataset com milhares de amostras cedidas pela UFSC, composto por históricos escolares digitalizados de diversas instituições de ensino e diplomas da universidade. Após coleta, realiza-se pré-processamento através de normalização de imagens e aplicação de OCR. Com o dataset formado, cada amostra passa pelo bloco de extração multimodal. Com base nas representações obtidas, emprega-se DBSCAN para identificar clusters de documentos similares, com seus parâmetros ($\epsilon$ e MinPts) otimizados por análise de k-distance graphs. Documentos marcados como outliers são descartados como ruído, restando apenas grupos de referência que estabelecem padrões de normalidade.

### 3.1.1. Extração Multimodal

O módulo de extração multimodal captura e combina características independentes e, no contexto deste estudo, complementares. Essa abordagem opera, em paralelo, três diferentes subprocessos de aprendizado profundo para a extração de features:

- Extração visual: utiliza visão computacional para extrair características de qualidade e consistência visual, incluindo textura, propriedades tipográficas, selos e padrões de cores;
- Extração textual: utiliza processamento de linguagem natural para extrair características linguísticas, como distribuição de termos e formatação de dados;
- Extração estrutural: semelhante à extração visual, no entanto extrai características ligadas à organização espacial e estrutural dos documentos, como a disposição geral dos elementos no documento.

Por fim, as características extraídas são normalizadas, submetidas a técnicas de redução dimensional e fundidas, o que resulta em uma representação completa, unificada e compacta de cada documento. Isso permite que o sistema detecte tanto fraudes grosseiras, como a presença de um selo ou logotipo claramente apócrifo, quanto inconsistências sutis presentes em contrafações bem elaboradas, como divergências estatísticas entre termos utilizados ou variações microtipográficas.

## 3.2. Classificação de Novos Documentos

O fluxo de classificação de um novo documento reutiliza o mesmo pipeline de pré-processamento e extração multimodal para garantir consistência na representação. O resultado é comparado contra todos os clusters de referência identificados na fase de treinamento. Calcula-se a distância mínima do documento aos clusters estabelecidos, utilizando o conceito de $\epsilon$-vizinhança do DBSCAN. Documentos cuja distância excede o limiar $\epsilon$ são categorizados como ”Suspeito”e encaminhados para revisão manual por especialistas do

domínio. Documentos próximos a algum cluster (distância $< \epsilon$) são classificados como ”Normal”e aceitos automaticamente.

## 4. Resultados Esperados

O desenvolvimento encontra-se em fase de implementação, com parametrização dos algoritmos em andamento. Como resultado do estudo, espera-se desenvolver um software prova-de-conceito capaz de distinguir documentos legítimos de falsificados, e que a arquitetura definida sirva como base para futuras expansões e comparação de métodos.

Em termos aplicados, espera-se aumentar a confiabilidade e segurança no registro, reduzindo a conferência manual para documentos legítimos, enquanto especialistas focam apenas em casos suspeitos, acelerando os processos de validação dentro da RR-DES. Como consequência, a solução visa fornecer apoio à tomada de decisões de órgãos reguladores, de forma a contribuir com a transparência e confiança no sistema educacional.

## Referências


Alameri, M., Mahmood, B., Ciylan, B., and Amged, A. (2023). Unsupervised forgery detection of documents: A network-inspired approach. *Electronics*, 12.

Boonkrong, S. (2024). Design of an academic document forgery detection system. *International Journal of Information Technology*, pages 1–13.

Dias, P. and Leal, A. (2022). Sites vendem diploma de curso superior para quem sequer pisou em sala de aula: ’documentação 100% original, emitida de dentro da universidade’, diz atendente. O Globo. Disponível em: *https://oglobo.globo.com/brasil/noticia/2022/11/sites-vendem-diploma-de-curso-superior-para-pessoas-que-nao-concluiram-ou-sequer-pisaram-em-uma-universidade.ghtml*. Acesso em: 05 abr. 2025.

INEP (2024). Censo da educação superior 2023: notas estatísticas.

Jain, R. and Wigington, C. (2019). Multimodal document image classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 71–77.

Jaiswal, G., Sharma, A., and Yadav, S. (2022). Deep feature extraction for document forgery detection with convolutional autoencoders. *Computers & Electrical Engineering*, 99:107770.

James, H., Gupta, O., and Raviv, D. (2020). Ocr graph features for manipulation detection in documents. *arXiv preprint arXiv:2009.05158*.

Mohammed, S., Nwobodo, L., and Ekene, N. (2024). Certificate fraud verification model using clustered-based classification approach. *Exploremática Journal of Innovative Engineering and Technology*, 5(1):60–72.

Palma, L. M., Vigil, M. A. G., Pereira, F. L., and Martina, J. E. (2019). Blockchain and smart contracts for higher education registry in brazil. *International Journal of Network Management*, 29.

RNP (2023). Blockchain da jornada acadêmica. Youtube. Disponível em: *https://www.youtube.com/watch?v=xqezMbjCeTM*. Acesso em: 13 mai. 2025.