UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Samuel Moreira Ransolin

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

Florianópolis
2025

Samuel Moreira Ransolin

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

Relatório de Trabalho de Conclusão de Curso 1 do Curso de Graduação em Ciências da Computação do Centro Tecnológico da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Ciências da Computação.
Orientadora: Giovana Nunes Inocêncio, M.a.

Florianópolis
2025

Samuel Moreira Ransolin

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

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Ciências da Computação" e aprovado em sua forma final pelo Curso de Graduação em Ciências da Computação.

Florianópolis, 11 de dezembro de 2025.

Banca Examinadora:

Giovana Nunes Inocêncio, M.a.
Universidade Federal de Santa Catarina

Prof. Jean Everson Martina, Dr.
Universidade Federal de Santa Catarina

Lucas Machado da Palma, M.e.
Universidade Federal de Santa Catarina

Gabriel Estevam de Oliveira, M.e.
Universidade Federal de Santa Catarina

RESUMO

Nos últimos anos, no Brasil, o expressivo aumento do número de ingressantes, formandos e de instituições de ensino superior, trouxe consigo desafios relacionados à validação da autenticidade de certificados acadêmicos — atualmente verificados de forma predominantemente manual, sujeita a erros e falhas, como a aceitação de documentos fraudulentos. Nesse contexto, surge a Jornada do Estudante, um sistema disponibilizado pelo Ministério da Educação (MEC), que oferece, entre outras funcionalidades, o acompanhamento de registros acadêmicos de estudantes por meio de uma rede distribuída, de forma a garantir que somente instituições de ensino digitalmente certificadas possam registrar créditos e certificações. O presente trabalho de conclusão de curso revisita o estado-da-arte em detecção via aprendizado de máquina de documentos falsificados, além de propor um protótipo de solução híbrida, que combina análise multimodal, *clustering*, detecção de anomalias e classificação de documentos de acordo com seu grau de legitimidade. Ao integrar esse sistema à Jornada do Estudante, é possível validar automaticamente os documentos antes de seu registro em *blockchain*, aumentando significativamente a segurança e a confiabilidade do processo de credenciamento.

Palavras-chave: segurança da informação, detecção de fraude, *machine learning*, *clustering*, detecção de anomalias, extração multimodal

ABSTRACT

In recent years, the significant increase in the number of higher education institutions, incoming students, and graduates in Brazil has brought challenges related to the validation of academic certificates' authenticity—currently performed predominantly manually and prone to errors and failures, such as the acceptance of fraudulent documents. In this context, the Jornada do Estudante, a system provided by the Ministry of Education (MEC), was introduced; among other features, it enables the tracking of students' academic records through a distributed network, ensuring that only digitally certified institutions can register credits and certifications. This undergraduate thesis revisits the state of the art in machine learning–based forgery detection for academic documents and proposes a hybrid prototype solution that combines multimodal analysis, clustering, anomaly detection, and document classification according to their degree of legitimacy. By integrating this system with the Jornada do Estudante, documents can be automatically validated before being recorded on blockchain, significantly enhancing the security and reliability of the credentialing process.

Keywords: information security, fraud detection, machine learning, clustering, anomaly detection, multimodal feature extraction

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No Brasil, entre 2013 e 2023, o número de matrículas de alunos na educação superior aumentou 36,2%, com uma média de crescimento anual de 3,2%. O número de concluintes acompanhou essa mesma tendência de crescimento, sendo que o ano de 2013 registrou cerca de 992 mil graduandos, enquanto 2023 terminou com mais de 1,3 milhões. Para acomodar essa demanda, existem 2580 instituições de ensino superior no país (BRASIL, 2024), das quais 87,8% são privadas. Essas estatísticas revelam um saldo extremamente positivo, mas também trazem à tona desafios que precisam ser superados.

Atualmente, a gerência, armazenamento e cuidado de documentos acadêmicos, como diplomas e históricos escolares, é responsabilidade da instituição de ensino que os emitem (MEC, 1978), além disso o próprio processo para a emissão desses documentos é burocrático, não computadorizado e sujeito a erros ou até mesmo fraudes, já que a validação desses atestados não possuem transparência ou redundância (Palma et al., 2019). Assim, a falta de modernização desses procedimentos deixam brechas que são conhecidas e utilizadas por instituições mal intencionadas — não é difícil encontrar portais de venda de diplomas falsos (Dias; Leal, 2022).

O comércio clandestino de diplomas falsos oferece certificados em diversas áreas e níveis, desde medicina até direito; desde a graduação até o doutorado, por valores que podem chegar a R$100.000, tornando essa prática altamente lucrativa e atraente para fraudadores (Palma et al., 2019). Investigações recentes demonstram que quadrilhas estruturadas conseguem emitir dezenas de milhares de documentos forjados, comercializados em sites especializados, com suposta publicação em diários oficiais (Fantástico, 2025). Para além da corrupção, esse tipo de fraude compromete a confiança pública nas instituições de ensino e no mercado de trabalho: indivíduos sem qualificação adequada podem assumir funções críticas, enquanto diplomas legítimos perdem valor diante da insegurança sobre sua autenticidade (Mohammed; Nwobodo; Ekene, 2024).

Neste cenário, o Ministério da Educação do Brasil (MEC), em parceria com o Ministério da Economia, disponibiliza e desenvolve o sistema da Jornada do Estudante junto a Universidade Federal de Santa Catarina (por meio do Laboratório Bridge e do Laboratório de Segurança em Computação), a Universidade Tecnológica Federal do Paraná e a Universidade Federal de Mato Grosso do Sul (MEC, 2022). Este sistema permite que estudantes acompanhem seus dados estudantis e disponibiliza documentos acadêmicos pertinentes à sua trajetória. Além disso, também existe uma iniciativa para que se torne uma plataforma conjunta para a emissão, além do registro, destes certificados, unificando diplomas, históricos escolares, currículos e até mesmo dados regulatórios das instituições de ensino superior (RNP, 2023).

Para armazenar estes dados, o sistema da Jornada do Estudante utiliza uma *blockchain Hyperledger Fabric*, que aproveita características como rastreabilidade às emissões e

descentralização da posse e imutabilidade dos registros. O projeto realiza o processamento dos dados em rede através de *smart contracts* e baseia-se em gestão de identidade forte, com certificados digitais que servem como base da identidade, de forma que somente entidades reconhecidas pelo projeto possam efetuar transações (Palma et al., 2019; RNP, 2023).

Ainda assim, hoje, a Jornada do Estudante não elimina o risco do registro de documentos falsificados, mas seu arcabouço permite o desenvolvimento de uma solução para este desafio. O presente Trabalho de Conclusão de Curso (TCC) trata da implementação e validação de um protótipo de *software* de inteligência artificial capaz de apontar documentos falsos antes de sua inserção neste ambiente. O algoritmo será baseado em uma abordagem híbrida de aprendizado de máquina não-supervisionado, que combina análise multimodal e técnicas de detecção de anomalias. Busca-se rotular documentos com base em um nível de probabilidade de fraudulência, para isso, utilizam-se extrações de características visuais (como textura, fonte, espaçamento, selos e assinaturas), textuais (como padrões linguísticos, formatação de números e distribuição de termos) e estruturais (como posição de campos, margens e tabelas), que serão refinadas conforme realização do TCC. Combinando essas *features* multidimensionais, é possível realizar o agrupamento dos documentos em *clusters* que representam padrões dominantes normais. Em sequência, modelos de detecção de anomalias são utilizados para a criação de detectores de referência a partir dos *clusters*, possibilitando a classificação de um novo documento submetido, em tempo real, através da avaliação do grau de desvio em relação aos padrões aprendidos — quanto maior o desvio e escore de anomalia, maior a probabilidade de que o documento seja falsificado. Finalmente, essa pontuação é mapeada para categorias discretas de suspeita, fornecendo um nível de probabilidade de fraude para cada inserção.

Ao integrar essa tecnologia à Jornada do Estudante, espera-se aprimorar o processo de registro e emissão de documentos acadêmicos no Brasil, apontando, em tempo real, a tentativa de inserção de um certificado fraudado na base de dados, garantindo segurança e confiabilidade a estes procedimentos.

1.1 OBJETIVOS

Esta seção apresenta o objetivo geral e objetivos específicos deste trabalho.

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver e integrar ao projeto da Jornada do Estudante um protótipo de software que categorize, com determinado nível de acurácia, certos documentos acadêmicos, como diplomas, históricos escolares e matrizes curriculares, por grau de probabilidade de fraudulência.

1.1.2 Objetivos Específicos

Para atingir o objetivo geral, os seguintes objetivos específicos devem ser cumpridos:

- Obter uma base de documentos acadêmicos para o treinamento e validação do *software*;
- Conforme necessário, projetar e desenvolver sistema para pré-processamento dos documentos, aplicando OCR e normalização dos documentos digitalizados para uniformizar formatos e qualidade;
- Projetar e desenvolver sistema para extração de características multimodais (visuais, textuais e estruturais);
- Projetar e desenvolver sistema para agrupamento dos documentos com base nas *features* extraídas, criando referenciais de comportamento padrão;
- Projetar e desenvolver sistema de detectores de anomalias que calculem escores de suspeita com base em métricas de similaridade adaptadas aos *clusters* e mapeá-los em categorias discretas de probabilidade de fraude;

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os conceitos teóricos necessários para a compreensão do presente trabalho.

2.1 APRENDIZADO DE MÁQUINA

O aprendizado de máquina é um subcampo da inteligência artificial que tem por objetivo o desenvolvimento de algoritmos capazes de aprender e tomar decisões a partir de um conjunto de dados, sem que seja necessária a programação explícita para essas tarefas específicas (Dietterich, 2003). Fundamentalmente, esses sistemas buscam aprender padrões em coleções de dados para, a partir da generalização desse conhecimento, realizar inferências sobre novas informações. Esse processo de aprendizado utiliza modelos matemáticos, principalmente estatísticos, que capturam relações complexas entre variáveis de entrada e saída através do ajuste de parâmetros internos, permitindo que a aplicação melhore seu desempenho conforme é exposta a mais dados (Sarker, 2021). Em geral, as técnicas de *machine learning* são categorizadas em três paradigmas principais: aprendizagem supervisionada, não supervisionada e por reforço.

O aprendizado supervisionado caracteriza-se pela utilização de conjuntos de dados rotulados, onde tanto as entradas quanto as saídas desejadas são conhecidas durante o treinamento. Nesse paradigma, o algoritmo aprende através de exemplos, de forma a possibilitar tarefas como classificação — a atribuição de classes discretas aos dados — e regressão — a predição de valores contínuos. Algoritmos clássicos dessa categoria incluem máquinas de vetores de suporte, redes neurais artificiais e métodos *ensemble* (Sarker, 2021). O aprendizado não-supervisionado, por sua vez, opera sobre dados não rotulados, sem o conhecimento das saídas desejadas, e busca compreender a organização natural de um dado conjunto a partir da identificação de padrões intrínsecos. Essa abordagem engloba técnicas como agrupamento (*clustering*) e detecção de anomalias (Sarker, 2021). Ainda, diferentes estratégias de aprendizado podem ser incorporadas, como algoritmos semi-supervisionados, utilizados quando um *dataset* tem poucos dados classificados, de forma a aproveitar a estrutura implícita do conjunto não categorizado para melhorar o desempenho do modelo (Sarker, 2021).

Adicionalmente, o aprendizado por reforço representa um paradigma distinto onde um modelo aprende através de interações com um ambiente, sendo recompensado ou penalizado com base em suas ações, de forma que gradualmente desenvolva estratégias ótimas. Essa abordagem é especialmente útil em áreas como teoria de jogos ou inteligência de enxame (Sarker, 2021).

Embora os conceitos fundamentais de aprendizado de máquina tenham sido estabelecidos há quase um século, com contribuições embrionárias nas décadas de 1950 e 1960 (Dietterich, 2003), essa área de estudo tem recebido grande destaque nas últimas décadas.

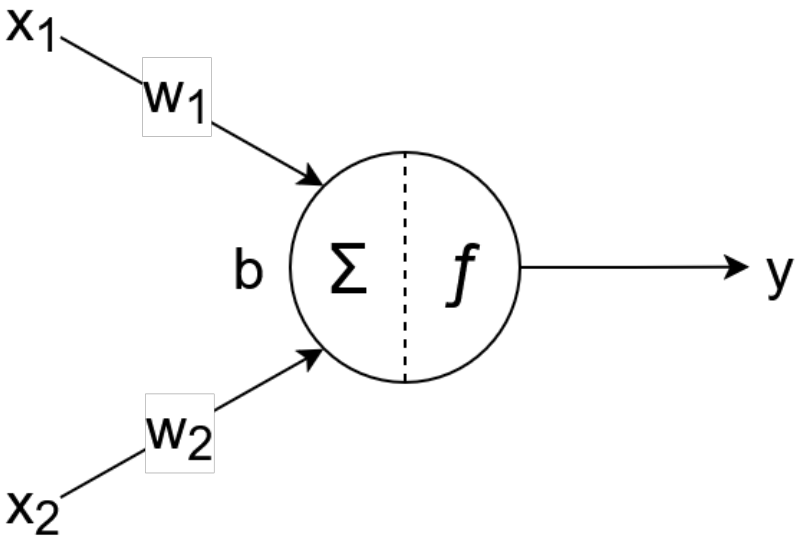
Esse ressurgimento deve-se principalmente ao aumento exponencial da capacidade computacional e a disponibilidade massiva de dados digitais. Também, a evolução do *hardware*, particularmente o advento de unidades de processamento gráfico de alto desempenho, possibilitou o treinamento de modelos complexos, antes computacionalmente intratáveis, transformando o aprendizado de máquina em uma tecnologia fundamental para aplicações modernas em diversas áreas (Sarker, 2021).

2.2 REDES NEURAIS PROFUNDAS

Redes neurais profundas, comumente utilizadas no paradigma de aprendizado supervisionado, são uma especialização de redes neurais artificiais. Diferentemente das técnicas tradicionais de *machine learning*, que requerem a engenharia manual de características, são capazes de autonomamente aprender representações complexas de um determinado conjunto de dados brutos. Isso é possível por sua estrutura multicamada, que permite a extração progressiva de características de baixo nível – em uma imagem, por exemplo, bordas e linhas – até padrões de alto nível – no mesmo exemplo, objetos e faces (Menghani, 2023).

A unidade mais básica de uma rede neural artificial é um neurônio artificial (que será referido aqui apenas como neurônio ou nó).

Figura 1 – Representação de um Neurônio Artificial



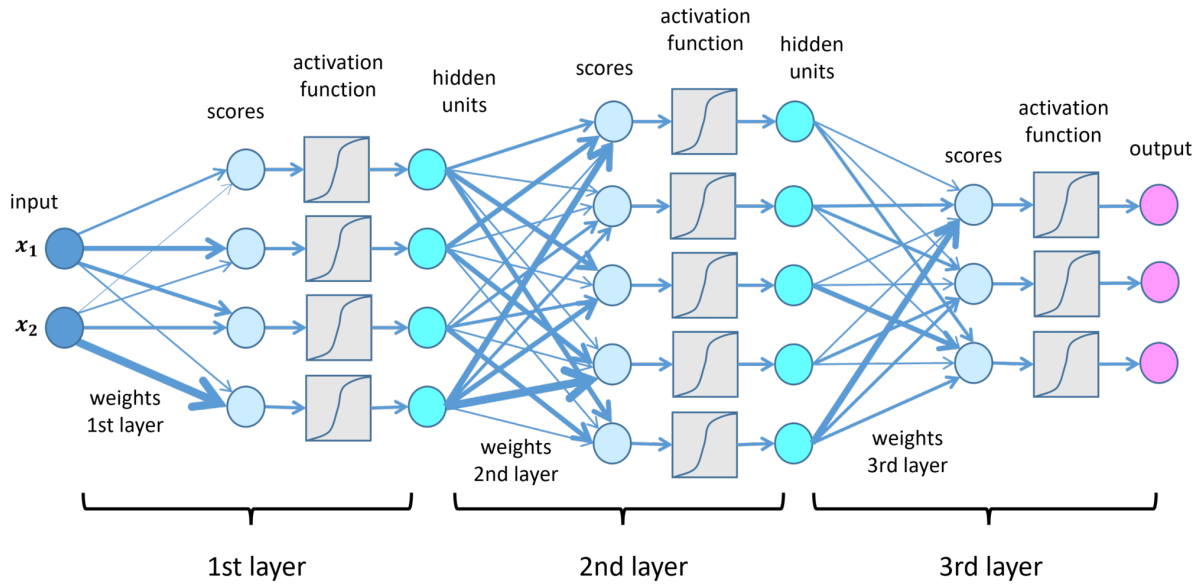
Fonte: o autor

No exemplo representado na figura 1, um neurônio recebe dados de entrada, x_1 e x_2, e produz uma saída y. Para isso, cada entrada é multiplicada por seu respectivo peso, w_1 e w_2, e somada junto a um termo de viés b – assim, a equação resultante é igual a

$\Sigma = x_1 * w_1 + x_2 * w_2 + b$. Finalmente, aplica-se uma função de ativação f sobre a soma para converter esse valor em um intervalo desejado – a tangente hiperbólica produziria um número dentro do intervalo $(-1, 1)$, enquanto a sigmoid produziria um intervalo entre $(0, 1)$, por exemplo – resultando na saída y (Goodfellow; Bengio; Courville, 2016). Em outras palavras, os pesos indicam a importância, ou força, da conexão entre a entrada e o neurônio; o viés atua como um limiar de ativação que independe das entradas; e a função de ativação transforma uma entrada linear em uma saída não linear, o que permite um mapeamento complexo entre entradas e saídas.

Uma rede neural artificial é formada pela interligação de neurônios, assim, uma camada da rede é denominada a partir de um grupo de nós interligados, que processam dados de uma maneira específica. Combinando uma camada de entrada, camadas intermediárias e uma camada de saída, obtém-se uma rede neural profunda, ilustrada na figura 2. Dessa forma, cada camada recebe entradas ponderadas a partir das camadas anteriores, aplicam uma função de ativação e propagam o resultado para as camadas subsequentes (Goodfellow; Bengio; Courville, 2016). Diferentes configurações destas, como a variação das conexões entre neurônios ou o emprego de funções de ativação distintas, tem por efeito especializações, ou habilidades de aprendizado específicas, assim, a utilização de múltiplas camadas permite a assimilação de representações hierárquicas complexas. (Alzubaidi et al., 2021).

Figura 2 – Representação de um Neurônio Artificial

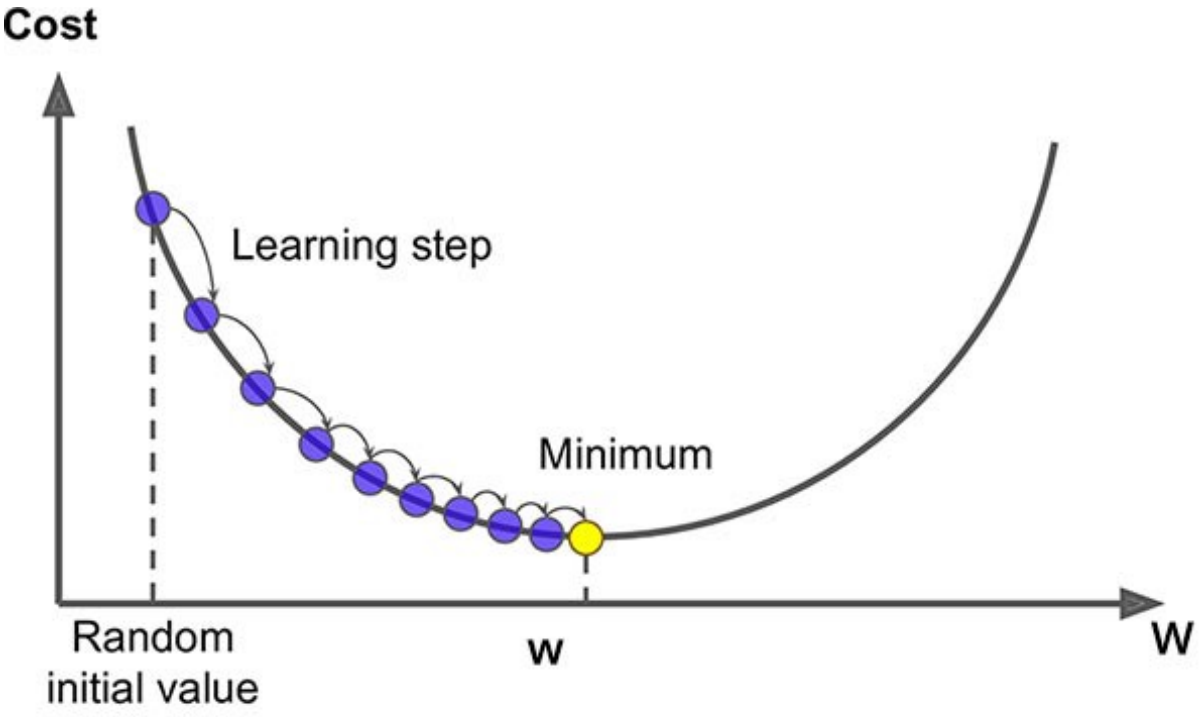


Fonte: `https://lamarr-institute.org/blog/deep-neural-networks/`. Acesso em: 25 junho 2025

O processo de aprendizado da rede é denominado treinamento e consiste na estimação e ajuste dos parâmetros através de um algoritmo de retropropagação. Seu objetivo é

calcular o gradiente de uma função que mede o erro entre o valor de saída computado e o esperado, além de ajustar os pesos e vieses dos neurônios na direção oposta ao gradiente, para minimizar o erro. Esse processo é executado em cada camada, propagando o erro desde a camada de saída até a de entrada, de forma iterativa, por múltiplas épocas, até que a rede converja para uma solução otimizada (Goodfellow; Bengio; Courville, 2016), como ilustrado na figura 3, em que o eixo y representa valores de erro e o eixo x valores de peso.

Figura 3 – Visualização do Gradiente Descendente



Fonte: `https://mlpills.dev/machine-learning/gradient-descent/`. Acesso em: 25 junho 2025

2.2.1 Redes Neurais Convolucionais

2.2.1.1 Reconhecimento Óptico de Caracteres

2.2.2 Processamento de Linguagem Natural

2.3 ANÁLISE MULTIMODAL

2.4 APRENDIZADO NÃO-SUPERVISIONADO

2.4.1 Algoritmos de Agrupamento

2.4.2 Detecção de Anomalias e Outliers

3 TRABALHOS CORRELATOS

As próximas seções apresentam os mais recentes trabalhos com temática ou abordagem semelhante.

3.1 VISÃO GERAL

É difícil encontrar bibliografia que trata especificamente do problema da identificação de documentos falsificados, sobretudo no âmbito acadêmico, já que a maioria dos trabalhos com temática similar aborda a classificação de fraudes. É importante fazer a distinção de que, enquanto a detecção de fraudes foca em adulterações de arquivos originais – como a mudança de notas, datas ou nomes –, a de documentos falsificados busca identificar aqueles completamente forjados desde sua criação. Isso não significa que as ideias e técnicas não possam ser aproveitadas e adaptadas entre um contexto e outro, pelo contrário, este trabalho de conclusão de curso tem como referência métodos nos dois domínios.

Nesta área, é predominante o emprego de estratégias de visão computacional, como o artigo de Jaiswal; Sharma; Yadav (2022), que utiliza *autoencoders* convolucionais para a extração de características em imagens hiperespectrais, focando em identificar incompatibilidade entre tintas. A análise de imagens é frequentemente combinada com técnicas complementares para melhorar a robustez da detecção: Alameri et al. (2023) propõem uma abordagem não supervisionada que utiliza correlações entre espectros de materiais dos documentos para gerar redes ponderadas, aplicando algoritmos de *clustering* para identificar padrões anômalos; James; Gupta; Raviv (2020) introduziram outra perspectiva ao reformular o problema como comparação de grafos, em que obtém, via *OCR*, caixas delimitadoras de tamanho entre caracteres, utilizando-as para o treinamento de classificadores que detectam a manipulação de *pixels*.

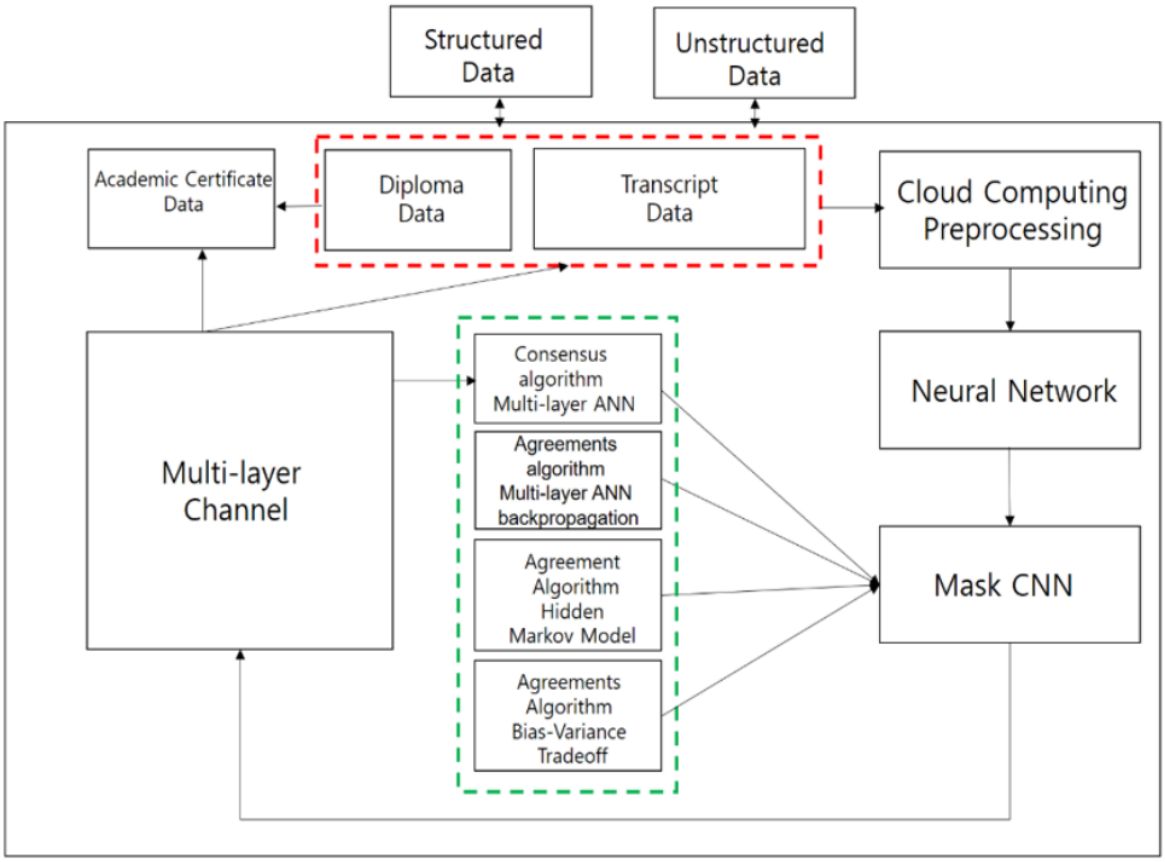
Alternativamente, também existem propostas que abordam a prevenção de fraudes através de outras tecnologias, como Boonkrong (2024), que propõe o emprego de funções criptográficas para detectar modificações em documentos previamente submetidos, em que são armazenados os valores de *hash* dos arquivos originais e legítimos, de forma que validações posteriores possam ser comparadas com o certificado primário. Contudo, essas abordagens preventivas não lidam com a classificação de documentos falsificados em sua concepção, representando uma lacuna pouco explorada, que o presente trabalho visa preencher. Em sequência seguem os trabalhos que guiaram a concepção da estratégia deste trabalho.

3.2 BLOCKCHAIN SMART CONTRACT TO PREVENT FORGERY OF DEGREE CERTIFICATES: ARTIFICIAL INTELLIGENCE CONSENSUS ALGORITHM

Para o problema de prevenção de fraudes de diplomas, o artigo de Kim (2022) propõe uma *blockchain* que incorpora algoritmos de aprendizado de máquina em diversas partes do processo de verificação de documentos e de consenso da rede. Além disso, de forma semelhante ao trabalho de Boonkrong (2024), quando um certificado é aceito, seu *hash* é calculado e integrado ao seu registro na rede, permitindo sua verificabilidade, de forma que qualquer adulteração seja facilmente detectada.

O autor apresenta o fluxo para submissão de um diploma na *blockchain* em quatro etapas principais, conforme figura 4.

Figura 4 – Representação da Arquitetura de Validação de Diplomas por Kim



Fonte: (Kim, 2022)

A primeira etapa, "*Cloud Computing Preprocessing*", consiste na entrada e pré-processamento do documento digitalizado e, através de um serviço em nuvem, normaliza os dados brutos da imagem, extrai metadados – como resolução, dimensões e contraste –, aplica correções de perspectiva e cria uma versão virtualizada semi-persistente do arquivo,

preparada para ser processada pelas etapas conseguintes, rejeitando entradas com formato e qualidade inconsistentes.

Em seguida, as imagens passam pela etapa "*Neural Network*", que utiliza *Faster R-CNN* para rapidamente filtrar e detectar artefatos visuais suspeitos. Através de sua rede de propostas regionais (*Region Proposal Network*), o algoritmo captura regiões de interesse – como selos da universidade, assinaturas, marcas d'água e outros padrões – e cria escores de confiança para cada uma, identificando áreas possivelmente adulteradas.

Os documentos passam então para a etapa "*Mask CNN*", que através de um modelo *Mask R-CNN* e das regiões de interesse previamente identificadas, segmenta a imagem, criando máscaras binárias a nível de *pixel*, ou seja, para cada região, o algoritmo estima quais *pixels* pertencem ao document legítimo e quais foram forjados. Essas máscaras são então encapsuladas como prova imutável dentro do bloco que será registrado na blockchain, além de servir como outra medida de escore de confiança do documento.

Por fim, ambas as pontuações de confiança são combinadas e passadas ao mecanismo de consenso dessa *blockchain*, representado pela etapa "*Multi-layer Channel*"que, ao invés de utilizar estratégias tradicionais como prova de trabalho ou prova de participação, combina múltiplos algoritmos de aprendizado de máquina. De forma geral, essa arquitetura é composta por quatro componentes principais:

- Um algoritmo baseado em uma rede neural multicamadas, que processa os escores de confiança fornecidos pelos processamentos anteriores e, a partir de certo limiar de confidência, imediatamente aprova o documento;
- Um algoritmo de aprendizado complementar à rede neural multicamadas, que cruza referências com padrões aprendidos de decisões anteriores – por exemplo, quando um diploma inicialmente validado como autêntico posteriormente se prova fraudulento – e ajusta os pesos da rede;
- Um algoritmo que utiliza um modelo oculto de Markov para, a partir de uma cadeia de regras, examinar padrões temporais e fornecer avaliações probabilísticas da autenticidade do documento;
- Um algoritmo que equilibra a complexidade (*overfitting*) com a generalização (*underfitting*) do modelo de rede neural, ou seja, procura encontrar *trade-off* ótimo entre viés e variância para decisões de consenso confiáveis.

Assim, cada nó da *blockchain* executa esses algoritmos e vota, com base na ponderação dos resultados, se devem incluir ou rejeitar o bloco com o diploma. Para qualquer decisão, exige-se quórum de pelo menos 2/3 de votos. Caso não seja atingido, seja por discordâncias das máscaras ou pela recusa de validadores de alta reputação, uma nova rodada de votação é iniciada com a reexecução das etapas "*Mask CNN*"e "*Multi-layer Channel*"com parâmetros ajustados. Esse ciclo se repete até obter consenso ou direcionar o

diploma a uma auditoria humana. Dessa forma, quando um documento é aprovado na rede, é classificado como autêntico no *ledger*; quando reprovado, é sinalizado como fraudulento.

Por fim, quando um terceiro – como empresa, universidade ou empregador – deseja verificar a validade de um diploma já registrado, basta a verificação do *hash* do documento já submetido.

3.3 MULTIMODAL DOCUMENT IMAGE CLASSIFICATION

O trabalho de Jain; Wigington (2019) não lida diretamente com a identificação de documentos falsificados ou fraudados, mas sim do problema geral de classificação de imagens. No entanto, a abordagem utilizada pelos autores é altamente relevante, pois mostra a eficácia da análise multimodal e pode ser aproveitada por este trabalho de conclusão de curso.

O *paper* propõe uma abordagem multimodal para a classificação de imagens de documentos diversos em dezesseis categorias utilizando o *dataset* RVL-CDIP. A proposta combina a fusão de características visuais e textuais para a rotulagem entre imagens e classes. Para isso, segue o *pipeline*:

1. Pré-processamento: normaliza e redimensiona as imagens para utilizações posteriores;
2. Extração de texto: utiliza OCR para extrair texto das páginas;
3. Extração multimodal, em paralelo:

   Modalidade textual: utiliza um modelos de linguagem para capturar informações semânticas dos textos extraídos;

   Modalidade visual: extrai características das imagens a partir de uma rede convolucional;
4. Fusão multimodal: combina as extrações textuais e visuais;
5. Classificação final: utiliza uma rede convolucional, que tem como entrada a fusão multimodal, para classificar os documentos.

Para a modalidade textual, os autores trazem à tona o problema de que texto extraído por OCR pode ser muito ruidoso, contendo erros a nível de caracteres ou até palavras. Por isso, capturam representações do conteúdo em três diferentes granularidades.

A nível de sequência, empregam ULMFiT (*Universal Language Model Fine-tuning*), um modelo que processa o documento como uma sequência de palavras e mantém uma "memória interna", que armazena informações contextuais conforme processa cada palavra sequencialmente. Isso permite que a rede neural capture sequências lógicas, dependências de longo prazo e contexto semântico entre palavras distantes. Como saída, o algoritmo

produz uma representação vetorial do texto que leva em consideração as características citadas.

A nível de palavra, para representar cada uma, empregam FastText *embeddings*. A técnica consiste em transformar os termos em vetores numéricos, de forma que palavras com significados similares fiquem próximas no espaço matemático. O vetor final do documento é calculado como a média dos *embeddings* de todas as palavras presentes.

A nível de caractere, para capturar padrões ortográficos, aplicam N-gramas de caracteres – sequências contínuas de n caracteres abduzidos de uma palavra – e criam um vetor numérico normalizado das ocorrências dos padrões obtidos.

Em resumo, as características de sequência preservam o contexto semântico geral do documento, as representações de palavra mantêm similaridades semânticas locais, e os N-gramas de caracteres oferecem robustez contra erros de OCR e palavras desconhecidas. Essas três representações são combinadas através de um método *ensemble*, que produz um vetor unificado que comporta essas *features* textuais.

Para a modalidade visual, o trabalho emprega a arquitetura de rede VGG-16 (*Visual Geometry Group*) para extrair características hierárquicas através de suas camadas convolucionais, capturando padrões de layout, tipografia, elementos gráficos e padrões de formatação presentes nos documentos. Como saída, a rede produz um vetor multidimensional que representa uma codificação densa e compacta de todas as informações visuais relevantes do documento.

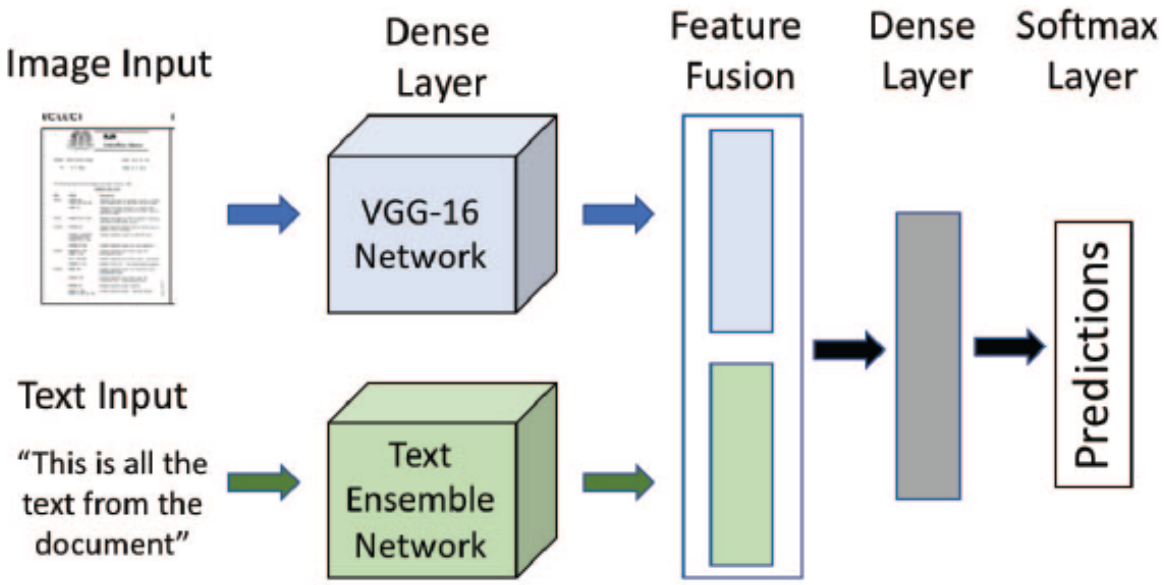
O principal interesse desse trabalho é a fusão das informações textuais e visuais, que tem por objetivo criar uma representação unificada, que preserve e potencialize as informações complementares de ambas as modalidades, e que permita que um modelo de classificação final explore sinergias entre essas diferentes características. Os autores propõem duas estratégias principais para combinar essas representações, das quais destaca-se a segunda, que combina os vetores anteriormente extraídos.

Essa abordagem explora quatro métodos distintos de junção. De forma geral, o primeiro é a concatenação simples, onde os vetores de características textuais e visuais são diretamente concatenados para formar um vetor unificado. O segundo método utiliza adição elemento a elemento, somando diretamente as representações de ambas as modalidades. O terceiro emprega *compact bilinear pooling*, uma técnica mais sofisticada que calcula o produto externo entre os vetores de características para capturar interações complexas entre as modalidades, permitindo que o modelo de classificação posterior aprenda correlações não-lineares entre informações visuais e textuais. Por fim, o quarto método implementa *multimodal gated units*, que utilizam mecanismos de atenção para aprender uma função de controle que determina automaticamente como ponderar e combinar as características de cada modalidade, o que permite que o modelo posterior adapte dinamicamente a importância relativa de informações visuais ou textuais dependendo do contexto específico do documento – para documentos altamente textuais como contratos ou rela-

tórios científicos, as características semânticas podem ser mais discriminativas, enquanto para documentos com layouts visuais distintivos como formulários ou apresentações, as características visuais podem ser mais relevantes.

Em sequência, para realizar a classificação final, os autores utilizam uma camada densa e uma camada final *softmax* para a predição, como ilustra a figura 5.

Figura 5 – Representação da Arquitetura de Fusão Multimodal por Jain; Wigington



Fonte: (Jain; Wigington, 2019)

Finalmente, o artigo compartilha as conclusões finais, onde explicita como essa abordagem superou consistentemente os métodos que utilizam apenas uma modalidade, com o método de adição elemento a elemento curiosamente alcançando os melhores resultados para a fusão multimodal, atingindo uma acurácia de $93,6\%$ no *dataset* RVL-CDIP.

3.4 CERTIFICATE FRAUD VERIFICATION MODEL USING CLUSTERED-BASED CLASSIFICATION APPROACH

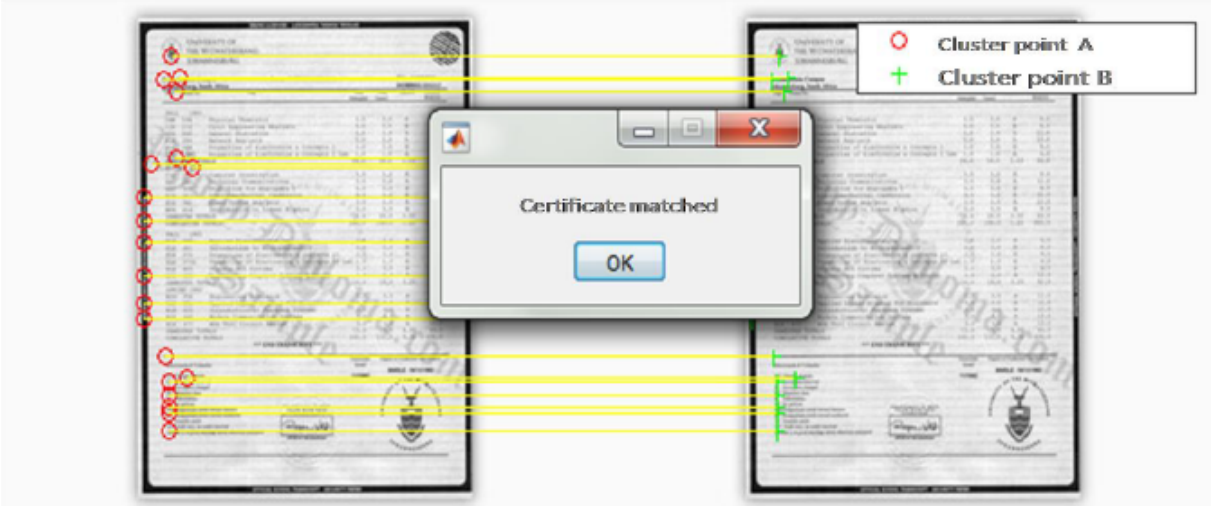
O trabalho de Mohammed; Nwobodo; Ekene (2024) lida diretamente com o problema de verificação da autenticidade de certificados acadêmicos. Para isso, propõe uma abordagem baseada em *clustering*, fundamentado na premissa de que documentos legítimos apresentam padrões consistentes de características que podem ser identificados através desses agrupamentos.

O *dataset* utilizado pelos autores consiste em mais de vinte e quatro mil amostras de documentos não rotulados, oficialmente emitidos por duas universidades: Enugu State University of Science and Technology e University of California Irvine.

A metodologia dos autores utiliza o algoritmo K-means como técnica principal para descobrir os padrões dominantes em certificados acadêmicos. Dessa forma, o processo de treinamento consiste na aplicação desse algoritmo sobre as características extraídas dos documentos do *dataset*, com o objetivo de agrupá-los em dezesseis grupos. O algoritmo inicializa centroides e atribui iterativamente cada arquivo ao centroide mais próximo usando um modelo de equidistância. Os centroides são atualizados através do cálculo das médias dos pontos pertencentes a cada *cluster* até atingir convergência. Para a verificação de documentos, o sistema extrai suas características e calcula distâncias em relação aos centroides estabelecidos, classificando-o como legítimo, quando próximo de algum padrão conhecido, ou suspeito, quando distante de todos os *clusters*.

Embora os autores não especifiquem os processos de extração de características, as figuras apresentadas sugerem fortemente o uso de *features* visuais. As imagens, como exemplificado na figura 6, mostram marcações circulares em pontos específicos dos certificados, destacando elementos como logos institucionais e aspectos estruturais dos documentos. Assume-se, portanto, que o sistema captura e converte essas informações em vetores numéricos compatíveis com o K-means.

Figura 6 – Resultado da Verificação de um Documento Autêntico



Fonte: (Mohammed; Nwobodo; Ekene, 2024)

O modelo foi validado experimentalmente com certificados reais e demonstrou capacidade de distinguir documentos autênticos de falsificados, já que os resultados reportaram uma acurácia geral de $86,53\%$. Os autores destacam como principal vantagem a capacidade do sistema de operar efetivamente com *datasets* limitados, característica importante em cenários reais onde a disponibilidade de dados rotulados é restrita por questões éticas e de privacidade.

4 METODOLOGIA PROPOSTA

Este capítulo define, de forma preliminar, a metodologia que será seguida durante o desenvolvimento do trabalho e está sujeita a mudanças conforme coleta e análise dos documentos acadêmicos.

4.1 VISÃO GERAL DA METODOLOGIA

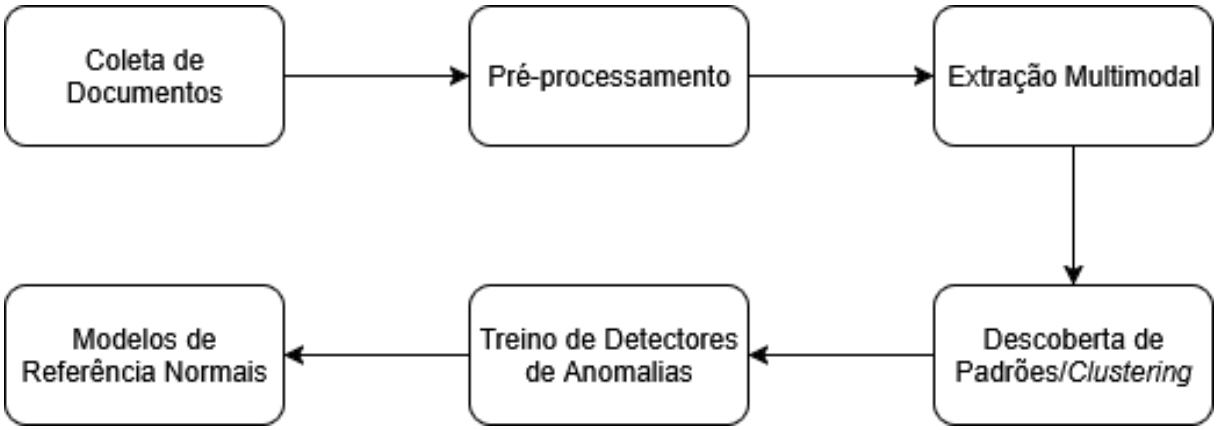
A metodologia proposta tem como base uma abordagem de aprendizado não-supervisionado e detecção de anomalias através da análise e extração multimodal. O processo completo consiste em duas etapas: treinamento dos modelos de referência e classificação de novos documentos.

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, sendo detectáveis através da análise multimodal das características extraídas de diversos contextos.

4.1.1 Treinamento dos Modelos de Referência

Representada na figura 7, a fase de treinamento inicia com a coleta de certificações acadêmicas diversas, seguida do pré-processamento através de técnicas de normalização de imagens e aplicação de OCR. Com o *dataset* formado, é realizada a extração e processamento multimodal de características visuais, textuais e estruturais dos documentos. Em sequência, com base nos dados obtidos na etapa anterior, é feita a descoberta de padrões utilizando algoritmos de *clustering* para identificar grupos de documentos com comportamentos similares, estabelecendo padrões dominantes de normalidade. Por fim, detectores de anomalias são treinados para cada padrão descoberto, gerando modelos de referência normais.

Figura 7 – Representação do Fluxo de Treino

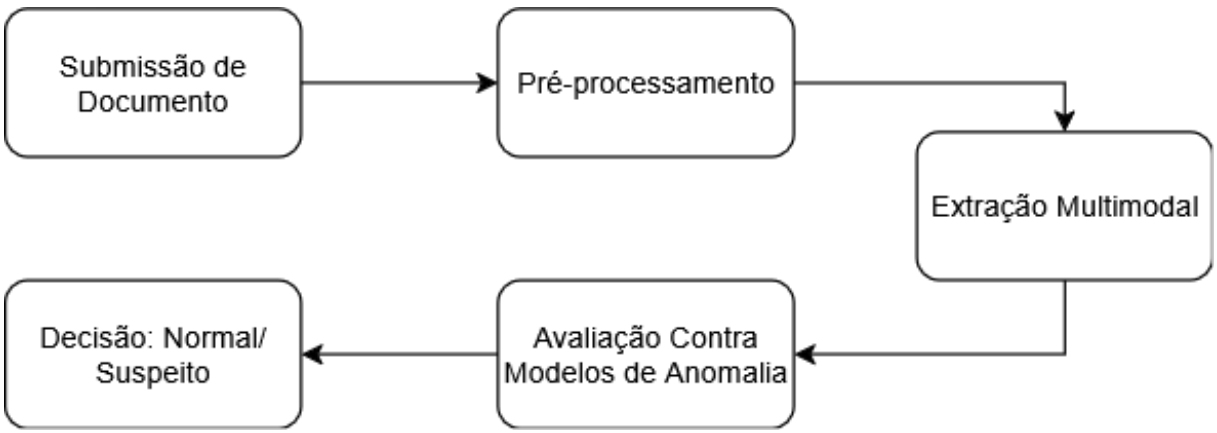


Fonte: o autor

4.1.2 Classificação de Novo Documento

Representada na figura 8, o processo de classificação de novo documento utiliza os modelos de referência estabelecidos na fase de treinamento para determinar a probabilidade de falsificação.

Figura 8 – Representação do Fluxo de Análise de Novo Documento



Fonte: o autor

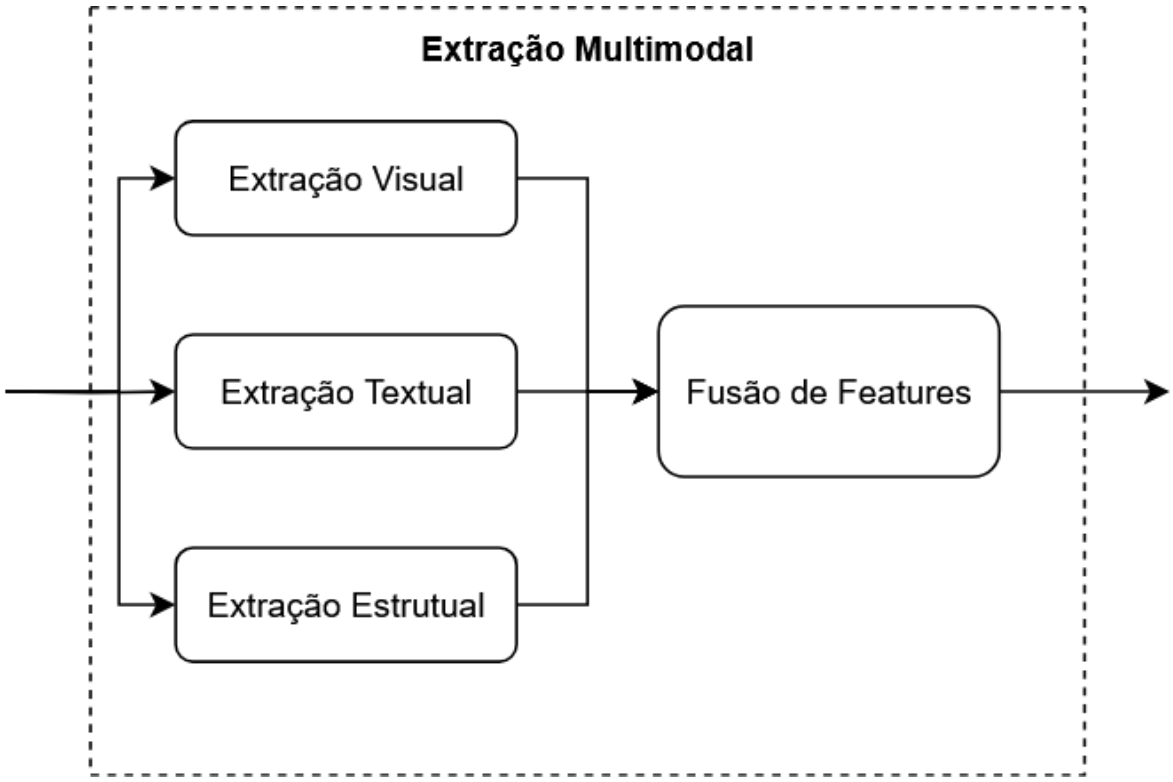
O processo inicia com a submissão de um novo certificado e, para garantir consistência na representação das características, passa pelo mesmo *pipeline* de pré-processamento e extração multimodal utilizado na fase de treinamento. Em seguida, os dados de representação do documento, obtidos na etapa anterior, são comparados contra todos os modelos de referência normal. Cada modelo calcula um escore de anomalia baseado na distância, ou similaridade, em relação ao padrão estabelecido pelo modelo. Essas pontuações representam, por fim, a probabilidade de falsificação do registro. Finalmente, utilizam-se métricas

de consenso para categorizar o arquivo, isto é, classificá-lo como normal ou suspeito a partir de determinado limiar de pontos.

4.1.3 Extração Multimodal de Características

O módulo de extração multimodal é responsável por capturar diferentes aspectos dos documentos. Essa abordagem permite aproveitar o mesmo *pipeline* de processamento combinando características independentes e, no contexto deste trabalho, complementares. Busca-se poder detectar tanto falsificações grosseiras quanto sofisticadas, uma vez que mesmo contrafações bem-feitas tendem a apresentar inconsistências sutis.

Figura 9 – Representação do Fluxo de Extração Multimodal de Características



Fonte: o autor

Ao invés de focar em características de domínios específicos, como nos trabalhos correlatos, essa abordagem combina três diferentes subprocessos de extração de *features* em paralelo, como representado na figura 9:

- Extração visual: extrai características ligadas ao layout, qualidade e consistência visual dos documentos. Inclui análise de textura, propriedades de fonte (espessura, tamanho, espaçamento), qualidade de assinaturas e selos, resolução de imagem, e padrões de cores e contrastes;

- Extração textual: utiliza modelos de processamento de linguagem natural para extrair características linguísticas e de formatação. Analisa padrões textuais, distribuição de termos, consistência na formatação de números, datas e códigos, além de verificar a coerência semântica do conteúdo;
- Extração estrutural: extrai características ligadas à organização espacial e estrutural dos documentos. Examina posicionamento de campos, formatação de tabelas, alinhamentos, margens, espaçamentos e a disposição geral dos elementos no documento.

Por fim, é realizada a fusão das características extraídas em todos os subprocessos. Para isso, os dados são normalizados e são aplicadas técnicas de redução dimensional, evitando a *maldição da dimensionalidade*, o que resulta em uma representação completa, unificada e compacta de cada documento.

5 PRÓXIMOS PASSOS

5.1 CRONOGRAMA

A seguir, segue o cronograma planejado para a próxima etapa do projeto e a disciplina de Trabalho de Conclusão de Curso 2.

Tabela 1 – Cronograma para TCC2.

Etapas	Meses				
	Ago	Set	Out	Nov	Dez
Obtenção e análise dos documentos acadêmicos	X	X			
Desenvolvimento do *software*		X	X	X	
Escrita da monografia				X	X
Entrega TCC II					X
Defesa pública					X

REFERÊNCIAS

ALAMERI, Mohammed et al. Unsupervised Forgery Detection of Documents: A Network-Inspired Approach. **Electronics**, v. 12, abr. 2023. DOI: 10.3390/electronics12071682.

ALZUBAIDI, Laith et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. **Journal of Big Data**, v. 8, n. 1, p. 53, mar. 2021. ISSN 2196-1115. DOI: 10.1186/s40537-021-00444-8. Acesso em: 27 jun. 2025.

BOONKRONG, Sirapat. Design of an academic document forgery detection system. **International Journal of Information Technology**, p. 1–13, jun. 2024. DOI: 10.1007/s41870-024-02006-6.

BRASIL. **Censo da Educação Superior 2023: notas estatísticas**. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), 2024.

DIAS, Pâmela; LEAL, Arthur. **Sites vendem diploma de curso superior para quem sequer pisou em sala de aula: 'Documentação 100% original, emitida de dentro da universidade', diz atendente**. 2022. 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: 5 abr. 2025.

DIETTERICH, Thomas G. Machine learning. In: ENCYCLOPEDIA of Computer Science. GBR: John Wiley e Sons Ltd., 2003. p. 1056–1059. ISBN 0470864125.

FANTÁSTICO. **Quatro pessoas são presas pela venda de 50 mil diplomas falsos e milhares carteirinhas de estudante**. 2025. 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: 5 abr. 2025.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. [S.l.]: MIT Press, 2016. http://www.deeplearningbook.org. Acesso em: 16 jun. 2025.

JAIN, Rajiv; WIGINGTON, Curtis. Multimodal Document Image Classification. **2019 International Conference on Document Analysis and Recognition (ICDAR)**, p. 71–77, 2019. DOI: 10.1109/ICDAR.2019.00021.

JAISWAL, Garima; SHARMA, Arun; YADAV, Sumit. Deep feature extraction for document forgery detection with convolutional autoencoders. **Computers & Electrical Engineering**, v. 99, p. 107770, abr. 2022. DOI: 10.1016/j.compeleceng.2022.107770.

JAMES, Hailey; GUPTA, Otkrist; RAVIV, Dan. OCR Graph Features for Manipulation Detection in Documents, set. 2020. DOI: 10.48550/arXiv.2009.05158.

KIM, Seong-Kyu. Blockchain Smart Contract to Prevent Forgery of Degree Certificates: Artificial Intelligence Consensus Algorithm. **Electronics**, v. 11, p. 2112, jul. 2022. DOI: 10.3390/electronics11142112.

MEC. **Aplicativo do MEC ganha prêmio de reconhecimento nacional**. 2022. Disponível em: https://www.gov.br/mec/pt-br/assuntos/noticias/2022/aplicativo-do-mec-ganha-premio-de-reconhecimento-nacional. Acesso em: 13 maio 2025.

MEC. **Portaria MEC/DAU n 33 de 2 de agosto de 1978**: Estabelece a sistemática para o registro de diplomas de curso superior. Brasília, DF: Ministério da Educação do Brasil, 1978.

MENGHANI, Gaurav. Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 55, n. 12, p. 1–37, mar. 2023. ISSN 1557-7341. DOI: 10.1145/3578938.

MOHAMMED, Shamsudeen; NWOBODO, Lois; EKENE, Njoku. Certificate Fraud Verification Model Using Clustered-Based Classification Approach. **Exploremetics Journal of Innovative Engineering and Technology**, v. 5, n. 1, p. 60–72, jun. 2024. ISSN 2636-590.

PALMA, Lucas M. et al. Blockchain and Smart Contracts for Higher Education Registry in Brazil. **International Journal of Network Management**, v. 29, 2019. DOI: https://doi.org/10.1002/nem.2061.

RNP, Rede. **Blockchain da jornada acadêmica**. 2023. Disponível em: https://www.youtube.com/watch?v=xqezMbjCeTM. Acesso em: 13 maio 2025.

SARKER, Iqbal. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**, v. 2, mar. 2021. DOI: 10.1007/s42979-021-00592-x.