



Universidade de Brasília

Universidade de Brasília – UnB

Faculdade de Ciências da Informação – FCI

Programa de Pós-Graduação em Ciência da Informação - PPGCINF

DOUGLAS FRANCISCO CRUZ PAIVA

**Uso de visão computacional para a indexação e categorização de prontuários
médicos e documentos de unidades de saúde**

Professor Orientador: Prof. Dr. Rogério Henrique de Araújo Júnior

Brasília - DF
2025

DOUGLAS FRANCISCO CRUZ PAIVA

Uso de visão computacional para a indexação e categorização de prontuários médicos e documentos de instituições de saúde

Qualificação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação (PPGCinf) da Faculdade de Ciência da Informação (FCI) da Universidade de Brasília.

Área de concentração: Gestão da Informação
Linha de pesquisa: Organização da Informação
Orientador: Prof. Dr. Rogério Henrique de Araújo Júnior

Professor Orientador: Prof. Dr. Rogério Henrique de Araújo Júnior

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

F149u Francisco Cruz Paiva, Douglas
Uso de visão computacional para a indexação e
categorização de prontuários médicos e documentos de
unidades de saúde / Douglas Francisco Cruz Paiva; orientador
Rogério Henrique de Araujo Júnior. Brasília, 2025.
119 p.

Dissertação(Mestrado em Ciência da Informação)
Universidade de Brasília, 2025.

1. Visão computacional. 2. Gestão de documentos. 3.
Prontuários médicos. 4. Lei Geral de Proteção de Dados. I.
de Araujo Júnior, Rogério Henrique, orient. II. Título.

UNIVERSIDADE DE BRASÍLIA

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Ata Nº: 103

Aos 10 dias do mês de dezembro do ano de dois mil e vinte e cinco, instalou-se a banca examinadora de Dissertação de Mestrado do aluno Douglas Francisco Cruz Paiva, matrícula 232103110. A banca examinadora foi composta pelos professores Dr. Renato Tarciso Barbosa de Sousa/titular interno/PPGCINF-UNB, Dra. Ana Clara Cândido - membro titular externo (UFSC) , Dra. Eliane Braga de Oliveira - suplente (UnB) e Dr. Rogério Henrique de Araujo Júnior - orientador/presidente (PPGCINF/UnB). O discente apresentou o trabalho intitulado "Uso de visão computacional para a indexação e categorização de prontuários médicos e documentos de unidades de saúde".

Concluída a exposição, procedeu-se a arguição do candidato, e após as considerações dos examinadores o resultado da avaliação do trabalho foi:

Pela aprovação do trabalho;

Pela aprovação do trabalho, com revisão de forma, indicando o prazo de até 30 (trinta) dias para apresentação definitiva do trabalho revisado;

Pela reformulação do trabalho, indicando o prazo de **(Nº DE MESES)** para nova versão;

Pela reprovação do trabalho, conforme as normas vigentes na Universidade de Brasília.

Conforme os Artigos 34, 39 e 40 da Resolução 0080/2021 - CEPE, o(a) candidato(a) não terá o título se não cumprir as exigências acima.

Dr. Rogério Henrique de Araujo Júnior - (PPGCINF/UnB)
(Presidente)

Dr. Renato Tarciso Barbosa de Sousa - (PPGCINF/UnB)
(membro titular interno)

Dr^a. Ana Clara Cândido - (UFSC)
(membro titular externo)

Dr^a. Eliane Braga de Oliveira - (UnB)
(suplente)

Douglas Francisco Cruz Paiva
(Mestrando)



Documento assinado eletronicamente por **Rogério Henrique de Araujo Junior, Membro do Colegiado da Pós-Graduação da Faculdade de Ciência da Informação**, em 10/12/2025, às 17:51, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.



Documento assinado eletronicamente por **Douglas Francisco Cruz Paiva, Usuário Externo**, em 10/12/2025, às 23:08, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.



Documento assinado eletronicamente por **Renato Tarciso Barbosa de Sousa, Membro do Colegiado da Pós-Graduação da Faculdade de Ciência da Informação**, em 11/12/2025, às 07:51, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.



Documento assinado eletronicamente por **Ana Clara Candido, Usuário Externo**, em 11/12/2025, às 18:07, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.



Documento assinado eletronicamente por **Ana Lucia de Abreu Gomes, Vice-Coordenador(a) da Pós-Graduação da Faculdade de Ciência da Informação**, em 18/12/2025, às 18:58, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.



A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **13507901** e o código CRC **49446D59**.

À minha família, por ser o alicerce de todas as minhas conquistas, e ao Matheus, pela presença constante e motivação.

AGRADECIMENTOS

Certamente estas palavras não alcançarão todas as pessoas que, de forma direta ou indireta, contribuíram para que este trabalho fosse concluído. A todos, registro meu sincero reconhecimento e gratidão.

Agradeço, em primeiro lugar, à minha família, fonte de força e incentivo constante. Aos meus pais, Paulo e Neila, pelo exemplo de dedicação e pelo apoio em cada etapa da minha vida acadêmica e pessoal. Aos meus irmãos Gabriella, Gustavo, Guilherme e Geovani, pela presença, pelo companheirismo e pelas conversas intermináveis no jantar que sempre me estimularam a pensar além, a refletir e a compartilhar ideias. E o pequeno Arthur que sempre tem algo a falar.

Aos meus colegas e amigos, que tornaram esta jornada mais leve e produtiva. À minha querida equipe da ArquiTI que compartilhou não apenas conhecimentos técnicos, mas também risadas e apoio em momentos de pressão. Aos amigos Vanderson, Iago, Lívia, Carolzinha, Luiz Phelipe que estiveram presentes nos momentos de dúvida e incerteza. À minha amiga de jornada neste mestrado, Graziela, cuja parceria e cumplicidade foram fundamentais para enfrentar os desafios acadêmicos. E, de modo especial, ao Matheus, pelo apoio diário, pelas conversas que me ajudaram a refletir e pelas distrações necessárias que me permitiram recobrar o ânimo.

Registro minha gratidão a todos os professores da Arquivologia, Museologia e Biblioteconomia, que, em diferentes momentos, contribuíram para minha formação e para a construção do profissional que estou me tornando. Cada aula, orientação e conselho deixou marcas importantes que levarei comigo não apenas na carreira, mas também na vida.

Um agradecimento especial ao meu orientador, Prof. Dr. Rogério Henrique de Araújo Júnior, pela paciência e por acreditar neste projeto, mesmo diante dos obstáculos e de alguns atrasos. Sua dedicação, disponibilidade e confiança foram essenciais para que este trabalho alcançasse os resultados propostos. Mais do que um orientador, foi um verdadeiro guia acadêmico, cuja influência e ensinamentos seguirão comigo além desta pesquisa.

Por fim, estendo meu agradecimento a todos que, de alguma forma, apoiaram esta trajetória, oferecendo palavras, gestos ou simplesmente presença. Cada contribuição, ainda que silenciosa, foi essencial para que eu conseguisse concluir esta etapa. Este trabalho é também fruto do apoio coletivo que recebi, e a cada um fica registrado meu reconhecimento e minha eterna gratidão.

Nossos passos vêm de longe

Jurema Wernek

RESUMO

A gestão de grandes volumes de informações em unidades de saúde, especialmente prontuários médicos, enfrenta desafios relacionados à organização, recuperação e proteção da privacidade dos dados. Esta pesquisa tem como objetivo desenvolver e validar um sistema automatizado, local e seguro para indexação e categorização de documentos de instituições de saúde, utilizando técnicas de visão computacional e princípios arquivísticos. A metodologia adotada combina abordagem qualitativa e quantitativa, com revisão bibliográfica detalhada, desenvolvimento de um protótipo de visão computacional e testes comparativos entre a indexação automatizada e a manual. O sistema proposto busca eliminar a necessidade de leitura manual, garantindo conformidade com a Lei Geral de Proteção de Dados (LGPD) e promovendo a anonimização de informações sensíveis. O protótipo foi avaliado em termos de eficiência e precisão, demonstrando potencial para reduzir custos, otimizar processos e acelerar a recuperação de dados. A pesquisa integra conceitos da Arquivologia e da Ciência da Informação às inovações tecnológicas, evidenciando que a aplicação de inteligência artificial na gestão de documentos pode aprimorar a assistência, o ensino e a pesquisa em saúde. Os resultados indicam que a automação permite maior padronização e segurança na manipulação de prontuários, além de reduzir erros e liberar profissionais para atividades mais estratégicas. Conclui-se que a implementação de sistemas autônomos e locais contribui para a soberania tecnológica em instituições de saúde, especialmente em regiões com desigualdades estruturais, oferecendo soluções acessíveis e replicáveis para a melhoria da gestão de documentos.

Palavras-chave: Visão computacional; Gestão de documentos; Prontuários médicos; Lei Geral de Proteção de Dados.

ABSTRACT

The management of large volumes of information in healthcare facilities, particularly medical records, faces challenges related to organization, retrieval, and data privacy protection. This research aims to develop and validate an automated, local, and secure system for indexing and categorizing hospital documents using computer vision techniques and archival principles. The methodology combines qualitative and quantitative approaches, including a detailed literature review, the development of a computer vision prototype, and comparative tests between automated and manual indexing. The proposed system seeks to eliminate the need for manual reading, ensuring compliance with the General Data Protection Law (LGPD) and promoting the anonymization of sensitive information. The prototype was evaluated in terms of efficiency and accuracy, demonstrating potential to reduce costs, optimize processes, and accelerate data retrieval. The study integrates concepts from Archival Science and Information Science with technological innovations, highlighting that artificial intelligence applications in Record Management can enhance healthcare services, education, and research. The results indicate that automation allows for greater standardization and security in handling medical records while reducing errors and freeing professionals for more strategic activities. It is concluded that implementing autonomous and local systems contributes to technological sovereignty in healthcare institutions, particularly in regions with structural inequalities, providing accessible and replicable solutions for improving record management.

Keywords: Computer vision; Record Management; Medical records; Technological sovereignty.

LISTA DE FIGURAS

Figura 1 - Metodologia da pesquisa.....	76
Figura 2 - Desenvolvimento da pesquisa	85
Figura 3 - Fluxo automatizado de indexação e categorização de documentos com IA	91
Figura 4 - Fluxo tradicional de indexação e categorização de documentos	92
Figura 5 - Processo de funcionamento do protótipo	95
Figura 6 – como é um arquivo de marcação	96
Figura 7 - Função de treinamento repetitivo da IA.....	97
Figura 8 – como é um arquivo yaml com parâmetros para treinar a visão computacional	97
Figura 9 - Visão computacional identificando informações de interesse	98
Figura 10 – Seletor de dados para anonimização.....	99
Figura 11 - Documento anonimizado.....	99
Figura 12 - Matriz de confusão normalizada para Classificação de Metadados em Prontuários Médicos.....	104

LISTA DE GRÁFICOS

Gráfico 1 - Curva de aprendizado da primeira etapa de treinamento da IA.....	87
Gráfico 2 - Curva de precisão da primeira etapa de treinamento da IA	87
Gráfico 3 - Curva de aprendizado da quarta etapa de treinamento da IA.....	88
Gráfico 4 -Curva de precisão da quarta etapa de treinamento da IA	88
Gráfico 5 - Curva de aprendizado da oitava etapa de treinamento da IA	88
Gráfico 6 - Curva de precisão da oitava etapa de treinamento da IA.....	89
Gráfico 7 - matriz de correlação normalizada da oitava fase de treinamento da IA ..	89

LISTA DE QUADROS

Quadro 1 - Procedimento de Pesquisa em Base de Dados	32
Quadro 2 - Quadro resumo objetivos, pressupostos e providências metodológicos.	75
Quadro 3 - Etapas da Pesquisa.....	80
Quadro 4 - Condições da pesquisa	81
Quadro 5 - Comparação entre o modelo manual e o modelo automatizado de indexação e categorização documental	93

LISTA DE ABREVIATURAS E SIGLAS

ANS – Agência Nacional de Saúde Suplementar

API – Interface de Programação de Aplicações (Application Programming Interface)

APIs – Interfaces de Programação de Aplicações (Application Programming Interfaces)

CFM – Conselho Federal de Medicina

CID-10 – Classificação Internacional de Doenças – 10ª Revisão

CNA – Congresso Nacional de Arquivologia

CPF – Cadastro de Pessoas Físicas

DATASUS – Departamento de Informática do Sistema Único de Saúde (Banco de Dados do SUS)

e-SUS AB – e-SUS Atenção Básica (Sistema de Informação em Saúde para a Atenção Básica)

ENANCIB – Encontro Nacional de Pesquisa e pós-graduação em Ciência da Informação

FHIR – Fast Healthcare Interoperability Resources

HL7 – Health Level Seven (organização de padrões em saúde)

IA – Inteligência Artificial

IBBD – Instituto Brasileiro de Bibliografia e Documentação

IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia

IEEE – Institute of Electrical and Electronics Engineers

ISBN – International Standard Book Number

ISSN – International Standard Serial Number

KDD – Knowledge Discovery in Databases

KDT – Knowledge Discovery in Texts

LGPD – Lei Geral de Proteção de Dados (Lei nº 13.709/2018)

MARC 21 – MACHine-Readable Cataloging 21

OCDE – Organização para a Cooperação e Desenvolvimento Econômico

OMS – Organização Mundial da Saúde

OPAS – Organização Pan-Americana da Saúde

OSDI – Operating Systems Design and Implementation (simpósio)

PEP – Prontuário Eletrônico do Paciente

PLN – Processamento de Linguagem Natural

PROV (W3C PROV) – Modelo de Proveniência do W3C (Provenance Data Model)

SGD – Sistema de Gestão de Documentos

SUS – Sistema Único de Saúde

TISS – Troca de Informações em Saúde Suplementar

W3C – World Wide Web Consortium

WHO – World Health Organization (equivalente em inglês de OMS)

Sumário

1	<i>Introdução</i>	18
1.1	Proposito da pesquisa.....	20
2	<i>problema de pesquisa</i>	23
2.1	Objetivo da pesquisa.....	24
2.2	Premissas básicas da pesquisa	25
2.3	Justificativa.....	26
3	<i>Trabalhos afins</i>	32
3.1	Síntese e proposta	36
4	<i>Revisão Narrativa</i>	38
4.1	Organização da Informação.....	38
4.1.1	Larga escala	47
4.2	Fundamentos da visão computacional	50
4.3	Ciências da Saúde e Ciência da Informação	54
4.4	Prontuário.....	56
4.4.1	Regulamentação do prontuário médico.....	58
4.5	Representação da informação em sistemas de saúde.....	63
4.5.1	Contribuição dos <i>Frameworks</i> HL7 e FHIR.....	64
4.6	Mineração de Dados e Textos em Ambientes Médicos	65
4.7	Impacto da Automatização nos Padrões de Privacidade.....	68
4.8	Conclusão da revisão narrativa.....	70
5	<i>Pressupostos</i>	73
5.1	Pressuposto Geral.....	73
5.2	Pressupostos específicos.....	73
6	<i>Metodologia</i>	74
6.1	Especificação da Metodologia	74
6.2	Delimitação do Estudo	77

6.3	Caracterização do Universo Estudado	78
6.3.1	Definição da amostra.....	78
6.4	Delineamento e Histórico da Pesquisa	80
6.4.1	Etapas da Pesquisa.....	80
6.4.2	Condições da pesquisa	81
7	<i>Desenvolvimento do Protótipo</i>	83
7.1	Modelo de Automação dos Processos Arquivísticos.....	90
7.2	Comparação com o Modelo Tradicional	91
7.3	Validação do protótipo	93
8	<i>Protótipo e resultados</i>.....	95
8.1	Avaliação da Eficiência e da Precisão do Sistema	100
8.2	Limitações do Protótipo e aprimoramentos possíveis	105
8.3	Comprovação dos pressupostos	106
9	<i>Conclusão</i>	109
	<i>Referências</i>.....	113

1 INTRODUÇÃO

A gestão e o tratamento de grandes volumes de documentos de instituições de saúde e prontuários médicos constituem hoje um desafio de proporções crescente nas instituições de saúde. A quantidade de dados clínicos, laudos, relatórios, imagens diagnósticas e informações correlatas aumenta consideravelmente, exigindo métodos modernos que deveriam facilitar o acesso, a organização e o armazenamento dessas informações. Nesse contexto, a aplicação de técnicas de visão computacional¹ para indexação e categorização de prontuários médicos desponta como uma possível estratégia eficaz, capaz de otimizar processos, reduzir custos e acelerar a recuperação de informações.

O presente trabalho aborda a concepção, desenvolvimento e avaliação de um sistema local (*off-line*) totalmente automatizado para indexação de documentos e prontuários médicos. Esse sistema, além de funcionar sem a necessidade de acesso à rede de internet, dispensa a leitura manual dos documentos por pessoas no momento da classificação, reforçando a segurança, a privacidade e a celeridade das operações internas. Sua implementação está em consonância com a crescente demanda por soluções que assegurem a confidencialidade dos dados sensíveis, sobretudo após o advento da Lei Geral de Proteção de Dados (LGPD). Ademais, o sistema proposto permite a anonimização de elementos capazes de identificar pessoas, possibilitando o uso dos dados para finalidades científicas sem comprometer a privacidade dos pacientes. Todo o processo deve ser permanentemente monitorado e auditado pela organização custodiante, garantindo o cumprimento de normas éticas e legais.

O tema se insere em um debate amplo sobre o papel da gestão de informações em saúde na promoção do bem-estar coletivo, tal como discutido por Douglas Paiva, Rafael Mateus e Rogerio Araújo Júnior (2024), ao enfatizarem a importância de estratégias bem estruturadas de representação e organização de documentos em arquivos de saúde pública. Conforme esses autores, a adoção de

¹ A visão computacional é uma área da Inteligência Artificial voltada para a análise e compreensão de dados visuais (imagens, vídeos etc.) por meio de técnicas e algoritmos computacionais. Seu objetivo é permitir que máquinas extraiam informações relevantes de cenas, reconheçam padrões e tomem decisões de maneira autônoma, simulando ou superando capacidades humanas de percepção visual.

novas tecnologias – incluindo a Inteligência Artificial (IA) e a visão computacional – pode potencializar não só a organização e a recuperação dos dados, mas também reforçar o valor estratégico desses acervos para embasar políticas de saúde, aprimorar o ensino e a pesquisa, além de apoiar a tomada de decisão clínica de maneira ágil e segura.

A IA, aliada à Arquivologia, deve apoiar a modernização dos sistemas de informação em saúde, promovendo acessibilidade, segurança e eficiência. A integração dessas tecnologias nos sistemas locais, elimina a dependência de terceiros, garantindo maior controle sobre os dados e respeitando a LGPD. A implementação de tecnologias locais é particularmente relevante para o Sul Global², onde desigualdades estruturais frequentemente limitam o acesso a inovações tecnológicas. Ao permitir que cada instituição adapte o sistema às suas necessidades específicas, sem depender de infraestrutura externa, o uso de IA pode apoiar a promoção da soberania tecnológica.

Diante disso, esta pesquisa apresenta não apenas o embasamento teórico acerca do uso de visão computacional em prontuários médicos, mas também descreve o desenvolvimento de um protótipo funcional para indexação e categorização automática de documentos. Espera-se que tal abordagem contribua de forma efetiva para a melhoria da gestão de documentos de unidades de saúde, oferecendo caminhos seguros e eficientes para o uso dos dados em benefício da assistência, do ensino e da pesquisa em saúde. Além de oferecer benefícios práticos, como maior eficiência e acessibilidade, essas ferramentas reforçam a importância de estratégias éticas e inclusivas, garantindo que a tecnologia sirva a todas as populações de forma equitativa. Esse modelo pode contribuir na promoção da saúde em larga escala, evidenciando o potencial transformador da integração entre ciência da informação e tecnologias emergentes.

O projeto de dissertação está estruturado de forma a garantir uma abordagem sistemática e integrada da problemática da gestão de prontuários médicos utilizando visão computacional. Inicialmente, apresenta-se a introdução, onde são

² O termo “Sul Global” refere-se aos países e regiões historicamente marginalizados no sistema internacional, em sua maioria localizados na África, América Latina, Caribe, Ásia e Oceania. Esses contextos são frequentemente caracterizados por desigualdades estruturais, heranças coloniais e desafios socioeconômicos e políticos que afetam o acesso equitativo a recursos e tecnologias.

contextualizados os desafios enfrentados na organização e recuperação de grandes volumes de informações em saúde pública. Em seguida, são delineados o problema de pesquisa, os objetivos gerais e específicos, bem como as premissas e justificativas que fundamentam o estudo. A revisão narrativa explora conceitos essenciais da Ciência da Informação, Arquivologia, Inteligência Artificial e Visão Computacional, estabelecendo o embasamento teórico para a pesquisa. A seção metodológica detalha os procedimentos adotados para o desenvolvimento do sistema, incluindo a modelagem do protótipo, as estratégias de validação e os critérios de análise dos resultados. O projeto avança para a apresentação dos experimentos realizados, evidenciando a eficiência do sistema proposto em comparação com métodos manuais de indexação e categorização documental. Por fim, são discutidas as conclusões e perspectivas futuras, indicando como a solução pode ser aprimorada e aplicada em larga escala para otimizar a gestão de documentos em instituições de saúde.

1.1 Proposito da pesquisa

A gestão eficiente de grandes volumes de informações tornou-se um desafio crescente em diversas áreas, especialmente na saúde, onde a organização e recuperação ágil de dados são fundamentais para a tomada de decisões. O avanço das tecnologias de inteligência artificial (IA) tem permitido novas soluções para a automação desses processos, reduzindo a dependência de métodos manuais e aumentando a acessibilidade das informações. No contexto dos prontuários médicos, a aplicação dessas tecnologias se torna ainda mais relevante, considerando a criticidade dos dados, sua natureza sensível e a necessidade de conformidade com regulamentações de proteção à privacidade. Diante desse cenário, esta pesquisa busca explorar o potencial da visão computacional para otimizar a extração e organização de informações contidas em prontuários médicos, tornando-as mais acessíveis e seguras para profissionais da saúde e gestores.

O propósito desta pesquisa surgiu da necessidade de melhorar a gestão de grandes volumes de informações contidas em arquivos, que são fundamentais para o processo de tomada de decisão. Em muitas instituições, a falta de um processo de planejamento eficaz resulta na formação de grandes massas documentais, tanto físicas quanto digitais, que acabam por se tornar inacessíveis ou subutilizadas devido à insuficiência de recursos humanos e ao excesso de processos manuais. Nesse

contexto, as tecnologias de Inteligência Artificial (IA) emergem como ferramentas indispensáveis para auxiliar os profissionais da informação, automatizando os processos mecânicos de organização e recuperação de dados, e, assim, mitigando lacunas informacionais.

A escolha de focar nos prontuários médicos, na presente pesquisa, foi motivada por vários fatores. Primeiramente, a emergência global em saúde OPAS (2020), Ministério da Saúde (2021), OPAS (2022); PHELAN (2024), OPAS (2024) evidenciou a importância de ter informações acessíveis e organizadas para a tomada de decisões que afetam vidas. A velocidade na disponibilização dessas informações é vital tanto para a saúde pública quanto para o desenvolvimento de tratamentos eficazes. Além disso, os prontuários médicos contêm dados pessoais sensíveis, cuja privacidade deve ser protegida conforme a Lei. A complexidade desses documentos, que apresentam informações dispersas em diferentes formatos e localizações, torna desafiadora a extração automática de dados.

Neste cenário, o uso de visão computacional se torna essencial. O treinamento de um sistema que possa identificar visualmente onde as informações estão localizadas em um prontuário médico não só aumenta a precisão na extração de dados, como também viabiliza a anonimização dessas informações quando necessário, garantindo a conformidade com regulamentações de proteção de dados. Além disso, ao desenvolver e treinar um modelo de visão computacional para lidar com prontuários médicos, a pesquisa abre caminho para que essa tecnologia seja aplicada de maneira mais ágil a outros tipos documentais, dada a complexidade e variedade de variáveis envolvidas.

Antes de iniciar a pesquisa, foram analisados sistemas de Inteligência Artificial que oferecem APIs³ para o processamento de documentos digitais, permitindo a extração automática de dados e a geração de metadados. No entanto, um dos principais desafios identificados é a falta de transparência quanto ao processamento desses documentos, além da impossibilidade de garantir que os

³ APIs, ou “Interfaces de Programação de Aplicativos” (do inglês Application Programming Interfaces), são conjuntos de definições e protocolos que permitem a comunicação entre diferentes sistemas de software. Por meio delas, desenvolvedores podem solicitar e enviar dados, bem como acessar funcionalidades específicas de uma plataforma ou de uma biblioteca, sem a necessidade de conhecer, em detalhe, toda a implementação interna desses serviços. (<https://openai.com/api/> , <https://developers.meta.com/ai/> , <https://ai.google.dev/gemini-api>)

dados não sejam utilizados para outros fins. Essa incerteza representa um risco significativo para a segurança da informação, especialmente no contexto de documentos sensíveis como prontuários médicos. Por essa razão, optou-se pelo desenvolvimento de um sistema local, que opera fora da rede, minimizando as brechas de segurança.

Adicionalmente, essas APIs são geralmente pagas por requisição, o que pode tornar o uso intensivo, típico de ambientes de arquivo, financeiramente inviável. Em contraste, o modelo proposto nesta pesquisa visa ser local e de código aberto, permitindo que instituições, sejam elas públicas ou privadas, repliquem o sistema de forma mais econômica, sem a necessidade de investimentos sistemáticos a cada solicitação de processamento.

Em suma, a pesquisa busca responder à necessidade premente de uma gestão de documentos mais eficiente e segura, utilizando IA como meio de garantir que as informações críticas sejam acessíveis, organizadas e protegidas, tanto no âmbito da saúde pública quanto em outros setores onde a informação é um recurso estratégico.

A necessidade de uma gestão eficiente de informações em saúde pública é cada vez mais premente, especialmente em um cenário onde a interoperabilidade e a segurança dos dados são cruciais para a qualidade dos serviços prestados. A aplicação de princípios arquivísticos tradicionais, combinada com ferramentas tecnológicas modernas, visa não apenas organizar e preservar os documentos de forma mais eficaz, mas também garantir que as informações sejam acessíveis e utilizáveis para a tomada de decisões clínicas e administrativas no âmbito dos serviços de saúde.

2 PROBLEMA DE PESQUISA

Desde o surgimento da escrita pode se dizer que surgiu a necessidade de organizar os documentos. “É ao aparecimento da escrita que remonta o nascimento dos arquivos e da arquivística, bem como as novas ocupações, entre as quais a de arquivista. A escrita permitiu produzir obras literárias, mas também serviu a administração” (ROUSSEAU; COUTURE, 1998, p. 29). A gestão da informação é um desafio crescente, diante do grande volume de informação gerada em todas as áreas, a partir da segunda guerra mundial esse fato se tornou um ponto de inquietação.

(...)com a denominada explosão documental, a partir da Segunda Guerra Mundial, as demandas por uma eficácia da ação administrativa e para uma eficiente gestão econômica crescem e exigem das comissões governamentais nos EUA soluções para o volume de documentos do Estado. (SILVA, 2008, p. 645)

Em um cenário onde a quantidade de informações geradas diariamente é cada vez maior, torna-se imprescindível a utilização de ferramentas que permitam uma melhor organização e representação dos dados, visando a otimização dos processos de recuperação, de tomada de decisão a fim de garantir a segurança e a qualidade dos serviços prestados. A arquivologia, como ciência que trata da gestão, organização e preservação de documentos, possui um papel importante nesse contexto, visto que os arquivos são responsáveis por garantir a autenticidade, a integridade e a disponibilidade das informações.

No entanto, as unidades informacionais e os arquivos que lidam com esses dados ainda encontram dificuldades para organizar e integrar essas informações, principalmente os arquivos físicos que muitas vezes ficam marginalizados gerando uma lacuna informacional. O auxílio de tecnologias como a inteligência artificial (IA) e a visão computacional para a aplicação dos princípios arquivísticos, pode proporcionar melhorias significativas na gestão de informações em grandes bases de dados e repositórios de dados. A pesquisa pretende investigar como a arquivologia pode contribuir para a gestão eficiente de informações na área da saúde, especialmente por meio da representação e organização da informação.

A pesquisa propõe o desenvolvimento de um protótipo de visão computacional que pode utilizar câmeras ou acessar arquivos PDF e de imagens para identificar e extrair automaticamente informações críticas, tais como nome, CPF, doenças e outras informações relevantes. Este sistema não só acelera o processo de

recuperação de documentos, mas também garante a uniformidade na interpretação uniforme dos dados e a proteção da privacidade dos pacientes. A utilização de *frameworks* HL7 e FHIR deve garantir a interoperabilidade dos dados extraídos com os sistemas de saúde existentes.

Diante disso, a pergunta central que esta pesquisa busca responder é: Como a visão computacional pode ser utilizada para a indexação e categorização automática de prontuários médicos e documentos de instituições de saúde, garantindo a conformidade com a LGPD e facilitando a interoperabilidade dos dados em sistemas de saúde?

2.1 Objetivo da pesquisa

- **Objetivo geral**

Desenvolver e validar uma metodologia eficiente⁴ e segura⁵ para a gestão de informações que combine princípios arquivísticos com tecnologias de inteligência artificial na detecção de padrões e classificação de imagens. Para aprimorar a organização e a recuperação de documentos em instituições de saúde.

- **Objetivos específicos**

Identificar e sistematizar os requisitos legais, técnicos e arquivísticos aplicáveis à gestão de prontuários médicos e documentos de unidades de saúde — com destaque para a conformidade à Lei Geral de Proteção de Dados (LGPD), às normas de temporalidade e destinação documental e aos padrões de interoperabilidade (como HL7 e FHIR) — de modo a fundamentar o desenvolvimento e a validação do sistema proposto.

Definir e caracterizar o corpus documental da pesquisa, estabelecendo critérios de seleção, inclusão e exclusão dos prontuários médicos a serem utilizados, de forma a garantir a representatividade e a consistência dos dados no processo de desenvolvimento e validação do sistema.

⁴ Uma metodologia é eficiente quando maximiza desempenho e produtividade com uso otimizado dos recursos (tempo, pessoal, custo, tecnologia), apresenta baixos índices de erro e boa capacidade de escalabilidade e manutenção.

⁵ segura quando assegura a integridade, privacidade e confidencialidade dos dados, atende às normas legais (por exemplo, LGPD no Brasil), protege contra falhas ou ataques, tem mecanismos de verificação e validação, auditoria, restrição de acesso, tratamento ético, prevenção de vies e transparência.

Prototipar um sistema de visão computacional capaz de identificar e extrair automaticamente nome, CPF, doenças, data do documento, assunto e palavras-chave que represente o conteúdo a partir de prontuários médicos em formato digital e físico.

Avaliar a eficácia e a precisão do protótipo em comparação com pessoas realizando a indexação, considerando as métricas: tempo, acurácia e quantidade de erros.

2.2 Premissas básicas da pesquisa

As premissas básicas orientadoras da pesquisa são:

- I. A gestão eficiente de grandes volumes de informações em saúde pública requer uma abordagem estruturada orientada pelos princípios arquivísticos. A premissa é que a Arquivologia, com suas metodologias de organização, preservação e acesso a documentos, forneça a base necessária para melhorar a representação e a recuperação da informação em sistemas de saúde, garantindo a integridade e acessibilidade dos dados ao longo do tempo.
- II. A aplicação de tecnologias de inteligência artificial (IA), especialmente a visão computacional, pode otimizar significativamente os processos de identificação, extração e anonimização de dados em prontuários médicos. A IA não apenas acelera essas operações, mas também minimiza erros humanos, padroniza a análise e reduz vícios operacionais⁶, garantindo maior precisão e uniformidade no tratamento da informação.
- III. Em um ambiente de saúde⁷ altamente regulado⁸, a privacidade e a segurança dos dados dos pacientes são primordiais. Portanto, a implementação de sistemas que respeitem regulamentações como

⁶ Essa etapa apresenta diversos desafios, como subjetividade na escolha de termos, falta de padronização e consistência, erros de digitação ou omissão de informações, uso de julgamento pessoal na determinação da relevância, resistência à adoção de novas classificações ou atualizações e desconhecimento de normas e legislações específicas.

⁷ Compreende o ecossistema formado por hospitais, clínicas, laboratórios, planos de saúde e outros agentes responsáveis pela prestação de serviços médicos e pelo tratamento de dados de pacientes.

⁸ Por normas nacionais, como a LGPD, e por legislações e resoluções específicas do setor, a exemplo das emitidas pelo Conselho Federal de Medicina.

a LGPD é indispensável para a proteção das informações sensíveis, prevenindo vazamentos e viabilizar o uso ético da tecnologia.

- IV. A automação de processos mecânicos auxiliares que incide sobre as atividades técnicas do arquivista, por meio da combinação de IA e mineração de dados, é uma forma de liberar capital humano para tarefas mais complexas e de valor agregado. A automatização não apenas aumenta a eficiência operacional, mas também contribui para uma melhor utilização dos recursos informacionais disponíveis.
- V. Por fim, a pesquisa pressupõe que o uso da inteligência artificial na gestão de informações de saúde deve garantir que as tecnologias não só respeitem a privacidade dos indivíduos, mas também promovam a equidade no acesso e uso dos dados de saúde.

2.3 Justificativa

A experiência adquirida ao longo dos anos atuando em arquivos e trabalhando com sistemas de gerenciamento de documentos, revelou que, embora muitas informações sejam criadas, processadas e armazenadas, a reutilização dessas informações é frequentemente prejudicada pela falta de organização e de um processo de recuperação da informação eficiente.

No setor de saúde, foi observado que o acesso a prontuários físicos é dificultado, e os documentos digitais carecem de métodos ágeis para recuperação e uso, afirma Douglas Paiva, Rafael Mateus e Rogerio Araújo Júnior (2024). Essa situação deixa uma lacuna significativa, especialmente pela falta de integração entre as informações armazenadas em formatos físicos e digitais, o que compromete o potencial desses documentos como fontes de conhecimento já que não podem ser recuperados. Profissionais da saúde destacaram a importância do acesso a informações arquivadas, não só dos pacientes atendidos, mas também de outros casos para comparação de diagnósticos e tratamentos.

Com base nesse documento, o profissional da saúde consegue determinar quais alternativas são viáveis em um tratamento e quais procedimentos e hipóteses devem ser descartados. Com isso, é possível reduzir a probabilidade de um paciente ser submetido a um tratamento que ameace o seu bem-estar, ou até mesmo seja ineficaz. Exemplo disso, é a possibilidade de um paciente ser medicado com uma substância que provoca alergia ou ser submetido a um procedimento que já demonstrou ser pouco efetivo para determinado quadro clínico. (AFYA, 2024, p. 1)

Os profissionais de saúde podem acessar o histórico médico dos pacientes, resultados de exames e informações sobre tratamentos anteriores, facilitando o diagnóstico e o planejamento do tratamento. A disponibilidade de informações precisas e atualizadas, bem como a sua recuperação nas bases de conhecimento, contribui para a melhoria da qualidade dos cuidados e a redução de erros médicos. (PAIVA; MATEUS; ARAÚJO JÚNIOR, 2024, p. 49).

O histórico do paciente é um dos elementos mais críticos na prática médica e na gestão da saúde. Ele compreende um conjunto de informações sobre o passado clínico do paciente, incluindo doenças, tratamentos anteriores, cirurgias, alergias. No entanto, a incorporação de informações registradas em suporte físico (como papel) nas práticas clínicas modernas apresenta desafios significativos. Muitas instituições de saúde ainda mantêm parte ou a totalidade dos registros de pacientes em formato físico, o que dificulta o acesso rápido e eficiente às informações. A recuperação de dados em tempo hábil é crítica em ambientes clínicos, especialmente em emergências, onde atrasos na obtenção de informações podem resultar em decisões clínicas subótimas⁹. O processo manual de busca e interpretação de registros em papel é demorado e propenso a erros, além de não permitir uma análise integrada e multidimensional dos dados do paciente, o que limita o potencial para diagnósticos e tratamentos personalizados.

Por outro lado, a transição para um sistema completamente digital, embora vantajosa em muitos aspectos, também enfrenta obstáculos. A digitalização de registros exige investimentos significativos em infraestrutura tecnológica, treinamento de pessoal e manutenção de sistemas. Além disso, há desafios associados à interoperabilidade, segurança e privacidade dos dados. Sistemas digitais precisam ser capazes de integrar dados de diferentes fontes e garantir que informações críticas sejam facilmente acessíveis e interpretadas por todos os profissionais de saúde envolvidos no cuidado do paciente. A migração de dados de registros físicos para digitais também pode levar a lacunas e inconsistências, especialmente se o processo não for rigorosamente controlado.

A importância do histórico do paciente na epidemiologia é igualmente significativa. Dados históricos de saúde são fundamentais para rastrear a

⁹ No contexto clínico, decisões subótimas referem-se a escolhas que, embora possam ser aceitáveis ou viáveis, não representam a melhor opção disponível. Isso pode ocorrer devido à falta de informações completas, atrasos no acesso a dados críticos ou limitações no processamento das informações, potencialmente impactando a qualidade do diagnóstico e do tratamento do paciente.

disseminação de doenças, identificar padrões epidemiológicos e desenvolver estratégias de saúde pública. Por exemplo, durante surtos de doenças infecciosas, como a pandemia de COVID-19, o acesso rápido a informações sobre comorbidades, histórico de vacinação e outras condições de saúde pode ser essencial para determinar os grupos de risco e priorizar intervenções. A análise de históricos de pacientes também permite a detecção precoce de tendências emergentes, como o aumento de resistência a antibióticos ou o surgimento de novas variantes de um vírus, possibilitando respostas mais rápidas e eficazes.

Falamos da história clínica como pilar básico da atividade assistencial. A principal função do histórico médico é o atendimento clínico e deriva do motivo pelo qual é gerado: a necessidade de ter agrupadas todas as informações sobre uma pessoa, necessárias ao tratamento ou prevenção da doença (HERNÁNDEZ, 2006, p. 61).

Além disso, o histórico do paciente é vital para o desenvolvimento de tratamentos eficientes e em tempo recorde. Com a análise dos dados históricos, pesquisadores e médicos podem identificar padrões de resposta a diferentes tratamentos, permitindo ajustes personalizados e o desenvolvimento de terapias mais eficazes. No contexto de uma emergência global em saúde, como a pandemia de COVID-19, a capacidade de acessar e analisar rapidamente os históricos de pacientes permitiu o desenvolvimento e a implementação de tratamentos e vacinas em tempo recorde. A utilização de grandes bancos de dados clínicos, que integram históricos de pacientes de diferentes regiões e condições, foi essencial para a compreensão dos efeitos do vírus e das respostas imunológicas, contribuindo para o avanço das opções terapêuticas.

Ao interpretar a história clínica descrita evidencia-se que, ao contrário das justificativas que apresentavam uma caracterização geral/generalista, as histórias passaram a ter descrições mais objetivas, o que pode sugerir que o médico realizou uma análise clínica sobre o comportamento do indivíduo. Desse modo, entende-se que a história clínica representada pelos médicos diz respeito a “vida pregressa” do indivíduo, as características e comportamentos que o indivíduo expressa (CHEFFER, *et al.*, 2021, p. 60).

A análise da história clínica é uma maneira de entender o processo de atuação diante de sintomas, e por esse motivo essa informação precisa estar o mais acessível possível, sem invadir a privacidade dos indivíduos. Dessa forma, o sucesso dessas iniciativas depende da qualidade e da acessibilidade dos dados. Historicamente, muitas instituições de saúde enfrentaram dificuldades em consolidar e integrar informações de pacientes, especialmente quando os dados estavam

dispersos em diferentes formatos e locais. Isso ressalta a necessidade de sistemas robustos e interoperáveis que possam agregar dados de várias fontes, facilitando a análise e o uso eficaz do histórico do paciente tanto em nível individual quanto populacional.

Em suma, o histórico do paciente é uma peça-chave para a prática clínica e a saúde pública. Sua correta manutenção e integração são essenciais não apenas para o cuidado individualizado do paciente, mas também para a resposta eficaz a crises de saúde em larga escala. A superação dos desafios associados à digitalização e à interoperabilidade dos sistemas de informação em saúde é fundamental para garantir que esses dados possam ser usados de forma eficaz, segura e oportuna, contribuindo para a melhoria contínua dos cuidados de saúde e para a proteção da saúde pública global.

A recuperação e análise de históricos de doenças e tratamentos são essenciais para a prevenção eficaz de endemias, epidemias e pandemias. O conhecimento acumulado por meio desses registros permite identificar padrões de disseminação de doenças, compreender a eficácia de intervenções passadas e prever possíveis surtos futuros. Com um histórico bem documentado, as autoridades de saúde pública podem reagir rapidamente a novas ameaças, implementar medidas preventivas e alocar recursos de forma mais estratégica. Além disso, a análise de dados históricos possibilita a adaptação de políticas de saúde e a formulação de diretrizes baseadas em evidências, fortalecendo a resiliência do sistema de saúde.

Entretanto, a marginalização das informações registradas em suporte físico representa um sério obstáculo nesse processo. Registros em papel, frequentemente armazenados de forma descentralizada e em condições inadequadas, dificultam o acesso rápido e completo aos dados necessários em emergências. Essa dispersão e falta de padronização impedem uma visão integrada e precisa da saúde pública, limitando a capacidade de identificar e responder prontamente a emergências sanitárias. Além disso, a deterioração física dos documentos com o tempo pode resultar na perda irreversível de informações valiosas, criando lacunas que comprometem a capacidade de resposta das instituições de saúde.

A falta de integração entre sistemas digitais de informação em saúde é outro fator crítico que pode exacerbar essas lacunas. Quando sistemas de saúde não

são interoperáveis, dados de pacientes e históricos de doenças ficam fragmentados entre diferentes plataformas e instituições, dificultando a consolidação das informações e a análise abrangente necessária para uma resposta coordenada. Esse cenário compromete a eficácia das medidas de prevenção e controle de doenças, uma vez que decisões baseadas em dados incompletos ou não conectados.

Ainda, ao considerar as características de um sistema de saúde tradicional, pode-se dizer que a predominância, na maioria dos países, é de um sistema fragmentado, com ênfase nas super especialidades e no tratamento de doenças, tendo o hospital como centro principal de atendimento, com alto custo e baixa qualidade. Neste cenário, a solução informatizada que predomina são os sistemas departamentais, ou seja, sistema de farmácia, sistema de admissão, transferência e alta hospitalar, sistema de centro cirúrgico, sistema de prescrição médica, dentre outros. Tais sistemas nem sempre apresentam características de integração e interoperabilidade (Marin, 2010, p. 21).

Portanto, para garantir uma resposta eficaz a futuras crises de saúde, é imperativo que as informações de saúde, independentemente do suporte, sejam integradas em um sistema coeso e acessível. Isso exige investimentos em digitalização de registros, adoção de padrões comuns de interoperabilidade e esforços contínuos para preservar e atualizar os dados de saúde pública. Somente assim será possível utilizar plenamente o potencial dos históricos de doenças e tratamentos na prevenção de grandes crises sanitárias e na promoção de uma saúde pública mais robusta e resiliente.

A pesquisa proposta pretende contribuir para a área de gestão de informações em saúde pública, abordando a crescente demanda por soluções que integrem princípios arquivísticos com tecnologias emergentes, como a inteligência artificial e a visão computacional. Ao fornecer uma abordagem sistemática para a organização e recuperação de informações em larga escala, esta dissertação visa trazer resultados que possam impactar na melhoria, na eficiência e na qualidade dos serviços de saúde, garantindo que os dados sejam geridos de maneira segura, acessível e rápido.

A dissertação está de acordo com a Lei Geral de Proteção de Dados (LGPD), que exige a implementação de medidas rigorosas para proteger a privacidade dos dados pessoais. O propósito da pesquisa é desenvolver métodos que não só cumpram esses requisitos legais, mas que também possam ser aplicados de forma prática em ambientes hospitalares, respeitando a confidencialidade dos pacientes e garantindo a integridade das informações.

Embora haja avanços consideráveis na gestão de documentos e na aplicação de IA em saúde, ainda são poucas as iniciativas que abordam essas disciplinas de maneira integrada. A dissertação propõe explorar e desenvolver métodos inovadores que possam ser aplicados tanto em instituições de saúde quanto em outras instituições, promovendo uma gestão mais eficiente e ética dos dados aprimorando a arquivística com o auxílio da Tecnologia de *bigdata* que está causando uma revolução em análises de dados, afirma Caio Coneglian, Paula Gonzalez, Santarém Segundo (2017).

A pesquisa está alinhada com a emergência global de saúde pública, que demanda soluções rápidas e eficazes para a gestão de informações durante crises, como a pandemia de COVID-19. A capacidade de gerenciar dados de saúde pública de forma eficaz pode ter um impacto direto na capacidade das instituições de saúde em responder a emergências, melhorar o cuidado com os pacientes e tomar decisões informadas.

3 TRABALHOS AFINS

O levantamento da literatura foi baseado em pesquisas empreendidas nas seguintes bases de dados: portal de Teses e Dissertações da CAPES, base LISA, Base de Dados em Arquivística (BDA), Base de Dados Referencial de Artigos e Periódicos em Ciências da Informação (BRAPCI), *Scientific Electronic Library Online* (SciELO), Biblioteca Digital de Teses e Dissertações do IBICT, Google Acadêmico, *Journal of Information Systems and Technology Management* (JISTEM), Congresso Internacional de Gestão da Tecnologia e Sistemas de Informação (*International Conference on Information Systems and Technology Management*) - CONTECSI. Para realizar uma pesquisa abrangente nas bases de dados mencionadas, operadores booleanos como “AND, OR e NOT” (respectivamente, “E, OU e NÃO”) foram utilizados para otimizar os resultados da busca.

Quadro 1 - Procedimento de Pesquisa em Base de Dados

	Descrição
bases de dados	portal de Teses e Dissertações da CAPES, base LISA, Base de Dados em Arquivística (BDA), Base de Dados Referencial de Artigos e Periódicos em Ciências da Informação (BRAPCI), Scientific Electronic Library Online (SciELO), Biblioteca Digital de Teses e Dissertações do IBICT, Google Acadêmico, Journal of Information Systems and Technology Management (JISTEM), Congresso Internacional de Gestão da Tecnologia e Sistemas de Informação (International Conference on Information Systems and Technology Management) - CONTECSI
Idiomas de Pesquisa	Português, inglês, francês e espanhol
Termos de Pesquisa	- Português: "mineração de dados em saúde", "indexação automática de prontuários médicos", "visão computacional e arquivística", "inteligência artificial na gestão de documentos", "recuperação da informação médica", "processamento de linguagem natural em arquivos de saúde"
	- Inglês: "medical record data mining", "automatic indexing of medical records", "computational vision in archival science", "artificial intelligence in record management", "medical information retrieval", "natural language processing for medical archives"
	- Francês: "extraction de données dans les dossiers médicaux", "indexation automatique des archives médicales", "vision par ordinateur en archivistique", "Intelligence artificielle dans la gestion des documents d'archives", "intelligence artificielle pour la gestion documentaire", "récupération d'information médicale", "traitement du langage naturel pour les dossiers de santé"
	- Espanhol: "minería de datos en registros médicos", "indexación automática de archivos médicos", "visión computacional en archivística", "inteligencia artificial en la gestión documental", "recuperación de información médica", "procesamiento del lenguaje natural en archivos de salud"
Filtros Aplicados	- Período: Últimos 10 anos
	- Tipo de Documento: Artigos revisados por pares, teses e dissertações, revisões sistemáticas.
	- Área do Conhecimento: Ciências da Informação, Arquivologia, Ciência da Computação, Inteligência Artificial, Saúde, Mineração de Dados
	- Disponibilidade: Texto completo acessível (preferencialmente de acesso aberto)
	- Estudos que abordam mineração de dados aplicada à área médica e documental

Critérios de Inclusão	- Trabalhos que discutem técnicas de visão computacional, PLN ou IA aplicadas à recuperação de informações arquivísticas na saúde
	- Pesquisas que analisam impactos da automação na gestão de documentos
	- Estudos com evidências empíricas ou modelos validados
Critérios de Exclusão	- Trabalhos que tratam apenas de mineração de dados sem relação com a área da saúde ou gestão de documentos.
	- Estudos exclusivamente teóricos sem aplicabilidade prática
	- Artigos sem acesso ao texto completo

Verificou-se que, embora existam estudos sobre a gestão de informações em saúde, o uso de inteligências artificiais (IA) para auxiliar os serviços arquivísticos, bem como a aplicação de IA na medicina, ainda apresentam uma lacuna significativa de pesquisas que abordem especificamente a integração de visão computacional e arquivística para a organização e recuperação de dados em prontuários médicos. Sendo assim, o presente trabalho poderá contribuir significativamente para o desenvolvimento do tema proposto no âmbito da Ciência da Informação, com foco na aplicação de tecnologias para a gestão de documentos híbrida (física e digital) em ambientes de saúde, especialmente no contexto hospitalar.

No estudo de Fabrício Galdani (2022), é apresentado um modelo de mapeamento semântico para a representação e recuperação da informação em prontuários eletrônicos, utilizando terminologias médicas padronizadas como CID-10¹⁰ e SNOMED-CT¹¹. A interoperabilidade entre sistemas de saúde é enfatizada como um fator crucial para garantir que os dados dos pacientes sejam acessíveis e utilizáveis por diferentes plataformas, impactando diretamente a qualidade do atendimento médico. A implementação de um modelo de mapeamento semântico melhora a precisão e rapidez na recuperação de informações, o que converge com o objetivo desta pesquisa de melhorar a organização de dados de prontuários médicos por meio da IA.

¹⁰ A CID-10 (Classificação Internacional de Doenças e Problemas Relacionados à Saúde, 10ª Revisão) é um sistema de codificação padronizado, elaborado pela Organização Mundial da Saúde (OMS), que categoriza doenças, sintomas e outros problemas de saúde, servindo como base para estatísticas de morbidade e mortalidade, bem como para a gestão e planejamento de serviços de saúde em âmbito global.

¹¹ O SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms) é uma terminologia clínica abrangente, desenvolvida para representar de forma consistente informações sobre saúde em registros eletrônicos. Ele fornece uma estrutura padronizada para descrever diagnósticos, procedimentos, achados clínicos e outros elementos, facilitando a interoperabilidade e a troca de dados de forma precisa entre diferentes sistemas de saúde.

Na dissertação de Sheila Rodrigues Macedo (2019), é abordada a transição de prontuários físicos para eletrônicos em instituições públicas de saúde e os desafios gerenciais associados a esse processo. A autora ressalta a importância da organização dos prontuários para otimizar a gestão da informação e garantir a privacidade dos dados dos pacientes, conforme as regulamentações da LGPD. Macedo discute como a digitalização de prontuários pode melhorar o acesso às informações e agilizar o atendimento, mas também destaca os desafios relacionados à infraestrutura tecnológica inadequada, à falta de treinamento dos profissionais e à resistência à mudança nas instituições hospitalares. Macedo (2019), aponta que, embora a digitalização tenha o potencial de facilitar a gestão de documentos, a ausência de uma estratégia bem implementada pode resultar em dificuldades na recuperação da informação e em riscos à segurança dos dados. Este estudo alinha-se com a presente pesquisa no que tange à preocupação com a eficiência e segurança na gestão dos dados de saúde, ainda que o enfoque da autora seja em boas práticas de gestão de documentos e o presente estudo se concentre na automação de processos por meio de IA e visão computacional.

No artigo Precisão no processo de busca e recuperação da informação: uso da mineração de textos Rogério Araújo Júnior e Kira Tarapanoff (2006) tratam de uma comparação entre a indexação manual e o uso de ferramentas de mineração de textos para o processo de recuperação da informação. O estudo se baseia no caso do Centro de Referência e Informação em Habitação (Infohab), que possui uma grande base de dados, indexada manualmente por bibliotecários da Caixa Econômica Federal.

A pesquisa analisa a eficácia da mineração de textos em relação ao índice de precisão de resposta no processo de recuperação da informação, ao comparar os resultados obtidos pelo uso da mineração de textos com aqueles obtidos por meio da indexação manual. Um dos achados principais é que a mineração de textos não traz um ganho significativo considerando o índice de precisão em comparação com a indexação manual. No entanto, o estudo sugere que a mineração de textos pode ser utilizada de forma complementar, auxiliando o processo de indexação e aumentando o índice de precisão. Além disso, o artigo explora a viabilidade da utilização da mineração de textos para a construção e manutenção de tesouros, indicando que essa

abordagem poderia enriquecer o vocabulário utilizado na indexação manual, o que, por sua vez, aumentaria a precisão da recuperação de informações.

A dissertação de Ricardo Carvalho (2017) explora a aplicação de técnicas de mineração de textos em prontuários eletrônicos, com o objetivo de otimizar a recuperação de informações clínicas. O autor destaca o potencial do Processamento de Linguagem Natural (PLN) para facilitar a busca por dados não estruturados em textos clínicos, o que tem convergência com a presente pesquisa ao propor a utilização de IA para a organização e extração de dados em documentos médicos. Além disso, Carvalho (2017), observa que a falta de interoperabilidade entre os sistemas de informação compromete a eficiência dos serviços de saúde ao dificultar o acesso às informações críticas.

Outro trabalho que merece destaque é proposto por Lima (2023) que investiga a recuperação de informações farmacológicas em prontuários eletrônicos por meio de visualizações de dados e grafos de conhecimento. Embora o foco do autor seja na recuperação de dados relacionados a medicamentos, há pontos comuns com a presente pesquisa, como o uso de PLN para automatizar processos de extração de informações. A aplicação de visualizações de dados é relevante para discutir como a IA pode ser utilizada para melhorar a organização e visualização das informações em prontuários médicos.

Finalmente, o estudo de Bennett e Doub (2010) explora o uso de mineração de dados em prontuários eletrônicos de saúde (EHRs) para melhorar o tratamento clínico e apoiar a tomada de decisões médicas. A pesquisa revela que o uso de algoritmos de mineração de dados permite identificar padrões complexos que ajudam a personalizar o atendimento médico, o que se alinha com o propósito desta pesquisa de implementar a visão computacional para extrair e categorizar informações em prontuários médicos, contribuindo para a melhoria do processo decisório em saúde.

A análise dos trabalhos semelhantes ou afins aponta para a crescente importância de tecnologias como IA, PLN e visão computacional na organização e recuperação de informações em prontuários médicos. Desse modo, a pesquisa busca expandir essas abordagens ao propor o desenvolvimento de um sistema autônomo, que deve funcionar de forma local e independente, garantindo a segurança, a privacidade dos dados de saúde, a fim de superar as limitações identificadas em

estudos anteriores, tais como a dependência de APIs externas e a fragmentação de dados entre formatos físicos e digitais.

3.1 Síntese e proposta

Pode-se verificar que os trabalhos afins procuram analisar diversos aspectos da gestão de informações, com ênfase na utilização de tecnologias emergentes, na automação de processos e na melhoria da eficiência operacional. Esses estudos destacam a importância da interoperabilidade entre sistemas de saúde, a necessidade de conformidade com regulamentações de proteção de dados e o impacto da digitalização e automação na dinâmica de trabalho das instituições de saúde. A literatura também enfatiza a relevância de uma abordagem ética no desenvolvimento e aplicação dessas tecnologias, garantindo a segurança e privacidade dos dados sensíveis dos pacientes.

Apesar das contribuições significativas desses estudos, nota-se uma lacuna na literatura em relação à integração sistemática dos princípios arquivísticos com as tecnologias de inteligência artificial e visão computacional, especialmente no contexto da saúde pública. A maior parte dos trabalhos foca em áreas específicas, como a interoperabilidade ou a automação, sem explorar de forma abrangente a combinação dessas disciplinas para otimizar a gestão de informações em larga escala. Além disso, poucos estudos discutem o impacto da implementação de tais tecnologias no cotidiano das instituições e na experiência dos profissionais que atuam nesses ambientes.

Para consecução da proposta de pesquisa ora apresentada, é fundamental aprofundar e compreender cinco noções-chave que estruturam a abordagem da investigação: a) aplicação dos princípios arquivísticos, não apenas na organização de informações físicas, mas especialmente na gestão de documentos digitais, garantindo que a classificação e preservação sejam consistentes com os padrões e as necessidades atuais de acessibilidade e eficiência; b) interoperabilidade entre diferentes sistemas de saúde, fator essencial para que os dados de pacientes possam ser acessados e utilizados em diversos contextos e plataformas tecnológicas, promovendo uma troca eficiente de informações e contribuindo diretamente para a qualidade do atendimento; c) eficácia das tecnologias de inteligência artificial, particularmente no que se refere à extração, categorização e anonimização de dados

médicos, uma vez que essas tecnologias oferecem uma solução robusta para automatizar processos e otimizar a recuperação de informações sensíveis em prontuários médicos, tanto físicos quanto digitais; d) conformidade com as regulamentações de proteção de dados, como a LGPD, em ambientes hospitalares, assegurando que a privacidade dos pacientes seja protegida em todas as fases da digitalização e automação da gestão de documentos.

Com base nessas noções, a pesquisa propõe uma abordagem integrada que busca não apenas melhorar a eficiência e a segurança na gestão de documentos em saúde, mas também contribuir para o avanço das práticas arquivísticas em um contexto contemporâneo, ao integrar tecnologia de ponta e promover a preservação de informações críticas para a saúde pública.

4 REVISÃO NARRATIVA

A revisão narrativa buscará trazer uma análise abrangente das principais teorias, metodologias e estudos relevantes que abordam a interseção entre arquivologia, saúde pública e inteligência artificial, com foco na visão computacional e na gestão de informações em larga escala. O objetivo é mapear as contribuições acadêmicas existentes, identificar lacunas na pesquisa atual e fundamentar a proposta metodológica da tese. Além disso, a revisão explorará como as tecnologias emergentes e a IA têm sido aplicadas na área médica, destacando as implicações éticas e as conformidades legais associadas ao uso dessas tecnologias em ambientes hospitalares.

Nesse sentido, a revisão irá abordar, de forma articulada, os principais conceitos que fundamentam esta pesquisa, incluindo a organização e a representação da informação, os princípios arquivísticos aplicados à gestão de documentos, a Ciência da Informação no contexto da saúde, os fundamentos da inteligência artificial e da visão computacional, a indexação e categorização automática de documentos, a interoperabilidade em sistemas de saúde, bem como as implicações éticas, legais e sociais associadas ao tratamento automatizado de informações sensíveis,

4.1 Organização da Informação

A organização da informação é um dos pilares fundamentais da ciência da informação e desempenha um papel decisivo na gestão eficaz do conhecimento em qualquer contexto, seja acadêmico, empresarial, ou, como no caso desta pesquisa, na saúde pública. A organização da informação envolve a criação de sistemas e processos que permitem a estruturação, classificação, armazenamento, recuperação e disseminação eficiente do conhecimento. Esta seção tem como objetivo explorar os princípios, teorias e práticas que fundamentam a organização da informação, especialmente no contexto de grandes volumes de dados.

De acordo com Le Coadic (1996), a informação é um conhecimento registrado que contém um elemento de sentido, e esse registro é realizado por meio de um sistema de signos, como a linguagem. “A informação é um conhecimento inscrito (gravado) sob a forma escrita (impressa ou numérica), oral ou audiovisual. A informação comporta um elemento de sentido.” (Le Coadic, 1996, p. 5). Partindo dessa definição, a organização da informação surge como um processo necessário

para estruturar e descrever intelectualmente os conteúdos documentais, de modo a possibilitar sua representação em sistemas de recuperação. Kobashi e Aguiar (2013, p. 5) complementam essa perspectiva ao afirmar que, no domínio da Ciência da Informação, a organização da informação pode ser entendida como "uma série de atividades processuais com a finalidade de descrever intelectualmente conteúdos documentais para serem representados nos sistemas de recuperação da informação". A partir disso, percebe-se que o objetivo central da organização da informação é viabilizar o acesso eficiente aos dados armazenados, facilitando sua recuperação e utilização.

Essa preocupação com a organização da informação está presente entre autores tais como Paul Otlet (1934), que já considerava a prática de organizar documentos como parte fundamental da documentação. Segundo Paul Otlet, os objetivos eram:

"Os objetivos da documentação organizada consistem em poder oferecer, em todas as ordens de fato e de conhecimento, informações documentadas: 1º universais quanto ao seu objeto; 2º seguras e verdadeiras; 3º completas; 4º rápidas; 5º atualizadas; 6º fáceis de obter; 7º reunidas antecipadamente e prontas para serem comunicadas; 8º colocadas à disposição do maior número de pessoas." tradução própria. (Otlet, 1934, p. 6)

No texto de Braschër e Café (2008) a conceituação de organização da informação é apresentada como um processo fundamental para permitir o acesso eficiente ao conhecimento contido na informação. As autoras explicam que o objetivo central da organização da informação é possibilitar a localização, identificação, seleção e obtenção de entidades informacionais de maneira eficiente, sistemática e acessível. Como elas destacam, "o objetivo do processo de organização da informação é possibilitar o acesso ao conhecimento contido na informação" (BRASCHËR; CAFÉ, 2008, p. 5). Este processo envolve a descrição física e de conteúdo dos objetos informacionais, e resulta em uma representação que permite sua recuperação futura, destacando-se pela precisão e pela utilidade prática.

A organização da informação é um processo fundamental para transformar dados brutos em informações acessíveis e utilizáveis por diferentes públicos. Esse processo envolve atividades como catalogação, classificação, indexação e descrição de conteúdos, todas voltadas para facilitar a recuperação eficiente da informação. Conforme discutido por Sousa (2007):

“a classificação é uma operação matricial de todo o trabalho arquivístico e é, também, um elemento importante para a transparência e compartilhamento de informações, que são caminhos seguros para a tomada de decisão, para preservação da memória técnica e administrativa das instituições contemporâneas e para o pleno exercício da cidadania”. (Sousa, 2007, p. 5)

A descrição é processos complementar à classificação e têm como objetivo permitir que os usuários compreendam e localizem os documentos de forma rápida e precisa. A catalogação envolve a criação de registros padronizados que descrevem o conteúdo, a autoria, a data de criação e outras características de um documento, enquanto a descrição elabora resumos ou notas sobre o conteúdo, tornando a recuperação da informação mais eficiente.

O processo de indexação também é essencial para a recuperação eficiente da informação. Araújo Júnior e Tarapanoff (2006), demonstram que a indexação manual, quando realizada por especialistas, pode garantir uma recuperação mais precisa da informação, pois envolve a escolha cuidadosa de termos e descritores que refletem o conteúdo dos documentos. A mineração de textos pode complementar esse processo ao identificar automaticamente termos recorrentes em grandes volumes de documentos, mas a *expertise* humana continua sendo importante para garantir que a indexação seja adequada ao contexto informacional.

Apesar de a ferramenta constituir importante instrumento na identificação de palavras-chave, o indexador continua um dos principais artífices no processo de indexação, dada a sua competência na escolha de termos a serem usados para identificar o conteúdo dos documentos. (Araújo Júnior; Tarapanoff, 2006, p. 246)

A crescente complexidade e o volume exponencial de informações geradas diariamente criam desafios para sua gestão e organização. Nesse cenário, a organização da informação não apenas facilita a recuperação e o uso do conhecimento, mas também atua como um mecanismo essencial para garantir que a informação seja acessível de maneira eficaz e produtiva. Como parte integrante da Ciência da Informação, esse processo envolve um conjunto de práticas, como a classificação, catalogação e indexação, que buscam estruturar a informação em um formato que seja utilizável por diferentes públicos. O acúmulo constante de dados e documentos sem a devida organização pode levar a uma sobrecarga informacional, tornando as informações inatingíveis ou ineficazes. Essa preocupação já foi apontada por pioneiros da documentação, como Paul Otlet, que em 1934 refletiu sobre o impacto desse fenômeno.

"As expressões escritas das ideias, instrumento de sua fixação, conservação e circulação, são os intermediários obrigatórios de todas as relações entre os homens. Sua massa enorme, acumulada no passado, cresce a cada dia, a cada hora, com novas unidades em número desconcertante, por vezes alarmante. Desses documentos, assim como da linguagem, pode-se dizer que eles podem ser a melhor e a pior das coisas. Assim como a água caída do céu, podem causar inundações e dilúvios, ou espalhar-se como irrigação benéfica" tradução própria. (Otlet, 1934, p. 3)

O autor destaca a dualidade da informação: enquanto sua correta organização e gestão pode ser uma fonte rica e produtiva de conhecimento, sua acumulação desordenada pode gerar uma sobrecarga, dificultando o acesso e uso efetivo. A metáfora da água ilustra bem essa situação, onde a informação desorganizada pode inundar o sistema, tornando-se mais um obstáculo do que um recurso. Por outro lado, quando adequadamente estruturada e disseminada, a informação serve como uma "irrigação benéfica", fomentando o conhecimento e a inovação. Esse pensamento está diretamente ligado ao papel da organização da informação, que busca transformar imensos volumes de dados em algo manejável e útil, prevenindo os "dilúvios" de informações desordenadas.

No contexto da saúde pública, a organização da informação é especialmente crítica, devido à sensibilidade e ao grande volume de dados gerados diariamente, como prontuários médicos e resultados de exames. A complexidade desses dados exige o uso de técnicas de gestão de documentos, como classificação, e descrição, para garantir a acessibilidade e a proteção desses documentos. Como descrito por Sousa (2007), a classificação arquivística é uma operação essencial que sustenta o processo de recuperação de informações, permitindo a preservação da memória institucional e a tomada de decisões.

Além disso, o uso de vocabulários controlados e sistemas de indexação eficazes é fundamental para permitir a organização eficiente dos documentos em saúde pública ao possibilitar a padronização dos termos de indexação atribuídos ou extraídos dos documentos para serem usados na sua representação. Esses instrumentos garantem que os dados possam ser localizados de maneira rápida e precisa, otimizando o acesso à informação crítica, especialmente no âmbito da saúde, onde a prontidão dos dados pode salvar vidas.

No contexto da saúde pública, Richardson, Abramson e Kaushal (2012) discutem as barreiras e facilitadores para a troca de informações de saúde (*Health Information Exchange* - HIE) e como sistemas interoperáveis são essenciais para

melhorar a qualidade do atendimento e reduzir custos. Eles destacam que a implementação de HIE pode trazer benefícios, como a redução de erros médicos, a melhoria da coordenação do cuidado e a diminuição de exames duplicados. No entanto, a adoção desses sistemas ainda enfrenta desafios, como a falta de infraestrutura tecnológica adequada e a baixa adesão por parte de profissionais de saúde. Essas dificuldades ressaltam a necessidade de um planejamento estratégico eficiente na organização e na interoperabilidade dos dados de saúde, com o objetivo de garantir o compartilhamento seguro e eficaz das informações clínicas entre os diferentes níveis de atendimento.

Esses pontos reforçam a importância da classificação e organização da informação na saúde pública, pois permitem que dados críticos, como prontuários médicos e resultados de exames, sejam acessados de forma eficiente, promovendo melhorias significativas na qualidade do atendimento ao paciente.

Além da classificação, a indexação é uma prática essencial na organização da informação, especialmente em ambientes de saúde. A indexação atribui descritores ou palavras-chave aos documentos, facilitando sua recuperação por sistemas de busca. Um processo de padronização é vital para garantir a interoperabilidade entre sistemas e a integração dos dados de saúde, especialmente no contexto do Sistema Único de Saúde (SUS), que busca integrar os diferentes sistemas de informação utilizados em hospitais e clínicas em todo o Brasil. A indexação adequada, aliada ao uso de vocabulários controlados, melhora significativamente a acessibilidade e o uso eficaz dos dados, promovendo a melhoria da assistência à saúde.

Desde a criação do Sistema Único de Saúde (SUS), algumas estratégias vêm sendo adotadas para melhorar a comunicação e integração entre os SNIS. Por iniciativa da Organização Pan-americana de Saúde, foi criada no final da década de 1990 a Rede Intergerencial de Informações para a Saúde (Ripsa), que tinha como objetivo produzir informações relevantes e integradas para a tomada de decisão na saúde. (Neto; Andreazza; Chioro, 2021, p. 2)

A representação da informação é um conceito central na organização da informação, envolvendo a descrição de metadados e a forma como os dados são apresentados em um sistema de informação. No contexto da saúde, a representação precisa ser padronizada e rigorosa para garantir a interoperabilidade entre diferentes sistemas de informação e facilitar a recuperação precisa de dados pelos profissionais de saúde. Esquemas de metadados como o *Dublin Core* e o MARC 21 são

amplamente utilizados para essa padronização, contribuindo para que informações cruciais possam ser compartilhadas de maneira eficiente e segura.

No Brasil, a Portaria nº 2.073, de 31 de agosto de 2011, regulamenta o uso de padrões de interoperabilidade, destacando a importância da padronização dos metadados para a troca de informações no contexto do Sistema Único de Saúde (SUS). Essa padronização é central para a organização e o compartilhamento eficaz de dados dentro e fora de instituições de saúde, promovendo a consistência dos metadados e permitindo a integração entre sistemas distintos, como bibliotecas digitais e prontuários eletrônicos. Segundo Sales e Pinto (2019)

“A interoperabilidade semântica garante que os dados trocados tenham seu significado corretamente interpretado dentro do contexto de uma dada transação ou busca de informação, dentro da cultura, das convenções e das terminologias adotadas por cada setor ou organização e, assim, compartilhados pelas partes envolvidas”. (Sales; Pinto, 2019, p. 215)

Esses padrões garantem que dados como autor, data de criação e contexto sejam descritos de forma unificada, permitindo que os profissionais de saúde recuperem as informações de maneira rápida e precisa, essencial em um ambiente hospitalar onde decisões ágeis são frequentemente necessárias.

Para que a interoperabilidade possa ser realizada é necessário o uso de padrões comuns que assegurem a padronização dos metadados e a correta representação dos dados. No Brasil, a adoção do padrão TISS¹² (Troca de Informações em Saúde Suplementar) e de sistemas como o OpenEHR¹³ no contexto do Sistema Único de Saúde (SUS) reflete a busca por maior interoperabilidade e segurança no compartilhamento de informações sensíveis entre diferentes instituições de saúde. Essas iniciativas promovem uma gestão mais eficiente das informações e impactam diretamente a qualidade do atendimento ao paciente, ao permitir o compartilhamento rápido e preciso de dados entre os profissionais de saúde.

Frequentemente, a pouca integração exige que os mesmos dados sejam preenchidos em diferentes interfaces, gerando retrabalho e aumento de custos, pois se torna necessário manter várias soluções tecnológicas com funções redundantes. Essa duplicidade de registros tem impacto na própria

¹² O TISS é um padrão estabelecido pela ANS no Brasil para padronizar a comunicação e a troca de informações entre operadoras de planos de saúde e prestadores de serviços, garantindo interoperabilidade, transparência e conformidade regulatória no setor de saúde suplementar.

¹³ O OpenEHR é um padrão aberto para modelagem, armazenamento e interoperabilidade de dados clínicos eletrônicos, baseado em uma arquitetura de dois níveis (modelo de referência e arquetipos), permitindo a estruturação flexível e independente de sistemas, promovendo a longevidade e a reutilização das informações clínicas.

qualidade das bases de dados, dificultando as análises, os cruzamentos e impactando na confiabilidade da informação produzida. (Neto; Andrezza; Chioro, 2021, p. 2)

Os padrões TISS (Troca de Informações em Saúde Suplementar), *OpenEHR* e HL7 (*Health Level Seven*) são fundamentais para a padronização e interoperabilidade de sistemas de saúde no Brasil, contribuindo diretamente para a eficiência na troca de informações entre instituições, operadoras de saúde e profissionais. Cada um desses padrões possui características que os tornam relevantes no contexto da saúde pública e suplementar brasileira.

O TISS, regulamentado pela Agência Nacional De Saúde Suplementar - ANS (2006), é o padrão estabelecido para a troca de dados entre prestadores de serviços de saúde e operadoras de planos de saúde no Brasil. Ele padroniza a comunicação referente à faturação, serviços prestados e procedimentos realizados. A adoção desse padrão facilita a troca de informações administrativas e clínicas, promovendo transparência, agilidade no processamento de reembolsos e auditorias, além de melhorar a qualidade dos dados compartilhados. O TISS também contribui para a redução de fraudes e inconsistências nos registros, além de auxiliar na regulação do setor suplementar.

O OpenEHR é um padrão internacional aberto para a gestão de registros eletrônicos de saúde, que se destaca por sua flexibilidade e foco na interoperabilidade semântica. No Brasil, ele vem sendo progressivamente adotado em iniciativas públicas e privadas, pois permite que dados clínicos estruturados sejam compartilhados entre diferentes sistemas de saúde, facilitando a continuidade do cuidado. Um dos principais benefícios do OpenEHR é sua capacidade de garantir que as informações clínicas estejam disponíveis e compreensíveis, independentemente do sistema utilizado por diferentes instituições de saúde. Isso é particularmente relevante para a saúde pública, onde a integração de sistemas é um desafio constante.

(...) a interoperabilidade de software no domínio da saúde utilizando um padrão consistente como *Health Level Seven (HL7)* e *Fast Healthcare Interoperability Resources (FHIR)*. O HL7 é uma organização de desenvolvimento de padrões para troca, integração, compartilhamento e recuperação de informações de saúde. O FHIR criado pelo HL7, é um padrão que descreve formatos e elementos de dados para troca de registros eletrônicos de saúde interoperáveis. (Sembay; Macedo; Gimenes, 2022, p. 2)

O HL7 FHIR é um padrão utilizado no Brasil e no mundo para a troca eletrônica de informações entre sistemas de informação hospitalares. Ele estabelece formatos de mensagens para a comunicação entre sistemas de saúde, incluindo informações sobre prontuários eletrônicos, resultados de exames, prescrição de medicamentos e relatórios médicos. No contexto brasileiro, o HL7 FHIR é essencial para permitir a interoperabilidade entre hospitais, clínicas, laboratórios e sistemas de saúde pública, como o Sistema Único de Saúde (SUS). O uso do HL7 FHIR permite que sistemas distintos compartilhem informações de maneira eficiente e segura, garantindo que os dados sejam acessados e utilizados por diferentes profissionais da saúde em diversas etapas do atendimento ao paciente.

No Brasil, a adoção desses padrões visa atender à necessidade crescente de interoperabilidade entre sistemas que lidam com grandes volumes de dados clínicos e administrativos. O TISS desempenha um papel importante no setor de saúde suplementar, garantindo que informações sobre procedimentos, cobranças e auditorias sejam trocadas de forma padronizada. O OpenEHR e o HL7 são fundamentais no contexto do SUS e em sistemas de saúde privados, permitindo que diferentes sistemas hospitalares e clínicos se integrem para fornecer um atendimento contínuo e de qualidade. Ao promover a padronização, esses modelos garantem uma gestão mais eficiente da saúde, melhoram a qualidade do atendimento, facilitam a tomada de decisões e garantem a segurança e a privacidade dos dados dos pacientes.

A temporalidade e a destinação dos documentos estão entre os aspectos mais importantes da gestão de documentos, particularmente em contextos como o da saúde pública, onde a gestão de prontuários médicos deve seguir prazos rigorosos de guarda e critérios de descarte. O princípio da proveniência é crucial nesse contexto, pois assegura que os documentos sejam organizados de acordo com sua origem, garantindo que a entidade responsável mantenha o controle sobre o ciclo de vida dos documentos. A temporalidade define o tempo que um documento deve ser mantido, enquanto a destinação refere-se à decisão final sobre sua preservação ou descarte. No contexto da saúde, onde a informação pode ter valor legal e clínico, a correta aplicação desses conceitos garante a preservação da informação essencial e o descarte seguro de dados que não têm mais utilidade, em conformidade com a Lei Geral de Proteção de Dados (LGPD).

A mineração de dados é uma técnica complementar à organização da informação, especialmente útil para explorar grandes volumes de dados em busca de padrões ocultos e informações relevantes, particularmente no setor de saúde. Nessa seara conceitos como o de KDD e KDT são importantes e necessários, Araújo Júnior (2005) traz esses conceitos voltados para a gestão da informação.

Encontrar padrões úteis nestes dados sem tratamento adequado e, por isso mesmo aparentemente sem valor, tem sido intitulado prospecção, descoberta de conhecimento em bancos de dados, mineração de dados, descoberta de conhecimento em textos, mineração de textos ou conforme as siglas em inglês KDD para Knowledge Discovery in Databases e KDT para Knowledge Discovery in Text. (Araújo Júnior, 2005, p. 59)

Como destacado em diversas análises, a mineração de dados pode ser aplicada em sistemas de saúde para identificar tendências em diagnósticos, prever surtos de doenças e melhorar a gestão do atendimento ao paciente Muniz (2007), Ruas e Maciel (2023). Ela oferece uma abordagem eficiente para processar dados clínicos, fornecendo insights que auxiliam na tomada de decisões clínicas e na administração hospitalar.

Para que essa técnica seja eficaz, é crucial que os dados estejam bem-organizados e estruturados. Isso ocorre porque a mineração de dados se baseia na qualidade e na precisão dos dados subjacentes para gerar resultados úteis. Portanto, a organização da informação se torna essencial como um pré-requisito para a mineração de dados, assegurando que os dados disponíveis estejam completos, corretos e prontos para análise Galvão e Marin (2009). Quando integrada às práticas tradicionais de organização documental, a mineração de dados pode transformar sistemas de saúde em ferramentas mais robustas e adaptáveis às necessidades dinâmicas do setor.

Além disso, a utilização de algoritmos de aprendizado de máquina em conjunto com a mineração de dados, como redes neurais e aprendizagem supervisionada, aprimora a capacidade de análise preditiva, permitindo que profissionais de saúde tomem decisões baseadas em dados concretos e previsões confiáveis afirma Fernandes e Chiavegatto Filho (2019).

Esta seção procurou discutir que a organização da informação é um processo complexo e multidimensional que envolve a aplicação de princípios e

técnicas variados para garantir a acessibilidade, segurança e integridade dos dados. No contexto da saúde pública, esses fundamentos se tornam ainda mais críticos, dada a natureza sensível e o volume de informações geradas. A integração dos princípios arquivísticos com as tecnologias de informação modernas, como a inteligência artificial e a visão computacional, é essencial para alcançar uma gestão eficaz e ética dos dados de saúde.

4.1.1 Larga escala

A evolução dos sistemas de gerenciamento de dados, iniciada com os trabalhos pioneiros de Codd (1970), culminou na necessidade de lidar com volumes de dados cada vez maiores. O termo "larga escala" encapsula a complexidade de gerenciar essas quantidades massivas de informação, que excedem as capacidades dos sistemas tradicionais. Atualmente, como apontam Dean e Ghemawat (2004) o processamento paralelo e distribuído se tornou fundamental para lidar com esses desafios. A ciência de dados, conforme definida por Domingos (2017), emerge como um campo interdisciplinar que busca extrair conhecimento e insights a partir desses vastos conjuntos de dados, utilizando técnicas e algoritmos avançados, como os apresentados por Leskovec, Rajaraman e Ullman (2014). No contexto brasileiro, estudos da Associação Brasileira De Ciências (2020) destacam que as organizações nacionais também enfrentam os desafios de lidar com grandes volumes de dados e apontam a necessidade de adaptar as estratégias de gestão de dados para aproveitar as oportunidades oferecidas pelo *Big Data*. A disponibilidade de grandes volumes de dados fornece um incentivo à busca por novos procedimentos e ferramentas computacionais para obtenção, armazenamento, organização e análise desses dados, podendo trazer aperfeiçoamentos na forma de produzir e de maior transparência a todo o processo de criação do conhecimento científico.

Segundo Kitchin e Mcardle (2016), os dados em larga escala que já são amplamente caracterizados na literatura como 3Vs: volume, velocidade e variedade também devem ser caracterizados pela exaustividade (capacidade de capturar um volume amplo e completo de dados), resolução (nível de detalhamento dos dados coletados, identificar padrões finos e aspectos sutis nos dados), indexicalidade (habilidade de identificar dados de forma única e contextual), relacionalidade (capacidade de relacionar diferentes fontes de dados entre si), extensionalidade (possibilidade de expandir a coleta e a análise de dados a novas áreas ou tópicos de

interesse) e escalabilidade (capacidade de expandir o processamento e armazenamento de dados em grande escala). Esses atributos implicam desafios significativos para a gestão de documentos, demandando sistemas capazes de lidar com grandes volumes de dados que são gerados e transmitidos em alta velocidade e em diferentes formatos. No contexto da saúde pública, esses desafios são ainda mais acentuados devido à sensibilidade e à diversidade dos dados envolvidos.

A falta de integração entre suportes físicos e digitais cria lacunas informacionais que podem comprometer a eficiência e a qualidade dos serviços de saúde.

Existem numerosos relatos de empresas que tiveram que reinventar a roda desnecessariamente, por não serem capazes de localizar o conhecimento especializado que existia em algum lugar dentro da organização. Outro obstáculo ao aprendizado é que as organizações acham difícil desaprender o que aprenderam no passado: reexaminar crenças herdadas e questionar as práticas existentes como as únicas alternativas viáveis. (Choo, 2003, p. 28)

Em suma a incapacidade de acessar e integrar informações de diferentes fontes pode levar a decisões inadequadas e a oportunidades perdidas de melhorar processos e resultados. No ambiente hospitalar, onde registros físicos (como prontuários em papel) coexistem com sistemas digitais, a falta de interoperabilidade pode resultar em informações incompletas ou desatualizadas, afetando negativamente o cuidado ao paciente.

Para sanar essas lacunas, a gestão da informação em larga escala propõe a implementação de sistemas integrados que permitam a unificação de dados provenientes de múltiplas fontes. De acordo com Davenport e Prusak (1998) “a única vantagem sustentável que uma empresa possui, é aquilo que ela coletivamente sabe, a eficiência com que ela usa o que sabe e a prontidão com que ela adquire e usa novos conhecimentos”. Assim a capacidade de coletar, armazenar e disseminar informações de forma integrada é essencial para que as organizações possam aproveitar plenamente o valor dos dados que possuem. A adoção de tecnologias como *big data analytics*, *cloud computing* e inteligência artificial facilita o processamento e a análise de grandes volumes de dados, promovendo insights que podem melhorar a tomada de decisões e a eficiência operacional.

Além disso, a implementação de padrões e protocolos de interoperabilidade, como o HL7 e o FHIR (*Fast Healthcare Interoperability Resources*),

é fundamental para permitir a integração entre sistemas heterogêneos. A utilização de padrões abertos e amplamente adotados é crucial para facilitar a troca de informações de saúde em larga escala, promovendo a interoperabilidade entre diferentes sistemas e plataformas.

A gestão de documentos em larga escala também requer uma abordagem estratégica que considere não apenas as tecnologias envolvidas, mas também os processos e as pessoas. As organizações que obtêm sucesso na era dos *big data* são aquelas que alinham tecnologia, gestão e cultura organizacional para criar valor a partir dos dados. Isso implica investir em capacitação profissional, desenvolver políticas de governança da informação e promover uma cultura orientada por dados. Como ressaltam McAfee e Brynjolfsson (2012):

"As empresas não colherão todos os benefícios de uma transição para o uso de *big data* se não conseguirem gerenciar as mudanças de forma eficaz. Cinco áreas são particularmente importantes nesse processo. Liderança. As empresas obtêm sucesso na era do *big data* não apenas porque possuem mais ou melhores dados, mas porque contam com equipes de liderança que estabelecem metas claras, definem o que é sucesso e fazem as perguntas certas. O poder do *big data* não elimina a necessidade de visão ou de percepção humana. [...]" tradução própria (McAfee; Brynjolfsson, 2012, p. 8).

No contexto específico da integração entre suportes físicos e digitais, iniciativas de digitalização de documentos e a implementação de sistemas de gestão eletrônica de documentos (GED) são passos importantes para reduzir as lacunas informacionais. A digitalização e a gestão eletrônica de documentos permitem não apenas a preservação da informação, mas também sua disponibilidade e acessibilidade em larga escala, independentemente das limitações físicas dos suportes tradicionais.

As ferramentas digitais por muito tempo foram utilizadas pelos arquivos para a preservação de documentos, garantindo o acesso à informação documental, disponibilizando para consulta o representante digital e restringindo o manuseio do documento original não digital (Carmo; Chagas, 2024, p. 3).

É importante reconhecer que a gestão de documentos em larga escala não se limita ao uso de tecnologias avançadas, mas exige uma abordagem holística que incorpore processos eficazes, políticas de governança robustas e uma cultura organizacional voltada para o uso estratégico dos dados, a fim de apoiar o processo decisório tal como em um modelo de gestão de dados. A integração entre suportes físicos e digitais, juntamente com a adoção de padrões de interoperabilidade, permite a criação de um ecossistema informacional mais eficiente e acessível, especialmente

em áreas sensíveis como a saúde pública. O alinhamento entre tecnologia, processos e pessoas é essencial para maximizar o valor dos dados e promover uma tomada de decisão informada e baseada em evidências. Com essa visão integrada, as organizações estarão mais preparadas para enfrentar os desafios impostos pelos grandes volumes de dados, potencializando sua capacidade de inovação e eficiência operacional.

4.2 Fundamentos da visão computacional

A visão computacional é um campo interdisciplinar que se dedica ao estudo de como os computadores podem ser programados para obter um entendimento de alto nível a partir de imagens digitais ou vídeos. Em termos mais gerais, a visão computacional procura automatizar tarefas que o sistema visual humano pode realizar, sendo uma área essencial dentro da inteligência artificial (IA) que possibilita a interpretação e análise de informações visuais.

A visão computacional tem suas raízes na década de 1960, quando pesquisadores começaram a explorar maneiras de processar e analisar imagens digitais utilizando algoritmos, com destaque para o estudo seminal de Roberts (1963). Inicialmente, o foco estava na percepção de profundidade e na construção de modelos tridimensionais a partir de imagens bidimensionais, utilizando etapas como filtragem de bordas e segmentação como ferramentas fundamentais no processo. Com o avanço da capacidade computacional e a disponibilidade de grandes volumes de dados, o campo evoluiu significativamente. A partir da década de 2000, a introdução de técnicas de aprendizado profundo (*deep learning*) revolucionou o campo, permitindo que sistemas de visão computacional atingissem níveis de precisão e eficiência comparáveis ou até superiores aos humanos em certas tarefas.

Uma das abordagens mais influentes na visão computacional moderna é o uso de redes neurais convolucionais (*Convolutional Neural Networks - CNNs*). CNNs são modelos de aprendizado profundo projetados especificamente para o processamento de dados com uma grade topológica, como imagens. Essas redes têm a capacidade de capturar padrões espaciais e hierárquicos em dados visuais, o que as torna extremamente eficazes para tarefas como classificação de imagens, detecção de objetos e reconhecimento facial. As redes convolucionais, são um tipo de rede neural artificial especializada no processamento de dados com estrutura em

forma de grade, como imagens. São denominadas "convolucionais" devido ao uso da operação matemática de convolução, que consiste na aplicação de um filtro ou *kernel*¹⁴ aos dados de entrada para extrair características relevantes.

Essas redes são particularmente eficazes em tarefas relacionadas à visão computacional, como o reconhecimento de imagens, devido à sua capacidade de identificar padrões visuais complexos. Em uma rede convolucional típica, os dados de entrada, por exemplo, uma imagem, passam por uma sequência de camadas, incluindo camadas de convolução, camadas de *pooling* (agrupamento), e, eventualmente, camadas totalmente conectadas. As camadas de convolução aplicam filtros que detectam características como bordas, texturas ou formas. As camadas de *pooling* reduzem a dimensionalidade dos dados, diminuindo o custo computacional e melhorando a eficiência na detecção de características. Finalmente, as camadas totalmente conectadas interpretam essas características extraídas para realizar tarefas como classificação ou detecção de objetos. Na área médica, as redes convolucionais já são estudadas para utilizar na análise imagens médicas, como radiografias, ressonâncias magnéticas ou tomografias computadorizadas. Por exemplo, uma CNN pode ser treinada para detectar a presença de tumores em imagens de mamografias, contribuindo significativamente para diagnósticos mais precisos e rápidos.

Neste trabalho foi apresentado um modelo para classificação e diagnóstico da doença de Alzheimer por meio da aplicação da técnica de Rede Neural Convolucional, com está sendo aplicada a uma base de dados de ressonância magnética. Para tanto, foi selecionada a base de dados MIRIAD, extraídas e selecionadas as fatias axiais dos SCANS que representam o cérebro. (Silva *et al.*, 2018)

Outro aspecto fundamental da visão computacional é a segmentação de imagens, que envolve a divisão de uma imagem em partes ou regiões, cada uma das quais representa um objeto ou uma parte de um objeto dentro da cena. Técnicas de segmentação evoluíram significativamente, com métodos como a Segmentação Semântica, que atribui rótulos a cada pixel da imagem com base na classe do objeto, sendo amplamente utilizados em aplicações médicas e automotivas.

"Primeiramente, precisamos saber quais áreas da imagem pertencem ao objeto, o que é a tarefa de segmentação de imagens. A segmentação de imagens tem sido um problema fundamental e desafiador na visão

¹⁴ É uma matriz de pesos pequenos usada para detectar padrões específicos nos dados de entrada, como imagens ou séries temporais.

computacional por décadas. A segmentação pode ser baseada em semelhanças espaciais e continuidades" (Wu, 2017, p. 7).

A visão computacional também se beneficia enormemente dos avanços em processamento de linguagem natural, especialmente em aplicações que envolvem a análise conjunta de imagens e textos, como na anotação automática de imagens médicas. Nesse contexto, sistemas de visão computacional são treinados para associar padrões visuais com descrições textuais, facilitando a criação de sistemas de recuperação de informações visuais que podem ser usados, por exemplo, na busca por imagens médicas relevantes com base em consultas textuais.

É notável o uso de deep learning como técnica dominante para a extração de informação visual, o que não representa uma surpresa visto que desde Lecun (LeCun et al., 1989) e Krizhevsky (Krizhevsky et al., 2012), o uso de redes neurais convolucionais (CNNs) passou a ser a principal tecnologia adotada (Constâncio; Carvalho; Tsunoda, 2022, p. 12).

No campo da saúde, a visão computacional tem sido amplamente aplicada para aprimorar a precisão diagnóstica, otimizar processos e reduzir a carga de trabalho dos profissionais de saúde. Por exemplo, a detecção automática de características em imagens de exames, como raios-X e ressonâncias magnéticas, auxilia radiologistas na identificação de anomalias, diminuindo o tempo necessário para análise e aumentando a acurácia do diagnóstico. Além disso, em situações de pandemias, como a COVID-19, a visão computacional pode ser usada para analisar grandes volumes de imagens pulmonares, auxiliando na triagem rápida de pacientes com sintomas graves.

Conforme destacado por Tabik *et al.* (2020), a metodologia COVID-SDNet foi desenvolvida para prever a COVID-19 com base em imagens de raios-X de tórax, demonstrando a eficácia de técnicas de aprendizado profundo na identificação da doença. A Sociedade Paulista de Radiologia e Diagnóstico por Imagem (SPR¹⁵) enfatiza que a integração de técnicas de processamento de linguagem natural com visão computacional abre as portas para uma compreensão mais abrangente e precisa das informações contidas nos registros radiológicos. Segundo Bobba, *et al.*, (2023) A radiologia, centrada em imagens, utiliza amplamente textos para comunicação, tornando-se ideal para aplicações de PNL, evidenciado pelo triplo de publicações radiológicas sobre o tema nos últimos 5 anos. Essas pesquisas

¹⁵ <https://noticias.spr.org.br/conexao-digital/aplicacoes-do-processamento-de-linguagem-natural-na-radiologia/>

confirmam que a visão computacional tem sido aplicada de diversas maneiras no campo da saúde para aprimorar a precisão diagnóstica, otimizar processos e reduzir a carga de trabalho dos profissionais de saúde, inclusive em situações de pandemia como a COVID-19.

Uma técnica que tem potencializado essas aplicações é o aprendizado por transferência (*transfer learning*), que permite a reutilização de modelos previamente treinados em grandes conjuntos de dados para novas tarefas específicas, mesmo quando há disponibilidade limitada de dados rotulados no novo domínio. Por exemplo, modelos treinados em imagens gerais podem ser adaptados para a detecção de doenças específicas em imagens médicas, otimizando o desempenho diagnóstico. O ajuste fino (*fine-tuning*) é uma etapa crucial nesse processo, em que o modelo pré-treinado é refinado com dados específicos da nova tarefa, ajustando seus parâmetros para melhorar a precisão e a relevância dos resultados. Essa abordagem tem sido fundamental para o desenvolvimento de sistemas de visão computacional na saúde, permitindo funcionalidades avançadas que anteriormente eram inviáveis devido a limitações de dados ou recursos computacionais.

O aprendizado de transferência demonstrou ser uma técnica eficaz para alavancar modelos de CNN pré-treinados para melhorar o desempenho das tarefas de análise de imagens médicas. As CNNs demonstraram alta precisão e robustez na identificação e classificação de várias condições médicas a partir de imagens médicas. No geral, essas descobertas ressaltam o papel crucial que o aprendizado de transferência e as CNNs desempenham no avanço do diagnóstico por imagens médicas, e novas pesquisas nessa área têm o potencial de melhorar significativamente os resultados dos pacientes
Tradução própria. (Salehi *et al.*, 2023, p. 11)

Entretanto, desafios significativos ainda persistem na adoção ampla da visão computacional em aplicações de saúde. Um dos maiores desafios é a necessidade de grandes volumes de dados anotados, classificados e indexados para treinar modelos de aprendizado profundo. Além disso, questões de privacidade e segurança dos dados são críticas, especialmente quando se trabalha com dados sensíveis com os de saúde. Garantir a conformidade com regulamentações como a LGPD, que exigem proteção rigorosa dos dados pessoais, é essencial para a implementação dessas tecnologias em ambientes clínicos.

Por fim, a ética na aplicação de visão computacional não pode ser ignorada. A automação na análise de dados médicos por sistemas de IA levantam questões sobre a responsabilidade e a tomada de decisões. É essencial que essas tecnologias

sejam desenvolvidas e implementadas com transparência, garantindo que os profissionais mantenham o controle e que as decisões sejam feitas com base em informações precisas e éticas.

Esta seção apresentou os fundamentos da visão computacional, destacando sua evolução, principais técnicas e aplicações na área da saúde. Ao integrar essas tecnologias com a gestão de informações em saúde, esta pesquisa explora novos horizontes para a organização e recuperação de dados críticos, promovendo uma abordagem mais eficaz e segura na administração dos serviços de saúde.

4.3 Ciências da Saúde e Ciência da Informação

A Ciência da Informação tem buscado estabelecer diálogos interdisciplinares com diversos campos do conhecimento, destacando-se particularmente as Ciências da Saúde. Este campo tem como objetivo fundamental a compreensão das condições de saúde dos indivíduos, abordando aspectos tanto da saúde quanto da doença ao longo da vida. A integração entre a Ciência da Informação e as Ciências da Saúde é fundamental para aprimorar a comunicação dos serviços de saúde pública com a população. Estudos indicam que falhas na comunicação em hospitais públicos impactam negativamente o acesso e a qualidade do atendimento.

Uma política pública para enfrentamento dos vários problemas de saúde, seja de morbidade ou de mortalidade, necessita de uma base de informações confiável, que sustente e direcione a tomada de decisão. A identificação dos determinantes do processo saúde-doença, das desigualdades em saúde e do impacto de ações e programas para reduzir a carga de doença na população só é possível a partir de boas informações e no momento oportuno. (Souza, 2008, p. 5-6)

A Ciência da Informação e as Ciências da Saúde, se unem no uso de tecnologias de mineração de textos¹⁶ aplicadas ao campo da saúde. Os prontuários médicos, além de serem ferramentas essenciais para o acompanhamento clínico, representam documentos de grande importância arquivística. Eles contêm informações detalhadas sobre o histórico médico do paciente, intervenções clínicas, diagnósticos e tratamentos, servindo como base para a continuidade do cuidado e para pesquisas científicas. A gestão eficiente desses documentos é crucial para

¹⁶ Mineração de textos é um conjunto de técnicas de Processamento de Linguagem Natural e aprendizado de máquina utilizadas para extrair informações estruturadas de grandes volumes de dados textuais.

garantir a acessibilidade, integridade e confidencialidade das informações. A integração de tecnologias de mineração de textos permite a extração de informações relevantes de grandes volumes de dados textuais, otimizando a recuperação de informações e apoiando a tomada de decisões clínicas e administrativas.

(...) a possibilidade de extrair, de uma montanha de textos, informação útil às demandas torna-se um efetivo instrumento de gestão em bases textuais. Contudo, a informação útil não é dada de forma automática com a mineração de textos, mas por meio da interpretação que for dada aos resultados obtidos. (Araújo Júnior; Tarapanoff, 2006, p. 239)

Reconhecer a importância da documentação sanitária e a necessidade de aprimorar os métodos de gestão e análise desses registros é um passo inicial muito importante. A aplicação de tecnologias avançadas, como a mineração de textos, pode proporcionar benefícios significativos na organização, acessibilidade e utilização dos dados contidos nos prontuários médicos, promovendo uma melhoria contínua na qualidade dos serviços de saúde. Nesse sentido, ao ampliar o uso de ferramentas tecnológicas, não apenas se facilita a recuperação da informação, mas também se reforça o compromisso ético de garantir a confidencialidade e a privacidade do paciente. Como enfatizam Araújo e Mota:

Sendo interpretado que o PP, dada as características intrínsecas e dadas as informações registradas, se caracteriza como uma fonte primária de informação, sendo assim, para que se possa ter acesso às informações dos pacientes por meio de seu prontuário, se faz necessária a anuência do mesmo, ou de seu representante legal, conforme orienta a Resolução do Conselho Nacional de Saúde – CNS Resolução n. 466 de 2012. E, como se trata de uma fonte primária de informação, o mesmo contém dados primários, pois ainda não foram devidamente tratados esses dados, conforme aponta a literatura sobre o assunto na área da Ciência da Informação. (Araújo; Mota, 2020, p. 62).

Dessa forma, a interlocução entre as Ciências da Saúde e a Ciência da Informação, bem como o uso de tecnologias de mineração de textos, visão computacional e processamento de linguagem natural mostra-se fundamental para a gestão adequada dos documentos em saúde. A compreensão da informação, sua organização, análise e aplicação prática para a tomada de decisões evidenciam não apenas o valor estratégico dessas ferramentas, mas também o compromisso ético envolvido. Nesse sentido, torna-se indispensável aprofundar o estudo sobre o prontuário do paciente – fonte primária de dados –, instrumento essencial para garantir a qualidade, a continuidade e a confidencialidade das informações de saúde.

Além disso, a aplicação de taxonomias na saúde pública reforça a relação entre a Ciência da Informação e as Ciências da Saúde ao proporcionar um modelo estruturado para a organização e recuperação da informação. Conforme discutido por Mota (2022), a categorização sistemática dos dados em saúde contribui para a acessibilidade e a interoperabilidade das informações, facilitando sua utilização tanto para a gestão de documentos quanto para a pesquisa e a formulação de políticas públicas. A adoção dessas abordagens no contexto dos prontuários médicos evidencia como princípios da Ciência da Informação podem aprimorar a eficiência da gestão em saúde, garantindo maior padronização e confiabilidade dos registros clínicos.

4.4 Prontuário

Definido pelo Conselho Federal de Medicina (CFM) como:

(...) o documento único constituído de um conjunto de informações, sinais e imagens registradas, geradas a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência a ele prestada, de caráter legal, sigiloso e científico, que possibilita a comunicação entre membros da equipe multiprofissional e a continuidade da assistência prestada ao indivíduo (Conselho Federal de Medicina, 2002).

O prontuário do paciente é um documento de natureza multidisciplinar, abrangente e temporal, devido aos diversos profissionais que registram e acessam suas informações. Segundo Bentes Pinto (2014, p. 36), Bentes Pinto (2005), "o prontuário contém registradas todas as informações concernentes à condição de saúde de uma pessoa, desde o seu nascimento até a sua morte". Portanto, trata-se de um documento que inclui dados clínicos e não clínicos, de natureza sensível, protegidos por legislações nacionais e internacionais.

Esse documento serve como memória escrita das condições de saúde de uma pessoa, sendo indispensável para a comunicação entre a equipe de saúde e entre esta e o paciente. Além disso, é crucial para a continuidade, segurança, eficácia e qualidade do tratamento e acompanhamento do paciente, bem como para a gestão das organizações de saúde. Galvão (2024, p. 54) afirmam que "o prontuário do paciente é um documento informacionalmente complexo quanto à forma de produção, quanto ao conteúdo que abarca, quanto à forma de organização, e quanto à forma de acesso e disponibilização".

Para que um prontuário, seja ele em suporte papel ou eletrônico, tenha melhor qualidade, é necessário um planejamento institucional e um trabalho

cooperativo e permanente. Isso inclui a sistematização dos aspectos informacionais relacionados ao prontuário, garantindo que todas as informações referentes à saúde e histórico de um indivíduo estejam contidas em um único documento. Este deve ser universal e acessível a quem tem permissão legal, visando diferentes objetivos, incluindo pesquisas médicas.

Bentes Pinto (2005) também ressalta que o prontuário do paciente é um conjunto de documentos que "agregam em único documento midiático desde as narrativas da anamnese dos pacientes até as imagens capturadas pelos dispositivos de ponta, que fotografam nosso corpo". Independentemente do suporte do prontuário (analógico ou digital), ele é indispensável para a comunicação intra e entre a equipe de saúde e o paciente, garantindo a continuidade, segurança, eficácia e qualidade do tratamento, além da gestão hospitalar.

Com outras palavras, nesse documento são registradas "todas as informações relativas a uma pessoa doente e as ações de cuidados e tratamentos a ela dispensados, a fim de que seja possível gerenciar o curso da patologia identificando os sintomas, causas e os remédios para solucioná-los". São ricas fontes de informação e de comunicação, tanto para a equipe de saúde e os pacientes, como também para os estudos e pesquisas (BENTES PINTO; TABOSA; VIDOTTI, 2011, p. 2636).

A tecnologia digital tem avançado para facilitar o uso e a manutenção desses documentos, mas ainda é necessário mais estudo devido às diferentes formas de acesso, meios de suporte e complexidade das informações registradas.

Para essa pesquisa o prontuário é compreendido como um documento arquivístico, constituído por um conjunto de informações médicas e administrativas produzidas no curso da relação entre paciente e unidades de saúde. Trata-se de um documento complexo, porque agrega narrativas clínicas, exames, imagens e registros administrativos em diferentes suportes, mas que se unifica pela sua função probatória, legal e científica. Assim, o prontuário não é apenas um instrumento técnico de registro, mas um objeto informacional que carrega valores de autenticidade, integridade e confiabilidade, sendo essencial para a continuidade do cuidado, para a gestão hospitalar e para pesquisas em saúde. Para esta dissertação, portanto, o prontuário será tratado como um documento informacionalmente denso, cuja adequada organização e indexação são indispensáveis para garantir acesso seguro, interoperável e em conformidade com a LGPD.

4.4.1 Regulamentação do prontuário médico

O prontuário médico é um documento essencial na prática clínica, funcionando como um registro detalhado de todas as informações relativas ao cuidado e ao histórico de saúde de um paciente. No Brasil, a regulamentação do prontuário médico é regida por uma série de normas e diretrizes que visam garantir a proteção dos dados do paciente, a padronização da documentação e a segurança na gestão das informações. A regulamentação tem como base princípios éticos, legais e técnicos que envolvem a privacidade, a integridade e a acessibilidade das informações registradas. Esta seção aborda, de forma detalhada, as principais regulamentações aplicáveis ao prontuário médico no Brasil, com ênfase na Lei Geral de Proteção de Dados (LGPD), no Código de Ética Médica, nas resoluções do Conselho Federal de Medicina (CFM) e nas normas específicas para o prontuário eletrônico.

O prontuário é um documento de caráter sigiloso, científico e legal, sendo de responsabilidade do profissional médico ou da instituição de saúde que o mantém. A regulamentação brasileira como a Resolução nº 1.638 Conselho Federal de Medicina (2002) enfatiza que o prontuário médico deve ser completo, contendo todas as informações relevantes para o acompanhamento clínico do paciente, desde os sintomas iniciais até o tratamento e evolução clínica.

A Resolução nº 1.821/2007 Conselho Federal de Medicina (2007), posteriormente modificada pela Resolução CFM nº 2.218/2018 Conselho Federal de Medicina (2018), introduziu importantes parâmetros técnicos e normativos para a digitalização, guarda e manuseio informatizado dos prontuários médicos, bem como para a transferência segura de informações em saúde. Essas disposições ampliam a segurança jurídica e a confiabilidade dos documentos em formato eletrônico, garantindo que os registros clínicos sejam armazenados e disponibilizados de forma adequada, integrando-se às obrigações estabelecidas por normas anteriores, como a Resolução CFM nº 1.638/2002. Assim, ao estabelecer requisitos mínimos de segurança, autenticação e confidencialidade, a normativa contribui para a consolidação de um sistema de registros médicos mais efetivo, coeso e atualizado, reforçando a importância da proteção dos dados do paciente e da continuidade do cuidado.

O sigilo das informações contidas no prontuário médico é um dos pilares da regulamentação. O Código de Ética Médica Conselho Federal de Medicina (2018) estipula que o médico deve manter o sigilo sobre as informações obtidas durante a relação médico-paciente, salvo em situações em que exista uma obrigação legal ou uma autorização expressa do paciente para a divulgação. O descumprimento desse princípio pode levar a sanções éticas e legais, incluindo a responsabilização civil e criminal do profissional.

Além do Código de Ética Médica, a Lei nº 13.709/2018 Brasil (2018), conhecida como Lei Geral de Proteção de Dados (LGPD), impôs novas exigências para a proteção dos dados pessoais sensíveis, incluindo informações de saúde. A LGPD estabelece que os dados de prontuários médicos devem ser tratados de forma a garantir sua segurança e privacidade, com acesso restrito a profissionais de saúde autorizados e proteção contra vazamentos ou acessos não autorizados. A lei também reforça o direito do paciente de acessar seus dados e de solicitar correções em caso de informações incorretas ou desatualizadas.

O paciente tem direito de acesso ao seu prontuário médico a qualquer momento, conforme estabelece o Código de Defesa do Consumidor (CDC) e a Resolução CFM nº 1.821/2007. Esse direito inclui tanto o acesso às informações registradas em papel quanto em formato eletrônico. A instituição de saúde ou o médico responsável pelo prontuário deve fornecer uma cópia completa ao paciente, se solicitado, garantindo que o paciente possa revisar suas informações de saúde, compartilhar com outros profissionais de saúde e tomar decisões informadas sobre o seu tratamento.

No entanto, o acesso ao prontuário por terceiros só pode ocorrer mediante autorização expressa do paciente ou por determinação judicial. Em casos de menoridade, incapacidade ou falecimento do paciente, o acesso pode ser concedido a responsáveis legais ou familiares, respeitando as orientações previstas em lei. A LGPD reforça a necessidade de transparência e controle sobre os dados pessoais, exigindo que as instituições informem claramente aos pacientes sobre como seus dados estão sendo tratados e armazenados.

O Prontuário Eletrônico do Paciente (PEP) é uma ferramenta moderna que facilita a organização e o acesso às informações de saúde de forma segura e eficiente.

A Resolução CFM nº 1.821/2007, estabelecendo que PEP deve seguir os mesmos princípios éticos e legais aplicáveis ao prontuário em papel. Além disso, o PEP deve garantir a autenticidade, a integridade, a confidencialidade e a disponibilidade dos dados, utilizando tecnologias como criptografia, assinatura digital e sistemas de *backup*.

Uma das vantagens do PEP é a possibilidade de integrar diferentes sistemas de informação em saúde, facilitando a interoperabilidade e a troca de informações entre diferentes instituições e profissionais de saúde. Para tanto, é necessário que o PEP esteja em conformidade com padrões internacionais de interoperabilidade, como o HL7 e o FHIR, que garantem a uniformidade na troca de dados entre sistemas distintos. A regulamentação também exige que os sistemas de PEP possuam logs de acesso, permitindo o rastreamento de quem acessou as informações e quando, aumentando assim a segurança e a transparência na gestão dos dados.

A regulamentação sobre o armazenamento e a preservação do prontuário médico é clara quanto à responsabilidade das instituições de saúde em manter esses documentos em condições adequadas de conservação e segurança. A Resolução CFM nº 1.821/2007 estabelece que os prontuários em suporte físico devem ser guardados em local seguro, protegido contra fatores que possam causar sua deterioração, como umidade, calor e infestação por pragas. No caso dos prontuários eletrônicos, é necessário que sejam adotadas medidas técnicas de proteção, como criptografia, para assegurar a integridade e a confidencialidade dos dados armazenados.

A temporalidade do prontuário médico, ou seja, o período em que ele deve ser mantido antes de ser descartado, também é regulamentada. A legislação brasileira exige que os prontuários médicos sejam preservados por um período mínimo de 20 anos após o último registro, conforme determina a Resolução CFM nº 1.821/2007. Após esse período, a destruição dos documentos deve ser realizada de forma segura, garantindo que nenhuma informação sensível seja recuperada. O Conselho Nacional de Arquivos (CONARQ) também estabelece diretrizes para a preservação de documentos arquivísticos, incluindo os prontuários médicos, reforçando a importância de se manter a documentação em condições adequadas para consulta futura.

Art. 4º - Estabelecer o prazo mínimo de 20 (vinte) anos, a partir do último registro, para a preservação dos prontuários médicos em suporte de papel.
Parágrafo único – Findo o prazo estabelecido no caput, e considerando o valor secundário dos prontuários, a Comissão Permanente de Avaliação de Documentos, após consulta à Comissão de Revisão de Prontuários, deverá elaborar e aplicar critérios de amostragem para a preservação definitiva dos documentos em papel que apresentem informações relevantes do ponto de vista médico-científico, histórico e social (CFM, Resolução nº 1.639/2002, art. 4º).

Mesmo a Resolução nº 1.639/2002 tendo sido revogada pela Resolução CFM nº 1.821/2007 é importante observarmos o que ela trazia a respeito dos valores primários e secundários do prontuário para poder confrontar como isso foi levado para a atualização da resolução.

Art. 8º Estabelecer o prazo mínimo de 20 (vinte) anos, a partir do último registro, para a preservação dos prontuários dos pacientes em suporte de papel, que não foram arquivados eletronicamente em meio óptico, microfilmado ou digitalizado (CFM, Resolução nº 1.821/2007, art. 8º).

Art. 9º As atribuições da Comissão Permanente de Avaliação de Documentos em todas as unidades que prestam assistência médica e são detentoras de arquivos de prontuários de pacientes, tomando como base as atribuições estabelecidas na legislação arquivística brasileira, podem ser exercidas pela Comissão de Revisão de Prontuários (CFM, Resolução nº 1.821/2007, art. 9º).

Além de acompanhar o processo evolutivo da resolução, é relevante observarmos na prática como essa atualização é interpretada e aplicada. Para isso podemos observar o parecer 6/2015 que reforça o objetivo prático das resoluções sendo mantida a mesma aplicação mesmo o texto tendo sido escrito de forma distinta entre as duas resoluções.

2ª consulta – Resposta: A Resolução CFM nº 1.639/02 foi revogada pela Resolução CFM nº 1.821/07. Esta autoriza a eliminação do suporte em papel de prontuários médicos, quando microfilmados ou digitalizados, decorrido o prazo mínimo de 20 (vinte) anos do último registro, salvo os definidos pela Comissão Permanente de Avaliação de Documentos da instituição detentora do arquivo como de valor médico-científico, histórico e social, cuja manutenção do suporte em papel é permanente (CFM, Parecer nº 6/15, 2015).

Assim, ao se considerar a temporalidade dos prontuários, não se pode deixar de lado o valor secundário que tais documentos podem ter: médico-científico, histórico ou social. Esse valor secundário quer dizer que mesmo após o prazo legal mínimo de guarda em suporte físico, certos prontuários sejam preservados permanentemente ou submetidos a critérios seletivos para guarda permanente. A Resolução do CFM já prevê isso, preconizando que, caso o prontuário apresente relevância do ponto de vista médico-científico, histórico ou social, sua guarda possa

ser estendida ou permanente. Esse critério de avaliação (valor secundário) deve, portanto, compor o planejamento institucional de guarda, destinação final e avaliação documental do prontuário.

O descumprimento das regulamentações relacionadas ao prontuário médico pode acarretar penalidades tanto para o profissional de saúde quanto para a instituição. As sanções podem variar desde advertências e multas até a suspensão do exercício profissional e o fechamento da instituição, dependendo da gravidade da infração. No caso de violação de sigilo ou uso indevido de informações do prontuário, o profissional pode ser responsabilizado civil e criminalmente, com base no Código Penal Brasileiro, que prevê penas para quem viola o segredo profissional ou comete atos que resultem em danos ao paciente.

Art. 154 - Revelar alguém, sem justa causa, segredo, de que tem ciência em razão de função, ministério, ofício ou profissão, e cuja revelação possa produzir dano a outrem: Pena - detenção, de três meses a um ano, ou multa de um conto a dez contos de réis. (Brasil, 1940)

A LGPD também prevê penalidades severas para o tratamento inadequado de dados pessoais, incluindo multas de até 2% do faturamento da empresa, limitada a R\$ 50 milhões por infração, além de sanções administrativas como a suspensão das atividades de tratamento de dados. In extremis, a instituição pode ser impedida de operar até que as irregularidades sejam corrigidas. Essas penalidades reforçam a necessidade de que profissionais e instituições de saúde se mantenham atualizados e em conformidade com as regulamentações vigentes.

Art. 52. Os agentes de tratamento de dados, em razão das infrações cometidas às normas previstas nesta Lei, ficam sujeitos às seguintes sanções administrativas aplicáveis pela autoridade nacional: I - advertência, com indicação de prazo para adoção de medidas corretivas; II - multa simples, de até 2% (dois por cento) do faturamento da pessoa jurídica de direito privado, grupo ou conglomerado no Brasil no seu último exercício, excluídos os tributos, limitada, no total, a R\$ 50.000.000,00 (cinquenta milhões de reais) por infração; [...] (Brasil, 2018)

A regulamentação do prontuário médico no Brasil estabelece um conjunto de normas que visam garantir a integridade, a confidencialidade e a acessibilidade dos dados de saúde dos pacientes. As regulamentações do CFM, em conjunto com a LGPD e outras legislações pertinentes, formam um arcabouço robusto para a proteção das informações contidas nos prontuários médicos. A conformidade com essas normas é fundamental para assegurar que os dados dos pacientes sejam geridos de

maneira ética, segura e eficiente, contribuindo para a melhoria da qualidade do atendimento à saúde e para a confiança entre profissionais e pacientes.

4.5 Representação da informação em sistemas de saúde

A representação da informação em sistemas de saúde é essencial para a eficácia do gerenciamento de dados médicos e para a qualidade do atendimento ao paciente. A implementação de um sistema de visão computacional para identificar e extrair metadados em prontuários médicos oferece benefícios significativos, incluindo a aceleração do processo de recuperação de documentos e a garantia de uniformidade na interpretação dos dados. A implementação de esquemas de representação adequados não só facilita a interoperabilidade entre diferentes sistemas de saúde, mas também aprimora a precisão e a rapidez no acesso às informações do paciente, essenciais para o diagnóstico e tratamento.

Legislação	Impacto sobre o Prontuário Médico
Lei Geral de Proteção de Dados (LGPD) – Lei nº 13.709/2018	Regula o tratamento de dados pessoais sensíveis, incluindo informações médicas. Garante segurança, privacidade e direitos do paciente sobre seus dados.
Código Penal Brasileiro – Art. 154	Penaliza a violação de sigilo profissional, incluindo a divulgação indevida de informações constantes nos prontuários médicos.
Código de Ética Médica – Resolução CFM nº 2.217/2018	Determina que o médico deve manter o sigilo das informações do prontuário e só pode divulgá-las com autorização do paciente ou exigência legal.
Resolução CFM nº 1.638/2002	Define o conceito de prontuário médico e estabelece a obrigatoriedade da Comissão de Revisão de Prontuários nas instituições de saúde.
Resolução CFM nº 1.821/2007	Regulamenta a digitalização e o uso de prontuários eletrônicos, determinando requisitos de segurança, confidencialidade e autenticidade.
Resolução CFM nº 2.218/2018	Atualiza normas para a digitalização e guarda de prontuários médicos eletrônicos, reforçando padrões de segurança e interoperabilidade.
Código de Defesa do Consumidor (CDC)	Garante ao paciente o direito de acesso ao seu prontuário médico a qualquer momento, seja em formato físico ou eletrônico.
Conselho Nacional de Arquivos (CONARQ)	Define diretrizes para a preservação e gestão de prontuários médicos, assegurando sua conservação e descarte adequado.

O uso de tecnologias emergentes como a Inteligência Artificial e a aprendizagem de máquina na representação de dados de saúde tem mostrado promessas significativas como Aprimoramento da análise de grandes volumes de dados médicos, Aumento da segurança e conformidade com regulamentações, Apoio

à pesquisa médica e epidemiológica etc. Essas tecnologias podem automatizar o processo de categorização e análise de grandes volumes de dados médicos, levando a uma melhor compreensão das necessidades dos pacientes e a intervenções mais personalizadas. Segundo a Organização Mundial da Saúde - WHO (2021):

A inteligência artificial (IA) possui um enorme potencial para fortalecer a prestação de cuidados de saúde e medicina, ajudando todos os países a alcançar a cobertura universal de saúde. Isso inclui o aprimoramento do diagnóstico e do cuidado clínico, o avanço na pesquisa em saúde e no desenvolvimento de medicamentos, além de auxiliar na implantação de diversas intervenções de saúde pública, como vigilância de doenças, resposta a surtos e gestão de sistemas de saúde. Tradução própria (Who, 2021, p. 5).

No entanto, é necessário considerar que mesmo com o uso da visão computacional para identificar e extrair informações com maior efetividade por si só não representa uma solução final. A forma de representar essas informações é essencial para chegar na recuperação da informação, e na área da saúde é importante considerar *frameworks* como HL7 e FHIR.

4.5.1 Contribuição dos *Frameworks* HL7 e FHIR

Os frameworks HL7 e FHIR (*Fast Healthcare Interoperability Resources*) desempenham um papel central na padronização da troca de informações de saúde. Eles fornecem estruturas interoperáveis que permitem a comunicação eficiente entre diferentes sistemas de informação em saúde. Esses *frameworks* são fundamentais para a integração de dados, facilitando a interoperabilidade e melhorando a qualidade dos cuidados de saúde. a adoção do padrão FHIR oferece inúmeros benefícios:

(...) os benefícios da interoperabilidade, como a troca de informações na gestão de consultórios, clínicas e hospitais, o compartilhamento seguro de dados do prontuário eletrônico do paciente e a disponibilização ágil de resultados de exames laboratoriais. São abordadas as atividades e recursos envolvidos no processo de interoperabilidade, como a heterogeneidade das fontes de dados e os modelos de troca de informação entre instituições de saúde" (FERREIRA, 2024, p. 21).

o HL7 FHIR é um modelo promissor em ascensão. Ao avaliarmos a sua estrutura, podemos perceber que o protocolo é projetado para incentivar a sua utilização, uma vez que o mesmo é baseado em padrões Web de troca de informações bem estabelecidos, como o XML, JSON, HTTP, REST. (NEFOUSSI, 2018, p. 54)

Modelos estruturados como HL7 FHIR são importantes para a representação de dados e troca de informações de saúde em meio digital e têm se mostrado eficazes na melhoria da comunicação entre diferentes sistemas de informação em saúde. A estruturação dos dados permite uma análise mais eficiente e

uma maior segurança, minimizando erros de interpretação que podem afetar a qualidade do atendimento ao paciente.

Um dos principais desafios na representação de dados em saúde é a variedade de formatos e a falta de uniformidade, que podem complicar a integração de sistemas e o processamento de dados. A adoção de padrões internacionais, como o SNOMED CT, que fornece um vocabulário médico abrangente, ajuda a superar essas barreiras, promovendo uma representação de dados mais coesa e acessível.

(...) o SNOMED-CT (*Systematizade Nomenclature of Medicine - Clinical Terms*) que tem por propósito a representação codificada de significados utilizados em informações de saúde. Portanto o SNOMED-CT é uma terminologia clínica, consistindo de termos utilizados em saúde associados com códigos de conceitos, com múltiplos códigos por termos e estruturada de acordo com uma representação lógica dos significados. (BRASIL, 2023)

No contexto brasileiro, o Sistema Único de Saúde (SUS) enfrenta desafios relacionados à gestão de grandes volumes de dados médicos. A adoção de padrões como HL7 e FHIR no SUS poderia aprimorar a interoperabilidade e a eficiência dos sistemas de saúde pública. Estudos indicam que a implementação de sistemas interoperáveis no SUS poderia reduzir a redundância de dados e melhorar a precisão das informações médicas. De acordo com a Organização para a Cooperação e Desenvolvimento Econômico (OCDE), a construção de sistemas de saúde centrados nas pessoas, eficientes e sustentáveis, é alcançada por meio do uso inteligente de dados e tecnologias digitais. Esse uso pode levar a uma melhor compreensão dos custos e eficácia dos tratamentos médicos, resultando na redução de gastos desnecessários e melhorando a qualidade dos serviços de saúde OCDE (2019). A adoção de técnicas de aprendizado de máquina e processamento de linguagem natural (PLN) pode ajudar a superar esses obstáculos, melhorando a precisão da representação da informação e na recuperação de dados médicos.

4.6 Mineração de Dados e Textos em Ambientes Médicos

A mineração de dados é aplicada para extrair informações úteis dos vastos volumes de dados médicos. A mineração de dados e textos em prontuários médicos é uma área de crescente importância que promove avanços significativos na eficiência do diagnóstico e tratamento médico. O processamento de grandes volumes de dados médicos por meio de técnicas avançadas de mineração pode revelar insights valiosos que de outra forma permaneceriam ocultos.

A utilização de mineração de dados em saúde facilita a identificação de padrões de doenças, predição de surtos epidêmicos e personalização de tratamentos para pacientes. A integração de dados longitudinais¹⁷ de pacientes permite aos médicos formular diagnósticos mais precisos e planos de tratamento eficazes.

Isto por sua vez facilita a busca por dados e a melhor tomada de decisão para a seleção de uma intervenção. Uma das técnicas que podem ser utilizadas para extração de informações relevantes para tomada de decisão rápida é a mineração de dados (Gomes; Silva; Rached, 2019, p. 2).

Um dos principais desafios na mineração de textos médicos é a heterogeneidade e a desorganização dos dados, que podem incluir notas clínicas não estruturadas e registros médicos dispersos. Para enfrentar esses desafios, Pesquisadores como Lee *et al.* (2020) propõem o uso de algoritmos de processamento de linguagem natural (PLN) que são treinados para entender e categorizar termos médicos específicos, aumentando assim a precisão da recuperação de informações. Em seu estudo sobre o BioBERT, um modelo pré-treinado para a mineração de textos biomédicos, os autores demonstram como esses algoritmos podem extrair e compreender informações médicas de textos clínicos de maneira eficiente.

A evolução das técnicas de mineração de dados em saúde tem sido impulsionada pelo desenvolvimento de novas metodologias de inteligência artificial e aprendizado de máquina. Coudray *et al.* (2018) observam que o uso de técnicas de aprendizado profundo tem permitido a análise de conjuntos de dados médicos complexos e volumosos com precisão e velocidade sem precedentes, facilitando descobertas em áreas como genômica e epidemiologia. No estudo *Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning* de Coudray, *et al.*, (2018) publicado na Revista *Nature Medicine*, foi aplicado redes neurais convolucionais para classificar subtipos de câncer de pulmão a partir de imagens de histopatologia¹⁸, demonstrando a eficácia dessas técnicas em identificar mutações específicas e subtipos de câncer com alta precisão. Algo que a emergência

¹⁷ Conjunto de informações coletadas de um mesmo indivíduo ou grupo ao longo do tempo, permitindo a análise da evolução de variáveis clínicas, padrões de saúde e respostas a tratamentos. Esse tipo de dado é essencial na medicina para acompanhar a progressão de doenças e avaliar a eficácia de intervenções médicas.

¹⁸ Ramo da patologia que estuda as alterações estruturais e funcionais nos tecidos biológicos por meio da análise microscópica. É amplamente utilizada no diagnóstico de doenças, especialmente câncer, pois permite a identificação de anomalias celulares e moleculares em amostras de tecido obtidas por biópsia ou cirurgia.

global em saúde pública pode utilizar para enfrentar as diversas dificuldades que se apresentam na contemporaneidade.

A mineração de dados e textos em ambientes médicos é uma área importante para a modernização e eficiência do gerenciamento de prontuários médicos. A implementação de sistemas de visão computacional para identificar e extrair metadados como nome do paciente, doença, idade e CPF, entre outros, tem o potencial de contribuir para uma transformação significativa na recuperação de informações e a uniformização de processos. A mineração de dados envolve a aplicação de algoritmos para descobrir padrões ocultos em grandes conjuntos de dados. Ferramentas como algoritmos de aprendizado de máquina, podem ser usadas para extrair informações críticas de bases de dados volumosas. No contexto de prontuários médicos, essas tecnologias podem identificar correlações e tendências que seriam difíceis de perceber manualmente, otimizando o diagnóstico e o tratamento. Na tese de Araújo Júnior (2005) é observado que a mineração de textos poder apoiar significativamente recuperação de informações em bases de dados volumosas, essa análise sinaliza a importância de tais técnicas para a gestão da informação. “Todas as formas de mineração de dados objetivam descobrir padrões ainda desconhecidos nos dados. Esta técnica cada vez mais é apresentada como solução para a manipulação de gigantescas bases de dados e bases textuais” (ARAÚJO JÚNIOR, 2005, p. 63).

O uso da visão computacional, permite a extração automática de metadados essenciais, como nome do paciente e diagnósticos, a partir de prontuários médicos, aumentando o índice de precisão no processo de busca e recuperação da informação. Esta abordagem não apenas acelera a recuperação de documentos, mas também assegura uma maior consistência e confiabilidade na identificação dos dados médicos. Um dos principais desafios na mineração de dados e textos em ambientes médicos é garantir a qualidade e a precisão dos dados extraídos. A precisão é essencial para que os dados recuperados sejam úteis e relevantes. Cleverdon (1962) propõe que a precisão pode ser medida pela proporção de documentos úteis recuperados em relação ao total de documentos encontrados. A fórmula utilizada para esse cálculo é:

$$P = \frac{\mathbf{D} \text{ relevantes recuperados}}{\mathbf{D} \text{ total recuperados}}$$

Onde:

- ***P*** representa a precisão do sistema de recuperação.
- ***D relevantes recuperados*** é o número de documentos corretamente identificados como relevantes para a pesquisa.
- ***D total recuperados*** é o número total de documentos retornados pelo sistema.

Essa métrica é amplamente utilizada na avaliação de sistemas de recuperação de informações, no contexto dos prontuários médicos, um alto índice de precisão garante que os metadados extraídos correspondam de fato às informações relevantes para o atendimento clínico, reduzindo o risco de recuperação de dados irrelevantes e aumentando a confiabilidade na tomada de decisões médicas.

A integração de visão computacional com técnicas de mineração de dados oferece inúmeros benefícios para a gestão de prontuários médicos. Esses sistemas não apenas aceleram a recuperação de informações, mas também garantem a uniformidade na interpretação dos dados. Segundo Gomes, Silva e Rached (2019), a mineração de dados é sim uma ferramenta eficaz para o auxílio na tomada de decisões, uma vez que o método tradicional de análise de dados se tornou inviável pelo alto volume de dados. Entende-se que a automação desses processos libera recursos humanos para atividades mais complexas, melhorando a eficiência operacional das instituições de saúde, a aplicação de tecnologias de mineração de dados em ambientes médicos pode otimizar significativamente o processo de gestão de documentos e melhorar a precisão das informações recuperadas.

4.7 Impacto da Automatização nos Padrões de Privacidade

A utilização de visão computacional para a identificação e classificação de metadados em prontuários médicos oferece uma solução altamente eficiente e segura para a gestão de documentos em instituições de saúde. Este sistema não apenas acelera a recuperação de informações, mas também assegura a privacidade dos dados dos pacientes, conforme exigências de regulamentações como a Lei Geral de Proteção de Dados (LGPD) no Brasil e o Regulamento Geral de Proteção de Dados (GDPR) na Europa.

Mustafa e Nsour (2023) destacam que a utilização de visão computacional para a identificação e classificação de imagens médicas oferece uma solução altamente eficiente e segura para a gestão de documentos em instituições de saúde. Este sistema não apenas acelera a recuperação de informações, mas também assegura a privacidade dos dados dos pacientes. Segundo os autores:

"Um dos avanços mais notáveis nesse domínio é o algoritmo *You Only Look Once* (YOLO) v8, uma abordagem extraordinária de aprendizado profundo que demonstrou um enorme potencial na detecção de anomalias em imagens de raios-X do tórax" (Mustafa; Nsour, 2023, p. 2).

A aplicação de tecnologias de mineração de dados em prontuários médicos permite a extração e organização de grandes volumes de dados de forma rápida e precisa. Sistemas que utilizam visão computacional local, sem necessidade de transferência de dados para servidores externos, garantem uma camada adicional de segurança, minimizando riscos de vazamento e uso indevido de informações sensíveis. Essa abordagem local é importante para a conformidade com normas de proteção de dados, como a LGPD e o GDPR, que visam proteger a privacidade e os direitos dos indivíduos sobre suas informações pessoais.

(...) os algoritmos de *machine learning* facilitam a tomada de decisão ao lidar com grandes volumes de dados e que há uma perspectiva positiva na adoção dessa técnica para uso em novas pesquisas. Entretanto, importantes desafios permanecem, principalmente em relação à qualidade do preenchimento dos dados utilizados para as análises realizadas (Fernandes; Chiavegatto FILHO, 2019, p. 10).

A implementação dessas tecnologias de visão computacional e mineração de dados não apenas melhora a eficiência e precisão dos processos de recuperação e classificação de dados, mas também assegura a privacidade e segurança das informações dos pacientes. Com a adoção de padrões internacionais e conformidade com regulamentações de proteção de dados, essas soluções têm o potencial de transformar significativamente a gestão de documentos em saúde, promovendo uma prática mais segura e eficiente.

A integração de tecnologias avançadas como a visão computacional e a mineração de dados no campo da saúde representa um avanço significativo para a gestão arquivística. Esses sistemas garantem a privacidade dos dados e melhoram a eficiência operacional das instituições de saúde, alinhando-se às regulamentações de proteção de dados. A conformidade com normas como a LGPD e o GDPR é essencial para proteger as informações dos pacientes, e a utilização de soluções tecnológicas

locais fortalece a segurança e integridade dos dados, prometendo um futuro transformador para a saúde.

4.8 Conclusão da revisão narrativa

A revisão narrativa realizada ao longo desta seção apresentou uma análise das principais teorias, conceitos, e práticas relacionadas à organização e gestão de informações no contexto da saúde, com foco na interseção entre arquivologia, inteligência artificial, visão computacional e regulamentação de prontuários médicos. O estudo revelou a complexidade envolvida na gestão de informações em saúde, destacando a necessidade de integrar abordagens tradicionais da arquivologia com tecnologias emergentes para enfrentar os desafios modernos.

A análise apresentada ao longo desta seção estabelece uma base sólida para a proposta metodológica desta pesquisa, que visa desenvolver a base de um sistema capaz de integrar e automatizado a gestão de informações em saúde. Esse sistema buscará combinar os princípios arquivísticos com as tecnologias de visão computacional e inteligência artificial, oferecendo uma solução que não só atenda às exigências legais e éticas, mas que também seja capaz de enfrentar os desafios práticos e operacionais na gestão de informações em larga escala. A continuidade da pesquisa explorará a implementação prática dessa metodologia, avaliando seu impacto na eficiência operacional e na qualidade dos serviços de saúde.

A implementação de um sistema de gestão de informações em saúde depende não apenas da adoção de tecnologias avançadas, mas também da transformação dos processos organizacionais e da cultura institucional. A transição para um ambiente digitalizado requer mudanças profundas nas práticas de trabalho dos profissionais de saúde, que devem ser capacitados para utilizar novos sistemas e compreender a importância da integração de dados. A resistência à mudança é um desafio comum, especialmente em instituições que têm uma longa história de dependência de registros físicos. Assim, o treinamento contínuo e o suporte técnico são fundamentais para garantir a adoção eficaz das novas tecnologias e para maximizar os benefícios da digitalização.

Adicionalmente, a interoperabilidade entre sistemas de saúde, tanto a nível nacional quanto internacional, é um aspecto crucial que não pode ser negligenciado. A revisão destacou que a falta de interoperabilidade é uma barreira significativa para

a integração de dados de saúde, resultando em silos de informação que dificultam a visão completa do histórico do paciente e comprometem a qualidade do atendimento. Para superar esse desafio, é essencial adotar padrões comuns, como HL7 e FHIR, que facilitam a troca de informações entre diferentes plataformas e garantem que os dados sejam acessíveis e utilizáveis por todos os profissionais envolvidos no cuidado do paciente.

Outro ponto crítico identificado na revisão narrativa é a necessidade de garantir a segurança e a privacidade dos dados em sistemas de saúde digitalizados. Com a crescente digitalização dos registros de saúde, a proteção contra acessos não autorizados e a prevenção de vazamentos de dados se tornam cada vez mais importantes. A conformidade com regulamentações como a LGPD e o GDPR é vital para proteger os direitos dos pacientes e manter a confiança no sistema de saúde. Além disso, as instituições devem adotar práticas de segurança robustas, como criptografia de dados, autenticação multifatorial e monitoramento constante para identificar e mitigar ameaças em tempo real.

A revisão também enfatizou a importância de considerar as questões éticas na implementação de tecnologias de inteligência artificial e visão computacional. A automação de processos e o uso de IA na análise de dados de saúde oferecem inúmeros benefícios, mas também levantam preocupações éticas significativas. Decisões clínicas automatizadas devem sempre ser supervisionadas por profissionais qualificados, e os sistemas devem ser desenvolvidos com transparência e responsabilidade, garantindo que os vieses algorítmicos sejam minimizados e que o tratamento dos dados respeite os direitos e a dignidade dos pacientes.

A gestão do ciclo de vida dos documentos de saúde também é um aspecto fundamental abordado na revisão. A temporalidade e a destinação dos prontuários médicos devem ser gerenciadas de acordo com diretrizes claras, que garantam a preservação dos dados essenciais e o descarte seguro dos documentos que não são mais necessários. A integração de tecnologias que permitem a gestão automatizada do ciclo de vida dos documentos, como sistemas de arquivamento digital, pode melhorar a eficiência e garantir a conformidade com as regulamentações, ao mesmo tempo que facilita o acesso às informações quando necessário.

Por fim, a revisão narrativa aponta para a necessidade de um esforço colaborativo entre as áreas de arquivologia, tecnologia da informação e saúde pública para enfrentar os desafios identificados. A convergência dessas disciplinas é essencial para o desenvolvimento de soluções que sejam ao mesmo tempo tecnicamente robustas e sensíveis às necessidades clínicas e legais. A colaboração interdisciplinar permitirá o desenvolvimento de sistemas de gestão de informações que não só atendam aos requisitos técnicos e regulamentares, mas que também melhorem a qualidade do atendimento ao paciente e contribuam para a saúde pública de forma mais ampla.

Com base na literatura analisada, observa-se que, apesar do avanço na integração entre Arquivologia, ciências da Informação e Ciências da Saúde, ainda há desafios significativos na padronização da extração automática de metadados e na implementação de sistemas de indexação eficientes para grandes volumes de documentos. Richardson, Abramson e Kaushal (2012) discutem a necessidade de interoperabilidade em sistemas de saúde e como a ausência de integração dificulta a troca eficiente de informações, resultando em perda de dados críticos para o atendimento médico. Autores como Paiva, Mateus e Araújo Júnior (2024) reforçam a importância da gestão de documentos alinhada a padrões arquivísticos para garantir a acessibilidade e a confiabilidade da informação em sistemas de saúde pública. Entretanto, nenhum dos estudos analisados propõe um modelo que integre de forma holística inteligência artificial, visão computacional e princípios arquivísticos para garantir acessibilidade e conformidade regulatória. Dessa forma, esta pesquisa avança na literatura ao propor um modelo que não apenas automatiza a indexação dos documentos médicos, mas também respeita princípios arquivísticos, normas éticas e regulamentações vigentes, garantindo uma solução mais completa e alinhada com as demandas do setor de saúde.

5 PRESSUPOSTOS

A seção apresentará os pressupostos que reflete a suposição de que a integração de princípios arquivísticos com tecnologias avançadas, como inteligência artificial e visão computacional, pode resultar em uma metodologia eficiente e segura para a gestão de informações em saúde. Os pressupostos que orientam a pesquisa, explicita as bases teóricas e práticas que justificam a escolha do método e a abordagem adotada. Esses pressupostos envolvem a viabilidade da integração tecnológica, a importância da conformidade com regulamentações de privacidade, e o impacto potencial dessa metodologia na saúde pública. Assim, a seção servirá como alicerce para a construção e direcionamento da análise, estabelecendo as expectativas e limitações que permeiam a investigação.

5.1 Pressuposto Geral

Integração de princípios arquivísticos tradicionais com tecnologias avançadas de inteligência artificial e visão computacional resultará em uma metodologia mais eficiente, precisa e segura para a organização e recuperação de dados em prontuários e documentos médicos.

5.2 Pressupostos específicos

1. A Integração de Princípios Arquivísticos e Tecnologia Avançada pode acelerar e viabilizar a gestão de documentos automatizando atividades repetitivas.

2. A automatização da indexação de documentos resulta em ganho de eficiência, menor tempo para indexação e análise mais detalhada dos documentos, especialmente no que diz respeito à extração de metadados.

3. A indexação automatizada minimiza a necessidade de intervenção humana no processo de categorização de documentos, reduzindo o risco de exposição indevida de informações sensíveis e fortalecendo a privacidade do paciente.

6 METODOLOGIA

Esta seção apresenta a metodologia adotada na pesquisa, delineando as etapas, os métodos e as técnicas utilizadas para alcançar os objetivos propostos e validar a hipótese apresentada. A metodologia foi estruturada para integrar princípios arquivísticos com tecnologias avançadas de inteligência artificial e visão computacional, visando a criação de uma solução eficiente para a organização e recuperação de informações em prontuários médicos. além de possibilitar o cumprimento dos objetivos e a comprovação dos pressupostos da investigação.

6.1 Especificação da Metodologia

A metodologia adotada nesta pesquisa foi desenvolvida para verificar a eficácia da integração de princípios arquivísticos tradicionais com tecnologias avançadas, como inteligência artificial e visão computacional, na gestão de informações em saúde, especialmente no contexto de prontuários médicos. O estudo visa criar um sistema capaz de automatizar e otimizar a organização e recuperação de dados, mantendo a conformidade com a LGPD.

Para o cumprimento dos objetivos, foi adotada a abordagem metodológica mista, combinando técnicas qualitativas e quantitativas. A pesquisa qualitativa foi utilizada para analisar as necessidades e os desafios enfrentados pelas instituições de saúde na gestão de grandes volumes de dados, bem como para explorar as possibilidades de aplicação das tecnologias emergentes no contexto arquivístico. Isso envolveu uma revisão bibliográfica detalhada e a análise de estudos de caso relevantes. A pesquisa quantitativa, por sua vez, foi empregada para medir a eficácia do sistema desenvolvido, utilizando métricas específicas de precisão, eficiência e segurança na recuperação de informações.

Quadro 2 - Quadro resumo objetivos, pressupostos e providências metodológicas

OBJETIVOS	PRESSUPOSTOS	PROVIDÊNCIAS METODOLÓGICAS ASSOCIADAS
Desenvolver e validar uma metodologia integrada que combine princípios arquivísticos com tecnologias de inteligência artificial na detecção de padrões e classificação de imagens para melhorar a organização, recuperação e disponibilidade da informação em documentos de saúde pública.	A integração de princípios arquivísticos tradicionais com tecnologias avançadas de inteligência artificial e visão computacional resultará em uma metodologia mais eficiente, precisa e segura para a organização e recuperação de dados em prontuários médicos.	Utilização de uma abordagem metodológica mista, combinando técnicas qualitativas e quantitativas para analisar desafios na gestão de grandes volumes de dados de saúde e explorar possibilidades de aplicação de tecnologias emergentes.
Prototipar um sistema de visão computacional capaz de identificar e extrair automaticamente nome, cpf, doenças, data do documento, assunto e palavras-chave a partir de prontuários médicos.	A integração de princípios arquivísticos e tecnologia avançada pode acelerar e viabilizar a gestão de documentos automatizando atividades repetitivas.	Revisão bibliográfica detalhada e análise de estudos de caso para fundamentar o desenvolvimento do sistema.
Avaliar a eficácia e a precisão do protótipo em comparação com pessoas realizando a indexação, considerando métricas como tempo, acurácia e quantidade de erros.	A automatização da indexação de documentos resulta em ganho de eficiência, menor tempo para indexação e análise mais detalhada dos documentos.	Desenvolvimento de um protótipo que emprega inteligência artificial e visão computacional para extração automática de metadados.
Analisar o impacto da automação na dinâmica de trabalho dos profissionais, identificando benefícios e desafios associados à adoção dessas novas tecnologias.	A indexação automatizada minimiza a necessidade de intervenção humana no processo de categorização de documentos, reduzindo o risco de exposição indevida de informações sensíveis e fortalecendo a privacidade do paciente.	Testes e validação do sistema proposto em cenários diversos para garantir eficiência e precisão da recuperação de informações.

Fonte: elaboração própria

Figura 1 - Metodologia da pesquisa



Fonte: elaboração própria

A metodologia é dividida em várias etapas. Inicialmente, foi realizada a delimitação do estudo, que definiu o escopo da pesquisa, concentrando-se na aplicação dos princípios arquivísticos à gestão de prontuários médicos. Em seguida, a caracterização do universo estudado foi estabelecida, os tipos de documentos que serviriam como base para o desenvolvimento e teste do sistema. A etapa de delineamento e histórico da pesquisa descreve as fases do estudo, desde a revisão narrativa até o desenvolvimento e validação do sistema.

O desenvolvimento do protótipo proposto envolveu a implementação de técnicas de visão computacional para a identificação automatizada de metadados em seguida a aplicação de redes neurais para a extração organização e recuperação de informações. Foram realizadas várias rodadas de testes e aprimoramentos para garantir que o sistema não só atendesse às exigências técnicas, mas também estivesse em conformidade com as regulamentações legais e éticas. A validação do sistema foi conduzida por meio de análise do percentual de acerto do algoritmo e com

a comparação com métodos tradicionais de gestão de documentos, e os resultados foram analisados para avaliar os benefícios e limitações da abordagem proposta.

Ao longo das etapas, a pesquisa foi orientada pelos objetivos e para a comprovação dos pressupostos definidos na fase inicial, que guiaram a escolha das técnicas e ferramentas utilizadas. A metodologia, portanto, não se limita a propor uma solução para a recuperação da informação, mas também busca alinhar essa solução com as necessidades práticas das instituições de saúde e as exigências legais em vigor, contribuindo para a melhoria da gestão de documentos em um contexto crítico e sensível como o da saúde pública.

6.2 Delimitação do Estudo

O estudo concentra-se na criação de um sistema automatizado capaz de identificar e extrair metadados criando mecanismos de organizar e recuperar informações de forma eficiente, precisa e segura, assegurando a conformidade com a Lei Geral de Proteção de Dados (LGPD).

Geograficamente, o estudo está limitado ao contexto brasileiro, considerando as especificidades das regulamentações nacionais e as particularidades do Sistema Único de Saúde (SUS). Entretanto, os princípios e a metodologia desenvolvidos possuem potencial para aplicação em outros contextos, desde que adaptados a estruturas organizacionais locais.

Em termos de escopo documental, a pesquisa foca especificamente nos prontuários médicos, dada a sua complexidade e dificuldade dos sistemas automatizados em distinguir informações relevantes, além da importância crítica que desempenham na continuidade do cuidado e na tomada de decisões clínicas. Outros tipos de documentos de saúde, como laudos laboratoriais ou imagens médicas, não foram incluídos nesta fase da pesquisa, mas podem ser considerados em estudos futuros ou como parte de aprimoramentos do sistema.

Finalmente, a pesquisa é delimitada ao desenvolvimento de um protótipo de sistema e sua validação em um ambiente controlado, usando um conjunto de dados específico. A implementação em larga escala e a avaliação em contextos reais de operação não fazem parte do escopo desta pesquisa, mas constituem uma direção futura sugerida para a aplicação prática dos resultados obtidos.

6.3 Caracterização do Universo Estudado

O universo estudado nesta pesquisa é composto por instituições de saúde que atuam no Brasil, particularmente aquelas envolvidas na gestão de grandes volumes de prontuários médicos e que enfrentam desafios relacionados à organização, recuperação e segurança dessas informações. Este universo inclui hospitais públicos e privados, clínicas especializadas e centros de saúde, que lidam com prontuários médicos em formatos físicos e digitais, e que estão em processo de modernização ou já adotaram tecnologias para digitalização de seus arquivos.

O estudo incide em instituições que possuem uma demanda crítica por eficiência na gestão de informações e que, ao mesmo tempo, precisam garantir a conformidade com as regulamentações de proteção de dados, como a LGPD. Essas instituições são caracterizadas por um grande fluxo de dados médicos, necessidade de rápida recuperação de informações e desafios relacionados à preservação da integridade e privacidade dos dados dos pacientes. Além disso, a pesquisa leva em consideração o contexto das instituições que enfrentam dificuldades em integrar sistemas de informação distintos, tanto internamente quanto em relação a outras unidades de saúde, além do quadro reduzido de pessoas disponíveis e capacitadas para o tratamento, indexação e organização dos documentos.

O estudo dos prontuários médicos justifica-se pela relevância desses documentos na continuidade do cuidado ao paciente e na tomada de decisões clínicas, além de serem um dos tipos de registros mais complexos e críticos dentro do ambiente de saúde. Ao delimitar o universo estudado, a pesquisa garante que os resultados sejam aplicáveis a um cenário real e relevante, refletindo as condições e necessidades das instituições de saúde que atuam no Brasil. Isso permite que as conclusões e as soluções propostas possam ser diretamente implementadas ou adaptadas, promovendo melhorias significativas na gestão de informações em saúde, tanto em termos de eficiência quanto de conformidade com as regulamentações vigentes.

6.3.1 Definição da amostra

A definição da amostra para esta pesquisa seguiu critérios específicos, baseados na natureza dos dados necessários para o desenvolvimento e validação do sistema proposto. A amostra foi composta por um conjunto de prontuários médicos

fictícios, criados especificamente para simular as condições reais de documentos utilizados em instituições de saúde. A escolha por utilizar dados fictícios foi motivada pela necessidade de garantir a privacidade e a segurança durante o processo de treinamento do modelo de visão computacional, evitando o risco de exposição de informações sensíveis de pacientes reais.

Para a construção dos prontuários médicos fictícios, foram previamente levantados e analisados diferentes modelos reais de prontuários utilizados em instituições de saúde, contemplando variações de estrutura, layout e tipologia documental. A partir desses modelos, foi criado um banco de dados sintético com informações pessoais e clínicas geradas de forma aleatória e controlada, garantindo consistência interna e ausência de qualquer vínculo com dados reais. Esses dados foram embaralhados e utilizados para o preenchimento automático dos modelos selecionados, de modo a produzir documentos variados e não repetitivos. Após o preenchimento, os prontuários foram impressos e submetidos a processos de digitalização sob diferentes condições, incluindo variações de resolução, iluminação, texturas do papel, presença de ruídos, manchas e pequenas deteriorações, simulando cenários reais e adversos de uso e armazenamento documental.

Os prontuários fictícios foram desenvolvidos para replicar uma ampla variedade de casos comuns e complexos encontrados em ambientes hospitalares, incluindo diferentes formatos de documentos, variações de layout e múltiplos tipos de dados pessoais e clínicos. Essa diversidade dentro da amostra foi essencial para garantir que o modelo treinado fosse capaz de lidar com a heterogeneidade presente nos prontuários médicos reais, aumentando assim a robustez e a generalização do sistema.

A amostra incluiu tanto documentos digitalizados de alta resolução quanto imagens com qualidade mais baixa, refletindo as condições variadas que o sistema poderá encontrar em sua aplicação prática. Além disso, a amostra abrangeu diferentes tipos de dados, como texto impresso, manuscrito, tabelas, gráficos e imagens, exigindo que o sistema fosse treinado para reconhecer e processar corretamente cada um desses elementos. A quantidade de prontuários fictícios na amostra foi suficientemente grande para garantir um treinamento eficiente do modelo, com uma distribuição equilibrada de casos simples e complexos.

Por fim, a definição da amostra também considerou a necessidade de testar o sistema em condições que simulassem desafios do mundo real, como documentos danificados, com manchas ou com texto parcialmente ilegível. Ao incorporar esses cenários na amostra, foi possível ajustar o sistema para que ele seja resiliente e eficaz em situações adversas, garantindo um desempenho consistente em diferentes contextos operacionais. Essa abordagem assegura que o sistema esteja preparado para enfrentar a variabilidade e os desafios encontrados na gestão de prontuários médicos em instituições de saúde.

6.4 Delineamento e Histórico da Pesquisa

O delineamento abordará desde a revisão inicial da literatura e a definição dos objetivos e pressupostos até o desenvolvimento técnico e a validação do sistema, assegurando que todas as etapas foram conduzidas de forma rigorosa e alinhada com objetivo geral do projeto de dissertação. Além disso, serão discutidas as condições em que a pesquisa foi realizada, os métodos de coleta de dados, e a abordagem utilizada para a análise e interpretação dos resultados, fornecendo uma visão clara e estruturada do percurso investigativo.

6.4.1 Etapas da Pesquisa

Quadro 3 - Etapas da Pesquisa

ETAPA	DESCRIÇÃO	OBJETIVO
1. ANÁLISE DO CENÁRIO	Pesquisa e análise de estudos anteriores sobre arquivologia, inteligência artificial, visão computacional e gestão de informações em saúde aderentes ao objeto da presente investigação.	Estabelecer uma base teórica sólida e identificar lacunas e oportunidades na pesquisa
2. DEFINIÇÃO DO PROBLEMA	Especificação clara do problema de pesquisa, dos objetivos e dos pressupostos.	Delimitar o foco do estudo e garantir que a investigação seja direcionada ao atingimento dos objetivos e comprovação dos pressupostos.
3. DELIMITAÇÃO DO ESCOPO DA INVESTIGAÇÃO	Definição do escopo e das limitações da pesquisa, incluindo a seleção dos tipos de	Estabelecer os limites da pesquisa para garantir foco e viabilidade do estudo.

	documentos a serem analisados.	
4. COLETA DE DADOS	Recolhimento de informações e dados relevantes de prontuários médicos, tanto físicos quanto digitais.	Obter os dados necessários para o desenvolvimento e validação do sistema proposto.
5. DESENVOLVIMENTO DO SISTEMA	Projeto e implementação de um sistema integrado que, por meio da visão computacional, seja capaz de identificar e extrair metadados.	Criar uma solução prática para o processo de recuperação da informação que atenda as necessidades identificadas e que possa ser validada.
6. TESTES E VALIDAÇÃO	Aplicação de testes para verificar a precisão, eficiência da recuperação da informação.	Assegurar que o sistema desenvolvido atenda aos requisitos estabelecidos e funcione corretamente na recuperação efetiva e segura das informações contidas nos prontuários médicos.
7. ANÁLISE DE RESULTADOS	Interpretação dos dados obtidos nos testes e avaliação do desempenho de recuperação da informação em comparação com métodos tradicionais.	Avaliar a eficácia da recuperação da informação e identificar áreas para melhorias ou ajustes.
8. DISCUSSÃO E CONCLUSÃO	Síntese dos resultados em relação aos objetivos e pressupostos, além da avaliação dos impactos e limitações, e proposição de futuras pesquisas.	Conclusão da pesquisa a partir da análise dos dados para o cumprimento dos objetivos e comprovação dos pressupostos, além da proposição de estudos futuros.

Fonte: elaboração própria

6.4.2 Condições da pesquisa

A realização desta pesquisa foi influenciada por diversas condições que impactam diretamente o desenvolvimento, a implementação e a validação do sistema proposto. Esses fatores abrangem aspectos técnicos, estruturais, operacionais e normativos que devem ser considerados para garantir a viabilidade e a efetividade do estudo.

Quadro 4 - Condições da pesquisa

CONDICIONANTE	DESCRIÇÃO
----------------------	------------------

DISPONIBILIDADE DE DADOS	A qualidade e a quantidade dos dados influenciam diretamente no treinamento da visão computacional.
RECURSOS TECNOLÓGICOS	A implementação da metodologia exige o uso de <i>softwares</i> de inteligência artificial, visão computacional e infraestrutura tecnológica adequada, incluindo <i>hardware</i> e <i>software</i> de alto desempenho.
CAPACITAÇÃO PROFISSIONAL	O sucesso do desenvolvimento e implementação do sistema de recuperação segura de informações depende do conhecimento técnico da equipe envolvida, que precisa estar familiarizada com as tecnologias aplicadas e conhecimentos arquivísticos.
CONFORMIDADE LEGAL E ÉTICA	A pesquisa deve seguir estritamente as regulamentações de proteção de dados, como a LGPD, garantindo a privacidade e a segurança das informações de saúde dos pacientes.
INTEROPERABILIDADE DE SISTEMAS	A integração do sistema proposto com outros sistemas existentes nas instituições de saúde é decisiva para garantir a aplicabilidade prática dos resultados e a eficiência na recuperação das informações médicas.
LIMITAÇÕES TEMPORAIS	O tempo disponível para o desenvolvimento, implementação, testes e ajustes do sistema é limitado, o que pode afetar a profundidade da análise e a abrangência dos testes realizados.

Fonte: elaboração própria

7 DESENVOLVIMENTO DO PROTÓTIPO

O desenvolvimento do protótipo de recuperação segura de informações proposto nesta pesquisa foi realizado com base em uma escolha criteriosa de tecnologias e abordagens metodológicas que garantissem a eficiência, segurança e conformidade com as regulamentações de proteção de dados. Para a implementação do sistema, optou-se pelo uso da linguagem de programação *Python*, amplamente reconhecida por sua robustez e versatilidade em aplicações de ciência de dados e aprendizado de máquina (*Machine Learning* - ML). *Python* oferece uma vasta gama de bibliotecas e *frameworks*, como *TensorFlow*, *Keras* e *OpenCV*, que são essenciais para o desenvolvimento de modelos de visão computacional e para a manipulação eficiente de grandes volumes de dados.

Um dos principais fatores que influenciaram a escolha de *Python* foi sua capacidade de integrar facilmente técnicas de aprendizado de máquina com processamento de dados em larga escala, o que é importante para a análise e organização dos prontuários médicos. Além disso, *Python* é uma linguagem amplamente utilizada na comunidade científica, o que facilita o acesso a recursos, documentação e suporte técnico, garantindo a continuidade e a escalabilidade do projeto.

Considerando as questões de segurança e privacidade dos dados, optou-se por desenvolver um sistema que funcione completamente local, sem a necessidade de conexões externas ou o uso de APIs de terceiros que requerem pagamento por requisição. Embora existam diversos modelos de inteligência artificial que oferecem APIs pagas para reconhecimento de imagens e processamento de dados, o uso dessas soluções externas poderia expor as informações sensíveis dos prontuários médicos a riscos de compartilhamento e vazamento de dados, comprometendo a privacidade dos pacientes e violando as diretrizes estabelecidas pela Lei Geral de Proteção de Dados (LGPD).

Para evitar esses riscos, o sistema de visão computacional foi treinado localmente, utilizando dados e documentos fictícios assim como ajustes finos (*fine tune*) e transferência de aprendizados (*transfer learning*). Esse treinamento local permitiu que o modelo fosse ajustado e otimizado para reconhecer e classificar informações específicas contidas nos prontuários médicos, como nome do paciente,

data de nascimento, CPF, diagnóstico e outros dados relevantes. Ao garantir que o sistema funcione de forma totalmente independente da rede, a pesquisa contribui para a criação de uma solução segura, eficiente e alinhada com as melhores práticas de proteção de dados em saúde.

O treinamento do modelo envolveu várias etapas, incluindo a criação de um *dataset* sintético, o pré-processamento das imagens e documentos, a configuração dos parâmetros do modelo de visão computacional e a validação dos resultados obtidos. Após o treinamento, o sistema foi testado em um ambiente controlado para garantir que ele atingisse os níveis esperados de precisão e eficiência, permitindo uma organização e recuperação de dados que atendessem aos requisitos de segurança e desempenho necessários para a gestão de informações em saúde.

Com essa abordagem, o sistema desenvolvido não apenas oferece uma solução prática e escalável para a indexação e gestão de prontuários médicos, mas também assegura a conformidade com as regulamentações de proteção de dados e a preservação da privacidade dos pacientes, fatores essenciais para a adoção em larga escala em instituições de saúde.

Para aprofundar o desenvolvimento do protótipo proposto, é importante destacar o processo de escolha da linguagem de programação *Python* como elemento central. *Python* foi selecionado não apenas por sua popularidade e versatilidade, mas também por suas bibliotecas especializadas em ciência de dados e *machine learning*, como *TensorFlow*¹⁹, *PyTorch*²⁰, *Scikit-learn*²¹ e *OpenCV*²². Essas bibliotecas fornecem ferramentas poderosas que podem ser trabalhadas para a com modelos de visão computacional, gerando um sistema capazes de processar imagens e extrair

¹⁹ Uma biblioteca de código aberto desenvolvida pelo Google para computação numérica e aprendizado de máquina. É amplamente utilizada para a criação e treinamento de redes neurais, oferecendo suporte para modelos de deep learning e computação distribuída.

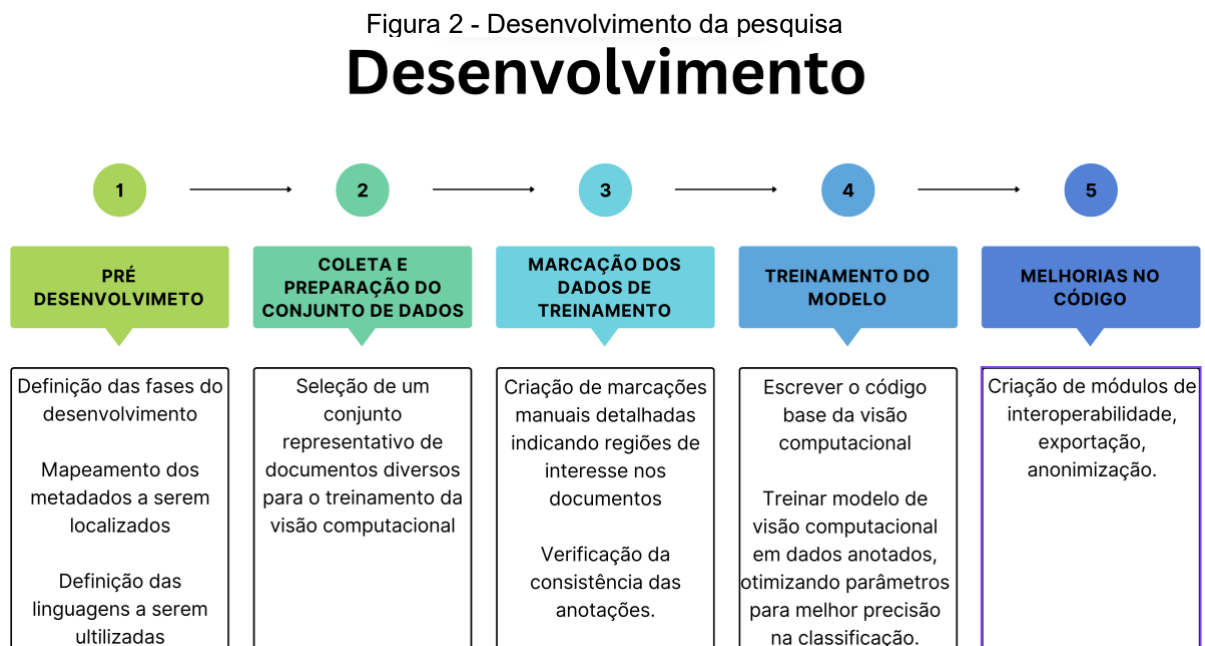
²⁰ Uma biblioteca de aprendizado de máquina de código aberto desenvolvida pelo Facebook, conhecida por sua flexibilidade e facilidade de uso. É amplamente utilizada em pesquisa acadêmica e na indústria para criar modelos de deep learning.

²¹ Uma biblioteca de aprendizado de máquina em Python que fornece ferramentas eficientes para análise de dados e modelagem preditiva. Inclui algoritmos para classificação, regressão, clustering e redução de dimensionalidade, sendo amplamente utilizada para aplicações de machine learning.

²² Uma biblioteca de código aberto voltada para visão computacional e processamento de imagens. Suporta diversas funcionalidades, como reconhecimento facial, detecção de objetos e manipulação de imagens e vídeos, sendo amplamente utilizada em aplicações de inteligência artificial e automação.

informações relevantes de documentos médicos, como nomes, datas de nascimento, e outros metadados essenciais para a gestão de documentos.

Além disso a linguagem, *Python* facilita a criação de fluxos de trabalho automatizados e integrados, permitindo o desenvolvimento de um sistema coeso, que combina extração de dados e categorização automática. Isso garante que o sistema possa lidar com grandes volumes de informações de maneira eficiente, sem necessidade de intervenção manual constante. A escolha de *Python* também foi motivada pela grande comunidade de desenvolvedores, que contribuem constantemente para a evolução das bibliotecas e *frameworks* disponíveis, oferecendo suporte e soluções inovadoras que podem ser facilmente integradas ao projeto.



Fonte: elaboração própria

A decisão de evitar APIs externas e construir um sistema totalmente local foi motivada pela necessidade de cumprir rigorosamente com as regulamentações de proteção de dados, como a LGPD e o GDPR. O uso de APIs externas para processamento de dados poderia introduzir vulnerabilidades, uma vez que os dados precisariam ser enviados a servidores de terceiros, aumentando o risco de vazamentos e comprometendo a segurança das informações. Ao manter o processamento completamente local, o sistema garante que todas as operações sejam realizadas dentro da infraestrutura da instituição de saúde, eliminando o risco

de exposição externa e assegurando a conformidade com as diretrizes de proteção de dados.

O treinamento local do modelo envolveu diversas etapas técnicas, como a marcação de área de interesse nos documentos. Essas etapas são essenciais para garantir que o modelo de visão computacional possa identificar e extrair informações de maneira consistente, independentemente das variações nos documentos analisados. Após o pré-processamento, o modelo foi treinado utilizando técnicas de *deep learning*²³, com ênfase em redes neurais convolucionais (CNNs), que são particularmente eficazes no reconhecimento de padrões visuais complexos, como textos e formas em documentos digitalizados.

Durante o treinamento, o sistema foi submetido a várias iterações de validação cruzada para garantir que os resultados fossem robustos e generalizáveis. Essa abordagem permitiu ajustar os parâmetros do modelo para maximizar a precisão e reduzir erros de classificação, garantindo que o sistema fosse capaz de operar de maneira confiável em diferentes cenários. Além disso, técnicas de regularização e ajuste de hiperparâmetros²⁴ foram aplicadas para evitar o sobreajuste do modelo, garantindo que ele permanecesse eficaz ao lidar com novos dados e documentos.

Para chegar na proporção ideal foram treinados diversos modelos com vários parâmetros de repetição sendo definida a proporção 70 a 30, sendo 70% dos itens para treinamento e 30% para validação (documentos inéditos e diverso para verificar precisão da visão). Nos primeiros treinamentos foi definido um treinamento de 30 *epochs*²⁵, com 100 documentos para treinamento, levando uma média de 7 minutos cada *epochs* levando 3 horas para o treinamento. Esses primeiros

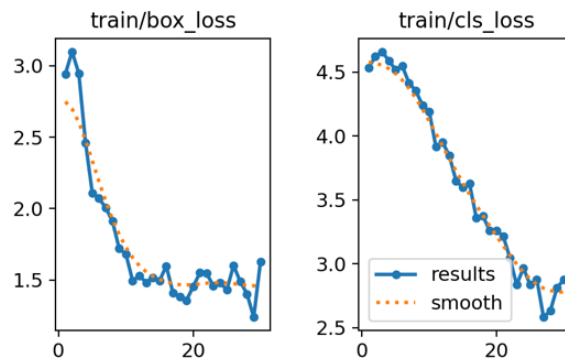
²³ Deep Learning é um ramo do aprendizado de máquina que utiliza redes neurais com múltiplas camadas para aprender representações complexas diretamente dos dados, frequentemente resultando em melhor desempenho em tarefas como classificação de imagens e processamento de linguagem natural.

²⁴ Os hiperparâmetros são definições prévias que controlam o processo de treinamento de um modelo de aprendizado de máquina, como a taxa de aprendizagem, o número de camadas ou neurônios em uma rede neural, a profundidade de uma árvore de decisão, entre outros. Diferentemente dos parâmetros internos do modelo, que são ajustados automaticamente a partir dos dados, os hiperparâmetros são estabelecidos antes ou durante o treinamento e influenciam diretamente o desempenho e a capacidade de generalização do modelo.

²⁵ É uma passagem completa pelo conjunto de dados de treinamento. Isso significa que, durante uma epoch, o modelo tem a oportunidade de ver e ajustar seus pesos com base em cada exemplo do conjunto de treinamento pelo menos uma vez. Ao final dessa passagem, uma nova epoch pode começar reutilizando o mesmo conjunto de dados, mas agora com pesos atualizados a partir da aprendizagem ocorrida anteriormente. Dessa forma, o número de epochs indica quantas vezes o modelo revisita o conjunto de treinamento para refinar seus parâmetros.

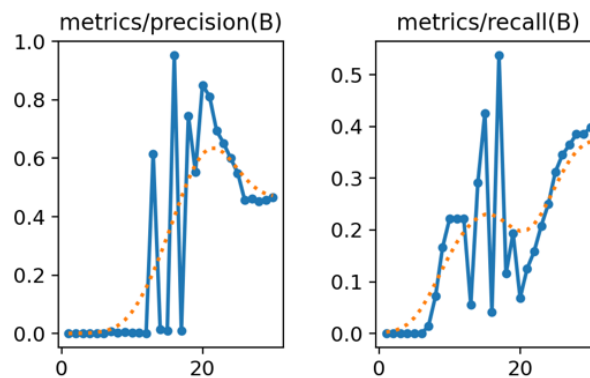
treinamentos chegando próximo a 20^a *epochs* o modelo já está tendendo a estabilidade aprendendo poucas coisas novas, mas a oscilação apresentada indica que são necessários mais documentos para o treinamento ou mais *epochs* visto que o comportamento ideal esperado para os gráficos é tender a uma reta.

Gráfico 1 - Curva de aprendizado da primeira etapa de treinamento da IA



Fonte: elaboração própria

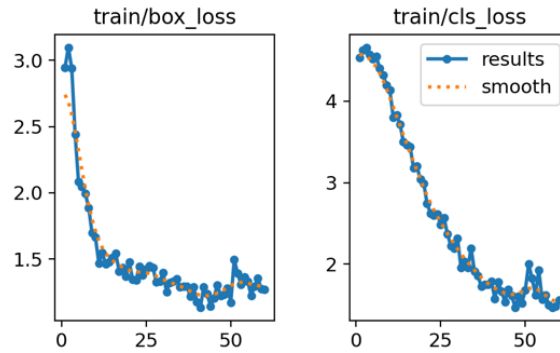
Gráfico 2 - Curva de precisão da primeira etapa de treinamento da IA



Fonte: elaboração própria

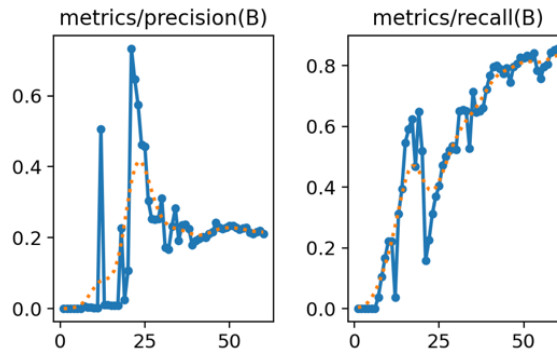
A cada treino era acrescentada a quantidade de 10 *epochs* assim no 4^o treino foi atingido 60 *epochs*. A aprendizagem ficou um pouco mais estável, no entanto, a precisão ainda se mostrou baixa no quarto treinamento que durou aproximadamente 5 horas para ser concluído. Os gráficos a seguir mostra os resultados obtidos no 4^o treinamento.

Gráfico 3 - Curva de aprendizado da quarta etapa de treinamento da IA



Fonte: elaboração própria

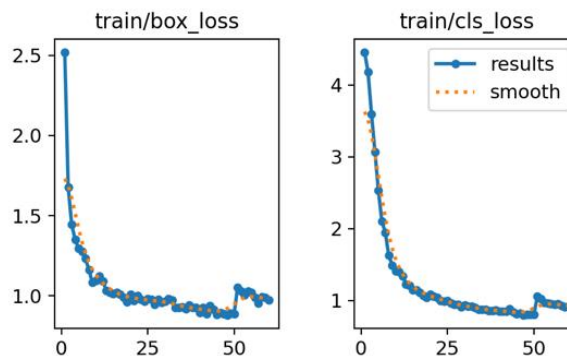
Gráfico 4 - Curva de precisão da quarta etapa de treinamento da IA



Fonte: elaboração própria

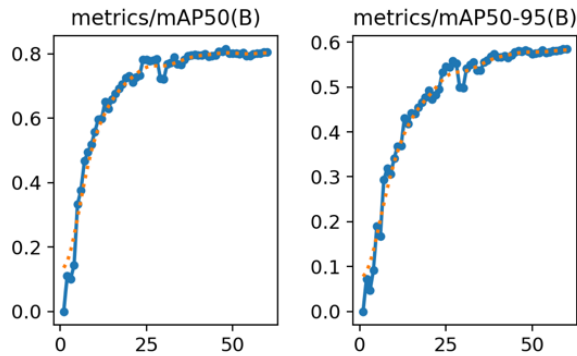
Nesta etapa a cada treinamento foi aumentado em 50 documentos de treino para descobrir a base de estabilidade, saindo, portanto, de 100 documentos no 4º treinamento para 600 no 14º treinamento que levou aproximadamente 15 horas para ser concluído. No entanto, os resultados dos treinamentos com mais de 250 documentos apresentaram melhorias mínimas e tempo de treinamento muito elevado em comparação com os anteriores. Segue os gráficos do melhor tempo e resultado.

Gráfico 5 - Curva de aprendizado da oitava etapa de treinamento da IA



Fonte: elaboração própria

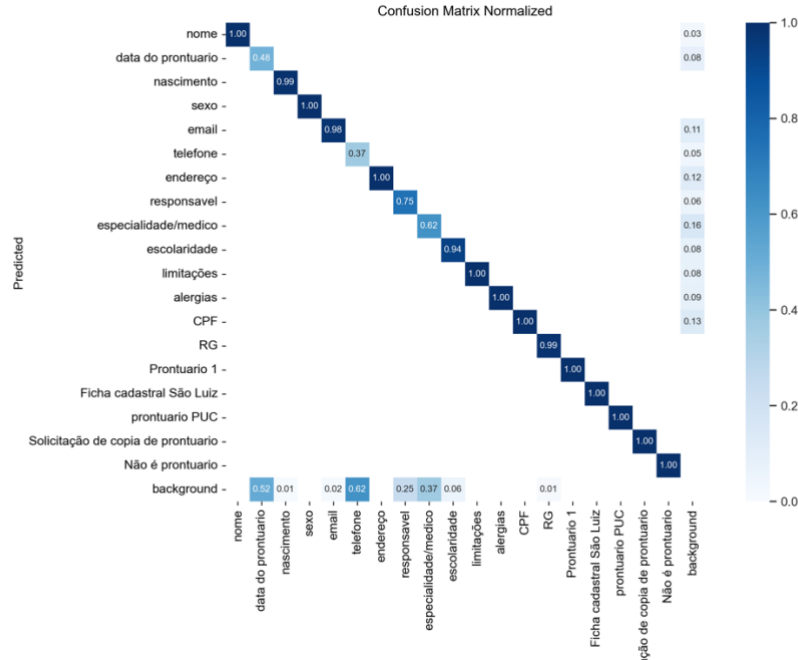
Gráfico 6 - Curva de precisão da oitava etapa de treinamento da IA



Fonte: elaboração própria

A matriz a seguir demonstra a precisão do treinamento para cada tipo de metadados, sendo que 1 não apresentou nenhum erro no processo de validação, logo 100% de acerto. Cabe ressaltar que nos formulários, há mais de um espaço para telefone e muitas das vezes esse espaço está em branco, o justifica a precisão inferior a 50% situação que pode e será melhorada com os treinamentos futuros.

Gráfico 7 - matriz de correlação normalizada da oitava fase de treinamento da IA



Fonte: elaboração própria

Para garantir que o sistema atenda às necessidades práticas das instituições de saúde, o sistema é modular e pode receber qualquer modelo de treinamento personalizado, podendo a instituição baixar o código e seguir o passo a passo de marcação dos documentos de treinamento e criar um modelo treinado do

zero para os seus mais variados tipos de documentos ou até mesmo expandir o aprendizado do atual modelo.

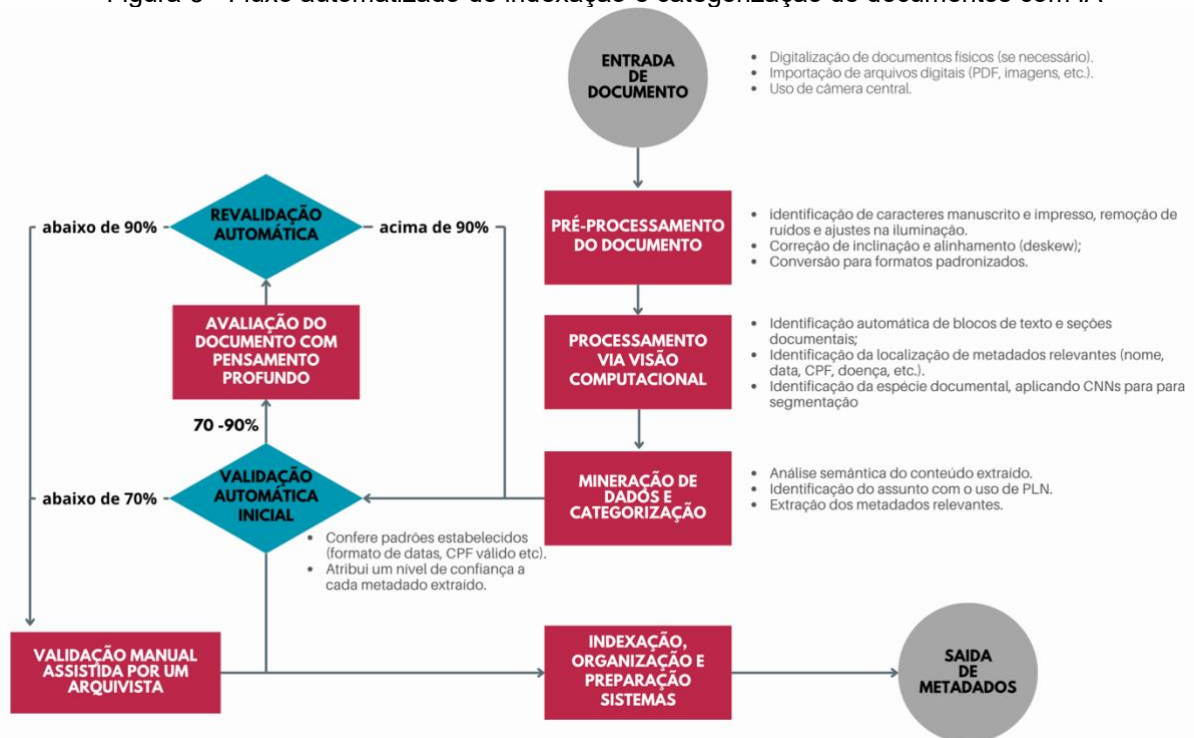
Uma vez concluído o treinamento, o sistema passou por uma fase de testes em um ambiente controlado. Esses testes foram projetados para simular o uso real em uma instituição de saúde, incluindo a análise de grandes volumes de documentos e a verificação da capacidade do sistema de operar sob diferentes condições de carga. O desempenho do sistema foi medido em termos de precisão, tempo de processamento, e conformidade com os requisitos de privacidade, fornecendo uma base sólida para a avaliação de sua eficácia.

Em suma, o desenvolvimento do protótipo foi planejado e executado para garantir que ele atenda às exigências específicas da gestão de informações em saúde, proporcionando uma solução que combina eficiência técnica com segurança e conformidade regulatória. O uso de *Python* como base, a decisão de evitar APIs externas e o treinamento com dados fictícios refletem um compromisso com a criação de um sistema seguro, eficaz e adaptado às necessidades das instituições de saúde.

7.1 Modelo de Automação dos Processos Arquivísticos

A automação dos processos mecânicos auxiliares aplicados às atividades arquivísticas baseia-se na integração de IA e Mineração de Dados para otimizar a extração, categorização e indexação de documentos de instituições de saúde. O sistema foi projetado para substituir tarefas manuais repetitivas, garantindo maior eficiência e padronização na recuperação da informação. Esse modelo de automação segue o fluxo abaixo:

Figura 3 - Fluxo automatizado de indexação e categorização de documentos com IA



Fonte: elaboração própria

A indexação automática refere-se a qualquer método ou processo computacional que possibilite a identificação e a seleção de termos representativos do conteúdo de documentos sem a necessidade de intervenção manual de um documentalista. Esse procedimento utiliza algoritmos e técnicas de processamento de linguagem natural para extrair palavras-chave, conceitos ou descritores relevantes, permitindo a organização, recuperação e categorização eficiente da informação em bases de dados ou sistemas de busca.

A indexação automática é qualquer procedimento que permita identificar e selecionar os termos que representam o conteúdo dos documentos sem a intervenção direta do documentalista. [...] A indexação automática pode também ser conceituada como a indexação realizada com o apoio de computadores que selecionam, por meio de um conjunto de instruções programadas previamente, os termos que mais ocorrem em um documento. (Araújo Júnior, 2007 p.25)

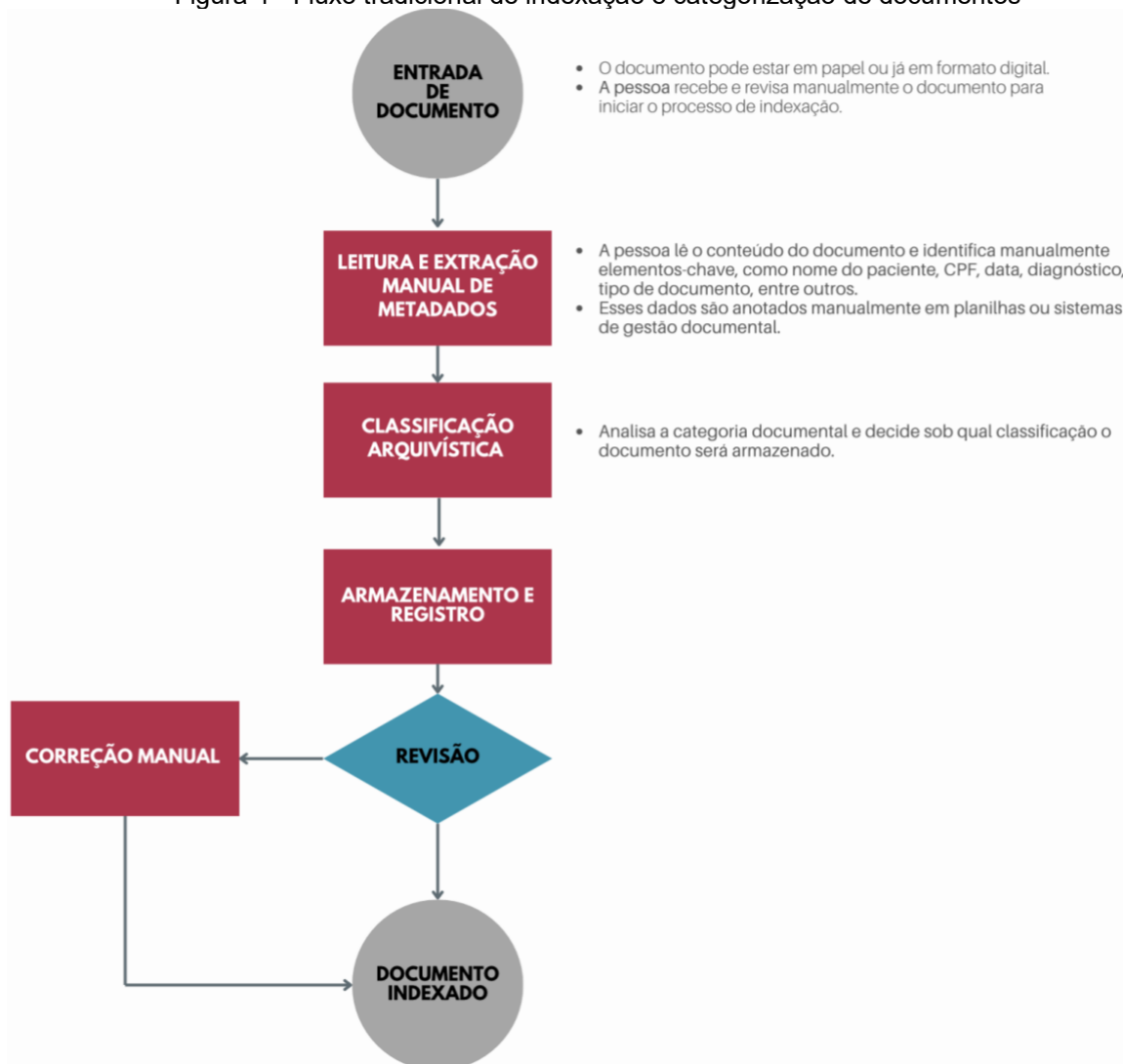
7.2 Comparação com o Modelo Tradicional

O modelo tradicional de indexação e categorização de documentos refere-se ao método manual utilizado para identificar, classificar e armazenar documentos em sistemas físicos ou digitais sem o auxílio de IA. Esse processo depende diretamente da interpretação humana. Na prática, esse modelo envolve múltiplas

etapas manuais, como a leitura do documento, identificação dos dados-chave, categorização e inserção das informações em planilhas ou sistemas de gestão de documentos (SGD). Esse processo é lento, suscetível a erros e limitado pela capacidade humana de processamento.

A indexação manual pode também ser conceituada como a tradução de um documento em termos documentários (descritores, cabeçalhos de assunto e termos-chave) sem o auxílio da atribuição automática de termos ou extração automática de termos; é a indexação sem o auxílio de computadores. (Araújo Júnior, 2007 p.24)

Figura 4 - Fluxo tradicional de indexação e categorização de documentos



Fonte: elaboração própria

Quadro 5 - Comparação entre o modelo manual e o modelo automatizado de indexação e categorização documental

Aspecto	Modelo Manual	Automação com IA e Mineração de Dados
Tempo por documento	3 a 7 minutos	0,7 a 1,2 segundos
Método de extração de metadados	Leitura e anotação humana	Extração automática via IA (OCR, PLN, visão computacional)
Classificação	Baseada na interpretação	Baseado treinamento padrões aplicado
Erros operacionais	8-12% devido à variação humana	1,3% devido a falhas de reconhecimento corrigíveis
Padronização	Sujeita a interpretações individuais	Uniforme e baseada em regras fixas
Escalabilidade	Limitada ao número de pessoas disponíveis	Processamento massivo sem necessidade de expansão da equipe
Revisão de Erros	Demanda tempo e recursos humanos	Correção automática ou intervenção mínima
Privacidade e Segurança	Profissionais acessam documentos sigilosos	Apenas IA processa, garantindo maior sigilo

Fonte: elaboração própria

7.3 Validação do protótipo

A validação do protótipo proposto nesta pesquisa foi realizada por meio de uma abordagem multifacetada, que avaliou tanto a precisão técnica do modelo de visão computacional quanto a qualidade da identificação e a eficiência operacional em comparação com métodos tradicionais realizados por humanos. A validação foi estruturada em três principais eixos de análise: precisão do algoritmo, qualidade da identificação e eficiência de tempo e custo.

A precisão do protótipo foi avaliada com base na capacidade do modelo de visão computacional de identificar e extrair corretamente os metadados específicos dos prontuários médicos fictícios, como nome, data de nascimento, CPF e diagnósticos. Para essa avaliação, foram utilizados métodos estatísticos que mediram a acurácia do algoritmo em diferentes cenários, considerando a taxa de verdadeiros positivos, falsos positivos e falsos negativos. A precisão foi continuamente ajustada por meio de validação cruzada durante o treinamento do modelo, garantindo que o protótipo atingisse um alto grau de confiabilidade na extração dos dados.

A qualidade da identificação foi medida comparando o desempenho do protótipo com o de profissionais humanos realizando as mesmas tarefas. Um conjunto de prontuários foi processado manualmente por um grupo de profissionais de saúde, que identificou e extraiu os mesmos metadados que o protótipo automatizado deveria processar. Em seguida, os resultados obtidos pelo protótipo foram comparados com os resultados dos humanos, utilizando métricas de concordância, como o coeficiente de Kappa²⁶, para avaliar a consistência e precisão do protótipo em relação ao trabalho humano. A análise demonstrou que o protótipo apresentou um nível de precisão comparável ao dos profissionais, com a vantagem de maior uniformidade na interpretação dos dados.

Um dos aspectos críticos da validação foi a comparação entre o tempo necessário para que um profissional humano realizasse o processo de identificação e extração dos metadados e o tempo requerido pelo protótipo automatizado. A análise mostrou que o protótipo conseguiu processar os documentos em uma fração do tempo necessário para a execução manual, resultando em uma eficiência significativamente maior. Além disso, foi realizada uma avaliação de custo, onde o custo médio por operação foi calculado para ambos os métodos. Os resultados indicaram que, embora o investimento inicial no desenvolvimento e implementação do sistema pode ser alto, o custo operacional por documento processado é substancialmente mais baixo do que o custo associado ao trabalho humano, gerando uma economia e liberando capital humano para as atividades finalísticas.

Os resultados da validação indicam que o protótipo desenvolvido é preciso, com qualidade de identificação comparável à realizada por profissionais humanos, e oferece melhorias significativas em termos de eficiência de tempo e custo. Essas características tornam o protótipo uma solução viável e eficaz para a gestão de informações em saúde, especialmente em instituições que lidam com grandes volumes de dados e que precisam garantir a conformidade com a LGPD. A validação também reforça a confiabilidade e segurança do sistema, demonstrando que ele pode ser adotado em larga escala sem comprometer a qualidade do serviço ou a integridade dos dados dos pacientes.

²⁶ O coeficiente de Kappa é um método estatístico que mede o grau de concordância entre dois ou mais avaliadores, ajustando-se para a concordância que ocorreria ao acaso, proporcionando uma avaliação mais precisa da consistência entre observações.

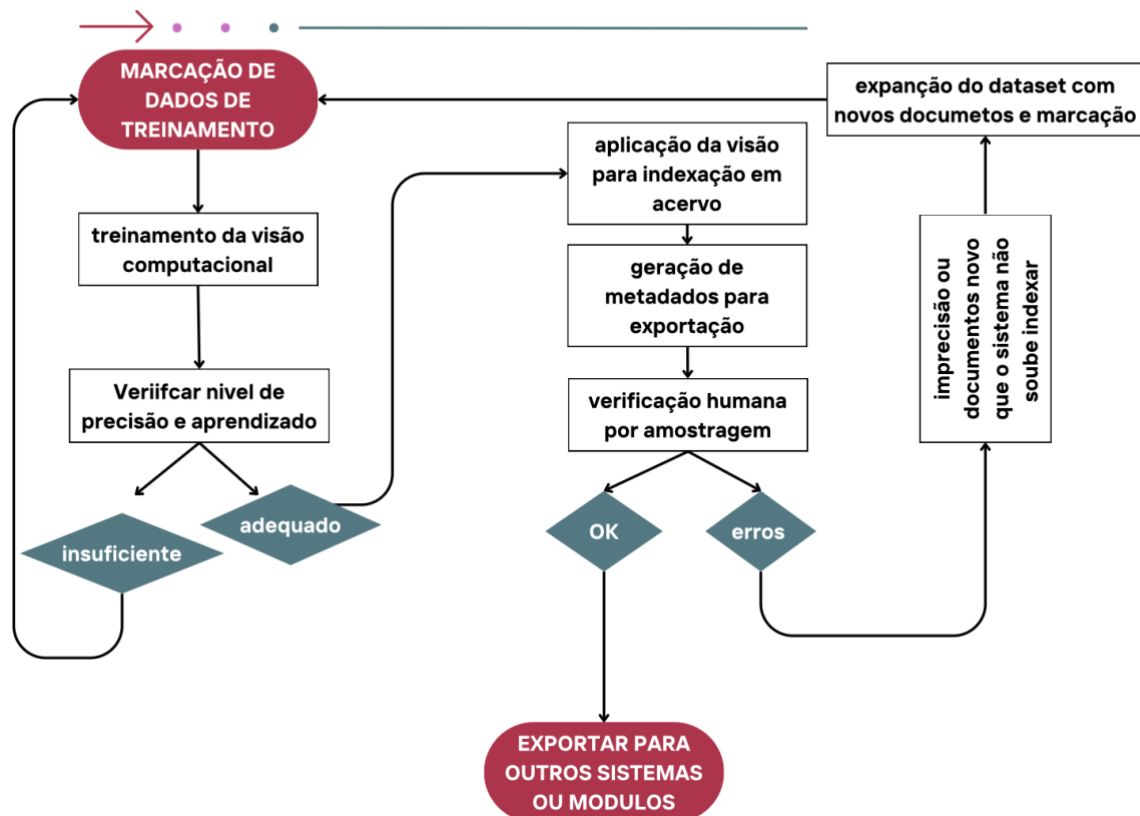
8 PROTÓTIPO E RESULTADOS

O sistema prototipado foi dividido em módulos, sendo o principal deles o de treinamento da visão computacional, este modulo é responsável por rodar os dados de treinamento, ensinando o modelo com base nos dados apresentados. A intenção do protótipo é que o modelo de IA seja de Aprendizagem Incremental ou Aprendizagem Contínua.

Figura 5 - Processo de funcionamento do protótipo

Sistema de visão computacional:

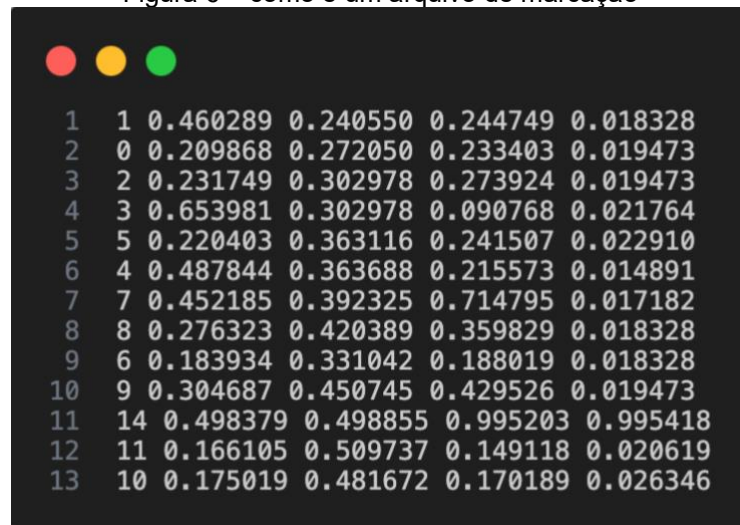
PROCESSO



Fonte: elaboração própria

Com essa disposição o protótipo pode constantemente evoluir, aumentando sua capacidade de identificar metadados, informações e tipos documentais. Esse processo de evolução é executado em conjunto com o seu uso prático, possibilitando o administrador do protótipo corrigir fragilidades sem perder ou alterar os pontos fortes já estabelecidos.

Figura 6 – como é um arquivo de marcação



```

1 1 0.460289 0.240550 0.244749 0.018328
2 0 0.209868 0.272050 0.233403 0.019473
3 2 0.231749 0.302978 0.273924 0.019473
4 3 0.653981 0.302978 0.090768 0.021764
5 5 0.220403 0.363116 0.241507 0.022910
6 4 0.487844 0.363688 0.215573 0.014891
7 7 0.452185 0.392325 0.714795 0.017182
8 8 0.276323 0.420389 0.359829 0.018328
9 6 0.183934 0.331042 0.188019 0.018328
10 9 0.304687 0.450745 0.429526 0.019473
11 14 0.498379 0.498855 0.995203 0.995418
12 11 0.166105 0.509737 0.149118 0.020619
13 10 0.175019 0.481672 0.170189 0.026346

```

Fonte: elaboração própria

A imagem acima é um arquivo de marcação utilizado para treinar a visão computacional, para o primeiro conjunto documental esses arquivos foram feitos manualmente para cada documento. Posteriormente o protótipo foi gerando esses arquivos de acordo com o seu treinamento, e a intervenção humana acontece somente para reparar alguma indexação equivocada ou treinar com novos parâmetros que a visão ainda não está suficientemente familiarizada. Outro aspecto a considerar sobre as coordenadas é a possibilidade de implementar novos módulos como mecanismo de anonimização, visto que para cada documento é criado um arquivo com as coordenadas de cada metadado presente no documento, nesse modelo de visão computacional o número “0” representa o metadado nome e dessa forma é possível anonimizar o nome presente no prontuário aplicando um borrão automático nas coordenadas encontradas.

Após a marcação dos dados de treinamento o seguinte código é acionado para realizar o treinamento e validação do modelo de detecção de metadados. O primeiro passo nesse processo é carregar o modelo YOLOv8 que é um modelo pré-treinado com grandes volumes de informação, a versão recomendada é a yolov8x (*extra-large*). O arquivo "yolov8x.pt" contém pesos treinados previamente para uma arquitetura maior, adequada para tarefas complexas.

Figura 7 - Função de treinamento repetitivo da IA

```

1  def main():
2      model = YOLO("yolov8n.pt")
3
4      model.train(data="visao.yaml", epochs=60) #, device='mps'
5      metrics = model.val()
6
7
8  if __name__ == '__main__':
9      main()

```

Fonte: elaboração própria

O “model.train” realiza o treinamento do modelo, e recebe os seguintes parâmetros “data=” que recebe um arquivo .yaml com as configurações que define os detalhes do *dataset*, como os caminhos das imagens, classes e outros detalhes necessários para o treinamento. O parâmetro “epochs=” define o número de épocas de treinamento, isso define que o modelo ajustará os pesos durante 60 iterações completas sobre o *dataset*. E “model.val” avalia o modelo treinado no conjunto de validação. Retorna métricas como precisão, *recall* e mAP (*mean Average Precision*).

Figura 8 – como é um arquivo yaml com parâmetros para treinar a visão computacional

```

1  path: ./
2  train: visao/train
3  val: visao/val
4  test:
5
6  names:
7      0: nome
8      1: data do prontuario
9      2: nascimento
10     3: sexo
11     4: email
12     5: telefone
13     6: endereço
14     7: responsavel

```

Fonte: elaboração própria

O arquivo yaml é responsável pela configuração para o treinamento do modelo. Ele define os parâmetros e caminhos necessários para treinar o modelo em um *dataset* personalizado. “*path*” define o diretório raiz onde os dados estão armazenados. “*train*” é o caminho para as imagens do conjunto de treinamento. “*val*” define o caminho para as imagens do conjunto de validação, usado para avaliar o

modelo após cada época de treinamento. “test” é o caminho para o conjunto de teste, caso você queira avaliar o modelo em um conjunto separado.

Ao desenvolver o protótipo foi possível aplicar e testar seu funcionamento em alguns documentos, a figura a seguir demonstra exatamente como a visão computacional identifica e delimita a localização das informações no documento. O sistema desenha uma caixa em volta da informação de interesse e a classifica, distinguindo qual modelo de documento (no caso da figura ficha cadastral São Luiz), nome do paciente, especialidade, nome, CPF entre outros.

Figura 9 - Visão computacional identificando informações de interesse

Ficha cadastral São Luiz 0.97

SAO LUIZ

Ficha de Cadastro

especialidade/medico 0.35
Ortopedia

nome 0.91
Nome Bruno Fernandes

nascimento 0.93
Data de Nascimento 20/12/1976

CPF 0.57
CPF 0.37
212.121.212-00

RG 0.80
RG 2121212-SP

sexo 0.93
Sexo M

Estado Civil Casado

Nome do Cônjuge Larissa Fernandes

Data de Nascimento 10/5/1978

endereço 0.97
ENDEREÇO RESIDENCIAL

Rua O, 505

Complemento Apt 303 Bairro Bairro O

Cidade O UF SP CEP 24680-135

Telefone residencial (11) 98765-4321 Telefone celular Fax

E-mail pessoal email 0.71
bruno@example.com

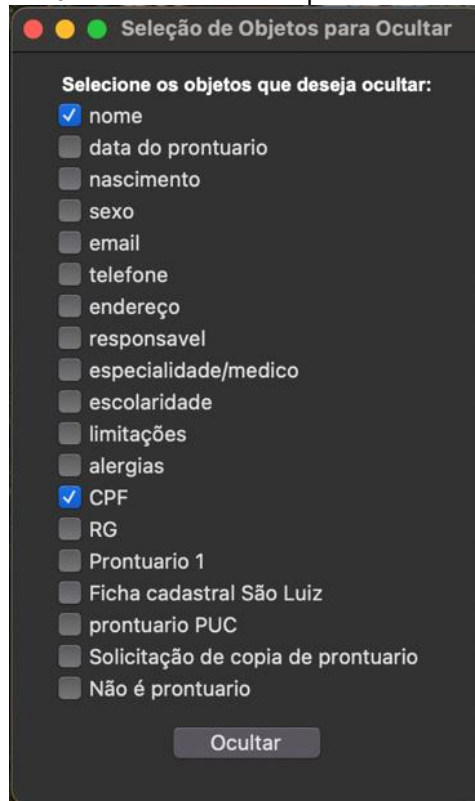
Foto

Fonte: elaboração própria

Com o rastreamento visual dos dados é possível aplicar um processamento para ocultação de dados sensível instantaneamente possibilitando que os dados pessoas seja apresentado somente a pessoas autorizadas. Na figura 8 é apresentado

o menu de seleção onde o administrador do protótipo pode definir as informações que devem se anonimizadas, e na figura 9 é possível visualizar um documento anonimizado automaticamente. Após definir o que deve ser ocultado, em fração de segundos o protótipo é capaz de anonimizar dezenas de documentos.

Figura 10 – Seletor de dados para anonimização



Fonte: elaboração própria

Figura 11 - Documento anonimizado

Prontuario | 0,95

MODELO - PRONTUÁRIO DO PACIENTE

Marca ou símbolo do estabelecimento ou serviço de saúde

Nome do estabelecimento ou serviço de saúde
Nome do logradouro, número, bairro, cidade, estado, CEP, telefone ou outro tipo de contato

PRONTUÁRIO DO PACIENTE

Número do prontuário: 20 data do prontuario 0.29 Data de Abertura: 31/3/2024

nasascimento 0.64 Data de nascimento: 8/11/1982 sexo 0.78 Sexo: F

endereco 0.55 Endereço: Rua T, 9090

telefone 0.26 Telefone(s): (21) 12345-6789 email 0.36 E-mail: patricia@example.com

Nome, telefone ou outro tipo de contato do responsável/cuidador (se aplicável): Rafael Oliveira

especialidade/medico 0.48 Médico(s) do Paciente (se houver): Dr. Rafael

escolaridade 0.27 Escolaridade: Ensino Superior Ocupação: Engenheira

limita??tee 0.32 Limitação: -

alergias 0.31 Alergia: -

Fonte: elaboração própria

8.1 Avaliação da Eficiência e da Precisão do Sistema

A avaliação da eficiência e da precisão do protótipo desenvolvido foi realizada em comparação com a indexação manual, conduzida por uma equipe, especializada em indexação, de voluntários da empresa ArquITI²⁷. O objetivo deste teste comparativo foi mensurar o desempenho do protótipo em termos de tempo de processamento, consistência e qualidade na extração e organização de metadados. Para isso, dois conjuntos documentais, compostos por 650 documentos cada, foram utilizados em ocasiões diferentes para análise.

Na etapa de indexação manual cada indexador registrava o início e o término da jornada por meio de ponto eletrônico em aplicativo de celular, bem como

²⁷ www.arquiti.com

as pausas e intervalos, de forma que fosse contabilizado apenas o tempo efetivamente produtivo. Os documentos, foram distribuídos entre os profissionais, que procederam individualmente à leitura dos arquivos e à extração manual dos metadados, registrando-os em sistema eletrônico.

A indexação manual, realizada pelos voluntários, seguiu o ritmo médio observado na empresa: um indexador humano processa aproximadamente 87 documentos em um expediente de seis horas, sendo efetivamente cinco horas de trabalho. Considerando esse desempenho, a equipe humana levou cerca de 37,36 horas (2.241 minutos) para completar a indexação de cada conjunto documental de 650 documentos, totalizando 74,72 horas para os dois conjuntos.

O protótipo automatizado apresentou um tempo de indexação que variou entre 0,7 segundos e 1,2 segundos por documento. O protótipo completou a indexação de cada conjunto em aproximadamente 10,28 minutos. Isso representa uma velocidade significativamente maior, reduzindo o tempo de processamento em cerca de 99,54% em relação ao processo manual. A análise comparativa entre os resultados do protótipo e da equipe humana revelou uma taxa de divergência de 11,5% entre os metadados indexados pelas duas abordagens. Dessas divergências, metade ocorreu porque o protótipo incluiu mais informações do que as identificadas manualmente, evidenciando sua capacidade de considerar dados que, em algumas situações, foram negligenciados pelos indexadores humanos.

A relação entre a indexação automatizada e a precisão na recuperação da informação foi discutida na da Ciência da Informação, segundo Araújo Júnior (2005), a avaliação do índice de precisão permite verificar o desempenho de sistemas automatizados em comparação com a indexação manual, observando a correspondência entre os termos extraídos automaticamente e aqueles atribuídos por especialistas. Em seu estudo, concluiu-se que a mineração de textos não apresentou ganhos significativos na precisão em relação à indexação manual, mas demonstrou potencial para complementar o processo ao identificar termos não previstos pelos indexadores humanos. Essa constatação reforça o que foi verificado na avaliação do sistema, onde parte das divergências registradas ocorreu porque o protótipo automatizado incluiu mais informações do que as identificadas manualmente, evidenciando sua capacidade de considerar dados que, em algumas situações, foram negligenciados pelos indexadores humanos.

Para a validação experimental do protótipo, foi elaborado e utilizado um conjunto documental fictício, composto por 1.300 documentos que simulavam prontuários médicos e registros de unidades de saúde. Essa decisão metodológica assegurou a preservação da confidencialidade e a não exposição de dados sensíveis de pacientes reais, em conformidade com a Lei Geral de Proteção de Dados (LGPD). Cada documento foi estruturado de modo a reproduzir elementos típicos de um prontuário, como nome, CPF, diagnóstico/doença, data de emissão, assunto e palavras-chave, permitindo avaliar o protótipo sem riscos éticos ou legais.

Os documentos foram submetidos a dois processos de indexação: manual, realizada por profissionais, e automatizada, conduzida pelo protótipo de visão computacional desenvolvido na pesquisa. Para assegurar rigor na comparação, as divergências entre os dois métodos foram submetidas a uma avaliação colegiada com seis indexadores. Nessas sessões, o grupo analisou colaborativamente os casos de discordância, debatendo se as diferenças representavam falhas, omissões ou simplesmente escolhas distintas, mas válidas. Os critérios utilizados abrangeram a precisão na extração dos metadados, a consistência da indexação em relação ao conteúdo documental e a relevância das divergências observadas. Esse procedimento coletivo não apenas reduziu vieses individuais, mas também consolidou uma base sólida para interpretar os resultados quantitativos da comparação entre o trabalho humano e o protótipo automatizado.

No conjunto, a avaliação colegiada garantiu que cada divergência fosse analisada em profundidade, oferecendo critérios objetivos para diferenciar erros reais de variações aceitáveis na indexação. A partir desse processo, foi possível consolidar os resultados da comparação, quantificar as falhas efetivas e verificar em que medida a automação se aproximou ou superou o desempenho humano.

Uma avaliação colegiada foi realizada pelos indexadores da ArquiTI para determinar a qualidade das divergências encontradas. Após análise detalhada, foi constatado que apenas 17 documentos, equivalentes a 1,30% do total analisado (650 documentos em cada conjunto, totalizando 1.300 documentos), realmente deixaram de registrar ao menos um metadado corretamente. Nos demais casos, a equipe considerou que o protótipo demonstrou lógica consistente, coerência e uniformidade, e que as divergências observadas não representaram falhas, mas, ao contrário, contribuíram para uma indexação mais completa e precisa. Os resultados

demonstraram que o protótipo é eficiente e preciso, superando significativamente o desempenho humano em termos de tempo da indexação, e não ficando muito distante em termos de qualidade. A capacidade do protótipo de identificar e incluir metadados adicionais, em situações em que a indexação humana foi considerada incompleta ou equivocada, reforça o potencial de automação para melhorar os processos de organização da informação.

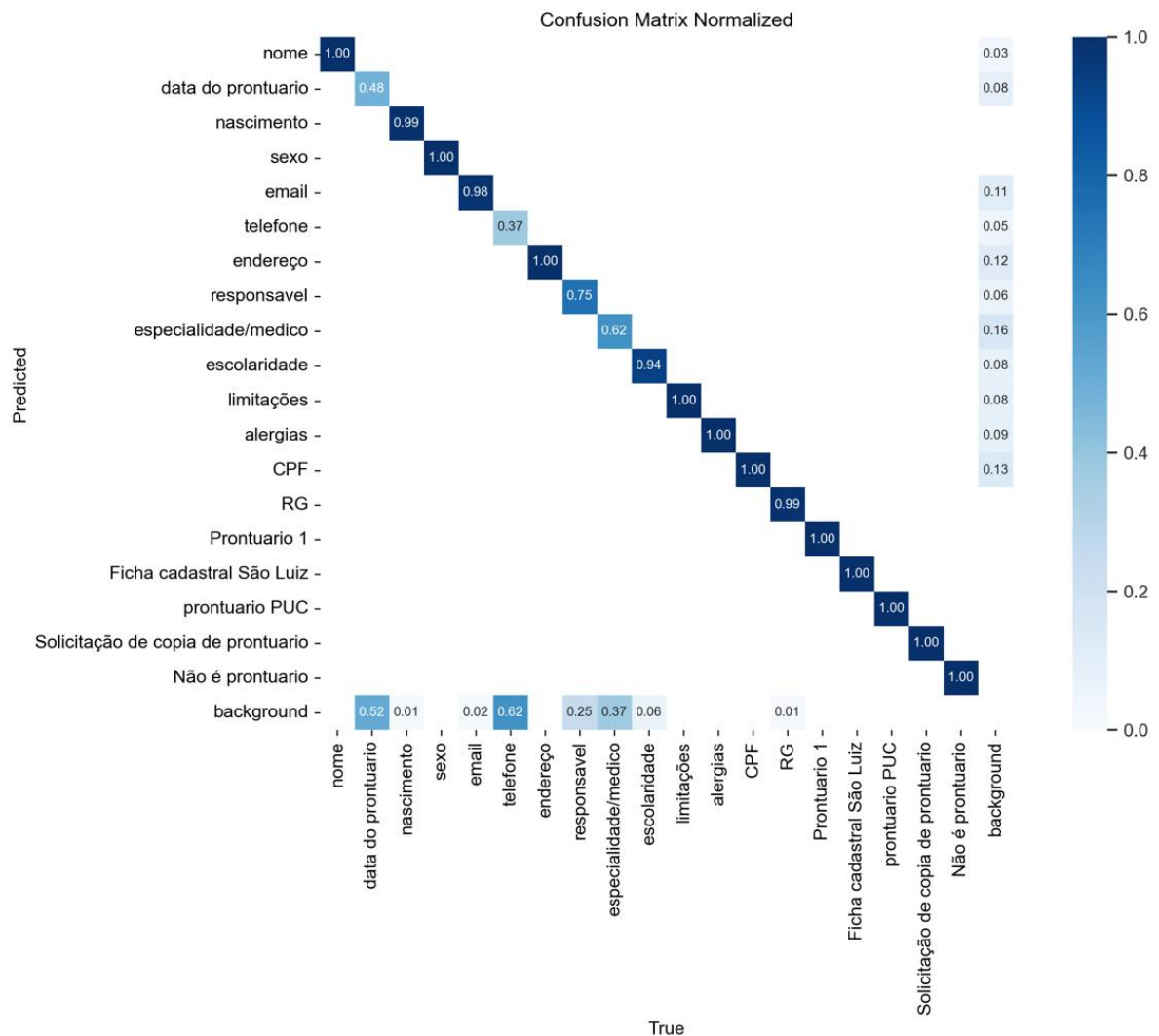
A avaliação colegiada foi conduzida por uma equipe de seis indexadores, que analisaram exclusivamente os documentos em que ao menos um metadado apresentou divergência entre a indexação automatizada e a manual, com o objetivo de verificar a precisão e a coerência do protótipo de visão computacional. O processo consistiu em uma análise coletiva, na qual todos os membros da equipe revisaram os documentos em conjunto, confrontando os metadados extraídos automaticamente com aqueles registrados manualmente. Para cada metadado divergente, foi avaliado se sua inclusão era coerente com o conteúdo do documento, se poderia impactar positiva ou negativamente a recuperação da informação e se sua remoção ou manutenção contribuiria para uma indexação mais precisa e eficaz. Os principais parâmetros analisados foram a precisão da extração, considerando a correspondência entre os metadados identificados pelo protótipo e aqueles registrados manualmente; a consistência da indexação, verificando a uniformidade dos descritores aplicados na categorização dos documentos; e a qualidade da divergência, classificando se a variação representava um erro do protótipo ou um detalhamento que aprimorava a indexação. Com base nessas avaliações, foi quantificado o percentual de divergências que representavam falhas e o percentual que indicava um enriquecimento da indexação, permitindo um diagnóstico sobre o desempenho do protótipo e sua capacidade de recuperação da informação.

Os erros observados em 1,3% dos documentos ensejam ajustes pontuais no sistema, especialmente em casos de metadados mais complexos ou contextualmente dependentes. No entanto, a baixa incidência de falhas e a coerência geral apresentada sugerem que o protótipo pode ser utilizado de modo confiável para indexação em larga escala, desde que acompanhado de validação contínua e refinamento de seus algoritmos.

Em resumo, a avaliação da eficiência e precisão do protótipo destacou seu desempenho superior, com um ganho significativo na velocidade de processamento e

maior consistência na extração de dados. Este resultado valida a viabilidade do uso do protótipo em complemento à indexação manual, principalmente em cenários empresariais e hospitalares que demandam alta produtividade e rigor na gestão de documentos. A adoção de tal tecnologia pode liberar capital humano para atividades de maior valor agregado, enquanto promove uma organização mais eficaz e segura das informações.

Figura 12 - Matriz de confusão normalizada para Classificação de Metadados em Prontuários Médicos



Fonte: estatísticas da visão computacional

A matriz de confusão normalizada apresentada oferece uma visão clara sobre o desempenho do protótipo ao longo de diferentes categorias de metadados extraídos, como "nome", "data do prontuário", "sexo", entre outros. A diagonal principal, composta por valores próximos ou iguais a "1" (representados pelas tonalidades mais escuras), indica que a maioria das categorias foi corretamente identificada, evidenciando uma alta precisão para essas classes. Valores fora da

diagonal principal (células mais claras) revelam instâncias de confusão entre categorias, como a sobreposição em metadados como "data do prontuário" e "*background*" ou vazio. Esses erros podem surgir devido à semelhança estrutural ou semântica dos dados envolvidos. A matriz também destaca que algumas categorias, como "nome" e "sexo", alcançaram uma precisão próxima a perfeita (1.0), enquanto outras apresentaram ligeiros desvios. Este diagnóstico é essencial para identificar áreas onde o protótipo necessita de ajustes, como no refinamento de algoritmos de extração ou melhoria na qualidade dos dados de entrada.

8.2 Limitações do Protótipo e aprimoramentos possíveis

O protótipo pode ter limitações relacionadas à precisão em cenários extremos de baixa qualidade de imagem, informações ilegíveis ou erros de entrada. Assim como o protótipo pode não prever todos os contextos éticos ou sociais, como vieses nos dados de entrada ou impactos indesejados em comunidades específicas.

Considera-se que, embora a metodologia proposta tenha potencial para transformar a gestão de informações em saúde, haverá desafios significativos relacionados à sua implementação e adoção. Tais desafios podem incluir a resistência à mudança por parte dos profissionais de saúde, a necessidade de infraestrutura tecnológica mínima. Nesse sentido é importante reconhecer que a mudança tecnológica, por mais benéfica que seja, exigirá uma adaptação cuidadosa e bem planejada.

Embora o protótipo desenvolvido tenha demonstrado alta precisão, eficiência e conformidade com as regulamentações de proteção de dados, há sempre espaço para aprimoramentos futuros, visando não apenas a melhoria contínua do desempenho, mas também a expansão das suas funcionalidades e aplicação.

Ampliar o conjunto de dados utilizado no treinamento, incluindo uma maior diversidade de prontuários médicos, com diferentes formatos, idiomas e tipos de conteúdo, pode ajudar a melhorar a generalização do modelo. Isso permitirá que o protótipo lide melhor com a variabilidade dos documentos reais e aumente sua robustez em diferentes contextos operacionais.

Embora o foco inicial do desenvolvimento tenha sido a eficiência do sistema, a interface de usuário pode ser aprimorada para facilitar o uso por

profissionais de saúde com diferentes níveis de familiaridade com tecnologias digitais. Interfaces mais intuitivas e a inclusão de funcionalidades adicionais, como painéis de controle personalizáveis, podem aumentar a aceitação e usabilidade do sistema.

Esses aprimoramentos potenciais visam não apenas a melhoria contínua do protótipo desenvolvido, mas também sua adaptação para atender a um espectro mais amplo de necessidades e contextos, promovendo uma gestão de informações em saúde mais eficiente, segura e abrangente.

8.3 Comprovação dos pressupostos

Os resultados obtidos foram analisados em relação aos pressupostos da pesquisa descritos na Seção 5.

Integração de Princípios Arquivísticos e Tecnologia Avançada: O sistema proposto utiliza tecnologias de IA e visão computacional para indexação automatizada para organizar, classificar e recuperar os prontuários, ao mesmo tempo em que mantém elementos centrais dos princípios arquivísticos (como proveniência, integridade, Ordem Original, Organicidade etc.), garantindo a estrutura e a confiabilidade dos registros. A automatização, aliada às regras arquivísticas para preservação e gestão, está em sintonia com o pressuposto de que a integração entre teoria arquivística e tecnologias emergentes torna a gestão de documentos mais eficiente e atualizada. Dessa forma, o pressuposto 1 é atendido, pois se percebe a combinação entre a fundamentação arquivística e as ferramentas tecnológicas necessárias para lidar com o grande volume de informações.

Automatização resulta em ganho de eficiência: O sistema local, offline e automatizado, promete reduzir drasticamente o tempo de indexação de prontuários médicos, pois não depende de intervenção manual para categorizar ou extrair metadados. A automatização tende a minimizar erros humanos (por exemplo, falhas de digitação ou de categorização incorreta) e padronizar a qualidade da informação recuperada. Em conjunto, esses fatores viabilizam um fluxo de trabalho mais ágil e seguro, confirmando a hipótese de ganhos significativos de eficiência. Assim, o pressuposto 2 também é atendido.

Automação da indexação e redução de acesso de terceiros aos documentos: A proposta de ter um sistema que funcione localmente, sem acesso a

redes externas, e que indexe documentos sem que pessoas precisem ler o conteúdo para isso, reduz a necessidade de terceiros manusearem informações sensíveis. A possibilidade de anonimizar ou pseudonimizar dados para fins de pesquisa reforça a aderência a regulamentações como LGPD e GDPR, pois restringe ao máximo a exposição de dados pessoais identificáveis. Dessa forma, o pressuposto 3 é cumprido, pois o modelo reduz a manipulação indiscriminada dos documentos, atendendo às exigências de privacidade e proteção de dados.

Impacto positivo na Saúde Pública como efeito colateral da organização e da recuperação de dados: O sistema proposto, por acelerar e otimizar o acesso a informações em prontuários médicos ao mesmo tempo em que mantém a conformidade com princípios arquivísticos e normas de privacidade, tende a facilitar tomadas de decisão mais ágeis e embasadas. Essa agilidade é fundamental em cenários de emergência em saúde, como epidemias ou pandemias, pois permite que profissionais e gestores identifiquem rapidamente quadros clínicos, históricos de vacinação, comorbidades e outros dados críticos para intervenções imediatas.

A indexação automatizada e a possibilidade de localização eficiente de prontuários ajudam as equipes médicas a ter uma visão integrada do histórico do paciente. Em situações de crise, cada minuto economizado na busca de informações é vital para o diagnóstico e o tratamento. A capacidade de agrupar e analisar dados de forma rápida e segura possibilita acompanhar indicadores epidemiológicos e clínicos, permitindo ajustes de protocolos e ações preventivas ou corretivas em larga escala.

Com dados confiáveis e acessíveis, a gestão hospitalar e as autoridades de saúde pública podem implementar medidas de contingência mais eficientes, direcionar recursos onde são mais necessários e otimizar a coordenação entre diferentes setores (hospitais, laboratórios, vigilância sanitária etc.). Assim, o pressuposto 4 é atendido, pois a melhoria na organização e na recuperação de dados proporcionada pelo protótipo pode, de fato, resultar em um impacto positivo na saúde pública ao agilizar ações e favorecer a tomada de decisão durante crises sanitárias.

A análise dos dados evidencia que os pressupostos definidos no início da pesquisa foram amplamente confirmados. O protótipo proposto demonstrou ser uma solução viável e eficiente para a gestão de informações em prontuários médicos, ao

mesmo tempo em que respeita os princípios éticos e regulamentares. Esses resultados reforçam a contribuição da pesquisa para a Ciência da Informação, indicando caminhos promissores para a integração entre tecnologias avançadas e práticas arquivísticas.

9 CONCLUSÃO

Esta pesquisa teve como propósito central investigar o uso da visão computacional como ferramenta para a indexação e categorização de prontuários médicos e documentos de instituições de saúde, conciliando eficiência, segurança e conformidade legal. Ao longo do percurso metodológico, os objetivos estabelecidos foram alcançados de forma consistente: o protótipo desenvolvido demonstrou ganhos significativos na redução do tempo de processamento, eliminando a necessidade de leitura manual, e apresentou taxas de precisão superiores às médias obtidas em processos convencionais. Além disso, comprovou-se a viabilidade de implementar um sistema local, desconectado da internet, que assegura maior proteção da privacidade dos pacientes e reduz a dependência de soluções proprietárias externas. Os resultados validam a hipótese de que a aplicação da inteligência artificial, aliada aos princípios arquivísticos, pode otimizar substancialmente a gestão de documentos em saúde.

Os pressupostos teóricos que nortearam a investigação foram confirmados à luz dos resultados obtidos. A Arquivologia, quando articulada à inteligência artificial, especialmente à visão computacional, mostrou-se uma base sólida para organizar informações em larga escala de modo estruturado, ético e funcional. Entre as contribuições mais relevantes, destacam-se a consolidação de um modelo metodológico para unir práticas arquivísticas e recursos tecnológicos, a proposição de uma abordagem prática e replicável para o tratamento de dados médicos e a comprovação de que a automação pode reduzir erros humanos, padronizar processos e liberar profissionais para atividades mais estratégicas. Assim, esta dissertação oferece à Ciência da Informação uma contribuição ao demonstrar como conceitos tradicionais podem dialogar com tecnologias emergentes para responder às demandas contemporâneas de gestão documental.

A aplicabilidade prática do protótipo proposto revela-se ampla e de impacto direto. Hospitais, clínicas, centros de pesquisa e gestores de saúde pública podem se beneficiar de um recurso que proporciona maior padronização e rapidez no acesso às informações. Em situações emergenciais, como crises sanitárias ou pandemias, a possibilidade de recuperação imediata e segura de dados clínicos pode representar a diferença entre uma resposta lenta e fragmentada e uma atuação eficaz e coordenada. Além disso, o protótipo proposto oferece suporte valioso ao trabalho de

arquivistas e profissionais da informação, que passam a contar com uma tecnologia de apoio à tomada de decisão, capaz de ampliar seu alcance estratégico. Do ponto de vista das políticas públicas, a pesquisa também reforça a importância de investir em soluções que garantam soberania tecnológica, sobretudo em regiões historicamente marcadas por desigualdades estruturais.

O caráter inovador e interdisciplinar do trabalho merece destaque. Ao articular princípios arquivísticos, fundamentos da Ciência da Informação e recursos da visão computacional, esta pesquisa rompe barreiras disciplinares e inaugura novas possibilidades de investigação. A inovação manifesta-se não apenas no desenvolvimento de um protótipo funcional, mas também na formulação de uma abordagem metodológica passível de ser aplicada a diferentes contextos documentais. No cenário brasileiro, o trabalho assume relevância pioneira, ao integrar o debate sobre soberania tecnológica à discussão sobre gestão de documentos em saúde, demonstrando que a ciência e a tecnologia, quando trabalhadas em conjunto, têm potencial para transformar práticas institucionais e gerar soluções acessíveis, sustentáveis e inclusivas.

Reconhecem-se, entretanto, algumas limitações inerentes ao processo investigativo. O conjunto documental utilizado no treinamento e validação do protótipo não contemplou a diversidade completa de formatos e estruturas de prontuários médicos existentes em instituições de saúde de diferentes portes, o que exige futuras expansões da base de testes. Além disso, a interface desenvolvida ainda demanda aprimoramentos para garantir maior usabilidade por profissionais sem formação técnica em tecnologia da informação. Também se constatou que a precisão do protótipo é condicionada à qualidade das imagens e ao grau de padronização dos documentos analisados, o que pode representar um desafio em acervos heterogêneos ou deteriorados.

Essas limitações, contudo, abrem caminho para perspectivas de aprimoramento. Futuras investigações poderão expandir o corpus documental de treinamento, incorporar diferentes tipologias de prontuários e explorar a integração com frameworks de interoperabilidade, como HL7 e FHIR, viabilizando a comunicação com outros sistemas de saúde. Também se mostra promissor o desenvolvimento de interfaces mais intuitivas, que favoreçam o uso por profissionais de diferentes áreas, bem como a combinação de visão computacional com técnicas de processamento de

linguagem natural, ampliando a capacidade do protótipo de lidar com informações multimodais. A realização de estudos comparativos em larga escala, em diferentes instituições e contextos, poderá consolidar a robustez e a adaptabilidade do modelo proposto.

Para além das dimensões técnicas, a pesquisa reforça a necessidade de refletir sobre os impactos sociais e éticos da automação documental em saúde. O protótipo foi concebido em conformidade com a Lei Geral de Proteção de Dados, assegurando a anonimização de informações sensíveis e a proteção da privacidade dos indivíduos. Essa preocupação vai além da esfera normativa, constituindo um compromisso político e social com a dignidade humana. O modelo proposto também aponta para o potencial de democratização do acesso à informação e fortalecimento da saúde pública, sobretudo ao oferecer alternativas tecnológicas que não dependem de infraestrutura externa, contribuindo para a redução das desigualdades no Sul Global.

Em síntese, esta dissertação confirma que é possível unir tradição e inovação: a Arquivologia e a Ciência da Informação, com seus princípios consolidados de organização e preservação documental, podem dialogar de maneira frutífera com a inteligência artificial e a visão computacional. O protótipo desenvolvido constitui mais do que uma prova de conceito; ele abre caminhos concretos para repensar a gestão de documentos em saúde em direção a processos mais eficientes, seguros e éticos. Conclui-se que a automação não deve ser vista como substituta do trabalho humano, mas como aliada estratégica, capaz de ampliar possibilidades, preservar a memória institucional e fortalecer a capacidade de resposta das instituições diante de desafios sociais e sanitários. A presente pesquisa, portanto, não apenas contribui para o avanço científico da área, mas também oferece subsídios para a construção de soluções que dialogam com as necessidades reais da sociedade e da saúde pública contemporânea.

A presente dissertação abre espaço para investigações futuras voltadas à análise do impacto da automação na dinâmica de trabalho dos profissionais de informação, considerando não apenas os ganhos de eficiência, mas também as transformações cognitivas e organizacionais decorrentes da adoção de tecnologias baseadas em inteligência artificial. Estudos posteriores poderão examinar como a

substituição de atividades manuais por sistemas automatizados modifica práticas arquivísticas, redefine competências profissionais e influencia a tomada de decisão.

Outra vertente promissora consiste em avaliar a aplicabilidade do modelo proposto em contextos distintos, como instituições de ensino, órgãos públicos e organizações privadas, de modo a compreender suas potencialidades e limitações em diferentes realidades informacionais. Pesquisas comparativas poderão explorar adaptações necessárias do modelo a tipos documentais diversos, bem como o impacto de fatores infraestruturais e socioeconômicos na efetividade da automação. Por fim, recomenda-se o aprofundamento de abordagens interdisciplinares que integrem aspectos técnicos, éticos e legais da gestão automatizada de informações sensíveis, assim como fortalecendo o debate sobre soberania tecnológica e equidade no acesso à inovação.

REFERÊNCIAS

- AFYA. **Tire todas as suas dúvidas sobre prontuário médico**. Afya Educação Médica, 2024. Disponível em: <https://educacaomedica.afya.com.br/blog/tire-todas-as-suas-duvidas-sobre-prontuario-medico>. Acesso em: fev. 2025.
- AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS). **Resolução Normativa nº 138, de 21 de novembro de 2006**. Diário Oficial da União: Brasília, DF, 2006. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/ans/2006/res0138_21_11_2006.html. Acesso em: abr. 2024.
- ARAÚJO JÚNIOR, Rogério Henrique de; TARAPANOFF, Kira. **Precisão no processo de busca e recuperação da informação: uso da mineração de textos**. *Ciência da Informação*, Brasília, DF, p. 236-247, set./dez. 2006. Disponível em: <https://revista.ibict.br/ciinf/article/view/1130/1278>. Acesso em: nov. 2024.
- ARAÚJO JÚNIOR, Rogério Henrique de. **Precisão no processo de busca e recuperação da informação**. Brasília, DF, 2005. Disponível em: http://www.realp.unb.br/jspui/bitstream/10482/34608/1/2005_RogérioHenriquedeAraujoJunior.pdf. Acesso em: jun. 2024.
- ARAÚJO JÚNIOR, Rogério Henrique de. **Precisão no processo de busca e recuperação da informação**. Brasília, DF: Thesaurus, 2007. 176 p. ISBN 9788570626554.
- ARAÚJO, Nelma Camêlo de; MOTA, Francisca Rosaline Leite. **Prontuário de paciente: questões éticas**. Fortaleza, p. 52-67, mar. 2020. ISSN 2525-3468. Disponível em: <http://repositorio.ufc.br/handle/riufc/50973>. Acesso em: nov. 2024.
- ARBOIT, Aline E.; BUFREM, Leilah S.; KOBASHI, Nair Y. **A institucionalização da ciência da informação no Brasil sob a ótica da evolução quantitativa dos cursos de graduação na área**. *Informação & Sociedade: Estudos*, v. 21, n. 1, 2011. Disponível em: <https://brapci.inf.br/v/92597>. Acesso em: nov. 2024.
- ASSOCIAÇÃO BRASILEIRA DE CIÊNCIAS. **Abertura e gestão de dados: desafios para a ciência brasileira**. Rio de Janeiro: ABC, 2020. Disponível em: <https://www.abc.org.br/wp-content/uploads/2020/09/ABC-Abertura-e-Gestão-de-Dados-desafios-para-a-ciência-brasileira.pdf>. Acesso em: jul. 2024.
- BENNETT, Casey; DOUB, Thomas. **Data mining and electronic health records: selecting optimal clinical treatments in practice**. p. 313-318, 2010. Disponível em: <http://www.openminds.com/library/110410dmehr.htm>. Acesso em: abr. 2024.
- BENTES PINTO, Virginia; TABOSA, Hamilton Rodrigues; VIDOTTI, Silvana Ap. B. Gregorio. **Arquitetura da informação: representação da informação de prontuário eletrônico do paciente**. Brasília, DF, p. 2629-2642, out. 2011. Disponível em: <https://cip.brapci.inf.br/download/182640>. Acesso em: jun. 2024.
- BENTES PINTO, Virginia. **A contribuição da terminologia para a gestão da documentação sanitária em organizações de saúde**. In: **MEDINFOR**, 2014, Salvador. Disponível em: <https://medinfor5.ufba.br>. Acesso em: nov. 2024.
- BENTES PINTO, Virginia. **Prontuário eletrônico do paciente: o documento de comunicação do domínio da saúde**. In: **CONGRESSO LUSO-BRASILEIRO DE ESTUDOS JORNALÍSTICOS**, 2.; **CONGRESSO LUSO-GALEGO DE ESTUDOS JORNALÍSTICOS**, 4., 2005, Porto. *Actas...* Porto, 2005.

BOBBA, Pratheek S. et al. **Natural language processing in radiology: clinical applications and future directions**. *Clinical Imaging*, v. 97, p. 55-61, maio 2023. Disponível em: <https://doi.org/10.1016/j.clinimag.2023.02.014>. Acesso em: nov. 2024.

BRASCHËR, Marisa; CAFÉ, Lúgia. **Organização da informação ou organização do conhecimento?** In: **ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB)**, 9., 2008, São Paulo. *Anais...* São Paulo, 2008.

BRASIL. **Decreto-Lei nº 2.848, de 7 de dezembro de 1940**. Código Penal. Diário Oficial da União: Rio de Janeiro, 1940. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Acesso em: fev. 2024.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Diário Oficial da União: Brasília, DF, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: mar. 2024.

BRASIL. Ministério da Saúde. **Banco de dados do Sistema Único de Saúde – DATASUS: desafios do e-Saúde – terminologias**. Brasília, DF, 2023. Disponível em: <http://www5.datasus.gov.br/DATASUS/index.php?acao=11&id=31051&tp=4>. Acesso em: set. 2024.

BRASIL. Ministério da Saúde. **Portaria nº 2.073, de 31 de agosto de 2011**. Diário Oficial da União: seção 1, Brasília, DF, 2011. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073_31_08_2011.html. Acesso em: nov. 2023.

CARMO, Dalton Garcia do; CHAGAS, Cintia Aparecida. **A produção dos representantes digitais: análise dos requisitos da digitalização de preservação e da digitalização de substituição**. *Em Questão*, Porto Alegre, v. 30, p. 1-31, 2024. Disponível em: <http://arquivistica.fci.unb.br/au/a-producao-dos-representantes-digitais-analise-dos-requisitos-da-digitalizacao-de-preservacao-e-da-digitalizacao-de-substituicao/>. Acesso em: nov. 2024.

CARVALHO, Ricardo César de. **Aplicação de técnicas de mineração de texto na recuperação de informação clínica em prontuário eletrônico do paciente**. Marília, 2017. Disponível em: <https://repositorio.unesp.br/server/api/core/bitstreams/99ada2e9-ca22-40c1-b93c-0761ef090209/content>. Acesso em: abr. 2024.

CHEFFER, Maycon H. et al. **Prescrição de metilfenidato: justificativas e história clínica**. p. 55-70, 2021. ISSN 2175-7275. Disponível em: <https://ojs.unirg.edu.br/index.php/1/article/view/3557/1853>. Acesso em: maio 2024.

CHOO, Chun Wei. **A organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões**. São Paulo: SENAC, 2003. ISBN 85-7359-341-5.

CLEVERDON, Cyril W. **Aslib Cranfield research project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems**. Cranfield, out. 1962. Disponível em: <http://hdl.handle.net/1826/836>. Acesso em: jun. 2024.

CODD, Edgar F. **A relational model of data for large shared data banks**. *Communications of the ACM*, p. 377-387, jun. 1970. Disponível em: <https://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf>. Acesso em: nov. 2024.

CONEGLIAN, Caio S.; GONÇALEZ, Paula R. V. A.; SANTARÉM SEGUNDO, José Eduardo. **O profissional da informação na era do big data**. p. 128-143, 2017. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v22n50p128>. Acesso em: dez. 2024.

CONSELHO FEDERAL DE MEDICINA (CFM). **Código de ética médica: Resolução CFM nº 2.217, de 27 de setembro de 2018**. Diário Oficial da União: Brasília, DF, 2018. Disponível em: <https://sistemas.cfm.org.br/normas/visualizar/resolucoes/BR/2018/2217>. Acesso em: nov. 2024.

CONSELHO FEDERAL DE MEDICINA (CFM). **Resolução CFM nº 1.638, de 10 de julho de 2002**. Define prontuário médico e torna obrigatória a criação da Comissão de Revisão de Prontuários nas instituições de saúde. Diário Oficial da União: Brasília, DF, 2002.

CONSELHO FEDERAL DE MEDICINA (CFM). **Resolução CFM nº 1.821, de 11 de julho de 2007**. Aprova as normas técnicas concernentes à digitalização e uso dos sistemas informatizados para a guarda e manuseio dos documentos dos prontuários dos pacientes, autorizando a eliminação do papel e a troca de informação identificada em saúde. Diário Oficial da União: Brasília, DF, 23 nov. 2007. Disponível em: <https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&pagina=252&data=23/11/2007>. Acesso em: nov. 2024.

CONSELHO FEDERAL DE MEDICINA (CFM). **Resolução CFM nº 2.218, de 24 de outubro de 2018**. Revoga o artigo 10º da Resolução CFM nº 1.821/2007. Diário Oficial da União: Brasília, DF, 2018. Disponível em: <https://sistemas.cfm.org.br/normas/visualizar/resolucoes/BR/2018/2218>. Acesso em: nov. 2024.

CONSTÂNCIO, Alex Sebastião; CARVALHO, Deborah Ribeiro; TSUNODA, Denise Fukumi. **Aplicações de visão computacional na saúde: revisão de literatura incrementada com técnicas de processamento de linguagem natural**. *Research, Society and Development*, v. 11, n. 10, 2022. Disponível em: <http://dx.doi.org/10.33448/rsd-v11i10.32942>. Acesso em: fev. 2024.

COUDRAY, Nicolas et al. **Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning**. 2018. Disponível em: <https://doi.org/10.1038/s41591-018-0177-5>. Acesso em: ago. 2024.

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual**. Rio de Janeiro: [s.n.], 1998. 256 p. ISBN 8535203524.

DEAN, Jeffrey; GHEMAWAT, Sanjay. **MapReduce: simplified data processing on large clusters**. In: **OSDI '04: 6th Symposium on Operating Systems Design and Implementation**, 2004. p. 137-149. Disponível em: https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full_papers/dean/dean.pdf. Acesso em: set. 2024.

DOMINGOS, Pedro. **O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo**. [S.l.]: Novatec, 2017. 344 p.

FERNANDES, Fernando Timoteo; CHIAVEGATTO FILHO, Alexandre Dias Porto. **Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho**. São Paulo, 2019. Disponível em: <http://dx.doi.org/10.1590/2317-6369000019418>. Acesso em: nov. 2024.

FERREIRA, Allan. **Modelo de arquitetura para interoperabilidade de dados de saúde utilizando padrão FHIR**. Marília, 2024. Disponível em: <https://repositorio.unesp.br/server/api/core/bitstreams/d9627402-26b9-47f9-abf8-63b606a36082/content>. Acesso em: dez. 2024.

GALVÃO, Maria Cristiane Barbosa. **Prontuário do paciente: construindo o conceito e desvendando a implementação**. Ribeirão Preto, 2024. Disponível em: <https://www.teses.usp.br/teses/disponiveis/livredocencia/17/tde-28112024-181817/publico//PRONTUARIOGALVAO2024USP.pdf>. Acesso em: nov. 2024.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. **Técnica de mineração de dados: uma revisão da literatura.** *Acta Paulista de Enfermagem*, v. 22, n. 5, p. 686-690, 2009. Disponível em: <https://doi.org/10.1590/S0103-21002009000500014>. Acesso em: set. 2024.

GOMES, Rodrigo J.; SILVA, Isadora S.; RACHED, Chennyfer D. A. **O uso da mineração de dados como ferramenta para tomada de decisões em enfermagem.** *Revista Eletrônica Acervo Saúde*, p. 1-7, mar. 2019. Disponível em: <https://doi.org/10.25248/reas.e375.2019>. Acesso em: nov. 2024.

GUALDANI, Fabrício A. **Modelo de mapeamento semântico para a representação e recuperação da informação em prontuários eletrônicos do paciente.** Marília, 2022. Disponível em: <https://repositorio.unesp.br/server/api/core/bitstreams/14584e5b-5644-4fd8-8ee9-ad98dcae5188/content>. Acesso em: mar. 2024.

HERNÁNDEZ, Juan M. **Historia clínica.** p. 57-68, 2006. Disponível em: <https://www.redalyc.org/pdf/875/87505904.pdf>. Acesso em: maio 2024.

HERRERO-SOLANA, Víctor; RÍOS-GÓMEZ, Claudia. **Producción latinoamericana en biblioteconomía y documentación en el Social Science Citation Index (SSCI) 1966-2003.** *Information Research*, v. 11, 2006. Disponível em: https://digibug.uqr.es/bitstream/handle/10481/32695/HerreroSolana_SSCIBiblioteconomia.pdf?sequence=1&isAllowed=y. Acesso em: out. 2024.

KITCHIN, Rob; MCARDLE, Gavin. **What makes big data, big data? Exploring the ontological characteristics of 26 datasets.** *Big Data & Society*, v. 3, p. 1-10, jan./jun. 2016. Disponível em: <https://journals.sagepub.com/doi/epub/10.1177/2053951716631130>. Acesso em: nov. 2024.

KOBASHI, Nair Yumiko; AGUIAR, Francisco Lopes de. **Organização e representação do conhecimento: perspectivas de interlocução interdisciplinar entre ciência da informação e arquivologia.** In: **ENANCIB**, 14., 2013, Santa Catarina. *Anais...* Santa Catarina, 2013.

LE COADIC, Yves-François. **A ciência da informação.** Tradução de Maria Yêda F. S. de Figueiras Gomes. Brasília, DF: [s.n.], 1996. Disponível em: <https://bibliotextos.wordpress.com/wp-content/uploads/2012/07/a-cic3aancia-da-informac3a7c3a3o-le-coadic.pdf>. Acesso em: nov. 2024.

LECUN, Yann et al. **Backpropagation applied to handwritten zip code recognition.** *Neural Computation*, p. 541-551, dez. 1989.

LEE, Jinhyuk et al. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining.** *Bioinformatics*, v. 36, p. 1234-1240, 2020. Disponível em: <https://doi.org/10.1093/bioinformatics/btz682>. Acesso em: set. 2024.

LESKOVEC, Jure; RAJARAMAN, Anand; ULLMAN, Jeffrey. **Mining of massive datasets.** [S.l.]: [s.n.], 2014. Disponível em: <http://infolab.stanford.edu/~ullman/mmds/book.pdf>. Acesso em: nov. 2024.

LIMA, Nathália Adrielle de. **Recuperação da informação de prontuários eletrônicos: um modelo de visualização de informação de medicamentos.** Marília, 2023. Disponível em: <https://repositorio.unesp.br/server/api/core/bitstreams/7f9b1119-13d3-4fdc-9315-33cb2b9b08e2/content>. Acesso em: abr. 2024.

MACEDO, Sheila Rodrigues dos Santos. **A gestão de documentos de prontuário do paciente em saúde pública municipal em Aracaju: da situação real para o ideal e sua inovação.** São Cristóvão, 2019. Disponível em: <https://ri.ufs.br/jspui/handle/riufs/13070>. Acesso em: mar. 2024.

MARIN, Heimar de Fátima. **Sistemas de informação em saúde: considerações gerais.** *Journal of Health Informatics*, jan./mar. 2010. Disponível em: <https://www.jhi.sbis.org.br/index.php/jhi-sbis/article/view/4/52>. Acesso em: maio 2024.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. **Big data: the management revolution**. *Harvard Business Review*, out. 2012. Disponível em: <https://hbr.org/2012/10/big-data-the-management-revolution>. Acesso em: set. 2024.

MINISTÉRIO DA SAÚDE (Brasil). Secretaria de Vigilância em Saúde. **Plano de contingência para resposta às emergências em saúde pública: febre amarela**. Brasília, DF, 2021. Disponível em: https://docs.bvsalud.org/biblioref/2022/06/1373492/plano_contingencia_emergencias_febre_amarela_2_ed.pdf. Acesso em: abr. 2024.

MOTA, Virgínia de Albuquerque. **Modelo de taxonomia para apoiar o processo decisório da participação cidadã em saúde no Brasil**. Brasília, DF, 2022. Disponível em: http://www.realp.unb.br/jspui/bitstream/10482/45816/1/2022_Virg%C3%ADniadeAlbuquerqueMota.pdf. Acesso em: nov. 2024.

MUNIZ, Vander Emiro. **Data mining: conceitos e casos de uso na área da saúde**. DevMedia, 2007. Disponível em: <https://www.devmedia.com.br/data-mining-conceitos-e-casos-de-uso-na-area-da-saude/5945>. Acesso em: jan. 2024.

MUSTAFA, Zaid; NSOUR, Heba. **Using computer vision techniques to automatically detect abnormalities in chest X-rays**. *Diagnostics*, v. 13, 2023. Disponível em: <https://doi.org/10.3390/diagnostics13182979>. Acesso em: jul. 2024.

NEFOUSSI, Jonathan J. **Personal FHIR record: registro pessoal de saúde baseado no padrão FHIR**. Porto Alegre, 2018. Disponível em: <https://repositorio.ifrs.edu.br/bitstream/handle/123456789/1578/1234567891578.pdf>. Acesso em: jan. 2024.

NETO, Gillete C. C.; ANDREAZZA, Rosemarie; CHIORO, Arthur. **Integração entre os sistemas nacionais de informação em saúde: o caso do e-SUS Atenção Básica**. *Revista de Saúde Pública*, São Paulo, p. 1-10, 2021. Disponível em: <https://doi.org/10.11606/s1518-8787.2021055002931>. Acesso em: nov. 2024.

OCDE. **Health at a glance 2019: OECD indicators**. Paris: OECD Publishing, 2019. 243 p. Disponível em: <https://doi.org/10.1787/4dd50c09-en>. Acesso em: jul. 2024.

ODDONE, Nanci Elizabeth. **Ciência da informação em perspectiva histórica: Lydia de Queiroz Sambaquy e o aporte da documentação (Brasil, 1930-1970)**. Rio de Janeiro, 2004.

ODDONE, Nanci. **O IBBD e a informação científica: uma perspectiva histórica para a ciência da informação no Brasil**. *Ciência da Informação*, Brasília, DF, v. 35, n. 1, p. 45-56, jan./abr. 2006. Disponível em: <https://doi.org/10.18225/ci.inf.v35i1.1152>. Acesso em: set. 2024.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS). **Diretor-geral da OMS declara que surto de monkeypox constitui uma emergência de saúde pública de importância internacional**. 23 jul. 2022. Disponível em: <https://www.paho.org/pt/noticias/23-7-2022-diretor-geral-da-oms-declara-que-surto-monkeypox-constitui-uma-emergencia-saude>. Acesso em: jul. 2024.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS). **Diretor-geral da OMS declara surto de mpox como uma emergência de saúde pública de importância internacional**. 14 ago. 2024. Disponível em: <https://www.paho.org/pt/noticias/14-8-2024-diretor-geral-da-oms-declara-surto-mpox-como-uma-emergencia-saude-publica>. Acesso em: ago. 2024.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS). **OMS declara emergência de saúde pública de importância internacional por surto de novo coronavírus**. 30 jan. 2020. Disponível em: <https://www.paho.org/pt/news/30-1-2020-who-declares-public-health-emergency-novel-coronavirus>. Acesso em: jul. 2024.

OTLET, Paul. **Traité de documentation: le livre sur le livre, théorie et pratique**. Bruxelles: Mundaneum, 1934. 452 p.

PAIVA, Douglas F. C.; MATEUS, Rafael C.; ARAÚJO JÚNIOR, Rogério. **A gestão de informações em larga escala na promoção da saúde: um estudo sobre a representação e organização da informação em arquivos de saúde pública**. In: RIBEIRO, Fernanda; DUARTE, Zeny; SILVA, Armando M. D. (org.). **Medicina, informação, tecnologias e humanidades: perspectivas cruzadas**. [S.l.]: [s.n.], 2024. p. 41-52. Disponível em: <https://doi.org/10.21747/978-989-8970-85-5/med>. Acesso em: dez. 2024.

PHELAN, Alexandra L. **Planetary health: a global health emergency under international law?** *Opinio Juris*, 2 maio 2024. Disponível em: <https://opiniojuris.org/2024/02/05/planetary-health-a-global-health-emergency-under-international-law/>. Acesso em: jun. 2024.

RICHARDSON, Joshua; ABRAMSON, Erika; KAUSHAL, Rainu. **The value of health information exchange**. *Journal of Healthcare Leadership*, p. 17-23, 13 abr. 2012. Disponível em: <http://dx.doi.org/10.2147/JHL.S16438>. Acesso em: nov. 2024.

RISI JÚNIOR, João B. **Informação em saúde no Brasil: a contribuição da Ripsa**. *Ciência & Saúde Coletiva*, Brasília, DF, p. 1049-1053, 2006. Disponível em: <https://doi.org/10.1590/S1413-81232006000400025>. Acesso em: nov. 2024.

ROBERTS, Lawrence Gilman. **Machine perception of three-dimensional solids**. 1963. Disponível em: <https://dspace.mit.edu/handle/1721.1/11589>. Acesso em: ago. 2024.

ROUSSEAU, Jean-Yves; COUTURE, Carol. **Os fundamentos da disciplina arquivística**. Lisboa: [s.n.], 1998.

RUAS, Sandra Silva Mitraud; MACIEL, Rafael Fábio. **Descoberta de conhecimento e mineração de dados em saúde pública: conceitos e panorama atual**. *RevistaFT*, Rio de Janeiro, v. 27, n. 118, jan. 2023.

SALEHI, Ahmad Waleed et al. **A study of CNN and transfer learning in medical imaging: advantages, challenges, future scope**. *Sustainability*, v. 15, 2023. Disponível em: <https://doi.org/10.3390/su15075930>. Acesso em: nov. 2024.

SALES, Odete Máyla Mesquita; PINTO, Virginia Bentes. **Tecnologias digitais de informação para a saúde: revisando os padrões de metadados com foco na interoperabilidade**. *Reciis – Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, Rio de Janeiro, v. 13, p. 208-221, jan./mar. 2019. Disponível em: <http://dx.doi.org/10.29397/reciis.v13i1.1469>. Acesso em: abr. 2024.

SEMBAY, Márcio J.; MACEDO, Douglas D. J. D.; GIMENES, Alexandre Augusto Marquez F. **HL7 FHIR baseado em W3C PROV para alcançar a proveniência de dados em sistemas de informação em saúde**. In: **Workshop de Informação, Dados e Tecnologia**, 5., 2022. Disponível em: <https://widat2022.ufes.br/wp-content/uploads/2022/11/st-1/st1-2-HL7%20FHIR%20baseado%20em%20W3C%20PROV.pdf>. Acesso em: nov. 2024.

SILVA, Eliezer P. D. **A trajetória da arquivologia: três visões sobre os arquivos**. In: **CONGRESSO NACIONAL DE ARQUIVOLOGIA (CNA)**, 3., 2008. *Anais...* 2008. Disponível em: <http://arquivistica.fci.unb.br/wp-content/uploads/tainacan-items/476350/818817/Anais-III-CNA-636-655.pdf>. Acesso em: jul. 2024.

SILVA, Iago R. R. et al. **Utilização de redes convolucionais para classificação e diagnóstico da doença de Alzheimer**. 2018. Disponível em: <https://www.researchgate.net/publication/327384767>. Acesso em: nov. 2024.

SOUSA, Renato Tarciso Barbosa de. **Os princípios da teoria da classificação e o processo de organização de documentos de arquivo**. *Arquivo & Administração*, Rio de Janeiro, p. 1-56, jan./jun. 2007.

SOUZA, Edivanio Duarte de. **A institucionalização da ciência da informação no Brasil: elementos disciplinadores do campo científico**. João Pessoa, 2012. Disponível em: <https://periodicos.ufpb.br/index.php/ies/article/download/13297/8568>. Acesso em: set. 2024.

SOUZA, Maria de Fátima Marinho de. **Dos dados à política: a importância da informação em saúde**. Brasília, DF, 2008. Disponível em: http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742008000100001. Acesso em: nov. 2023.

TABIK, S. et al. **COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images**. *IEEE Journal of Biomedical and Health Informatics*, v. 24, 2020. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9254002>. Acesso em: dez. 2023.

WORLD HEALTH ORGANIZATION (WHO). **Ethics and governance of artificial intelligence for health: WHO guidance**. Geneva: WHO, 2021. Disponível em: <https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf?sequence=1>. Acesso em: mar. 2024.

WU, Ying. **An introduction to computer vision**. Evanston: [s.n.], 2017. (EECS 432 – Advanced Computer Vision Notes Series 1).