Desenvolvimento de Sistema para Realização de Metanálises com a Utilização de Técnicas de Mineração de Textos: Uma Aplicação em Genética Forense

Camila Siqueira Lins[1], Flávio Rosendo da Silva Olivera[1]

Curso Tecnológico em Análise e Desenvolvimento de Sistemas
[1]Instituto Federal de Ensino, Ciência e Tecnologia de Pernambuco (IFPE)
Paulista - PE - Brasil

csl1@discente.ifpe.edu.br
flavio.oliveira@paulista.ifpe.edu.br

Resumo. *A Metanálise é uma poderosa ferramenta de revisão sistemática, fundamental para validações estatísticas em estudos científicos. Na área biomédica, o principal repositório acessado é o PubMed, onde sua concentração de leitores é voltada principalmente para os títulos e resumos dos artigos, que chegam a ser visualizados 2,5 vezes mais do que os textos completos. Uma importante área do conhecimento é a Genética Forense, que é essencial para analisar evidências biológicas em investigações criminais e questões legais. Nesse contexto, esse trabalho visa a criação de um sistema automatizado, capaz de extrair informações do PubMed. Nesta aplicação visa-se gerar uma base de dados da literatura em genética forense, por onde serão gerados gráficos, agrupamentos e análises, facilitando e automatizando a criação de revisões sistemáticas e metanálises. gerando inferências e permitindo a criação de trabalhos científicos de grande impacto de forma otimizada e com maior confiabilidade pela sua capacidade de replicação e uso de ferramentas estatísticas.*
Palavras-chaves: *genética forense; mineração de dados; revisão sistemática; metanálise.*

Abstract: *Meta-analysis is a powerful systematic review tool, essential for statistical validations in scientific studies. In the biomedical field, the primary repository accessed is PubMed, where the focus of readers is mainly on the titles and abstracts of articles, which are viewed 2.5 times more than the full texts. An important area of knowledge is Forensic Genetics, which is essential for analyzing biological evidence in criminal investigations and legal issues. In this context, this work aims to create an automated system capable of extracting information from PubMed. This application aims to generate a database of literature in forensic genetics, from which graphs, clusters, and analyses will be generated, facilitating and automating the creation of systematic reviews and meta-analyses. This will generate inferences and allow the creation of high-impact scientific works in an optimized way, with greater reliability due to its ability to replicate and use statistical tools.*
Keywords: *forensic genetics; data mining; systematic review; meta-analysis.*

Instituto Federal de Educação, Ciências e Tecnologia de Pernambuco. Campus Paulista.
Curso de Análise e Desenvolvimento de Sistemas. 24 de julho de 2024.

1 . Introdução

A ciência forense, também conhecida como ciência legal, utiliza métodos científicos para investigar casos criminais. Tudo descoberto em uma cena de crime como impressões digitais e amostras de DNA, podem ser úteis para gerar evidências (Husan, 2022). Genética Forense é um ramo derivado da ciência legal e criminalística, e envolve a análise de variações dentro de populações através do estudo de características herdadas. O resultado dessas análises fornecem evidências cujo propósito é resolver conflitos legais. A Genética Forense se estabelece como uma área interdisciplinar que combina princípios da Genética Molecular, Biologia Forense, Bioinformática e Estatística para analisar evidências biológicas em investigações criminais e questões legais (Arenas *et al*., 2017). A Ciência Forense é muito importante pois sem a sua ajuda, alguns casos não poderiam ser comprovados. (Husan, 2022).

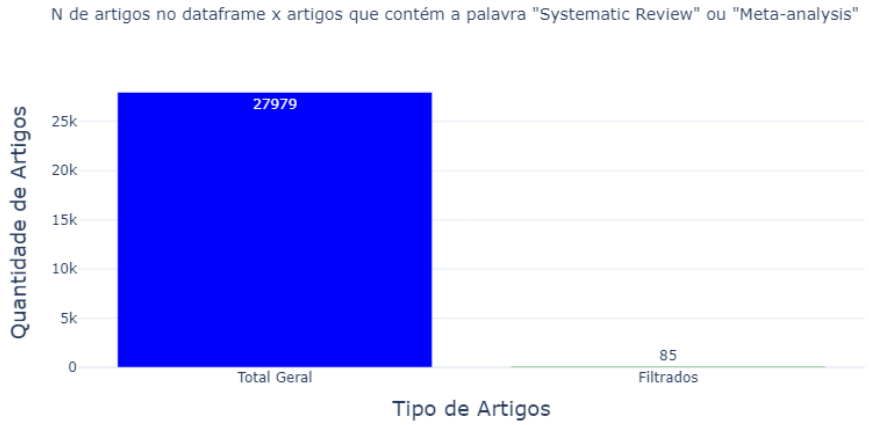
Ao adentrar nas publicações científicas existentes, é evidenciado que a revisão da literatura é essencial para compreender o estado atual de qualquer área científica. Inicialmente desenvolvida na medicina devido à evidência de que conselhos baseados em opiniões de especialistas são menos confiáveis do que aqueles fundamentados em resultados experimentais, a revisão da literatura é um processo complexo e propenso a erros e subjetividade. Para mitigar esses desafios, abordagens como a Revisão Sistemática (RSL) têm sido adotadas, embora ainda enfrentem problemas significativos de tempo, erros e subjetividade (Stefanovic *et al,* 2021).

Por outro lado, a Metanálise surge como uma ferramenta importante para o desenvolvimento de pesquisas. Ela se destaca como uma abordagem sistemática objetiva, ou seja, ela é uma forma de SLR fundamentada em análises estatísticas replicáveis, que garante resultados baseados em evidência. Esta metodologia oferece muita flexibilidade em sua utilização, abrangendo uma ampla gama de aplicações que se adaptam de acordo com o objetivo da pesquisa e da disponibilidade dos dados (Borenstein *et al*, 2009).

No entanto, pesquisas indicam que atividades de SLR são intensivas e manuais, sugerindo que a aplicação de tecnologia poderia otimizar esse processo. Existem diversos estudos que propõem o uso de ferramentas de análise de texto e extração de dados para enfrentar esse desafio (Scott *et al*, 2021). Os pesquisadores buscam essas ferramentas para tornar suas pesquisas mais eficientes, aprimorar a colaboração, ter um melhor controle do protocolo e histórico, além de melhorar a qualidade das diretrizes e padronizações. A extração de dados é a etapa em que os pesquisadores mais procuram auxílio desses recursos (Munn *et al*, 2020). É necessário ferramentas que visam auxiliar a seleção dos artigos dentro dos critérios de inclusão e exclusão de forma automática, eliminando duplicatas, aprimorando os processos de busca, criando subgrupos, estabelecendo correlações e expandindo a coleta de dados para uma maior gama de análises estatísticas (Kitchenham *et al*, 2009).

A análise dos dados levantados por esta pesquisa, apresentada na figura 1, revelou que apenas 0,3% das publicações na área forense mencionava a expressão "revisão sistemática" ou "meta-análise" em seus títulos ou resumos. Essa constatação motivou a aplicação do sistema desenvolvido nesta área de estudo.

Figura 1. Gráfico indicando o número de artigos na base de dados que são uma metanálise ou uma revisão sistemática



Fonte: Elaborado pela autora

Os dados usados nesta pesquisa tem origem do PubMed. Gerido pelo *National Center for Biotechnology Information* (NCBI), é um repositório gratuito de literatura disponível na web. Com milhões de consultas feitas diariamente, o PubMed tem sido, desde a sua criação, a principal fonte de acesso à literatura biomédica (Lu Z., 2011). Durante as pesquisas de artigos no PubMed, a maior parte dos leitores concentra sua atenção apenas no títulos, e quando interagem com algum artigo para obter mais informações, é comum que os resumos sejam visualizados 2,5 mais vezes do que os textos completos, tornando-os a parte mais acessada de um artigo biomédico (Beller *et al*, 2013).

Segundo Kitchenham e Brereton (2013), a proposta de ferramentas de análise textual foi amplamente discutida, com o objetivo de identificar palavras ou frases que descrevem artigos individuais e contar a frequência de termos importantes em cada artigo. Ferramentas de exibição visual também são recomendadas para identificar se artigos semelhantes em termos de frequência dessas palavras estão sendo tratados de forma consistente. Essas abordagens visam solucionar problemas de tempo, subjetividade na seleção e reduzir o risco de erros e omissões.

Nesse contexto, o propósito deste estudo é sugerir um método mais veloz e eficiente para a extração de informações destinadas à criação de metanálises. Propondo a implementação de um sistema automatizado capaz de extrair informações de pesquisas disponíveis no *PubMed*, permitindo a geração automática de gráficos e facilitando análises estatísticas. Com essa ferramenta, esperamos que as metanálises sejam realizadas de forma

mais ágil, com maior precisão e menor risco de erros, contribuindo para a produção de resultados mais confiáveis e consistentes.

Este trabalho está organizado da seguinte forma: na Seção 2 será apresentada a fundamentação teórica, abordando temas como Genética Forense, Metanálise, Mineração e Visualização de Dados. Na seção 3 serão apresentados os trabalhos relacionados, que abordam outras ferramentas de auxílio a revisão sistemática. Na Seção 4 serão evidenciados os materiais e métodos de estudo utilizados, detalhando a base de dados, as ferramentas e tecnologias aplicadas. Na Seção 5 serão exibidos os experimentos realizados e seus resultados, por fim, na Seção 6 serão realizadas as considerações finais sobre o trabalho desenvolvido, destacando as conclusões principais e sugestões para futuras pesquisas.

2. Referencial Teórico

2.1. Genética Forense

A genética é fundamental para compreender a humanidade de diversas maneiras. Trata-se do estudo dos genes, que podem ser analisados em níveis celulares, evolutivos, populacionais, entre outros. Uma das principais ferramentas dessa área é a análise de DNA e suas variantes. O estudo nessa área é crucial porque ajuda a entender o que define uma espécie e o que causa variações dentro da mesma espécie (Griffiths *et al*, 2015).

Os avanços nos estudos de genética permitiu que a ciência forense se transformasse no que é hoje, permitindo a identificação de indivíduos através das análises de DNA, o que foi chamado de "impressão digital de DNA". Essa inovação tornou possível que evidências de DNA fossem usadas na resolução de casos criminais. Hoje em dia, a genética forense desempenha um papel crucial na investigação de crimes e na realização de testes de parentesco, como os testes de paternidade, sendo uma área de importância global (Goodwin *et al*, 2010).

2.2. Metanálise

A revisão sistemática é um processo conduzido com base em critérios de inclusão pré-definidos. Devido ao seu elevado rigor metodológico, tornou-se uma ferramenta essencial para a área da saúde, sendo amplamente utilizada para apoiar o desenvolvimento de práticas e decisões clínicas (Moher *et al*, 2015).

A meta-análise é uma parte integrante da revisão sistemática, que utiliza métodos estatísticos, tanto descritivos quanto inferenciais, para resumir dados de vários estudos sobre um tema específico. O processo mais comum de condução de uma meta-análise envolve quatro etapas básicas: (1) busca, que inclui a definição da sequência de buscas e das bases de dados a serem utilizadas; (2) avaliação, que aplica critérios de inclusão, exclusão e qualidade; (3) síntese, que consiste na extração e categorização dos dados; e (4) análise, que narra os

resultados e chega a conclusões que respondem à pergunta de pesquisa. Essas técnicas auxiliam na geração de conhecimento a partir de múltiplos estudos, de maneira tanto quantitativa quanto qualitativa (Mengist *et al*, 2020).

2.3. Mineração e Visualização de Dados

A mineração de dados é uma metodologia que permite a descoberta de informações valiosas, conhecimentos ou padrões ocultos em grandes conjuntos de dados, utilizando diversas abordagens estatísticas. Em contraste com os métodos tradicionais que transformam dados em conhecimento através de análise e interpretação manual, a mineração de dados oferece várias vantagens. Essas abordagens são mais rápidas, vantajosas, economizam tempo e são objetivas (Yu *et al*, 2021).

A mineração de texto é uma subárea da mineração de dados que busca extrair novas informações valiosas de fontes não estruturadas (ou semi-estruturadas). Ela extrai informações de dentro dos documentos e agrega essas peças extraídas em toda a coleção de documentos-fonte para descobrir ou derivar novas informações. Assim, dado um conjunto de documentos como entrada, os métodos de mineração de texto buscam descobrir novos padrões, relacionamentos e tendências contidos nos documentos. Ao realizar essas tarefas, métodos de mineração de dados, como classificação ou aprendizado estatístico e geração de gráficos são frequentemente integrados para lidar e interpretar o que está contido no texto (Gonzalez *et al*, 2016).

A visualização de dados é aliada a mineração de dados, enquanto a mineração descobre padrões significativos em grandes repositórios de dados, a visualização de dados é a representação gráfica desses dados usando formas, cores e imagens para uma melhor compreensão. Essas técnicas têm sido utilizadas há muito tempo em diversos campos para aprimorar a percepção dos dados (Grinstein, 1995). Apresentar dados de forma visual facilita a interpretação, exploração de estruturas, detecção de tendências e padrões, além de exibir os resultados de maneira clara (Suwignyo, 2022).

3. Trabalhos relacionados

Esta seção abordará alguns dos sistemas existentes, suas funcionalidades e as limitações que esta proposta visa superar.

Na área de criação de revisões sistemáticas, diversos sistemas e ferramentas têm sido desenvolvidos para auxiliar pesquisadores a otimizar suas tarefas, que vão desde a coleta e categorização de dados até a geração de gráficos estatísticos e análise temporal (Stefanovic *et al,* 2021). Um exemplo disso é o *Mendeley*, um gerenciador bibliográfico que combina um aplicativo desktop e um site para ajudar pesquisadores a compartilhar e acessar dados de pesquisas (Kusumaningsih *et al*, 2023).

Embora existam muitas aplicações que auxiliam etapas específicas de uma revisão sistemática, nesta subseção destacamos três sistemas populares que oferecem funcionalidades

distintas entre si, mas que carecem de algumas características que este estudo pretende adicionar.

O RevMan (*Review Manager*) é um software de código aberto desenvolvido pela Cochrane para a criação e manutenção de revisões sistemáticas. Ele ajuda a compilar e marcar textos, realizar análises estatísticas e gerar gráficos baseados nos dados inseridos pelo usuário. Porém a exportação de dados do RevMan é restrita a formatos específicos, o que dificulta a reutilização dos dados em outras plataformas ou para novas análises. Outra limitação é que ele é específico para publicações na *Cochrane Library*, tornando-o menos flexível para pesquisadores que desejam reutilizar os dados de revisões sistemáticas para novas pesquisas ou análises detalhadas fora do ambiente *Cochrane* (Schmidt *et al*, 2019).

O *VOSviewer* (van Eck & Waltman, 2010) é uma ferramenta poderosa para criar e explorar mapas baseados em dados de redes. Ele permite a exploração de vínculos de coautoria, coocorrência, citação, acoplamento bibliográfico e cocitação. Seu diferencial então é a criação de mapas e imagens para a visualização dos dados, o que pode enriquecer muito o valor de uma revisão. Essas análises podem ser visualizadas de três formas distintas: rede, sobreposição ou densidade (Arruda *et al*, 2022).

O *Parsifal* é uma ferramenta gratuita para organização de revisões sistemáticas de literatura, facilitando seu monitoramento e disponível como aplicação *web*. o *Parsifal* oferece várias funcionalidades através do preenchimento de formulários e geração de documentos. Ele também suporta relatórios técnicos e a classificação de publicações através de tabelas, e tem o foco maior na sua possibilidade de convidar colaboradores para todos trabalharem no mesmo ambiente. No entanto, apresenta algumas limitações, como estar disponível apenas em inglês, dificuldades na gestão de colaboradores e relatos de problemas técnicos, e páginas de erro durante o convite para colaboração em projetos (Stefanovic *et al,* 2021).

Após a leitura dos arquivos de entrada, o sistema automatiza a montagem de um arquivo CSV, criando assim uma base de dados pronta para análises. O sistema proposto neste trabalho, apesar de ser acessado através dos scripts desenvolvidos em *Python*, possui uma interface gráfica para a visualização de seus resultados. Essa interface permite que os usuários interajam com os resultados de maneira intuitiva, e facilita o seu compartilhamento com os colaboradores de pesquisa.

A Tabela 1 ilustra as funcionalidades abrangidas pela proposta e compara se os três sistemas mencionados também oferecem essas funcionalidades ou não.

4. Materiais e Métodos

A ferramenta de mineração de textos desenvolvida para esta pesquisa apresenta uma solução inovadora para facilitar e otimizar a realização de revisões sistemáticas. Ao automatizar a extração e análise de dados de artigos publicados no *PubMed*, a ferramenta não só economiza

tempo e esforço dos pesquisadores, mas também garante uma abordagem mais abrangente e precisa na identificação de artigos relevantes.

Nas subseções 4.1 e 4.2 são apresentados os requisitos necessários para o desenvolvimento da aplicação. Em seguida, na Seção 3.3, são apresentados as ferramentas e tecnologias usadas para o desenvolvimento.

4.1. Requisitos Funcionais - RF

Os requisitos do sistema foram definidos para atender as necessidades de análise e processamento de dados científicos provenientes do PubMed. A separação em requisitos funcionais e não funcionais visa organizar as funcionalidades essenciais do sistema e garantir que ele atenda aos critérios de qualidade, desempenho e usabilidade esperados.

Começando com requisitos funcionais (RF), dez requisitos foram levantados, e descrevem as funcionalidades que o sistema deve implementar para realizar as tarefas de importação, processamento e análise de dados científicos. Foi levantado que, para iniciar a análise de dados, é fundamental a capacidade de importar os dados brutos do PubMed.

[RF001] O sistema deve permitir a importação de arquivos CSV e TXT contendo dados de artigos do PubMed: Este requisito garante que o sistema possa ingerir os dados na sua forma original, permitindo assim uma análise completa e abrangente.

[RF002] O sistema deve consolidar os dados dos arquivos importados em um único dataframe: A consolidação dos dados em um único *dataframe* é um passo crucial para a eficiência da análise.Ao unificar as informações de diferentes arquivos, o sistema facilita a aplicação de técnicas estatísticas e de mineração de dados.

[RF003] O sistema deve aplicar expressões regulares para extrair informações textuais: A extração de informações textuais através de expressões regulares permite que o sistema identifique e isole os elementos de interesse nos textos dos artigos, como autores, palavras-chave e afiliações.

[RF004] O sistema deve gerar gráficos estatísticos automaticamente a partir dos dados processados: A visualização gráfica é uma ferramenta poderosa para a interpretação de dados. Ao gerar gráficos estatísticos automaticamente, o sistema permite que o usuário identifique padrões e tendências de forma rápida e intuitiva.

[RF005] O sistema deve identificar e destacar os autores mais frequentes: A identificação dos autores mais frequentes é útil para identificar os principais pesquisadores em um determinado campo e suas colaborações.

[RF006] O sistema deve realizar análises temporais das publicações: A análise temporal permite identificar a evolução de um campo de pesquisa ao longo do tempo, destacando períodos de maior atividade e mudanças de paradigma.

[RF007] O sistema deve permitir a identificação de termos mais frequentes nos textos analisados: A identificação de termos frequentes é essencial para entender o vocabulário utilizado em um determinado campo e para identificar os principais conceitos e tópicos.

[RF008] O sistema deve realizar análises da produção científica por países e regiões específicas: A análise da produção científica por países e regiões permite identificar os principais centros de pesquisa e as colaborações internacionais.

[RF009] O sistema deve exportar os resultados das análises em formatos adequados para visualização (e.g., JSON, CSV, PNG): A exportação dos resultados em diferentes formatos permite que os usuários compartilhem os resultados de suas análises com outros pesquisadores e utilizem os dados em outras ferramentas.

[RF010] A interface gráfica deve permitir a visualização dos gráficos de forma interativa para ter acesso aos números e exportação da imagem: Uma interface gráfica interativa facilita a exploração dos dados e permite que o usuário personalize a análise de acordo com seus interesses.

Já nos requisitos não funcionais (RNF), cinco foram levantados, e eles especificam as características técnicas e operacionais do sistema, garantindo que ele seja fácil de usar, expansível e eficiente.

[RNF001] O sistema deve ser desenvolvido utilizando o Google Colab para facilitar a colaboração e execução do código: O Google Colab oferece um ambiente de desenvolvimento colaborativo e gratuito, ideal para projetos de análise de dados. Além disso, sua integração com outras ferramentas do Google facilita o compartilhamento e a colaboração.

[RNF002] O sistema deve utilizar bibliotecas de código aberto, como pandas, regex, Flask e Plotly, para minimizar custos: O uso de bibliotecas de código aberto garante a acessibilidade, a extensibilidade e a comunidade de usuários, além de reduzir os custos de desenvolvimento.

[RNF003] O sistema deve permitir a adição de novas funcionalidades sem necessidade de grandes reestruturações: Um sistema modular facilita a manutenção e a expansão do

sistema, permitindo que novas funcionalidades sejam adicionadas sem comprometer a estabilidade do sistema como um todo.

[RNF004] O usuário deve personalizar as buscas, escolhendo os termos mais relevantes para seu estudo: A personalização das buscas permite que o usuário adapte a análise aos seus objetivos específicos, aumentando a relevância dos resultados.

[RNF005] O usuário deve escolher quais gráficos rodar de acordo com a especificação de sua pesquisa:
A possibilidade de escolher quais gráficos gerar permite que o usuário foque nas informações mais relevantes para sua pesquisa, evitando a geração de gráficos desnecessários.

Tabela 1. Comparação entre funcionalidades dos sistemas

Problema	Funcionalidade	RevMan	VosViewer	Parsifal	Proposta
Categorizar trabalhos manualmente é algo trabalhoso	Análise de subgrupo	Sim	Não	Sim	Sim
Achar os termos mais frequentes no texto demanda muito tempo	Algoritmos de busca de termos	Não	Não	Não	Sim
A geração de gráficos estatísticos demanda muito tempo	Geração de gráficos automática	Sim	Não	Não	Sim
Identificar as instituições que trabalham em determinada área é um processo demorado	Identificação de países e Instituições dos artigos de forma automática	Não	Sim (pode variar)*	Não	Sim
Identificar a relevância de um tema ao longo dos anos manualmente é algo trabalhoso	Análise temporal das publicações	Não	Sim	Não	Sim

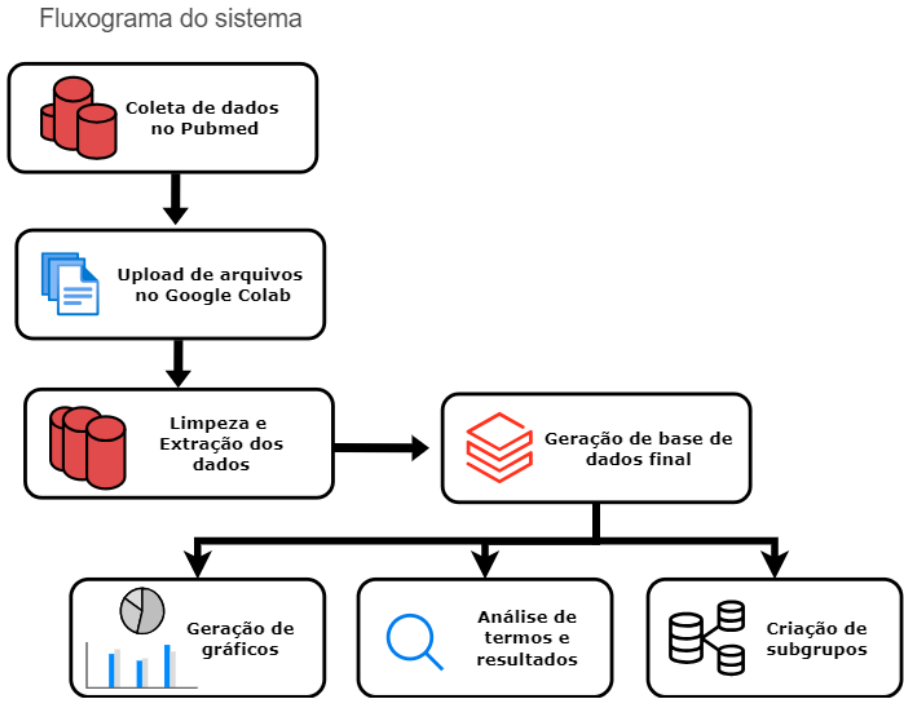
*dependendo das configurações e bases de dados acessadas.

Fonte: Elaborado pela autora

Na Figura 2, é apresentado um fluxograma que ilustra o processo de funcionamento do sistema proposto. Inicialmente, os arquivos coletados do *PubMed* são carregados no ambiente do *Google Colab*, onde o código está localizado. O código então realiza automaticamente a limpeza dos dados, removendo informações desnecessárias e também extraindo dados relevantes, como os resumos dos artigos. Este processo gera a base de dados final, a partir da

qual serão realizadas todas as análises necessárias para a revisão e metanálise. As análises incluem a criação de gráficos, a formação de subgrupos, e a verificação de termos, tudo adaptado às necessidades específicas da pesquisa do usuário.

Figura 2. Tela do fluxograma do sistema, de sua coleta até a geração de dados final



Fonte: Elaborado pela autora

4.3 Ferramentas e Tecnologias

Toda a extração e manipulação dos dados foi feita através da plataforma de desenvolvimento *Google Colab*. Após o tratamento da base de dados, as análises foram conduzidas utilizando expressões regulares, juntamente com bibliotecas clássicas de manipulação de dados como pandas para aplicar filtros. A geração de gráficos foi realizada utilizando a biblioteca *Plotly*, uma vez que esta biblioteca converte as figuras em formato *JSON*, facilitando sua manipulação e incorporação em aplicações web. No fim, os gráficos gerados poderão ser acessados através de uma interface gráfica.

A interface deste sistema foi concebida com o objetivo de proporcionar acesso simplificado aos gráficos, ou seja, ela não se destina a fornecer acesso direto ao sistema desenvolvido, mas sim a permitir o acesso aos resultados gerados por ele para profissionais e pesquisadores da área forense.

Ela foi desenvolvida utilizando uma combinação de tecnologias e ferramentas modernas e gratuitas, especialmente adequadas para a criação de aplicações web eficientes.

Para o desenvolvimento do *back-end*, foi utilizado *Flask*, um *microframework Python* que se destaca pela sua simplicidade e flexibilidade. O *Flask* facilitou a criação de rotas, permitindo a definição clara dos caminhos de acesso da aplicação, além de proporcionar uma gestão eficiente de *templates*, que são essenciais para a renderização dinâmica das páginas.

No *front-end*, foi empregado *JavaScript* para implementar a lógica e a interatividade da aplicação, conferindo-lhe dinamismo e capacidade de resposta às ações do usuário. O *HTML* foi utilizado para estruturar as páginas da aplicação, definindo a disposição dos elementos e garantindo a correta semântica dos conteúdos apresentados. Complementando essas tecnologias, o *CSS* foi utilizado para a estilização das páginas.

5. Aplicação Prática e Resultados

Esta seção apresenta a aplicação prática da ferramenta desenvolvida ao longo deste estudo. Na sua aplicação prática, foi realizada uma metanálise, detalhando o uso da ferramenta no contexto da Genética Forense, analisando sua eficácia e os resultados obtidos. Este experimento foi projetado para validar a funcionalidade e o desempenho da solução proposta, demonstrando seu impacto e suas contribuições para a área de estudo.

Os dados foram coletados da base de dados *PubMed* utilizando os descritores "*Forensic AND DNA*", resultando em um total de 27.979 artigos publicados entre os anos de 1963 e 2024. Devido à limitação do *PubMed* de selecionar no máximo 10.000 artigos por vez, a coleta foi realizada em três etapas distintas:

1. **Primeira Etapa**: Abrangendo os anos de 1963 a 2004, foram selecionados 9.009 artigos.
2. **Segunda Etapa**: Abrangendo os anos de 2005 a 2013, foram selecionados 9.601 artigos.
3. **Terceira Etapa**: Abrangendo os anos de 2014 a 2024, foram selecionados 9.369 artigos.

Os artigos foram baixados em dois formatos: *CSV* e *TXT*. Pois os arquivos possuem informações diferentes O *CSV* já vem tabelado com informações importantes como o Título do artigo, *DOI*, autores, entre outros. Porém ele carece de informações importantes para a realização de uma revisão, como por exemplo o resumo do artigo. Essa informação e de países e instituições de cada artigo foram então coletadas dos metadados do arquivo *TXT*, identificando padrões na estrutura dos textos para extrair suas informações. Isso resultou em um total de seis arquivos, dois de cada formato para cada etapa de coleta. Posteriormente, esses arquivos foram combinados em uma única base de dados para análise subsequente. Essa metodologia permitiu a organização e a gestão eficiente de um grande volume de dados, mesmo com a limitação da seleção presente no site.

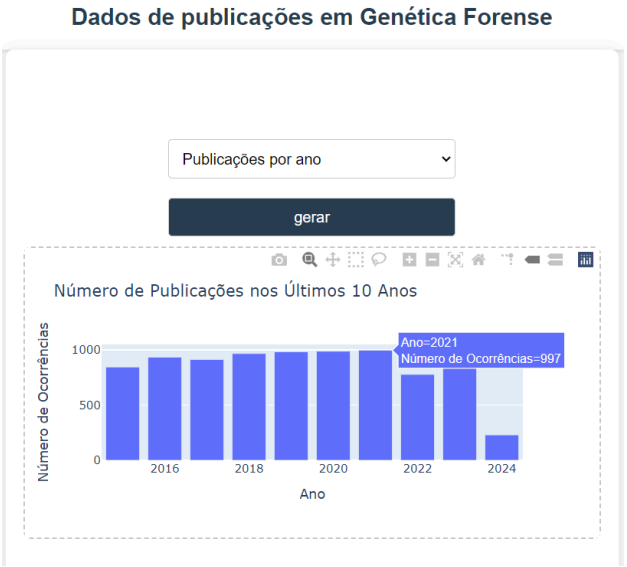
Dos arquivos *csv*, foram excluídas as colunas desnecessárias para as análises deste trabalho, como por exemplo a "PMID", restando colunas como Título, Autores, Ano de Publicação e similares. Para complementar os dados do *CSV*, foram utilizadas expressões regulares com a biblioteca do *Python regex* nos arquivos *TXT*. Esse processo foi realizado para todos os arquivos. Foi também desenvolvido um código para identificar artigos que não

possuíam informações sobre instituições ou resumos, evitando a inclusão de dados incompletos no *dataframe*. O resultado final foi um *csv* consolidado contendo 27.979 linhas e 9 colunas, integrando todas as informações coletadas e processadas. A base de dados foi processada de acordo com as funcionalidades descritas na Tabela 1, abrangendo desde a categorização automática de trabalhos até a geração de análises temporais, metanálises de autores influentes, identificação dos termos mais utilizados em pesquisas, países com maior produção científica global, e análise das publicações por áreas específicas.

5.1 Análise Temporal

Uma das funcionalidades centrais do sistema foi a análise temporal das publicações. Utilizando técnicas avançadas de visualização de dados, foram gerados gráficos que demonstram a evolução das pesquisas ao longo dos anos. Isso permitiu identificar tendências e padrões significativos na área estudada. De 2015 a 2021, o número de publicações manteve-se relativamente estável, indicando um interesse consistente na área estudada. No entanto, houve uma queda no número de publicações em 2022. Esta diminuição pode refletir vários fatores, como mudanças no foco de pesquisa, questões de financiamento, ou impactos externos como a pandemia de COVID-19, que podem ter afetado a capacidade dos pesquisadores de conduzir e publicar seus trabalhos. A identificação dessas tendências é crucial para compreender o desenvolvimento e a direção futura das pesquisas na área.

Figura 3. Tela do gráfico de publicações nos últimos 10 anos



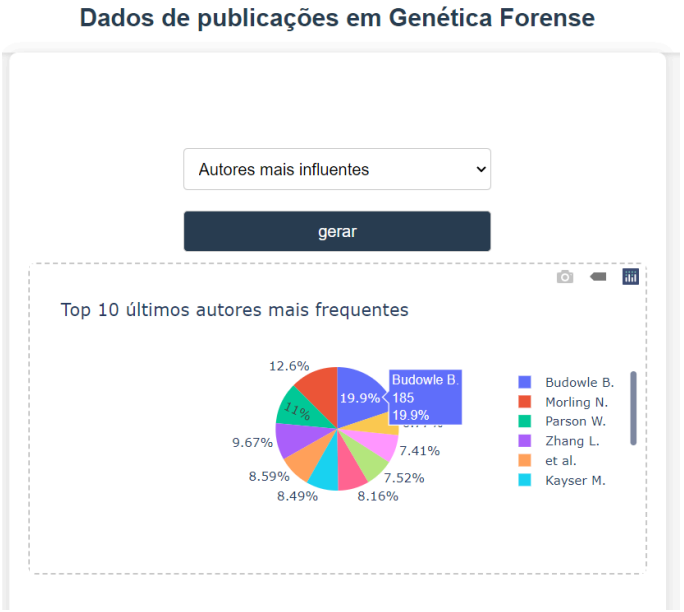
Fonte: Elaborado pela autora

5.2 Autores Mais Influentes

O sistema também gerou gráficos para identificar os autores mais influentes no campo de estudo. Isso foi alcançado através da frequência de cada autor. Conforme ilustrado no gráfico de pizza, Bruce Budowle emerge como o autor mais frequente, com 168 publicações em que

atua como orientador. Esse destaque indica a influência significativa de Budowle na área, possivelmente devido a sua extensa experiência e contribuições substanciais ao campo da Genética Forense. Outros autores notáveis incluem Morling N. e Parson W., que também aparecem com frequência. A identificação desses pesquisadores proeminentes pode ajudar a entender as principais influências e tendências nas colaborações científicas atuais, bem como a dominância de autores em áreas específicas e direcionar parcerias na área.

Figura 4. Tela do gráfico dos 10 autores que mais aparecem como último autor



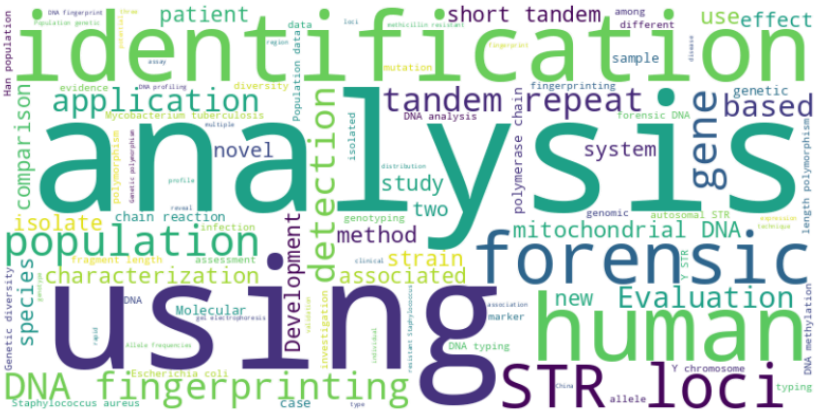
Fonte: Elaborado pela autora

5.3 Análise dos Termos Mais Usados

Utilizando técnicas de processamento textual, foram identificados os termos mais frequentemente utilizados nas pesquisas analisadas. Conforme mostrado na nuvem de palavras, termos como "*identification*", "*analysis*", "*using*", "*forensic*", "*human*", e "*DNA fingerprinting*" aparecem com destaque. Esses termos indicam que a maioria das pesquisas está focada em métodos de identificação e análise utilizando ferramentas forenses em amostras humanas, com um forte ênfase em técnicas de impressão digital de DNA.

Além disso, palavras como "*STR loci*", "*gene*", "*population*", e "*tandem repeat*" sugerem que as discussões sobre genética populacional e a utilização de marcadores de STR (*Short Tandem Repeats*) são temas recorrentes e de importância crescente no campo. O uso frequente desses termos aponta para a relevância das tecnologias de análise genética no avanço das investigações forenses e suas aplicações na identificação de indivíduos e no estudo da diversidade genética. Essa análise dos termos mais utilizados oferece uma visão abrangente sobre os tópicos prioritários e emergentes na literatura científica recente, permitindo uma melhor compreensão das áreas de interesse e desenvolvimento na pesquisa forense.

Figura 5. Nuvem de palavras dos termos mais frequentes no resumos dos artigos

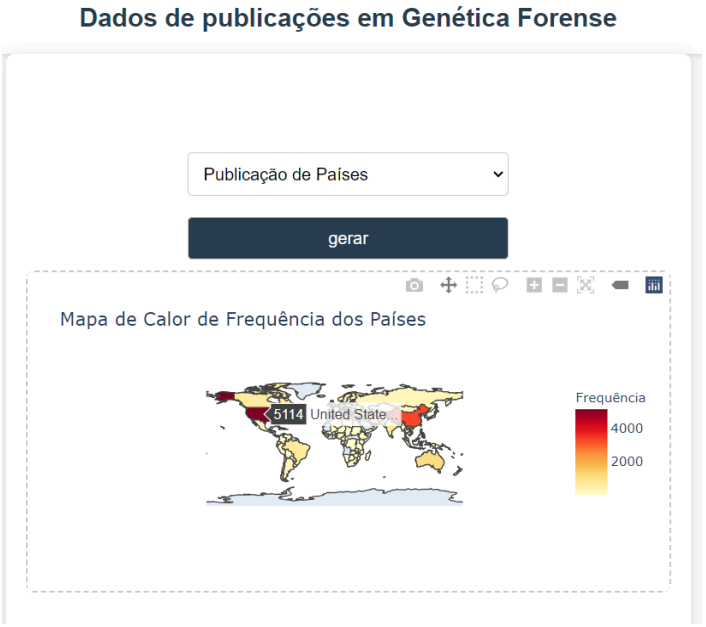


Fonte: Elaborado pela autora

5.4 Produção Científica por Países

O sistema permitiu analisar e comparar a produção científica entre diferentes países, destacando aqueles que mais contribuem para o avanço do conhecimento na área estudada. Essa análise foi crucial para entender as dinâmicas globais de pesquisa. Conforme mostrado no mapa de calor gerado, os Estados Unidos aparecem em disparado como o país com mais publicações científicas, com a China aparecendo logo em seguida. O domínio dos Estados Unidos e a ascensão da China em publicações científicas mostram a importância de políticas governamentais que incentivam a pesquisa e inovação, além de destacar a competição e colaboração saudável que impulsionam avanços científicos. Essa análise evidencia a importância das contribuições de ambos os países para o cenário global de pesquisa.

Figura 6. Tela do gráfico de número de autores por país

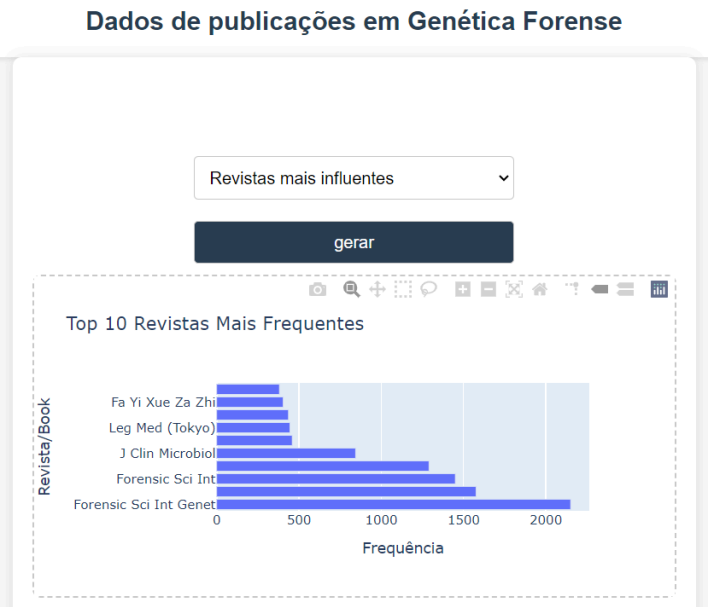


Fonte: Elaborado pela autora

5.5 Análise de Revistas mais frequentes na área

Foi realizada a contagem de frequência das revistas de publicação de genética forense. Destaca-se a revista Forensic Science International: Genetics como a mais proeminente, com mais de 2.000 publicações, evidenciando sua liderança e influência significativa no campo. A presença de revistas como a Journal of Clinical Microbiology, que não é especificamente voltada para a genética forense, indica que essa área de pesquisa muitas vezes se entrelaça com outros campos, como a microbiologia clínica. Isso pode sinalizar tanto a interdisciplinaridade e a aplicabilidade ampla dos métodos de genética forense quanto uma possível lacuna na existência de revistas especializadas o suficiente para cobrir todos os aspectos dessa área. Essas revistas não apenas fornecem uma plataforma para a disseminação de pesquisas de ponta, mas também estabelecem padrões para metodologias e práticas na área de genética forense. A alta frequência de publicações nessas revistas sugere que são fontes confiáveis e respeitadas dentro da comunidade científica.

Figura 7. Tela do gráfico de frequência de revistas



Fonte: Elaborado pela autora

6. Conclusão

O sistema implementado neste projeto foi projetado para otimizar a criação de metanálises e revisões sistemáticas a partir de dados coletados no *PubMed*. Utilizando um ambiente de execução no *Google Colab*, o código processa os dados, limpando e organizando informações relevantes como os resumos dos artigos, e gera uma base de dados consolidada. A partir dessa base, o sistema realiza diversas análises, como a criação de gráficos, a formação de subgrupos, e a verificação de termos mais frequentes. A interface gráfica desenvolvida para

visualização dos gráficos permite o acesso fácil e interativo aos resultados dessas análises, proporcionando uma ferramenta eficiente para pesquisadores na área de Genética Forense.

Os resultados dos experimentos demonstram que o sistema proposto é capaz de realizar uma ampla gama de análises e gerar informações valiosas a partir de grandes conjuntos de dados. A integração de técnicas avançadas de processamento de dados e visualização permitiu explorar profundamente a literatura científica, identificando padrões e tendências que seriam difíceis de obter manualmente.

Este projeto tem o potencial de causar um grande impacto não apenas na área da genética forense, mas em diversas outras áreas de pesquisa científica e acadêmica. O sistema automatizado desenvolvido facilitará significativamente o acesso e análise da vasta quantidade de literatura disponível no *PubMed*. Isso não apenas promoverá uma pesquisa mais eficiente e eficaz na genética forense, mas também em áreas como medicina, biologia, ciências sociais, entre outras.

Embora o sistema proposto para a criação automatizada de metanálises e revisões sistemáticas apresenta avanços significativos na análise de dados científicos, ele possui algumas limitações que devem ser reconhecidas e discutidas.

Primeiramente, o algoritmo desenvolvido para o estudo foi especificamente projetado para trabalhar com dados provenientes do *PubMed*. Apesar de o *PubMed* ser uma base de dados extensiva e de alta relevância na área biomédica, existem outras bases de dados igualmente importantes, como Google Acadêmico, *Scopus*, *LILACS* e *Web of Science*, que não foram incluídas na análise. Cada uma dessas plataformas possui suas particularidades e pode oferecer uma gama diversa de publicações e perspectivas que não são capturadas pelo sistema atual. A integração com essas outras bases de dados poderia enriquecer ainda mais as análises e fornecer uma visão mais abrangente do panorama científico.

Outra limitação significativa é a restrição do *PubMed*, que permite a exportação de no máximo 10.000 artigos por arquivo. Esta limitação requer que a coleta de dados seja dividida em múltiplas etapas e posteriormente combinada, o que pode introduzir complexidade adicional no processo e aumentar o tempo necessário para a preparação dos dados.

A proposta também não inclui uma análise aprofundada de todos os possíveis parâmetros e filtros que poderiam ser aplicados para refinar ainda mais a revisão sistemática e a metanálise, por não lidar com o conteúdo de um artigo completo, mas apenas do resumo. Além disso, o programa está atualmente operando localmente, o que permite a realização de testes detalhados e ajustes necessários antes de seu lançamento em um ambiente de produção. No futuro, a interface também será disponibilizada em um ambiente de hospedagem adequado, proporcionando acesso mais amplo e facilitado aos usuários finais.

Por último, embora o sistema facilite a automação e a eficiência na geração de metanálises, ele não substitui a necessidade de uma revisão crítica e interpretação dos resultados por parte dos pesquisadores. A análise automatizada pode fornecer *insights*

valiosos, mas a interpretação e contextualização dos dados ainda dependem do conhecimento e da experiência dos especialistas na área.

Em resumo, o sistema proposto oferece uma solução avançada para a criação de metanálises e revisões sistemáticas, mas é importante estar ciente dessas limitações para orientar futuras melhorias e garantir que o trabalho seja complementado por análises e abordagens adicionais conforme necessário.

Agradecimentos

A realização deste trabalho só foi possível graças ao apoio e incentivo de várias pessoas e instituições, às quais gostaria de expressar minha mais sincera gratidão.

Em primeiro lugar, agradeço à minha família, que sempre esteve ao meu lado, oferecendo amor e compreensão. Um agradecimento especial a Gabriel Lins pelo suporte incondicional ao longo desta jornada acadêmica.

Agradeço também ao Laboratório de Bioinformática e Biologia Evolutiva (LABBE) da Universidade Federal de Pernambuco (UFPE) por proporcionar os recursos necessários e um ambiente acolhedor e propício para a pesquisa e desenvolvimento deste trabalho. O apoio técnico e acadêmico recebido pelos professores Valdir Balbino e Sérgio Paiva foi indispensável para a conclusão deste projeto.

Às minhas queridas amigas, que sempre estiveram ao meu lado, oferecendo palavras de encorajamento. Obrigada por tornarem essa jornada mais leve e alegre.

Por fim, agradeço a todos os professores e colegas que, de alguma forma, contribuíram para o meu crescimento pessoal e acadêmico. A experiência compartilhada com todos vocês foi enriquecedora e inesquecível.

Referências

ARENAS, Miguel et al. Forensic genetics and genomics: Much more than just a human affair. **PLoS Genetics**, v. 13, n 9, p. e1006960, 2017.

ARRUDA, Humberto et al. VOSviewer and bibliometrix. **Journal of the Medical Library Association: JMLA,** v. 110, n. 3, p. 392, 2022.

BELLER, Elaine M. et al. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. **PLoS medicine**, v. 10, n. 4, p. e1001419, 2013.

BUSELLATO, Stefano. Zaratustra versus Parsifal. **Cadernos Nietzsche**, v. 38, n. 1, p. 84-105, 2017.

BORENSTEIN, Michael et al. **Introduction to meta-analysis.** Chichester, West Sussex, United Kingdom: John Wiley & Sons, 2009.

GOODWIN, William; LINACRE, Adrian; HADI, Sibte. **An introduction to forensic genetics.** Hoboken, New Jersey, EUA: John Wiley & Sons, 2010.

GONZALEZ, Graciela H. et al. Recent advances and emerging applications in text and data mining for biomedical discovery. **Briefings in bioinformatics,** v. 17, n. 1, p. 33-42, 2016.

GRIFFITHS, A. J. F., WESSLER, S. R., CARROLL, S. B., & DOEBLEY, J. **An Introduction to Genetic Analysis.** W.H. Freeman. USA: W.H. Freeman, 2015.

GRINSTEIN, Georges; THURAISINGHAM, Bhavani. Data mining and data visualization: Position paper for the second IEEE workshop on database issues for data visualization. In: WORKSHOP ON DATABASE ISSUES FOR DATA VISUALIZATION. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995. p. 54-56.

SUWIGNYO, Heri et al. Robust data mining visualization for learning outcomes. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND EDUCATION (ICIT&E), 2, Malang, Indonesia, 2022, pp. 54-59.

Husan, Sayedul. (2022). Role of Forensic Evidence in the Criminal Investigation: A Legal Analysis in Bangladesh Perspective. 1. 181-192.

KITCHENHAM, Barbara; BRERETON, Pearl. A systematic review of systematic review process research in software engineering. **Information and software technology**, v. 55, n. 12, p. 2049-2075, 2013.

KITCHENHAM, Barbara et al. Systematic literature reviews in software engineering–a systematic literature review. **Information and software technology**, v. 51, n. 1, p. 7-15, 2009.

KUSUMANINGSIH, Dewi; DARMAYANTI, Rani; LATIPUN, Latipun. Mendeley Software improves students' scientific writing: Mentorship and training. **Jurnal Inovasi Dan Pengembangan Hasil Pengabdian Masyarakat**, v. 2, n. 1, 2024.

LU, Zhiyong. PubMed and beyond: a survey of web tools for searching biomedical literature. **Database**, v. 2011, p. 1-13, 2011.

MENGIST, Wondimagegn; SOROMESSA, Teshome; LEGESE, Gudina. Method for conducting systematic literature review and meta-analysis for environmental science research. **MethodsX**, v. 7, p. 100777, 2020.

MOHER, David et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. **Systematic reviews**, v. 4, p. 1-9, 2015.

MUNN, Zachary et al. Are systematic review and guideline development tools useful? A Guidelines International Network survey of user preferences. **JBI Evidence Implementation**, v. 18, n. 3, p. 345-352, 2020.

SCHMIDT, Lena et al. Introducing RAPTOR: RevMan parsing tool for reviewers. **Systematic Reviews**, v. 8, p. 1-4, 2019.

SCOTT, Anna Mae et al. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. **Journal of clinical epidemiology**, v. 138, p. 80-94, 2021.

STEFANOVIC, Darko et al. Analysis of the tools to support systematic literature review in software engineering. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2021. p. 012013.

VAN ECK, Nees; WALTMAN, Ludo. Software survey: VOSviewer, a computer program for bibliometric mapping. **scientometrics,** v. 84, n. 2, p. 523-538, 2010.

VOSVIEWER Visualizing scientific landscapes. Centre for Science and Technology Studies, Leiden University, The Netherlands. Disponível em: https://www.vosviewer.com; free, donations accepted. Acesso em: 03 jun 2024.

YU, Bin et al. A survey on federated learning in data mining. **Wiley Interdisciplinary Reviews:Data Mining and Knowledge Discovery**, v. 12, n. 1, p. e1443, 2022.