

O uso de *Machine Learning* no Diagnóstico e Prognóstico do Câncer: uma Revisão Bibliográfica

Dora Yovana Barrios-Leal, Danilo Jordão Xavier, Rodrigo de Oliveira Plotze

Faculdade de Tecnologia de Ribeirão Preto (FATEC)

genyovana@gmail.com, danjordan55@gmail.com, rodrigoplotze@gmail.com

Resumo. O câncer foi responsável por cerca de 10 milhões de mortes em 2020. Devido a sua complexidade e o grande volume de dados disponível, esta revisão teve como objetivo descrever os principais algoritmos de machine learning usados no estudo da doença. Os resultados mostram que o câncer de mama, pulmão e cervical são os mais frequentes, e os algoritmos mais utilizados foram Support Vector Machine, Decision Tree, Random Forest, e k-Nearest Neighbors. As linguagens de programação mais utilizadas foram Python e R. A análise dos resultados deixa evidente o estabelecimento de técnicas de machine learning como ferramentas promissoras no diagnóstico e prognóstico do câncer.

Abstract. Cancer was responsible for about 10 million deaths in 2020. Due to its complexity and large data volume available, this review aimed to describe the main machine learning algorithms used to study the disease. The results show that breast, lung and cervical cancer are the most frequent. The most used algorithms were Support Vector Machine, Decision Tree, Random Forest, and k-Nearest Neighbors, while Python and R were the programming languages preferably used. The analysis of the results evidences the establishment of machine learning techniques as promising tools for cancer diagnosis and prognosis.

1. Introdução

Segundo a GLOBOCAN 2020, um banco de dados on-line que fornece estatísticas globais de câncer e estimativas de incidência e mortalidade em 185 países para 36 tipos de câncer (GLOBOCAN, 2020), em todo o mundo surgiram 19,3 milhões de novos casos de câncer (18,1 milhões excluindo câncer de pele não melanoma) e foram registradas quase 10,0 milhões de mortes por câncer (9,9 milhões excluindo câncer de pele não melanoma) no ano de 2020. O câncer de mama feminino ultrapassou o câncer de pulmão como o tipo mais diagnosticado, com uma estimativa de 2,3 milhões de novos casos (11,7%), seguido pelo de pulmão (11,4%), colorretal (10,0%), próstata (7,3%) e câncer de estômago (5,6%). O câncer de pulmão permaneceu como a principal causa de morte por câncer, com uma estimativa de 1,8 milhão de mortes (18%), seguido pelo câncer colorretal (9,4%), fígado (8,3%), estômago (7,7%) e mama feminina (6,9%) (SUNG et al., 2021).

O câncer é uma doença multifacetada capaz de se auto sustentar e interagir dinamicamente com o microambiente ao seu redor, no organismo do portador da doença. Por consequência, consegue responder e se adaptar de forma muito rápida a diferentes tratamentos, o que, até os dias de hoje frustra tanto pacientes, quanto pesquisadores e médicos, apesar do avanço significativo no entendimento dos processos relacionados a doença (BI et al., 2019).

Devido a sua complexidade, várias questões surgem durante cada estágio do tratamento, como um método confiável de detecção da doença, melhor detecção das lesões realmente cancerosas e correta delimitação das margens do tumor em uma intervenção cirúrgica. Além disso, é de grande importância também rastrear a evolução do câncer e o eventual desenvolvimento de resistência ao tratamento no decorrer do tempo, assim como a agressividade da doença (BI et al., 2019).

Em todos os campos indicados acima, uma série de avanços tem ocorrido nos últimos anos e um grande volume de dados tem sido documentado, como por exemplo, relacionado a avanços no campo de imagens utilizadas em diagnóstico e progressão de doenças. A avaliação de perfis moleculares de diferentes tipos de cânceres também tem gerado um volume enorme de dados. Esse grande volume de dados é o que atualmente está gerando a maior dificuldade em utilizar de maneira eficaz e eficiente todo o avanço que tem ocorrido até os dias de hoje. São tantos dados, que sua interpretação acaba se tornando complicada e o que deveria facilitar o entendimento da doença, acabou se tornando uma barreira (BI et al., 2019).

À medida que aprendemos mais sobre a doença em si, aprendemos mais sobre o poder das ferramentas que já estão disponíveis. Avanços recentes em metodologias de inteligência artificial (IA) fizeram grandes avanços na quantificação automática de padrões radiográficos em dados de imagens médicas. Por exemplo, o *deep learning* é um método especialmente promissor que pode aprender diferentes padrões apresentados em imagens, principalmente quando há a disponibilidade de um grande volume de dados. Foi demonstrado que tal abordagem pode igualar e até superar o desempenho humano em tarefas específicas. Apesar de exigir grandes conjuntos de dados para treinamento, o *Deep learning* demonstrou relativa robustez (LECUN; BENGIO; HINTON, 2015; HOSNY et al., 2018). De forma geral o *machine learning* também é um ramo da IA que emprega uma variedade de técnicas estatísticas, probabilísticas e de otimização que permitem que os computadores “aprendam” com exemplos passados e detectem padrões difíceis de discernir de conjuntos de dados grandes, ruidosos ou complexos e tem sido usado principalmente como auxílio para diagnóstico e detecção de câncer (CRUZ; WISHART, 2006; KOUROU et al., 2015)

Esses são exemplos da aplicação da IA em várias áreas relacionadas ao câncer. No entanto, para ajudar a entender melhor como, onde e por quais razões a IA pode ser utilizada, é necessário também um melhor entendimento dos principais processos relacionados ao surgimento e ao desenvolvimento do câncer. Da mesma forma que existem diferentes áreas do câncer onde a IA pode ser aplicada, existem um grande número de abordagens de IA que podem ser utilizadas (Figura 1), gerando inúmeras possibilidades. Um estudo detalhado do problema deve ser realizado para selecionar qual delas é a mais apropriada.

Com o grande número de possibilidades e aplicações da IA como ferramenta em várias áreas relacionadas ao câncer, o objetivo do presente trabalho é realizar uma revisão bibliográfica, visando buscar quais são os principais algoritmos de *machine learning* utilizados para estudos de câncer, as principais linguagens de programação ou programas, assim como as metodologias de estudo empregadas, principalmente com relação as amostras utilizadas.

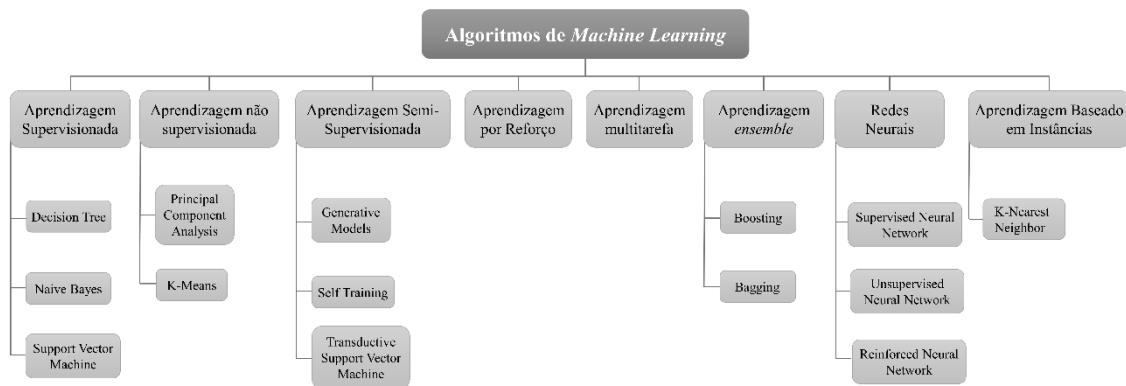


Figura 1. Diferentes tipos de algoritmos usados em *machine learning*.

Modificado de Batta, 2020.

2. Material e Métodos

A pesquisa foi realizada a partir de uma consulta sistemática nas bases de dados de Google Acadêmico e PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) utilizando-se os seguintes termos (descritores): *machine learning algorithm cancer*. Foi utilizada uma restrição na busca com relação a data de publicação, sendo utilizados aqueles artigos publicados no intervalo de tempo entre 2010 até abril de 2022. Os critérios adotados para a inclusão dos artigos foram os seguintes: descrição do tipo de câncer, tipo de amostra, base de dados, os diferentes algoritmos e as diferentes linguagens de programação ou programas utilizados. E como critério de exclusão: falta de informação sobre os algoritmos utilizados, tipo de câncer estudado e revisões bibliográficas. Todos os gráficos aqui representados foram gerados utilizando os pacotes básicos do programa R (R CORE TEAM, 2020).

3. Resultados e Discussão

Na consulta pelos descritores “*machine learning algorithm cancer*” foram recuperados aproximadamente 107.000 resultados na base de dados de Google Acadêmico e 1071 na PubMed. No total, foram analisados 70 artigos e somente 38 destes cumpriam com os critérios de inclusão adotados nesta revisão (Tabela 1).

3.1 Câncer e Base de Dados

Os tipos de câncer encontrados de forma mais frequente nesta revisão foram o de mama, pulmão e cervical (41%, 15% e 10% respectivamente) (Tabela 1 e Figura 2), que

são concordantes com a alta incidência e mortalidade no mundo, principalmente os de mama e pulmão. Por conta desses dados, é de grande importância que maiores esforços sejam empregados em pesquisas para melhorar o diagnóstico e prognóstico destes tipos de cânceres, assim como a maior disponibilidade de dados para estudos futuros.

O câncer de mama é o câncer mais comum entre as mulheres, registrando 11,7% dos novos casos de câncer, superando os 11,4% novos casos de câncer de pulmão, que é mais comum entre os homens. No entanto, o primeiro representa 6,9% das mortes mundiais, porcentagem menor que os 18,0% de mortes relacionadas ao câncer de pulmão. O câncer cervical está entre os nove cânceres mais comuns, com 3.1% dos novos casos e 3.4% das mortes mundiais (SUNG et al., 2021). No Brasil o câncer de mama também é o mais comum entre as mulheres com 29,7% dos novos casos. Já entre os homens, o mais comum é o câncer de próstata, com 29,2% e em terceiro lugar estão o de traqueia, brônquio e pulmão, com 7,9% (INCA, 2020).

Dentre os estudos avaliados no presente trabalho, a base de dados mais utilizada foi a *Breast Cancer Wisconsin Data Set* (WDBC - 26%). Esta base de dados proporciona imagens histológicas, com as características dos núcleos celulares presentes nas imagens digitalizadas de um aspirado por agulha fina (PAAF) proveniente de uma massa mamária e conta com um número de total de 32 atributos celulares (ver Tabela 1) (Disponível em: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))). Em 24% (BDPE) dos artigos avaliados, a base de dados utilizada foi obtida durante o próprio estudo, geralmente de diferentes hospitais. Entre os tipos de dados mais frequentes estavam as características demográficas (sexo e idade), tabagismo, alcoolismo, gravidez, diagnóstico de HPV e características histoclinicas (classificação e tamanho do tumor, entre outras) (Anexo Tabela 1). Outros tipos de dados interessantes que foram produzidos nos estudos foram a inclusão de análises de perfis genéticos e de expressão gênica (18%).

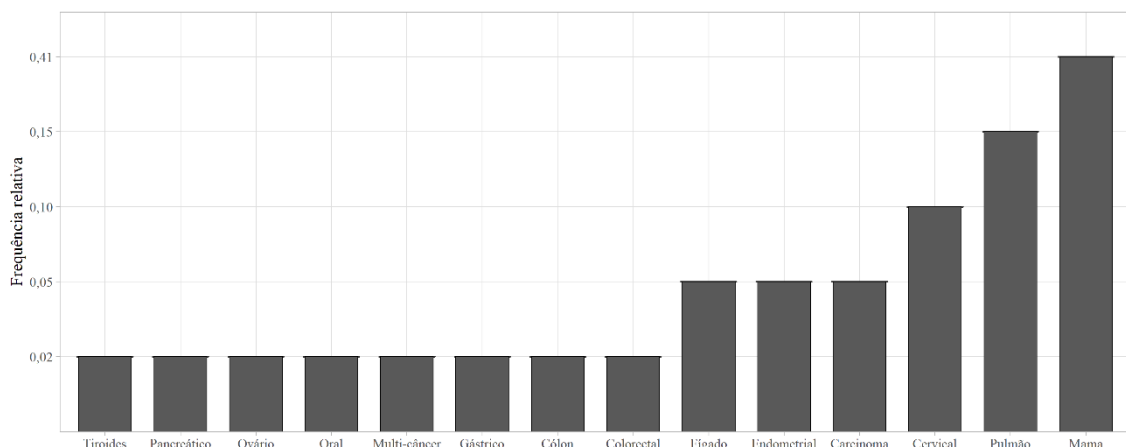


Figura 2. Frequência relativa do tipo de cânceres encontrados nesta revisão.

Fonte: (Autoria Pessoal).

3.2 Algoritmos de *machine learning*

A metodologia de *machine learning* utiliza diferentes algoritmos para resolver problemas com diferentes tipos de dados. No entanto, não existe um único tipo de

algoritmo capaz de resolver todos os problemas assim como não existe um problema que possa ser resolvido por apenas um algoritmo. Portanto, o tipo de algoritmo empregado depende do tipo de problema que você deseja resolver, do número de variáveis, do tipo de modelo que melhor se adequa a ele e assim por diante (BATTÁ, 2020). Nesta revisão, 85% dos estudos empregou mais de um algoritmo na base de dados utilizada, com o intuito de encontrar o melhor algoritmo que explicasse os dados. A utilização de 3 ou 4 algoritmos foi a porcentagem mais alta encontrada (21% e 24% respectivamente) (Figura 3) para análise dos dados onde foram utilizados algoritmos de diferentes tipos, como algoritmos de aprendizagem supervisionada e não supervisionada, redes neurais ou *ensemble*.

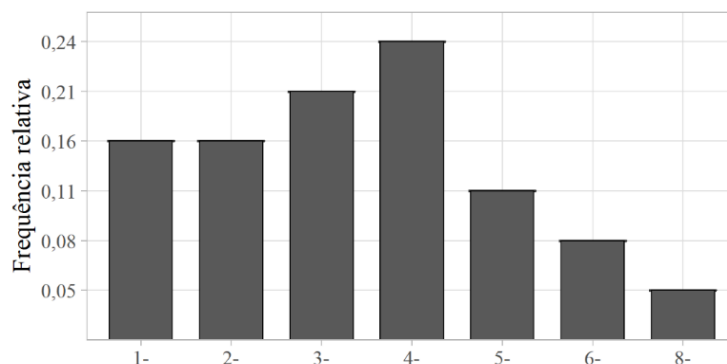


Figura 3. Frequência relativa da quantidade de algoritmos utilizados para analisar os dados nos artigos incluídos nesta revisão.

Fonte: (Autoria Pessoal)

No total, foram utilizados 34 diferentes algoritmos, sendo que o mais utilizado nos artigos desta revisão foi o *Support Vector Machine* (SVM) com 17%, seguido do *Decision Tree* (DT), *Random Forest* (RF) e *k-Nearest Neighbors* (KNN) com 12%, 10% e 10% respectivamente (Tabela 1 e Figura 4). Na categoria “Outros” estão agrupados 20 algoritmos diferentes que tiveram uma única ocorrência (Tabela 1), entre eles o ROMA (*Ovarian Malignancy Algorithm*) criado especificamente para o câncer de ovário ou algoritmos híbridos como o K-SVM (Híbrido entre KNN e SVM).

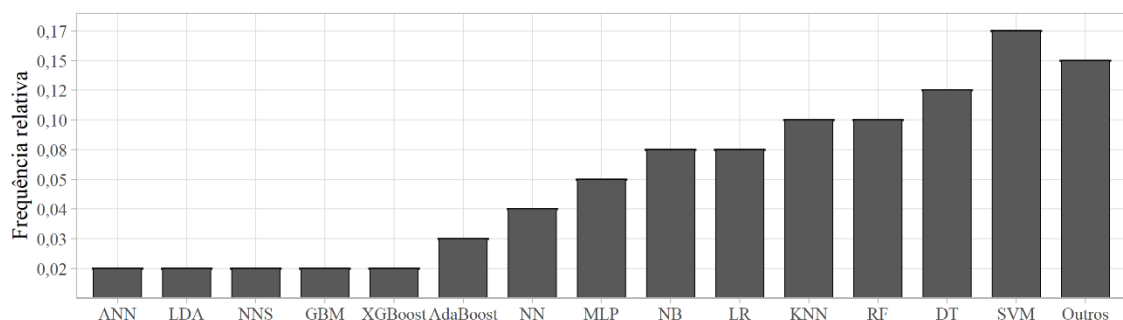


Figura 4. Frequência relativa dos algoritmos utilizados nos diferentes estudos.

Fonte: (Autoria Pessoal)

Os algoritmos SVM e DT são de aprendizagem supervisionada, o que significa que precisam de assistência externa, onde um conjunto de dados de entrada é dividido em dados de treinamento e teste. O conjunto de dados de treinamento tem uma variável de saída que precisa ser prevista ou classificada, ou seja, o algoritmo tem conhecimento

prévio de quais devem ser os valores de saída (por exemplo, paciente com câncer versus controle saudável) (VIEIRA; LOPEZ PINAYA; MECHELLI, 2019; BATTA, 2020). O SVM tornou-se atualmente o mais poderoso algoritmo de classificação em termos de precisão preditiva, pela sua capacidade de aprender padrões de classificação de dados com precisão e reprodutibilidade equilibradas. Além de sua alta versatilidade, (KHARYA; DUBEY; SONI, 2013; PISNER; SCHNYER, 2019) que se estende por vários cenários da ciência de dados, incluindo estudos de diferentes tipos de câncer como foi observado nesta revisão, pesquisa de distúrbios cerebrais (PISNER e SCHNYER, 2019), previsão de diagnóstico de diabetes (VILORIA et al., 2020), modelagem de nicho ecológico (BARRIOS-LEAL; NEVES-DA-ROCHA; MANFRIN, 2019) e discriminação entre seleção de varredura e neutralidade em genomas de populações humanas (SCHRIDER e KERN, 2018). No presente trabalho, o algoritmo SVM mostrou uma acurácia maior que 70% entre os estudos avaliados.

O DT é uma classe de algoritmos de classificação autoritativa e a técnica consiste em separar recursivamente observações em ramos de decisão para construir uma árvore com o propósito de melhorar a precisão da previsão. A estrutura em árvore é usada em classificação de registros de dados desconhecidos e são usualmente usados em aplicações mais práticas (KHARYA; DUBEY; SONI, 2013; BATTA, 2020). Estes algoritmos também têm se tornando muito valorizados, sendo considerados os mais úteis e poderosos em mineração de dados (KHARYA; DUBEY; SONI, 2013; SOMVANSHI et al., 2017). A maior vantagem desta classe de algoritmos é a capacidade de lidar com o número de dados de entrada como nominais, numéricos ou alfabéticos e o processamento de dados que podem conter valores ausentes e com erros. Esta última característica é particularmente interessante já que os dados podem variar nas diferentes bases de dados disponíveis. Assim, ao usar as regras de decisão, os DT são usados para extrair dados de uma grande quantidade de conjuntos de dados disponíveis, mesmo com considerável variação entre as bases. O DT simplesmente classifica os dados que podem ser facilmente armazenados e, além disso, possibilitando uma nova análise e classificação dos dados tratados (SOMVANSHI et al., 2017). Nesta revisão só foram encontrados três estudos que utilizaram diferentes bases de dados para o diagnóstico ou prognóstico do câncer e usaram o algoritmo DT.

O RF foi o terceiro algoritmo mais frequente e é um classificador que consiste em um conjunto de classificadores estruturados em árvore com vetores aleatórios independentes, que são distribuídos de forma idêntica para votar em uma classificação geral para um determinado conjunto de entradas. Entre as vantagens, está uma maior precisão e eficácia no trabalho com grande quantidade de dados, gerenciamento de milhares de variáveis de entrada de forma rápida e eficaz, além do fornecimento de informações sobre variáveis que são importantes e de técnicas para estimar dados incompletos. Entre as desvantagens está o sobreajuste de um único conjunto de dados, principalmente nas tarefas de regressão (ABDULKAREEM e ABDULAZEEZ, 2021).

Por outro lado, o KNN que foi o quarto algoritmo mais frequente nesta revisão é de aprendizagem baseado em instâncias, que se refere a uma família de técnicas para classificação e regressão, que produz um rótulo/predicação de classe com base na similaridade da consulta com seu(s) vizinho(s) mais próximo(s) no conjunto de treinamento. Este algoritmo é de fácil implementação e entendimento, mas sua principal desvantagem é que se torna significativamente mais lento à medida que o tamanho dos

dados em uso aumenta (BATTÀ, 2020). Outros algoritmos encontrados nesta revisão são variantes dos algoritmos mais frequentes (SVM, DT, RF, KNN) ou híbridos entre esses algoritmos.

3.3 Métodos de Validação

Nos artigos revisados, o método de validação mais frequente foi a Validação Cruzada (VC) denominada *k-fold* e com diferentes valores de *k* ($k=5$, $k=10$, $k=15$). Entre outros métodos utilizados estavam o *Leave-One-Out Cross-Validation* (LOOCV) e validação interna (VI). Finalmente, em seis dos artigos avaliados, não possível recuperar o método de validação utilizado (NA) (Figura 5)

A ideia da validação cruzada surgiu na década de 1930 e a ideia foi se desenvolvendo posteriormente com a participação de outros pesquisadores (LARSON, 1931; MOSTELLER e WALLACE, 1963), e tem dois objetivos possíveis: 1) estimar o desempenho do modelo de aprendizado a partir de dados disponíveis usando um algoritmo, ou seja, avaliar a generalização de um algoritmo, e 2) comparar o desempenho de dois ou mais algoritmos e avaliar o melhor para os dados disponíveis. Assim, vários procedimentos têm sido propostos para atingir esses dois objetivos: *Resubstitution validation*, *Hold-Out Validation*, *K-Fold Cross-Validation*, *Leave-One-Out Cross-Validation* e *Repeated k-fold cross-validation* (REFAEILZADEH et al., 2020).

Atualmente, o método da validação cruzada é amplamente aceito nas comunidades de mineração de dados e aprendizagem de máquina, sendo utilizada principalmente como um procedimento padrão para a estimativa de desempenho e seleção do melhor modelo. A validação cruzada de dez vezes ($k D 10$) é a mais comum (REFAEILZADEH et al., 2020), como mostram nossos resultados onde *10-fold cross validation* foi utilizado em 41% entre todos os métodos encontrados de validação. Em conjunto, essa e outras variações de validação cruzada somam 82% entre os métodos de validação utilizados (Figura 5).

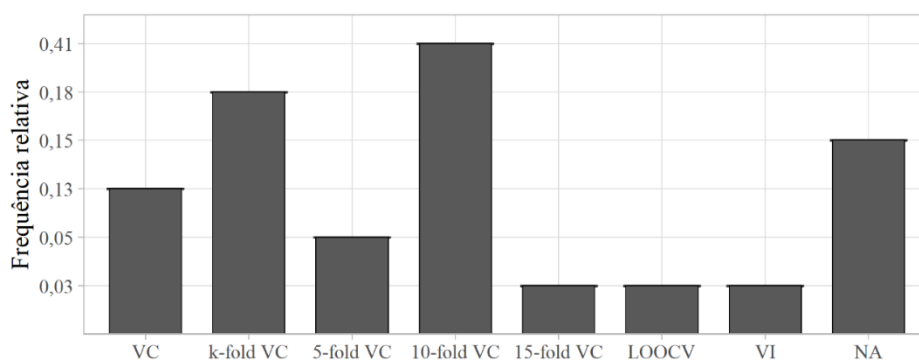


Figura 5. Frequência relativa dos métodos de validação encontradas nos artigos avaliados.

Fonte: (Autoria Pessoal)

3.4 Linguagens e Programas

No 32% dos artigos revisados não foi descrita nenhuma informação (NA) sobre as linguagens ou programas utilizados na aplicação dos algoritmos (Figura 6). As linguagens de programação *open source* mais utilizadas foram o Python (17%)

(PYTHON SOFTWARE FOUNDATION, 2020) e o R (12%) (R CORE TEAM, 2020). Ambos com a utilização de diferentes pacotes (Python: “Spark ML”, “MLlib”, “libsvm”, “scikit-learn”, “scikit-feature” e R: “e1071”). Os programas mais utilizados foram o Weka 3: *Machine Learning Software in Java* (15%), que é uma coleção de algoritmos de aprendizado de máquina (FRANK; HALL; WITTEN, 2016) e o MATLAB (12%) que é uma plataforma de programação e computação numérica usada para analisar dados, desenvolver algoritmos e criar modelos (THE MATHWORKS INC, 2010) (Figura 6). Entre outras ferramentas menos utilizados com uma ou duas ocorrências estão a TensorFlow que é uma plataforma de código aberto para *machine learning* (MARTÍN et al., 2015), o XGBoost, um software de código aberto que fornece um *framework* de *gradient boosting* para soluções de aprendizado de máquina em tarefas de classificação e regressão em diferentes linguagens (CHEN e GUESTRIN, 2016). O Orange, que é um kit de ferramentas de visualização de dados, aprendizado de máquina e mineração de dados de código aberto (DEMSAR et al., 2013) e o JUPYTER, que é o mais recente ambiente de desenvolvimento interativo baseado na Web para notebooks, código e dados que permite configurar e organizar fluxos de trabalho em ciência de dados, computação científica, jornalismo computacional e aprendizado de máquina (KLUYVER et al., 2016).

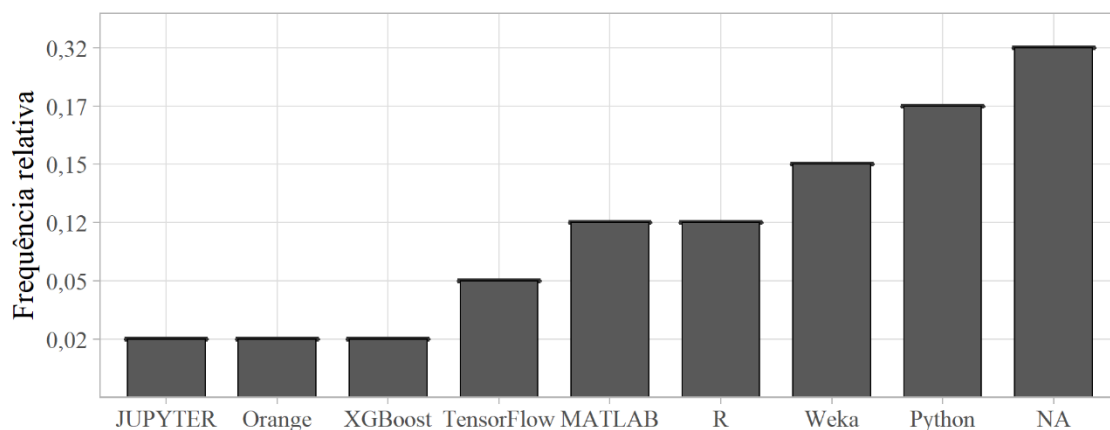


Figura 6. Frequência relativa das linguagens e programas utilizados para os experimentos com os diferentes algoritmos nos artigos avaliados.

Fonte: (Autoria Pessoal)

4. Conclusões

Com a aumento significativo no volume de dados gerados em diferentes estudos na medicina e mais especificamente na área do câncer, surgiu a necessidade de utilizar a inteligência artificial para avaliar e analisar esses dados. Nos últimos anos, apesar de se observar que o nicho dos algoritmos está crescendo rapidamente, alguns deles têm sido utilizados com maior frequência na análise de dados provenientes de estudos de câncer, por exemplo como o *Support Vector Machine* (SVM), *Decision Tree* (DT), *Random Forest* (RF) e *k-Nearest Neighbors* (KNN). Provavelmente, a adaptação desses algoritmos para dados e problemas mais específicos pode aumentar sua performance, podendo esta ser uma tendência, como observado em alguns trabalhos avaliados no presente estudo. Finalmente, com base na análise dos resultados desta revisão, fica evidente um aumento da integração de dados multidimensionais heterogêneos que, combinados com a aplicação de diferentes técnicas de classificação, podem fornecer ferramentas promissoras para o diagnóstico e prognóstico do câncer.

Tabela 1. Descrição dos artigos incluídos nesta revisão com a referência do artigo, tipo de câncer, o número de amostras incluídas (N), tipo de dados, algumas características importantes analisadas, banco de dados (BD), algoritmos e acurácia.

ID	Publicação	Câncer	N	Tipo Dados	Características importantes	BD	Algoritmos	Acurácia
1	(AGARAP, 2018)	Mama	569	Imagens digitalizadas da PAAF**	Características dos núcleos celulares**	WDBC	GRU-SVM, MLP, NNS, SR, SVM.	Todos > 90%
2	(AHMAD et al., 2013)	Mama	547	Características populacionais e histoclínicas	História de câncer familiar, idade do diagnóstico, infertilidade, local e tamanho do tumor entre outros.	ICBC	DT, ANN, SVM	Todos > 0,93
3	(AKTER et al., 2021)	Cervical	72	Características comportamentais	Comportamento na higiene pessoal, percepção de severidade e vulnerabilidade, suporte social e conhecimento entre outras.	CCBR	DT, RF, XGBoost	Todos > 93%
4	(ALGHUNAIM e AL-BAITY, 2019)	Mama	254	Microarrays	Expressão gênica e metilação de DNA	CGA	SVM, DT, RF	Todos > 88%
5	(AMRANE et al., 2021)	Mama	699	Características clínicas e demográficas	Histoclínicas	BCD	NBC, KNN	Todos > 96%
6	(ASRI et al., 2016)	Mama	699	Imagens digitalizadas da PAAF**	Características dos núcleos celulares**	WDBC	SVM, DT, NB, KNN	Todos > 95%
7	(CHU et al., 2020)	Carcinoma oral	408	Características clínicopatológicas, demográficas e estilo de vida	Sexo, idade, tabagismos, alcoolismo, HPV, historial de câncer, mucosa bucal, entre outros	BDPE	LR, DT, SVM, KNN	Todos > 70%
8	(FERNANDES et al., 2021)	Colorretal	20	Tecido de mucosa <i>ex-vivo</i>	Medições espectrais	BDPE	SLP, KNN, RFR, NN DTFMR, LRFMO	NA
9	(GÜNAKAN et al., 2019)	Endometrial	762	Tecido tumoral e características demográficas	Idade, tamanho do tumor, invasão miometrial e do espaço linfovascular	BDPE	NB	Variou entre 87% e 97%
10	(HE et al., 2020)	Carcinoma (13 canceres)	4909	Mutações somáticas	600 genes	ICGC	RF	0,8822

11	(HYUN et al., 2019)	Pulmão	396	Características clínicas e imagens de tomografias por emissão de pósitrons	Idade, sexo, fumante, tamanho do tumor e características radiômicas	BDPE	RF, NN, NB, LR, SVM	Variou entre 0,671 e 0,772
12	(JABBAR, 2021)	Mama	699	Imagens digitalizadas da PAAF**	Características dos núcleos celulares**	WDBC	Ensemble de BN, RBF	97%
13	(JEREZ et al., 2010)	Mama	3.679	Características histoclínicas, demográficas e terapêuticas	Idade, tamanho tumor, tipo de tratamento, receptores hormonais	El Álamo-I	MLP, SOM, KNN	0,77
14	(KALAFI et al., 2019)	Mama	8.066	Características demográficas e clínicas	Idade, radioterapia, quimioterapia, história de câncer familiar, métodos de diagnóstico, classificação do câncer, receptores de progesterona	UMMC BCR	MLP, SVM, DT, RF	Variou entre 80,5% e 88,2%
15	(KO et al., 2017)	Pancreático	34	Exosomes	12 genes	BDPE	LDA	NA
16	(SUMAN e HOODA, 2019)	Cervical	858	Características demográficas, hábitos e histórico médico	Gravidez, tabagismo, alterações hormonais, uso de anticoncepcional e citologias	HCU	RF, NN, SVM, DT, BNet, AdaBoost	Todos > 94%
17	(LU et al., 2020a)	Cervical	472 858	Características demográficas, historial diagnóstico, história da vida sexual e gravidez, histológicas, genéticas	Idade, tabagismo, alcoolismo, número de gravidez, diagnóstico de HPV e citologias, sequenciamento genômico.	BDPE HCU	LR, DT, SVM, MLP, KNN.	Variou entre 77% e 82%
18	(LU et al., 2020b)	Ovário	472	Características demográficas, histológicas, sangue e marcadores de tumores.	Idade, menopausa, células germinativas de ovário, teratoma, hemoglobina, plaquetas, antígenos	HSU,	DT, ROMA, LogR	Variou entre 0,79 e 0,97
19	(LYNCH et al., 2017)	Pulmão	10.442	Características demográficas, histológicas	Idade, forma das células cancerígenas, números de tumores, tamanho do tumor	SEER	Ensemble LR, DT, GBM, SVM,	--
20	(MEHMOOD et al., 2021)	Cervical	737	Características demográficas e historial clínico e pessoal	Idade, fumante, gravidez, anticoncepcional, doenças de transmissão sexual,	UCI	RF, ANN, SNNA	ANN: 93,6%

21	(MOHAMMED et al., 2020)	Mama	286 699	Imagens PAAF**	digitalizadas da	Características celulares**	dos	núcleos	WDBC BCD	DT, NB, SMO	Variou entre 75% e 99%
22	(NEMATZADEH; IBRAHIM; SELAMAT, 2015)	Mama	699	Imagens PAAF**	digitalizadas da	Características celulares**	dos	núcleos	WDBC	DT, NB, NN, SVM	Variou entre 79% e 98%
23	(PODOLSKY et al., 2016)	Pulmão	220	Expressão genica de tumores		12,533 genes			DFIHM	KNN, NB, SVM, DT	Alta
24	(PRAISS et al., 2020)	Endometrial	46.773	Características demográficas, histológicas		Idade, forma das células cancerígenas, números de tumores, tamanho do tumor			SEER	EACCD	Alta
25	(MANDEEP, 2015)	Mama	699 NA	Imagens PAAF**	digitalizadas da	Características celulares**	dos	núcleos	WDBC	SVM, KNN, LR, NB	Variou entre 64% e 100%
26	(RAWAL, 2020)	Mama	699 NA	Imagens PAAF**	digitalizadas da	Características celulares**	dos	núcleos	WDBC	SVM, LR, RF, KNN	Variou entre 0,96 e 1,0
27	(SADHUKHAN; UPADHYAY; CHAKRABORTY, 2020)	Mama	NA	Imagens digitalizadas da punção aspirativa por agulha fina e citologia		Contornos das células			NA	KNN, SVM	Todos > 97%
28	(SHARMA; KULSHRESTHA; DANIEL, 2018)	Mama	699	Imagens PAAF**	digitalizadas da	Características celulares**	dos	núcleos	WDBC	LR, NNS, SVM	Variou entre 89% e 96%
29	(SIDEY-GIBBONS e SIDEY-GIBBONS, 2021)	Mama	683	Imagens PAAF**	digitalizadas da	Características celulares**	dos	núcleos	WDBC	GLMs, SVM	Todos > 0,94
30	(TABARES-SOTO et al., 2020)	Multi-câncer	174	Microarrays		12.533 genes			11-TD	KNN, SVM, LR, LDA, NB, MLP, RF, DT	Vario entre 10% e 90%
31	(TANINAGA et al., 2019)	Gástrico	1431	Características demográficas, testes de sangue, endoscopias e presença de <i>H. pilory</i>		Idade, sexo, gastrite crônica, úlceras			BDPE	LR, XGBoost	0,63 e 0,94
32	(TURGUT; DAGTEKIN; ENSARI, 2018)	Mama	133	Microarrays		1919 proteínas			BDPE	SVM, KNN, MLP, DT, RF, LR, Adaboost, GBM	NA

33	(TURKI, 2018)	Tiroides Colón Fígado	25 21 21	Microarrays	1146, 5607, 5607 genes	GEO	DeepBoost, xgboost, Boost I, SVM	Variou entre 0,50 e 0,89
34	(WANG et al., 2021)	Fígado	535	Características demográficas e parâmetros clínico- patológicos	Idade, sexo, tabagismo, alcoolismo, função hepática, escore de Child-Pugh e características do tumor	BS	DT, LR, RF, GBM, GBDT	Variou 0,59 e 0,73
35	(WU; ZHAO, 2018)	Pulmão	12	Imagens de tomografia computorizadas		NCI	NN-EDM	77.8%
36	(XIE et al., 2021)	Pulmão	153	Caraterísticas histopatológicas	Classificação, estagio do tumor	BDPE	KNN, NB, AdaBoost, SVM, RF, NN	Varia entre 0,63 e 1,00
37	(YUAN; LU; ZOU, 2020)	Pulmão	150	Perfis de expressão gênica	150 perfis de expressão gênica	GEO	SVM, RF	“Razoável”
38	(ZHENG; YOON; LAM, 2014)	Mama	569	Imagens digitalizadas da PAAF**	Características dos núcleos celulares**	WDBC	K-SVM	Todos > 95%

Fonte: (Autoria Pessoal)

** WDBC: Punção Aspirativa por Agulha Fina (PAAF). Características dos núcleos celulares como raio, textura, perímetro, área, suavidade, compacidade, concavidade, pontos côncavos, simetria.

Abreviações dos algoritmos: ANN - Artificial Neural Network; BN - Bayesian Network; BNet -Bayes Net; DT - Decision Tree; DTFMR - Decision Tree for Multioutput Regression; EACCD - Ensemble Algorithm for Clustering Cancer Data; GLMs - General Linear Model regression; GRU-SVM Neural Network Architecture Combine; GBM – LightGBM; GBDT - GradientBoosting Decision Tree Classification; GBM- Gradient Boosting Machines; KNN - k-Nearest Neighbors; K-SVM – Hybrid of K-means and support vector machine; LDA - Linear Discriminant Analysis; LR - Linear Regression; LogR - Logistic Regression; LRFMO - Linear Regression for Multioutput; MLP - Multilayer Perceptron; NBC - Naïve Bayesian; NNS - Nearest Neighbor Search; NN - Neural network; NNEDM - Neural network - Entropy Degradation Method; RBF - Radial Basis Function; RF - Random Forest; RFR - Random Forest Regression; ROMA - Ovarian Malignancy Algorithm; SMO - Sequential Minimal Optimization; SNNA - Shallow Neural Network; Architecture; SOM - Self Organisation Maps; SLP - Single Layer Perceptron; SR - Softmax Regression; SVM - Support Vector Machine; AdaBoost; XGBoost; DeepBoost.

Abreviações das Bases de Dados: BDPE - Base de Dados do Propria Estudo; WDBC - Wisconsin Diagnostic Breast Cancer; ICBC - The National Cancer Institute of Tehran; CCBP - Cervical Cancer Behavior Risk - UCI Machine Learning Repository; CGA - The Cancer Genome Atlas; BCD - The Breast Cancer Dataset - University of California, Irvine (UCI); ICGC - The International Cancer Genome Consortium; UMMCBCR - The University Malaya Medical Centre Breast Cancer Registry; HCU - Hospital of Caracas University; HSU - Hospital of Soochow University; SEER - Surveillance, Epidemiology, and End Results Program; UCI - The University of California’s database at Irvine; DFIHM -Dana-Farber Cancer Institute, Harvard Medical school, University of Michigan, University of Toronto, Brigham and Women’s Hospital, Harvard Medical school; 11-TD 11_Tumors database; GEO - Gene Expression Omnibus; BS – BioStudies; NCI - National Cancer Institute.

5. Referências

- ABDULKAREEM, N. M. AND ABDULAZEEZ, A. M. (2021). Machine learning classification based on Random Forest Algorithm: A review. *Journal of Science and Business*, v. 27, n. January, p. 128–142, Disponível em: <<https://ideas.repec.org/a/aif/journal/v5y2021i2p128-142.html>>.
- AGARAP, A. F. M. (2018). On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset. *ACM International Conference Proceeding Series*, n. 1, p. 5–9.
- AHMAD, L. G. et al. (2013). Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics*, v. 04, n. 02.
- AKTER, L. et al. (2021). Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques. *SN Computer Science*, v. 2, n. 3. Disponível em: <<https://doi.org/10.1007/s42979-021-00551-6>>.
- ALGHUNAIM, S.; AL-BAITY, H. H. (2019). On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context. *IEEE Access*, v. 7, p. 91535–91546.
- AMRANE, M. et al. (2021). Breast Cancer Classification Using Machine Learning Algorithms. *Lecture Notes in Networks and Systems*, v. 141, p. 571–578.
- ASRI, H. et al. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, v. 83, n. Fams, p. 1064–1069. Disponível em: <<http://dx.doi.org/10.1016/j.procs.2016.04.224>>.
- BARRIOS-LEAL, D. Y. D. Y.; NEVES-DA-ROCHA, J.; MANFRIN, M. H. (2019). Genetics and Distribution Modeling: The Demographic History of the Cactophilic *Drosophila buzzatii* Species Cluster in Open Areas of South America. *The Journal of heredity*, v. 110, n. 1, p. 22–33.
- BATTA, M. (2020). Machine Learning Algorithms - A Review . *International Journal of Science and Research*. v. 9, n. 1, p. 381.
- BI, W. L. et al. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*.
- CHEN, T. AND GUESTRIN, C. (2016). XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA. Anais. New York, NY, USA: ACM. Disponível em: <<http://doi.acm.org/10.1145/2939672.2939785>>.
- CHU, C. S. et al. (2020). Machine learning and treatment outcome prediction for oral cancer. *Journal of Oral Pathology and Medicine*, v. 49, n. 10, p. 977–985.
- CRUZ, J. A. AND WISHART, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, v. 2, p. 59–77.
- DEMŠAR, J. et al. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, v. 14, p. 2349–2353. Disponível em: <<http://jmlr.org/papers/v14/demsar13a.html>>.

- FERNANDES, L. et al. (2021). Diffuse reflectance and machine learning techniques to differentiate colorectal cancer ex vivo. *Chaos*, v. 31, n. 5. Disponível em: <<https://doi.org/10.1063/5.0052088>>.
- FRANK, E.; HALL, M. A.; WITTEN, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition.
- GLOBOCAN (2020). GLOBOCAN 2020: New Global Cancer Data. Disponível em: <<https://www.uicc.org/news/globocan-2020-new-global-cancer-data>>. Acesso em: 11 maio 2022.
- GÜNAKAN, E. et al. (2019). A novel prediction method for lymph node involvement in endometrial cancer: Machi learning. *International Journal of Gynecological Cancer*, v. 29, n. 2, p. 320–324.
- HE, B. et al. (2020). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, v. 1866, n. 11, p. 165916. Disponível em: <<https://doi.org/10.1016/j.bbadis.2020.165916>>.
- HOSNY, A. et al. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, v. 18, n. 8, p. 500–510.
- HYUN, S. H. et al. (2019). A Machine-Learning Approach Using PET-Based Radiomics to Predict the Histological Subtypes of Lung Cancer. *Clinical Nuclear Medicine*, v. 44, n. 12, p. 956–960.
- INCA, (2020). Estimativa 2020. Incidência de Câncer no Brasil. Disponível em: <https://www.inca.gov.br/>
- JABBAR, M. A. (2021). Breast cancer data classification using ensemble machine learning. *Engineering and Applied Science Research*, v. 48, n. 1, p. 65–72.
- JEREZ, J. M. et al. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, v. 50, n. 2, p. 105–115.
- KALAFI, E. Y. et al. (2019). Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. *Folia Biologica (Czech Republic)*, v. 65, n. 5–6, p. 212–220.
- KHARYA, S., DUBEY, D., SONI, S. (2013). Predictive Machine Learning Techniques for Breast Cancer Detection. *International Journal of Computer Science and Information Technologies*, v. 4, n. 6, p. 1023–1028. Disponível em: <www.ijcsit.com>.
- KLUYVER, T. et al. (2016). Jupyter Notebooks -- a publishing format for reproducible computational workflows. (F. Loizides, B. Schmidt, Eds.) In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, Anais.
- KO, J. et al. (2017). Combining Machine Learning and Nanofluidic Technology to Diagnose Pancreatic Cancer Using Exosomes. *ACS Nano*, v. 11, n. 11, p. 11182–11193.

- KOUROU, K. et al. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, v. 13, p. 8–17. Disponível em: <<http://dx.doi.org/10.1016/j.csbj.2014.11.005>>.
- LARSON, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, v. 22, n. 1, p. 45–55.
- LECUN, Y., BENGIO, Y., HINTON, G. (2015). Deep learning. *Nature*, v. 521, n. 7553, p. 436–444.
- LU, J. et al. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Generation Computer Systems*, v. 106, p. 199–205. Disponível em: <<https://doi.org/10.1016/j.future.2019.12.033>>.
- LU, M. et al. (2020). Using machine learning to predict ovarian cancer. *International Journal of Medical Informatics*, v. 141.
- LYNCH, C. M. et al. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Physiology & behavior*, v. 176, n. 5, p. 139–148.
- MANDEEP, R. (2015). Breast Cancer Diagnosis and Recurrence Prediction Using Machine Learning Techniques. *International Journal of Research in Engineering and Technology*, v. 04, n. 04, p. 372–376.
- MARTÍN, A. et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Disponível em: <<https://www.tensorflow.org/>>.
- MEHMOOD, M. et al. (2021). Machine Learning Assisted Cervical Cancer Detection. *Frontiers in Public Health*, v. 9, n. December, p. 1–14.
- MOHAMMED, S. A. et al. (2020). Analysis of breast cancer detection using different machine learning techniques. [s.l.] Springer Singapore. v. 1234 CCIS
- MOSTELLER, F. AND WALLACE, D. L. (1963). Inference in an authorship problem - a comparative-study of discrimination methods applied to authorship of disputed Federalist Papers. *Journal of the American Statistical Association*, v. 58, n. 302, p. 275.
- NEMATZADEH, Z., IBRAHIM, R., SELAMAT, A. (2015). Comparative Studies on Breast Cancer Machine Learning Techniques. 2015 10th Asian Control Conference (ASCC), p. 1–6.
- PISNER, D. A. AND SCHNYER, D. M. (2019). Support vector machine. Elsevier Inc.
- PODOLSKY, M. D. et al. (2016). Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific Journal of Cancer Prevention*, v. 17, n. 2, p. 835–838.
- PRAISS, A. M. et al. (2020). Using machine learning to create prognostic systems for endometrial cancer. *Gynecologic Oncology*, v. 159, n. 3, p. 744–750. Disponível em: <<https://doi.org/10.1016/j.ygyno.2020.09.047>>.

- PYTHON SOFTWARE FOUNDATION. Python, 2020. Disponível em: <<https://www.python.org/psf/>>.
- R CORE TEAM. (2020). R: A Language and Environment for Statistical Computing Vienna, Austria. R Foundation for Statistical. Disponível em: <<https://www.r-project.org/>>.
- RADAKOVICH, N., NAGY, M., NAZHA, A. (2020). Machine learning in haematological malignancies. *The Lancet Haematology*, v. 7, n. 7, p. e541–e550. Disponível em: <[http://dx.doi.org/10.1016/S2352-3026\(20\)30121-6](http://dx.doi.org/10.1016/S2352-3026(20)30121-6)>.
- RAWAL, R. (2020). Breast Cancer Prediction using Classification Techniques. *International Journal of Emerging Trends in Engineering Research*, v. 8, n. 9, p. 6074–6079.
- REFAELZADEH, P. et al. (2020). Encyclopedia of Database Systems. Encyclopedia of Database Systems.
- SADHUKHAN, S., UPADHYAY, N., CHAKRABORTY, P. (2020). Breast Cancer Diagnosis Using Image Processing and Machine Learning. Springer Singapore. v. 937
- SCHRIDER, D. R. AND KERN, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, v. 34, n. 4, p. 301–312. Disponível em: <<http://dx.doi.org/10.1016/j.tig.2017.12.005>>.
- SHARMA, A., KULSHRESTHA, S., DANIEL, S. (2018). Machine learning approaches for breast cancer diagnosis and prognosis. 2017 International Conference on Soft Computing and its Engineering Applications: Harnessing Soft Computing Techniques for Smart and Better World, *icSoftComp 2017*, v. 2018- Janua, p. 1–5.
- SIDEY-GIBBONS, J. A. AND SIDEY-GIBBONS, C. J. (2021). Machine learning in medicine: a practical introduction to natural language processing. *BMC Medical Research Methodology*, v. 21, n. 1, p. 1–18.
- SOMVANSHI, M. et al. (2017). A review of machine learning techniques using decision tree and support vector machine. *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*.
- SUMAN, S. K. AND HOODA, N. (2019). Predicting risk of Cervical Cancer : A case study of machine learning. *Journal of Statistics and Management Systems*, v. 22, n. 4, p. 689–696.
- SUNG, H. et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, v. 71, n. 3, p. 209–249.
- TABARES-SOTO, R. et al. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, v. 2020, n. 4, p. 1–22.
- TANINAGA, J. et al. (2019). Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. *Scientific Reports*, v. 9, n. 1, p. 1–9. Disponível em: <<http://dx.doi.org/10.1038/s41598-019-48769-y>>.

THE MATHWORKS INC. (2010). MATLABNatick, Massachusetts.

TURGUT, S., DAGTEKIN, M., ENSARI, T. (2018). Microarray breast cancer data classification using machine learning methods. 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018, p. 1–3.

TURKI, T. (2018). An empirical study of machine learning algorithms for cancer identification. ICNSC 2018 - 15th IEEE International Conference on Networking, Sensing and Control, p. 1–5.

VIEIRA, S., LOPEZ PINAYA, W. H., MECHELLI, A. (2019). Introduction to machine learning. [s.l.] Elsevier Inc.

VILORIA, A. et al. (2020). Diabetes Diagnostic Prediction Using Vector Support Machines. *Procedia Computer Science*, v. 170, p. 376–381. Disponível em: <<https://doi.org/10.1016/j.procs.2020.03.065>>.

WANG, Y. et al. (2021). Predicting postoperative liver cancer death outcomes with machine learning. *Current Medical Research and Opinion*, v. 37, n. 4, p. 629–634. Disponível em: <<https://doi.org/10.1080/03007995.2021.1885361>>.

WU, Q. AND ZHAO, W. (2018). Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm. *Proceedings - 2017 International Symposium on Computer Science and Intelligent Controls, ISCSIC 2017*, v. 2018- Febru, p. 88–91.

XIE, Y. et al. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational Oncology*, v. 14, n. 1.

YUAN, F., LU, L., ZOU, Q. (2020). Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, v. 1866, n. 8, p. 165822. Disponível em: <<https://doi.org/10.1016/j.bbadis.2020.165822>>.

ZHENG, B., YOON, S. W., LAM, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, v. 41, n. 4 PART 1, p. 1476–1482. Disponível em: <<http://dx.doi.org/10.1016/j.eswa.2013.08.044>>.