UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto n°17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXV SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2021

USO DE MACHINE LEARNING NA IDENTIFICAÇÃO E CLASSIFICAÇÃO DE SALMONELLA SPP.

Vinicius Pereira de Santana[1]; Raquel Guimarães Benevides[2]; Danitza Romero-Calle[3] e Glen Jasper Yupanqui-García[4]

1. Bolsista PIBIC/FAPESB, Graduando em Bacharelado em Ciências Biológicas, Universidade Estadual de Feira de Santana, e-mail: vpdesantana@gmail.com
2. Orientador, Departamento de Ciências Biológicas, Universidade Estadual de Feira de Santana, e-mail: raquelgb@gmail.com

PALAVRAS-CHAVE: tecnologias de diagnóstico, inteligência artificial, bioinformática

INTRODUÇÃO

Segundo um informativo da OMS – Organização Mundial da Saúde – (2021), *Salmonella* é um gênero de bactérias bacilares gram-negativas pertencente à família Enterobacteriaceae. Tal gênero possui apenas duas espécies, *Salmonella bongori* e *Salmonella enterica*, porém mais de 2500 sorotipos diferentes identificados até o momento (OMS, 2021). Avanços no campo da biologia molecular têm gerado novas técnicas para o diagnóstico de infecções e contaminações, como por exemplo o PCR (Polymerase Chain Reaction). Embora estes métodos forneçam uma análise precisa e confiável de amostras bacterianas, eles normalmente não são usados em ambientes clínicos devido a seus altos custos e baixa velocidade, além de desvantagens em relação a sensibilidade e robustez (MANZOOR et al., 2014; WATTIAU et al., 2011). O uso de *machine learning* pode se apresentar como um mecanismo auxiliar para métodos de identificação e classificação de espécies e sorotipagem de bactérias.

De acordo com a International Business Machines Corporation - IBM (2020), *machine learning* é um ramo da inteligência artificial (IA) e da ciência da computação que foca no uso de dados e algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão (IBM, 2020). Métodos de *machine learning* têm se tornado mais comuns em aplicações biológicas à medida que a disponibilidade de vários tipos de dados ômicos tem aumentado (WEBB, 2018). Tais métodos podem ser úteis na identificação e classificação de bactérias patogênicas, através de características particularmente difíceis de discernir utilizando os métodos convencionais disponíveis atualmente (XU; JACKSON, 2019). Esse estudo apresenta uma revisão sistemática com a finalidade de descrever e analisar os documentos científicos que empregam abordagens de *machine learning* na caracterização de *Salmonella* spp.

MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)

A seleção dos arquivos foi conduzida nos bancos de dados Scopus, WoS e PubMed, com critérios de inclusão e exclusão aplicados por processos automatizados (scripts) e manuais, onde 2294 registros foram identificados. Foi realizada uma revisão manual dos arquivos, a partir dos critérios de inclusão e de exclusão, onde arquivos que não possuíam as palavras-chave no título e no resumo foram descartados. Após a revisão, restaram 164 arquivos elegíveis. Posteriormente, foi realizada uma revisão manual conduzida por três revisores independentes, examinando os textos por completo. Após a exclusão dos arquivos considerados inelegíveis pelos três revisores, restaram 109 arquivos. Dentre os 109 arquivos, 35 artigos foram considerados elegíveis por unanimidade dos três revisores. O presente trabalho foi conduzido apenas com os dados dos artigos considerados elegíveis por unanimidade (n=35).

RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

Dentre os arquivos estudados, foram identificados o uso de 16 algoritmos de *machine learning* diferentes. Dentre os 16 algoritmos, 12 são abordagens de aprendizado supervisionado, enquanto apenas quatro são abordagens de aprendizado não-supervisionado. Além disso, as abordagens de aprendizado supervisionado foram mais estudadas, contabilizando 55 usos distintos entre todos os arquivos, enquanto foram contabilizados apenas quatro usos de abordagens não-supervisionadas entre os arquivos. Na Figura 1 é possível visualizar a proporção do uso de cada algoritmo.

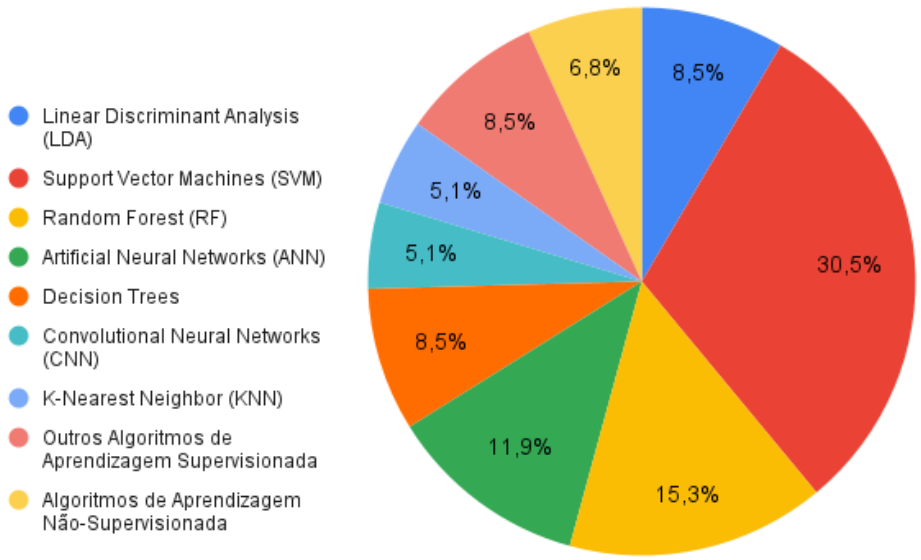

Figura 1 - Proporção do uso de cada algoritmo de *machine learning* dentre os arquivos

Foi realizada uma análise temporal dos estudos que utilizam *machine learning* na identificação e classificação *Salmonella* spp. Os arquivos achados estão distribuídos ao longo de 15 anos, sendo o mais antigo publicado em 2006 e o mais recente em Maio de 2021. Também foi realizada uma análise minuciosa em cada texto com o propósito de entender quais sorotipos de *Salmonella* spp. foram os mais estudados, representada visualmente na Figura 2.

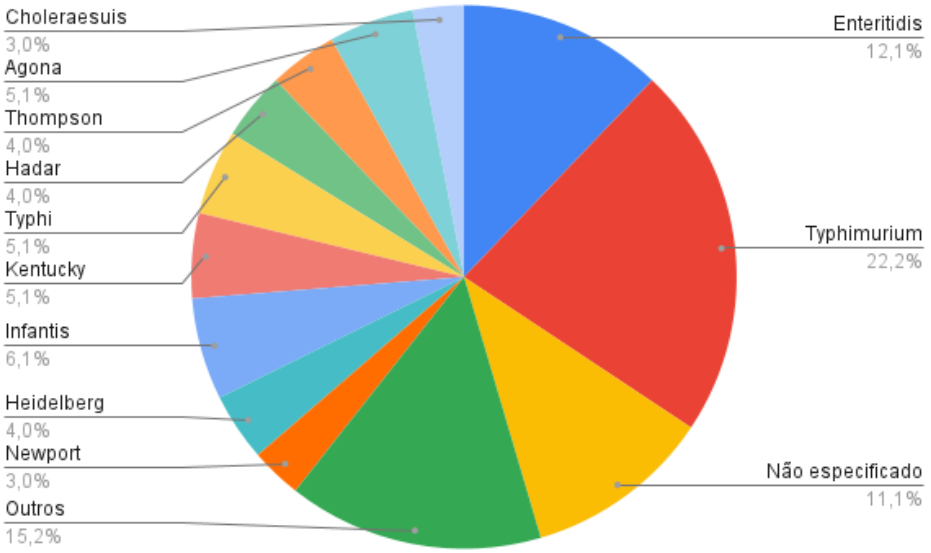

Figura 2 - Sorotipos de *Salmonella* spp. mais estudados dentre os arquivos

Ao levar em conta que em problemas de identificação e classificação, onde os dados são relativamente estruturados e bem categorizados, como é o caso dos estudos analisados, os algoritmos supervisionados tendem a superar as abordagens não supervisionadas, além do fato das abordagens supervisionadas requererem menos dados (INGEDATA, 2018). O uso de algoritmos supervisionados na caracterização de *Salmonella* spp. foi bastante variado e numeroso, mas se destacam algumas aplicações como a detecção dos hospedeiros zoonóticos através das sequências genômicas (LUPOLOVA; LYCETT; GALLY, 2019b; FORTINO et al., 2014; WANG et al., 2011) e também a caracterização através de padrões presentes em imagens e odores (MBELWA; MBELWA; MACHUVE, 2021; RAHMAYUNA et al., 2018b; SEO et al., 2018; BONAH et al., 2019). Como citado anteriormente, os métodos de aprendizado não-supervisionado tiveram pouca representatividade entre os arquivos analisados. As aplicações desse tipo de abordagem se restringiram apenas a atribuição dos hospedeiros através de sequências genômicas (LUPOLOVA; LYCETT; GALLY, 2019b) e classificação de sorotipos (CHEN et al., 2014).

CONSIDERAÇÕES FINAIS (ou Conclusão)

O presente estudo sinaliza a possibilidade de aplicação de algoritmos de *machine learning* na caracterização de organismos do gênero *Salmonella* spp. Embora a aplicação de algoritmos supervisionados já esteja bem estabelecida nesse contexto, fica evidente que restam muitos desafios e problemas pendentes para otimizar o uso desses mecanismos, principalmente no que diz respeito aos algoritmos de natureza não-supervisionada. Entretanto, há o potencial de que a utilização de algoritmos de *machine learning* futuramente se tornem os componentes essenciais para a detecção de *Salmonella* spp., devido a capacidade variada de aplicações em uma pluralidade de problemas, além da possibilidade de serem empregados em bases de dados mais complexas.

REFERÊNCIAS

BONAH, E. et al. Electronic nose classification and differentiation of bacterial foodborne pathogens based on support vector machine optimized with particle swarm optimization algorithm. Journal of Food Process Engineering, v. 42, n. 6, 1 out. 2019.

CHEN, H. C. et al. A composite model for subgroup identification and prediction via bicluster analysis. PLoS ONE, v. 9, n. 10, 27 out. 2014.

FORTINO, V. et al. Transcriptome dynamics-based operon prediction in prokaryotes. BMC Bioinformatics, v. 15, n. 1, 16 maio 2014.

INGEDATA. Why Supervised Learning still often beats Unsupervised Learning? Disponível em: <https://www.ingedata.net/blog/supervised-learning-vs-unsupervised-learning>. Acesso em: 1 out. 2021.

INTERNATIONAL BUSINESS MACHINES CORPORATION. What is Machine Learning? Disponível em: <https://www.ibm.com/cloud/learn/machine-learning>. Acesso em: 26 set. 2021.

LUPOLOVA, N.; LYCETT, S. J.; GALLY, D. L. A guide to machine learning for bacterial host attribution using genome sequence data. Microbial Genomics, v. 5, n. 12, 1 dez. 2019b.

MANZOOR, S. et al. Rapid identification and discrimination of bacterial strains by laser induced breakdown spectroscopy and neural networks. Talanta, v. 121, p. 65–70, 2014.

MBELWA, H.; MBELWA, J.; MACHUVE, D. Deep Convolutional Neural Network for Chicken Diseases Detection. International Journal of Advanced Computer Science and Applications, v. 12, n. 2, 2021.

ORGANIZAÇÃO MUNDIAL DE SAÚDE. Foodborne diseases. Disponível em: <https://www.who.int/health-topics/foodborne-diseases#tab=tab_1>. Acesso em: 27 set. 2021.

RAHMAYUNA, N. et al. Pathogenic Bacteria Genus Classification using Support Vector Machine. 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). Anais...IEEE, nov. 2018b.

SEO, Y. et al. Morphological image analysis for foodborne bacteria classification. Transactions of the ASABE, v. 61, n. 1, p. 5–13, 2018.

WANG, H. et al. An Integrative Approach for Genomic Island Prediction in Prokaryotic Genomes. 2011.

WATTIAU, P.; BOLAND, C.; BERTRAND, S. Methodologies for Salmonella enterica subsp. enterica Subtyping: Gold Standards and Alternatives. Applied and Environmental Microbiology, v. 77, n. 22, p. 7877, nov. 2011.

WEBB, S. Deep learning for biology. Nature, v. 554, n. 7693, p. 555–557, 22 fev. 2018.

XU, C.; JACKSON, S. A. Machine learning and complex biological data. Genome Biology 2019 20:1, v. 20, n. 1, p. 1–4, 16 abr. 2019.