Preliminary Results of an Artificial Intelligence model for Gastric Cancer Classification using Helicobacter pylori virulence markers and IL-17 SNPs

BRUNO MARI FREDI[1], MIKAELA NAGAHARA[2], VINÍCIUS CASADO MORAES[1], ELIZABETH SUCHI CHEN[3], MÔNICA SANTIAGO BARBOSA[4], SPENCER LUIZ MARQUES PAYÃO[2], LUCAS TREVIZANI RASMUSSEN[2,5], JOSÉ CELSO ROCHA[1]

[1.] SÃO PAULO STATE UNIVERSITY , SCHOOL OF SCIENCES, HUMANITIES AND LANGUAGES, ASSIS/SÃO PAULO, BRAZIL
[2.] MARÍLIA MEDICAL SCHOOL, MARÍLIA/SÃO PAULO, BRAZIL
[3.] FEDERAL UNIVERSITY OF SÃO PAULO, PAULISTA SCHOOL OF MEDICINE, SÃO PAULO/SÃO PAULO, BRAZIL
[4.] FEDERAL UNIVERSITY OF GOIAS, INSTITUTE OF TROPICAL PATHOLOGY AND PUBLIC HEALTH, GOIÂNIA/GOIAS, BRAZIL
[5.] UNIVERSITY CENTER OF THE INTEGRATED COLLEGES OF OURINHOS, OURINHOS/SÃO PAULO, BRAZIL

Introduction: Gastric cancer is a multifactorial disease influenced by genetic factors and Helicobacter pylori (H. pylori) infection. Disease severity is associated with H. pylori virulence markers that promote chronic inflammation and tissue damage. IL-17 plays a key role in this inflammatory response, and several Single Nucleotide Polymorphisms (SNPs) are associated with gastric cancer. Artificial Neural Networks are promising tools for predicting and molecularly characterizing diseases, including gastric cancer risk, based on molecular variables. Objective: This study aimed to develop an Artificial Neural Networks Multilayer Perceptron (ANN-MLP) algorithm for classifying and predicting the risk of gastric cancer based on the presence of nine IL-17 SNPs; eight H. pylori virulence markers and demographic factor (age and sex). Materials and Methods: A total of 305 gastric biopsy samples (95 control, 115 gastritis, 95 gastric cancer) were analyzed. Genotyping of three IL17A, four IL17F, and two IL17RA SNPs, H. pylori detection, and assessment of eight virulence markers (cagA, cagT, cagE, cagG, cagM, virB11, dupA, oipA) were performed by qPCR. The dataset was split into 80% for model development and 20% for external validation. The development set was further divided into training, validation, and internal testing (70%, 15%, 15%) using stratified sampling. Categorical variables were one-hot encoded. An ANN-MLP with feedforward architecture was implemented for multiclass classification. Training was performed using learning algorithm backpropagation with categorical cross-entropy loss. Model architecture and hyperparameters were empirically optimized. Performance was evaluated using Accuracy, Precision, Recall, F1-score, and AUC. Results: The proposed ANN-MLP achieved an accuracy of 66.13% on the external validation set, with Precision, Recall, and F1-score of 65.83%, 65.72%, and 65.75%, respectively. Class-wise analysis showed higher performance for gastritis (F1 = 70.83%), followed by gastric cancer (F1 = 64.86%) and control (F1 = 61.54%). The model achieved an AUC of 0.72, with class-wise AUC values of 0.684 (control), 0.703 (gastritis), and 0.771 (gastric cancer). Discussion: The model showed moderate performance in external validation, with a higher F1 score for gastritis, possibly reflecting the inflammatory signature associated with IL-17 and H. pylori markers. On the other hand, gastric cancer showed the highest

AUC, despite a lower F1-score, indicating better discrimination. These findings support the potential of integrating host genetic and bacterial factors for gastric cancer classification, although further model optimization and exploration of alternative machine learning approaches may improve predictive performance. Conclusion: The proposed approach demonstrates the feasibility of integrating host genetic and bacterial features for multiclass classification of gastric disease, with potential to identify key molecular factors associated with gastric cancer development.

Agradecimentos: São Paulo Research Foundation (FAPESP). Process Number: 2024-21796-2 and 2023/16191-1