Nº 179615

Enchancing proactive patient outreach with medical message classification using synthetic data from ChatGPT

Adriana Camargo Brito
Adriano Galindo Leal
Gustavo Torres Custódio
Edilsin José Rodrigues
Michele Marcia Viana Martins
Vinicius Monteiro de Paula Guirado

Pôster apresentado no IEE SYMPOSIUM SERIES ON COMPUTATINAL INTELLIGENCE, 2025, Trondheim, Noruega. 1 slide.

2025 IEEE SYMPOSIUM ON CI IN HEALTH AND MEDICINE (IEEE CIHM)

PAPER 34

ENHANCING PROACTIVE PATIENT OUTREACH WITH MEDICAL MESSAGE CLASSIFICATION USING SYNTHETIC DATA FROM CHATGPT

AUTHORS

Adriana Camargo de Brito (1)
Adriano Galindo Leal (1)
Gustavo Torres Custodio (1)
Edilson José Rodrigues (1)
Michele Marcia Viana Martins (2)
Vinícius Monteiro de Paulo Guirado (3)

AFFILIATIONS

(1) Institute for Technological Research, São Paulo, Brazil
(2) Federal University of Viçosa, Minas Gerais, Brazil
(3) University of São Paulo Medical School, São Paulo, Brazil

INTRODUCTION

The growing need for proactive patient outreach has driven the use of automated message classification. While advanced tools exist, their high costs limit small companies. Traditional, cost-effective methods remain a viable alternative.

OBJECTIVE

This article aims to identify the effect of data augmentation with synthetic data generated by ChatGPT 3.5 on the performance of traditional and Transformer models for the binary classification of health insurance user messages regarding the presence or absence of medical needs.

AI-generated image

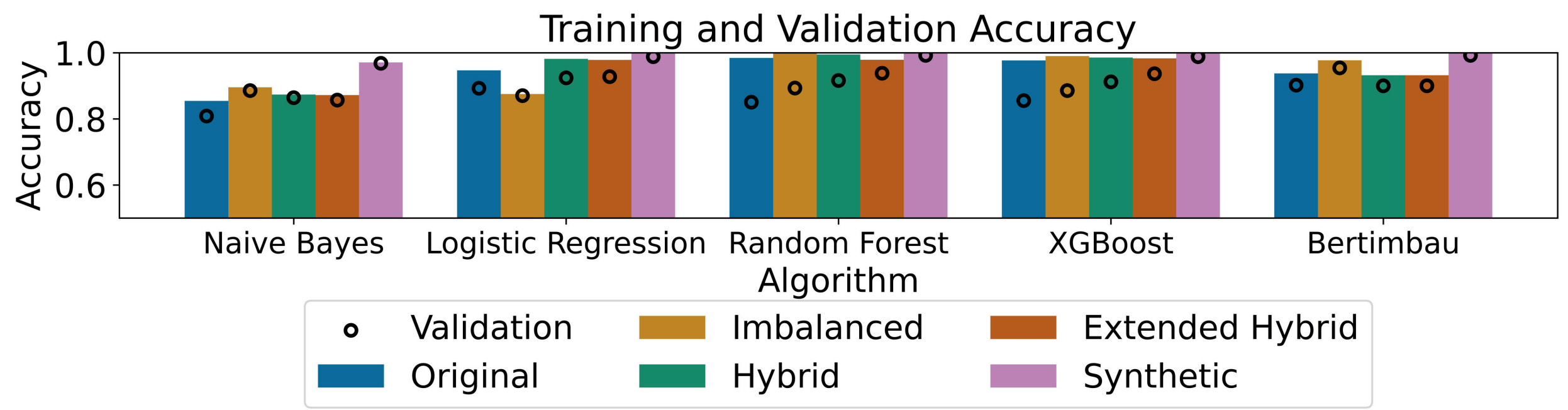
METHODOLOGY

The primary aim is to classify messages that indicate medical needs, aiding in patient outreach and efficient resource allocation. We applied Natural Language Processing (NLP) techniques, including Logistic Regression, Naive Bayes, Random Forest, XGBoost, and BERTimbau, a Portuguese transformer-based model. Models were trained on both original and augmented datasets.

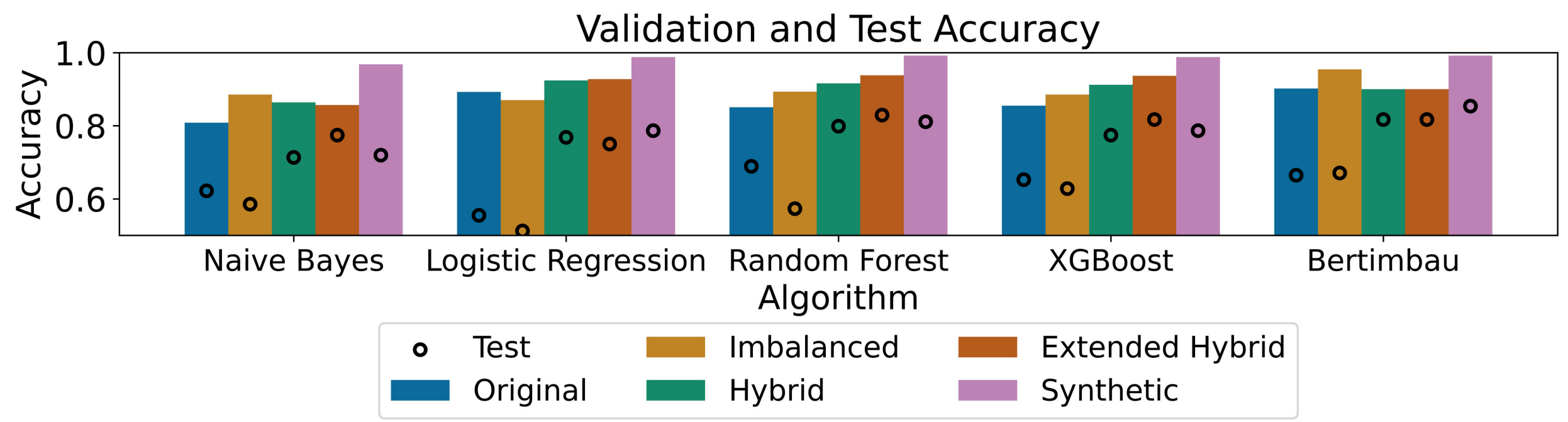
DATASET

A. Camargo de Brito, A. Galindo Leal, G. Torres Custódio, E. José Rodrigues, M. Marcia Viana Martins, and V. Monteiro de Paula Guirado, "Dataset for the IEEECIHM2025 paper "Enhancing Proactive Patient Outreach with Medical Message Classification Using Synthetic Data from ChatGPT"," Dec. 2024. [Online]. Available: **https://github.com/IPT-TD-SIAA/IEEECIHM2025.git**

01 Training and Validation Accuracy

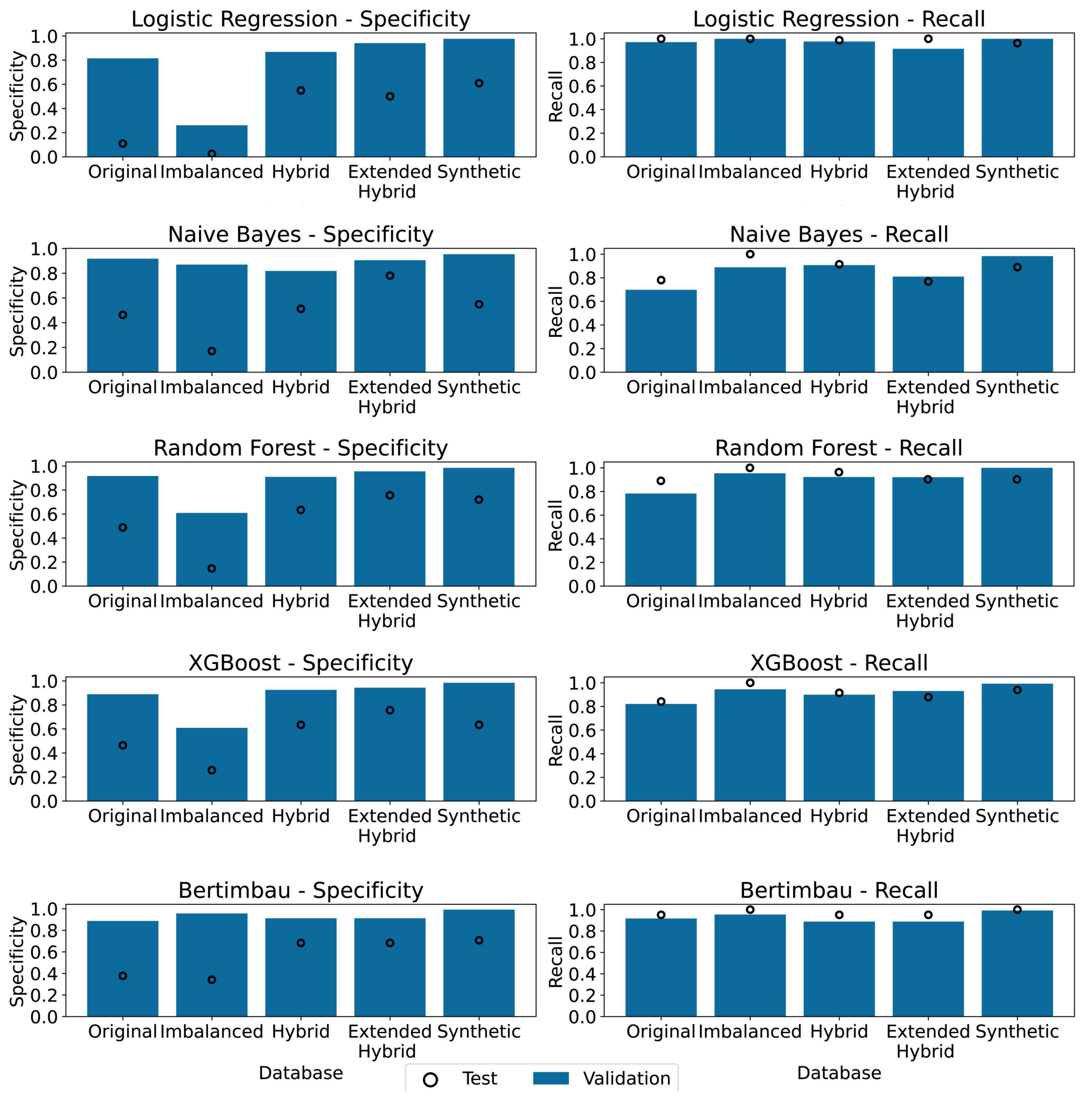


02 Validation and Test Accuracy



03 Specificity and Recall

In binary classification, messages with medical needs are labeled as "one" and without as "zero."
Models trained on original or imbalanced datasets struggled to correctly identify messages without medical needs, especially with simpler algorithms. Hybrid and synthetic datasets improved adaptability, though capturing subtle nuances in patient communication remained challenging.



ANALYSIS

Models trained on all datasets exhibited similar training and validation accuracies, with some overfitting observed in traditional approaches. The use of synthetic data helped mitigate accuracy gaps, enhancing model generalization. Hybrid and synthetic datasets resulted in the highest validation accuracy, particularly in models addressing imbalanced data.
Models trained on hybrid and synthetic datasets demonstrated superior generalization, maintaining higher accuracy on unseen data. Specificity and recall varied across models, with synthetic data contributing to improved classification performance. Overall, hybrid and synthetic datasets enhanced model robustness.

CONCLUSION

This study demonstrated that GPT-generated synthetic data improves medical message classification by enhancing model accuracy, reducing overfitting, and improving generalization. The BERTimbau model showed significant performance gains with synthetic data, making it a valuable strategy for addressing class imbalance. These findings highlight the potential of synthetic data to optimize automated medical message processing and enhance operational efficiency. Future research should explore its application across different languages and medical subfields to further validate its effectiveness.