Measles Outbreak Prediction Using Machine Learning Models: A Regional Surveillance Study from South-West Romania

Ana-Maria BOLDEA[a], Alexandra-Daniela ROTARU-ZAVALEANU[b], Andrei-Florentin BĂIAŞU[a,*], and Mircea-Sebastian ŞERBĂNESCU[c]

[a] Doctoral School, University of Medicine and Pharmacy of Craiova, Romania.
[b] Department of Epidemiology, University of Medicine and Pharmacy of Craiova, Romania.
[c] Department of Medical Informatics and Biostatistics, University of Medicine and Pharmacy of Craiova, Romania.
E-mails: calugaruanamaria11@yahoo.ro; alexandra.zavaleanu@gmail.com; (*) andrei.baiasu@umfcv.ro; mircea_serbanescu@yahoo.com

* Author to whom correspondence should be addressed;

Abstract

Measles remains a major public health concern in Eastern Europe, despite the availability of effective vaccination programs. This study aimed to evaluate the applicability of machine learning models for case-level classification and prediction of measles outbreak dynamics using regional surveillance data from five adjacent counties in South-West Romania. A retrospective dataset provided by the Regional Center for Public Health Craiova was analyzed, comprising 625 confirmed measles cases reported in 2023 from Dolj, Vâlcea, Gorj, Mehedinţi, and Olt counties. The dataset incorporated demographic variables (age, sex, residence type), temporal features (month, epidemic wave), and vaccination-related information (vaccination status, number of doses). Two supervised machine learning algorithms—Random Forest and Logistic Regression—were developed to classify cases as outbreak-associated versus sporadic, and 5-fold cross-validation was applied to assess model robustness and generalizability. Random Forest demonstrated superior predictive performance compared to Logistic Regression across all evaluation metrics. Random Forest achieved a mean accuracy of 84.4% (95% CI: 81.9–86.9%) and a ROC–AUC of 87.6% (95% CI: 85.1–90.1%), with sensitivity of 82.2% (95% CI: 78.0–86.4%) and specificity of 86.9% (95% CI: 83.2–90.6%). Logistic Regression showed moderate performance with accuracy of 79.0% (95% CI: 76.0–82.0%) and ROC–AUC of 79.0% (95% CI: 75.1–82.9%). The diagnostic odds ratio was substantially higher for Random Forest (36.0; 95% CI: 18.4–53.6) compared to Logistic Regression (16.2; 95% CI: 8.6–23.8), indicating stronger discriminative capacity. These findings suggest that machine learning-based approaches can enhance epidemiological surveillance by providing reproducible, data-driven insights into measles transmission patterns, particularly in regions lacking robust real-time outbreak monitoring systems. Integrating such models into public health surveillance frameworks may improve early detection capabilities and support evidence-based decision-making in outbreak prevention and control.

Keywords: Measles; Machine Learning (ML); Random forest.