Unraveling the drivers of leptospirosis risk in Thailand using machine learning

*Pikkanet Suttirat, Sudarat Chadsuthi, **Charin Modchang**, and Joacim Rocklöv*

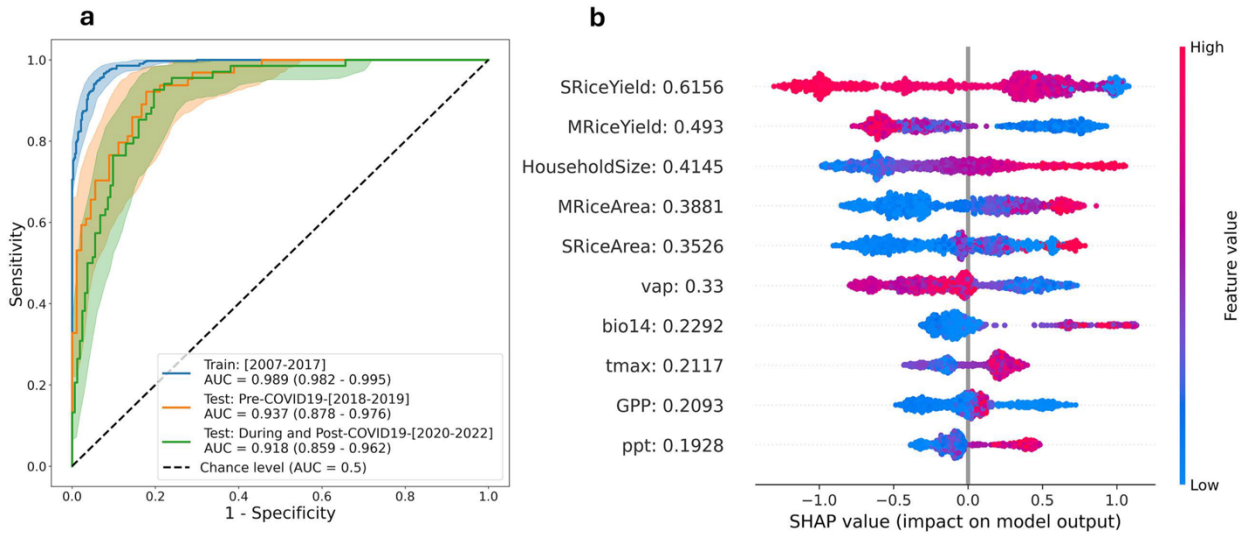
Rationale and objectives: Leptospirosis is a serious bacterial disease that poses a major public health threat in Thailand, with over 51,000 cases reported between 2007 and 2022. The bacteria thrive in contaminated water and soil, particularly in rice farming areas, and human infection occurs through contact with these contaminated environments. Since the disease results from complex interactions between weather patterns, farming practices, and socioeconomic conditions, researchers aimed to develop a machine learning model that could predict which provinces in Thailand are at high risk for leptospirosis outbreaks by analyzing these multiple factors together.

Summary: The research team built an artificial intelligence model using a technique called XGBoost to analyze 16 years of disease surveillance data alongside information about rainfall, temperature, rice farming activities, household size, and economic indicators across Thailand's provinces. They trained the model on data from 2007-2017 to distinguish between provinces with high versus low disease risk, then tested its accuracy on data from 2018-2022. To understand which factors were most important for predictions, they used a method called SHAP analysis that reveals how each factor influences the model's decisions.

Outcome: The model successfully predicted leptospirosis risk with high accuracy (93.7% before the COVID-19 pandemic and 95.1% after). Rice production patterns emerged as the strongest predictor of disease risk - areas with traditional, less mechanized rice farming practices showed higher risk. Household size and certain climate factors like vapor pressure also played important roles. Interestingly, the model's performance dropped during the COVID-19 pandemic years (2020-2021), likely due to disrupted disease surveillance and underreporting when healthcare systems were overwhelmed. This research provides a valuable framework for developing early warning systems that could help Thai health authorities target prevention efforts to the most vulnerable communities.

Research grant and acknowledgement: The Rockefeller Foundation under the Strengthening the Early Warning and Outbreak Detection Systems through Nation-wide Event-based and Syndromic Surveillance (STONES) project (Grant No. 2021PPI005).

Related SDGs goal: 3. Good health and well-being.



Graphical summary: XGBoost model performance metrics and SHAP-based feature importance analysis. **(a)** Receiver operating characteristic (ROC) curves showing model discrimination ability across different datasets, with corresponding area under the curve (AUC) values. Results are derived from the optimal hyperparameters identified through 5-fold cross-validation. The diagonal line represents random classification (AUC = 0.5), while AUC = 1.0 indicates perfect prediction. The shaded area represents 95% confidence intervals (95% CIs) based on 1,000 bootstraps. 95% CIs of AUC were given in parentheses. **(b)** SHAP (SHapley Additive exPlanations) analysis of the top 10 predictive features, ranked by mean absolute SHAP values (2007–2022). Individual observations are represented as dots, with color intensity indicating feature magnitude (red = high, blue = low). SHAP values on the x-axis show each feature's contribution to the predicted probability, centered at log odds = 0 (corresponding to a predicted probability of 0.5).

Related publication:

Suttirat P, Chadsuthi S, **Modchang C**, and Rocklöv J. Unraveling the drivers of leptospirosis risk in Thailand using machine learning. *PLOS Neglected Tropical Diseases*.
https://doi.org/10.1371/journal.pntd.0013618