Abstract

Foot and Mouth Disease (FMD) has affected Uganda for over 60 years, causing significant economic losses in the livestock industry, with a 23% decrease in income for stakeholders and market values for bulls and cows reducing by 83% and 88%, respectively. Previous research utilized machine learning (ML) models to predict FMD occurrences for enhanced preparedness in resource-constrained endemic countries. However, these models were validated in stationary environments, which do not account for real-world distribution shifts where the underlying statistical properties of the dataset change over time, leading to degradation in predictive performance. This study aimed to improve ML-based predictive performance for FMD outbreaks under distribution shifts to enhance preparedness in Uganda.

The research employed an experimental design across six phases. First, a literature review identified key risk factors for FMD outbreaks, guiding data source selection. Second, historical data was collected retrospectively from various sources. The third phase focused on data preprocessing, using techniques including mean imputation, duplicate removal, and visualization, along with a two-way statistical approach to detect distribution shifts. In the fourth phase, seven ML algorithms including Random Forest, Support Vector Machine, Classification and Regression Tree, Gradient Boosting Machine, Logistic Regression, k-Nearest Neighbor, and AdaBoost were trained while class imbalance was concurrently addressed using data augmentation techniques. The fifth and sixth phases involved testing and validating these models to evaluate their performance under distribution shifts. Two approaches were explored to enhance model performance: a data-centric approach that integrated techniques including borderline-SMOTE, active learning, probabilistic calibration and pseudo-labeling, and a model-centric approach that involved tuning and stacking Random Forest, Gradient Boosting Machine, and AdaBoost.

The findings indicated significant distribution shifts in rainfall and maximum temperature. Rainfall yielded a test value of 0.1215 with a p-value of 0.0000, while maximum temperature yielded a test value of 0.0833 with a p-value of 0.0024. In comparative performance analysis, Random Forest with borderline-SMOTE was the top model under stationary conditions, achieving 92% accuracy, an Area Under the Curve (AUC) of 0.97, recall of 0.94, precision of 0.90, and an F1-score of 0.92. However, under varying distributions, its accuracy dropped to 46%, with declines in other performance metrics. The data-centric approach yielded a weighted average performance score of 84.3%, while the model-centric approach achieved a score of 87.4%.

In this study, the model-centric approach, Random Forest Boosting (RFB), was identified as the top performer, outperforming the data-centric approach, Calibrated Uncertainty Prediction (CUP), in predicting Foot and Mouth Disease outbreaks in Uganda. Both approaches present promising solutions for tackling distribution shifts in outbreak prediction under varying conditions, surpassing existing methods. The findings provide valuable insights for proactive measures to mitigate the impact of outbreaks, enhancing preparedness through early detection and targeted resource allocation.

Keywords: Foot and Mouth Disease, machine learning, data augmentation, distribution shifts, active learning, probabilistic calibration, pseudo annotation, predictive performance