Integrating GeoVisualization, Machine Learning, and Deep Learning for Classification and Prediction of Rocky Mountain Spotted Fever in Arizona

THE UNIVERSITY OF NEW MEXICO

Al Ekram Elahee Hridoy[1,2], BS; Yan Lin[1,2], PhD; Zhuoming Liu[2,3], MS

[1]Department of Geography and Environmental Studies, University of New Mexico

[2]UNM Advanced Spatial Informatics Research and Education(ASPIRE)

[3]Department of Computer Science, University of New Mexico

PROBLEM STATEMENT

- Rocky Mountain Spotted Fever (RMSF) is a potentially fatal Tick Disease (TD) caused by the bacterium Rickettsia and has disproportionally affected communities in southwest US (e.g., Arizona, and Navajo Nation). To address this problem, we have developed a machine and deep learning classification and prediction system that uses geovisualization techniques to identify high-risk areas for RMSF in Arizona.
- Our system includes a first-of-its-kind climate-based classification model with a Receiver Operating Characteristic Area Under the Curve (ROC AUC) of 0.97, which can accurately predict which counties in Arizona are most likely to have a high incidence of RMSF. Additionally, we have implemented a deep learning method known as Long-Short Term Neural Networks(LSTM) to predict the possible occurrence of RMSF in each county.
- Our approach to predicting the spread of RMSF will enable public health officials to take timely preventive measures and allocate resources to the areas at greatest risk. Our machine learning system has the potential to significantly reduce the incidence of this potentially fatal disease in the southwestern United States.

OBJECTIVES

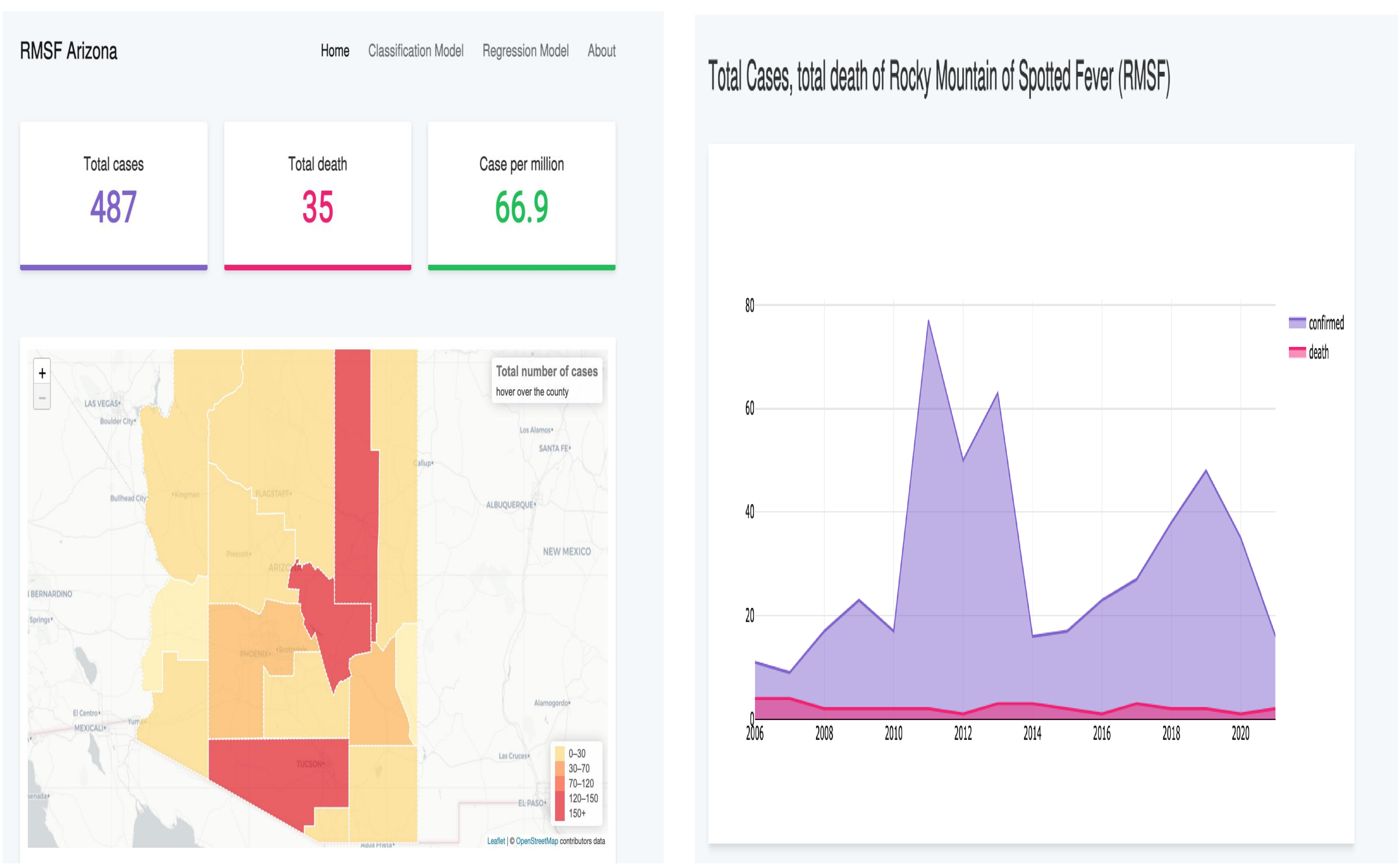
- To visualize RMSF disease statistics in a web application.
- To implement climate-based classification system to map most likely to have a high incidence of RMSF.
- To implement deep learning based prediction of RMSF incidence system for each county.

DATA AND DATA SOURCES

We used yearly time-series data from 2006 to 2021.

1. RMSF cases. Source: ARIZONA DEPARTMENT OF HEALTH SERVICES
2. Land Surface and Temperature (Degree Celsius). Source: MODIS Land Surface Temperature and Emissivity (MOD11
3. Relative Humidity (In Percentage) and Precipitation (mm, daily total). Source: GRIDMET: University of Idaho Gridded Surface Meteorological Dataset
4. NDVI (Unit Less). Source: MODIS Terra Daily NDVI
5. CDC Social Vulnerability index (SVC). Source: CDC/ ATSDR Social Vulnerability Index
6. Total population. Source: CDC/ ATSDR Social Vulnerability Index

GEOVISUALIZATION WEB INTERFACE



WEBSITE

- https://alhridoy.github.io/RMSF-dashboard

METHODS

- We used several machine learning methods for classification problems and built a Random Forest algorithm pipeline which gave 0.95 accuracy with ROC AUC 0.82 before tuning and after hyper parameter tuning we achieved ROC AUC of 0.974 with 5 k-fold cross validation. Class imbalance of high. Incidence and low incidence was addressed by RandomOverSampler technique.
- For prediction of RMSF, we used a deep learning technique, Long Short Term Neural Networks(LSTMs). The architecture of the model consists of two LSTM layers, each with 64 units and a ReLU activation function, followed by a single Dense layer with one unit, and the model is compiled with the mean absolute error (MAE) loss function and the Adam optimizer .
- For classification we followed the CDC's definition of "high incidence" as a county with more than 10 cases per 100,000 population, the data was binned for classification models by assigning a value of 1 to counties with an incidence rate greater than or equal to 10/100000, and 0 to those with a lower incidence rate.

RESULTS AND DISCUSSION

- After GridSearch we adopted Random Forest algorithm for classification model with ROC AUC of 0.974 , recall of 1.0 and precision 0.42. We tested our model with 2021 data and found that Gila County 94.06% and Navajo county had 36.05 % of having a high incidence of RMSF. Also we found that LST and population are the most important predictor variables.
- LSTM results suggest that test Root Mean Square Error (RMSE) values ranging from 0.006 to 9.055.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
ExtraTreesClassifier	0.95	0.97	0.97	0.96	0.12
GaussianNB	0.83	0.91	0.91	0.87	0.01
Perceptron	0.79	0.88	0.88	0.84	0.01
LinearSVC	0.95	0.82	0.82	0.95	0.03
RandomForestClassifier	0.95	0.82	0.82	0.95	0.16
LogisticRegression	0.95	0.82	0.82	0.95	0.04
LinearDiscriminantAnalysis	0.93	0.81	0.81	0.93	0.04
BaggingClassifier	0.93	0.81	0.81	0.93	0.03
XGBClassifier	0.90	0.79	0.79	0.92	0.39
PassiveAggressiveClassifier	0.90	0.79	0.79	0.92	0.01
QuadraticDiscriminantAnalysis	0.88	0.78	0.78	0.90	0.01
AdaBoostClassifier	0.88	0.78	0.78	0.90	0.10
LabelSpreading	0.88	0.78	0.78	0.90	0.03
LabelPropagation	0.88	0.78	0.78	0.90	0.02
ExtraTreeClassifier	0.88	0.78	0.78	0.90	0.01
DecisionTreeClassifier	0.88	0.78	0.78	0.90	0.01
NearestCentroid	0.67	0.67	0.67	0.75	0.01
KNeighborsClassifier	0.90	0.64	0.64	0.90	0.02
SVC	0.90	0.64	0.64	0.90	0.01

Fig. 2: Machine Learning algorithms for classification of RMSF incidence

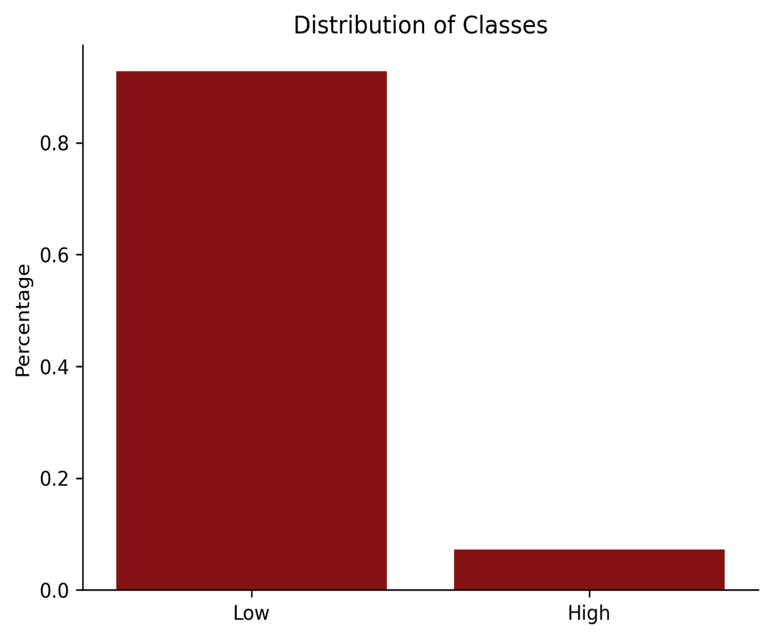


Fig. 3: Distribution of RMSF classes

```
Pipeline(steps=[('scaler', StandardScaler()),
                ('ros', RandomOverSampler(random_state=42)),
                ('RF',
                 RandomForestClassifier(min_samples_leaf=2, n_estimators=1500,
                                        n_jobs=-1, random_state=42))])
```

Fig. 4: Final Random Forest pipeline with parameters

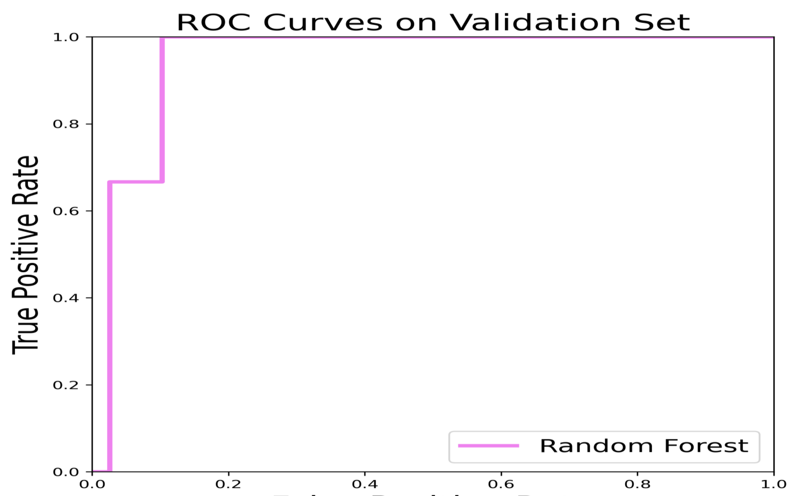


Fig 5: ROC Curve on Validation Set

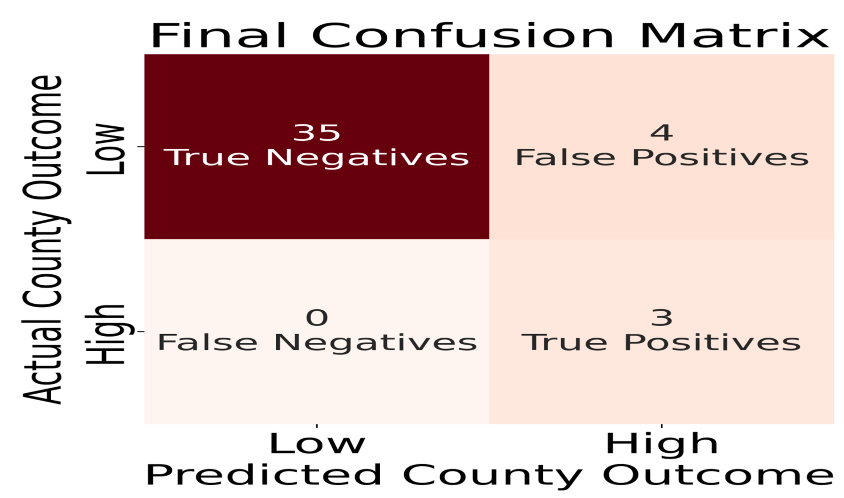


Fig 6: Final Confusion Matrix

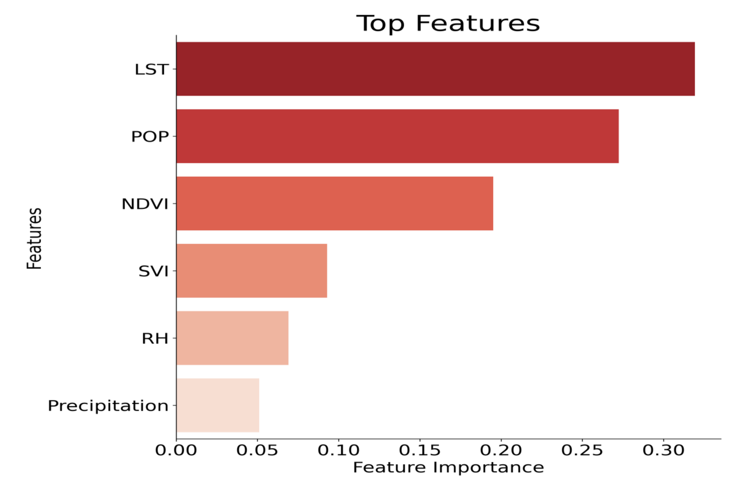


Fig 7: Feature importance of RF

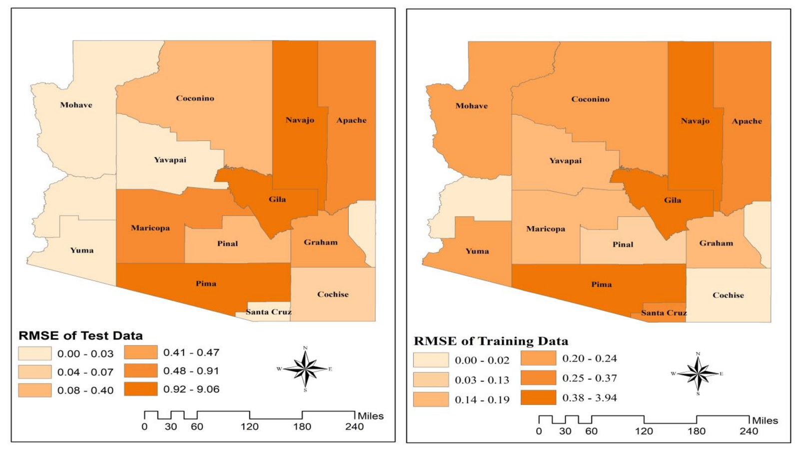


Fig. 8 : RMSE of test and training data

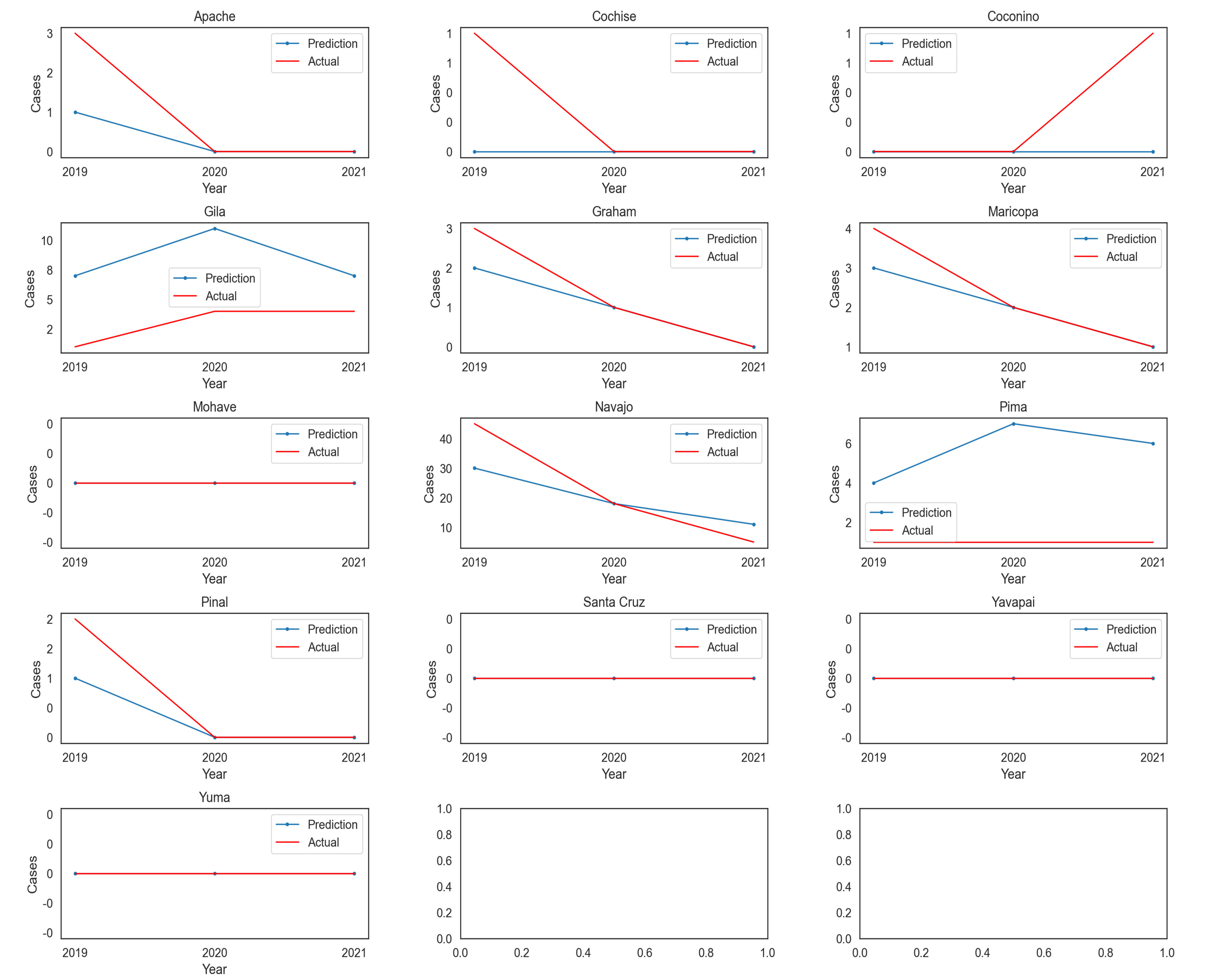


Fig. 9 : LSTM prediction on 3 years of test data 2019-2021

LIMITATION AND FUTURE WORK

- The data is satellite derived ;thus, temporal and spatial resolution of data is a limitation of this study.
- We had county-level RMSF cases. More, granular-level data will facilitate more accurate modelling of the disease dynamics.
- In future work, authors want to integrate more granular RMSF data at zip-code level to improve the prediction model.

ACKNOWLEDGMENTS

- This work was facilitated with Google Earth Engine and ADHS, we thank to GEE team and ADHS.
- Funding for this project provided by the National Science Foundation, Award Number 201960.
- This work has not been formally reviewed by NSF. The views expressed are solely those of the authors and do not necessarily reflect those of these Agencies.