USING CLIMATIC VARIABLES TO PREDICT THE SPREAD OF LYME DISEASE

Sean Iredell

Abstract

Lyme disease, the most common vector-borne disease in the United States, is driven by the spread of black-legged ticks, whose range is expanding due to climate change. This research examines the spatial and climatic factors influencing the spread of Lyme disease, utilizing machine learning models to predict future disease patterns. A comprehensive approach integrating CDC Lyme disease incidence data and NOAA climate data was employed, with challenges addressed through extensive data wrangling to merge disparate datasets. Initial exploratory data analysis revealed negligible correlations between climate variables and Lyme disease across the entire U.S., but regional analysis highlighted significant climatic influences in the Midwest.

Two models were tested: a Long Short-Term Memory (LSTM) model and a Convolutional Neural Network (CNN). The CNN, was selected for its ability to capture spatial heterogeneity, achieved promising results with an R^2 of 91.46% on training data and 63.35% on testing data, though overfitting limited its generalizability. The study found that climate variables alone cannot fully explain the spread of Lyme disease, underscoring the need for more granular, seasonal data and additional factors such as land cover, wildlife hosts, and human activity. Despite its limitations, this research provides a baseline for modeling tick-borne disease dynamics and offers a tool for public health officials to predict and prepare for future Lyme disease trends. Future work should refine temporal data aggregation, incorporate additional environmental and socio-ecological variables, and explore the broader impacts of climate change on vector-borne diseases.

Literature Review

Lyme disease is "the most common vector borne disease in the United States", and has shown massive signs of growth in recent years as the habitat for native tick populations has expanded (Stevenson, 2019). Utilizing spatial health data and mapping the spread of disease is one of the most effective ways to combat its spread, but especially in the case of lyme disease which is exclusively spread by ticks. This makes the mapping of the affected area very important, because lyme disease cannot spread at a faster rate being that it is a tick borne illness. Thus, mapping the affected area, and predicting future areas based on changes in climatic conditions should be all but certain.

Lyme disease is spread by the *Borrelia burgdorferi* bacteria and the black-legged tick, mostly in the eastern half of the United States. With the disease exclusively spread by ticks, its virality is almost entirely determined by the ability for these animals to survive and reproduce. This leads to a necessity for climatic conditions which support the growth and spread of ticks. These conditions have become enhanced with increasing temperatures and moderate climatic conditions in the midwest and northeast. Climate change was classified as a "key driver of the northern spread of the disease", with an expansion rate of "250-500 km by 2050" (Simon, Julie A., et al., 2014). Climate change provides the necessity for a multifunction model that accounts for spatial distribution of the current ticks, reported cases of lyme disease, as well as, predictive modeling that better understands the climatic conditions of the future that are driving the increased range of these ticks. Climate change is something that has already been observed in Scandinavia with climatic factors causing "prolonged vegetation period causing an increase in both [tick] range and abundance" (Jung et al., 2019).

Currently, in the field of research there is a significant amount of research being conducted on spatio-temporal forecasting into the spread of black-footed ticks, and thus the spread of lyme disease. However, there exists a major gap in research when it comes to predictive modeling of the future movement of lyme disease as the black-footed tick's habitat in North America grows with warming temperatures in Canada and the northern United States. Currently, the main statistical model that has been used in testing in North America was a standard regression approach, which ignores the effects of non-linear variables and collinearity between variables (Couper et al. 2021). There has been significant success in mainland Europe and Scandinavia with Boosted Regression Trees, and Neural Network models (Jung et al., 2019), (Chumachenko et al., 2022), which were able to predict lyme disease abundance with high accuracy, at as little as 4% error.

This gap in current literature will prove to be significant to explore because it will give us insight into the role that climate change is playing in the expanding territory of black footed-ticks and the spread of lyme disease. This research will help epidemiologists and public health officials identify future hotspots where ticks may become prevalent based on climate features. Combining these climatic variables with other predictive variables including vegetation coverage and prevalence of other wildlife hosts that ticks feed on, will provide a comprehensive multi-faceted approach that will further bolster existing research on the subject.

Our research team will create a multi-variable machine learning model to predict the spread of ticks, and present its findings to help scientists better understand the effects of climate change. The proposal is to use a neural network model to weight and evaluate several variables in the spread of lyme disease through black-legged ticks. These variables will include climatic conditions in addition to current tick distribution and growth patterns. We believe this model will be successful because of its ability to handle non-linear relationships between multiple variables, as opposed to many of the papers in the lit review which consisted of simple logistic regressions.

Data Wrangling

When exploring data options for this project there were two main sources that were used. The first was a lyme disease incidence tracker that was created by the CDC. This data source was in a CSV format and provided county level data for every county in the contiguous United States. The county level data was how many lyme disease cases per year per county and even labeled each county as "low incidence" and "high incidence." The second data source consisted of monthly climate data from NOAA, retrieved via an API from a climate database that provided monthly data for every county in the United States. This data features climatic variables such as average temperature, temperature range, and precipitation. The first step in data wrangling was creating yearly averages for all the climate variables, so that they could be compared to the lyme disease incidence data which was already in a yearly format. This consisted of calculating yearly averages for temperature, a yearly range for temperature created by subtracting the min and max temperature for the year, and precipitation averages from the year. These global statistics provide the basis for a standard exploratory data analysis that will help us clean the data and further understand underlying trends that will help with creating as accurate a study as possible.

The next part of the data wrangling proved to be the hardest with the lyme disease data and weather data being labeled with different column attributes. Lyme disease data is labeled with the county name (ie. Los Angeles County), and the weather data being labeled with its FIPS county code (ie. 60001). This mismatch leads to the necessity for a key that can be used to match up the values. The key the research team chose to use was data from the U.S. census that provided both the county name and county code together. This allowed the researcher to append the county name to the weather data based on the county code, and eventually create a merged data frame that contains both lyme disease data and climate data. There was some initial trouble when merging the two data frames because of a mismatch in the climate data state codes and the FIPS codes that represented the counties leading to several counties not having appropriate state attached, however this only affected roughly 1/3 of the data set. Since the error was not all encompassing we were still able to conduct a basic exploratory data analysis. The last step I had to take before merging the data frames was reshaping the lyme disease dataframe so that the yearly data was moved into columns while the county data was moved into rows. All of these steps allowed the research team to conduct a cursory exploratory data analysis to determine if the entire study area would be appropriate for research, or if the study area must be segmented into smaller areas.

Exploratory Data Analysis Part I

With the conclusions from the data wrangling analysis a small sample exploratory data analysis was performed. The first step was to visualize the datasets to ensure that there were no immediately obvious outliers. The first dataset that was called under question was the Delaware average temperature which was significantly above the average for the region which included average yearly temperatures all below 65 degrees fahrenheit.

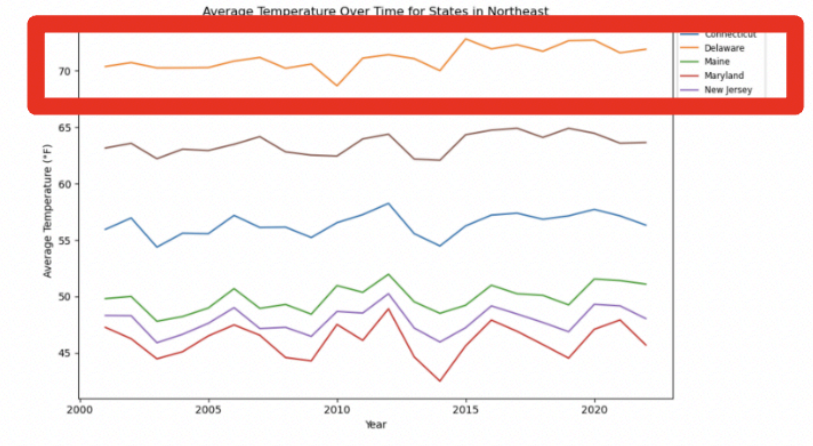

Figure 1. Delaware Average Temperature

This seemed to be the only state level dataset that was significantly out of proportion, the other datasets seemed to align with state and region averages, which was a positive achievement for the other climatic variables. The next step after exploring the climatic variables for basic viability was understanding their relationship with the lyme disease case study. This EDA came with a basic correlation matrix to better understand the relationship between climate variables and the spread of lyme disease. The first correlation matrix featured an almost negligible effect of climatic variables on the eastern half of the United States.

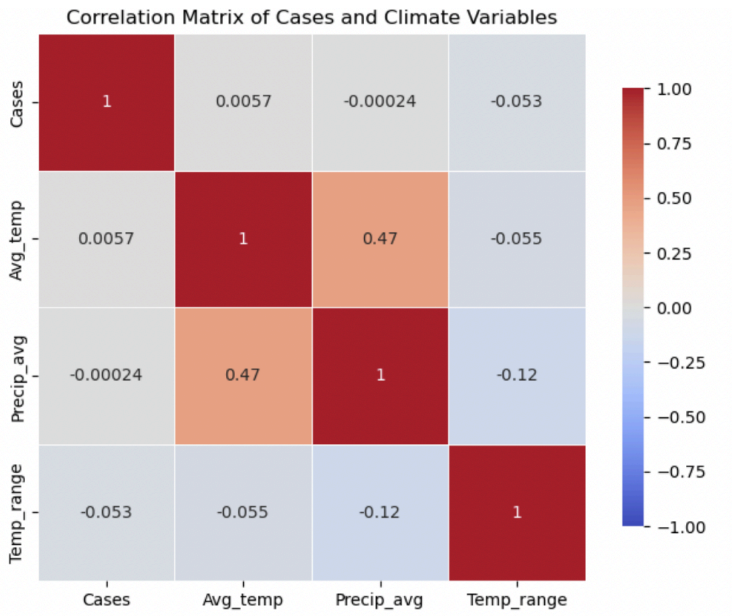

Figure 2. Climatic Variables and their effects on Lyme Disease

As pictured in the above correlation matrix there was little to no correlation determined by climatic variables across the whole dataset. However, with the significant increase in lyme disease instances that was observed in the overall dataset the research team did not give up and decided to take a more regional approach to lyme disease cases and their spread. The states that were selected for analysis were then broken up into regions to determine if there were any regional differences that were not accounted for in the overall correlation matrix. In the second running of the correlation matrix the midwest region provided extremely encouraging results with regard to climatic conditions being correlated with lyme disease spread as shown in figure three. In the matrix, climatic variables provide at least somewhat of a correlation in terms of lyme disease spread across the midwest. This is a major departure from the other two regions where there was little to no correlation between climatic variables and the spread of lyme disease. Overall, this finding suggests that for the future of the project the research team should focus on exploring other climatic variables and their influence on the midwest. Based on current data it could be presumed that the midwest is the most affected region both because of their prevalence of ticks and because they are the most susceptible to climate change variables because of the rise in temperatures in that area of the world.

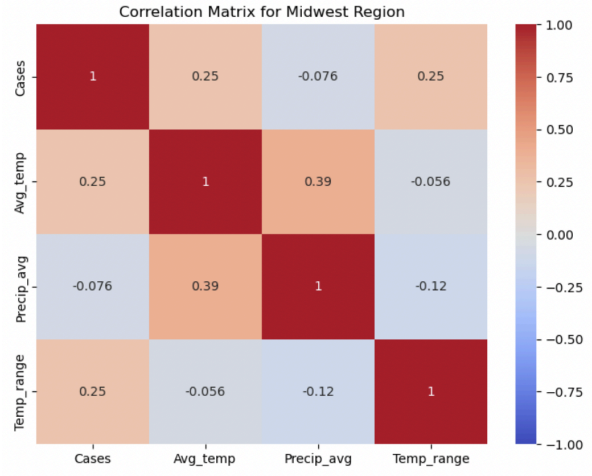

Figure 3. Correlation matrix for the midwest

Exploratory Data Analysis Part II

During early states of model building it was determined that a second round of exploratory data analysis was required to be able to make a better fitting model. Two important considerations that were not considered in the first model were temporal lag and land cover data. Because of tick growing cycles it was important to account for at least one year of lag within the data. This led to a second EDA of correlation matrices to help determine how spatial lag would affect different regions of the country. Additionally, the research team wanted to understand how changes in the granularity of the data would change correlations. Thus the team made correlation matrices that compared results from different states. This was important because when training the model I wanted to see which states would have the best baseline comparisons for the climate variables and investigate those states first. These states can be pictured in figure four.

```
State: Arizona
  Cases vs. Precip_avg: 0.18
  Cases vs. Avg_temp: -0.32
  Cases vs. Temp_range: -0.05

State: Delaware
  Cases vs. Precip_avg: 0.00
  Cases vs. Avg_temp: 0.10
  Cases vs. Temp_range: -0.29

State: New Hampshire
  Cases vs. Precip_avg: -0.02
  Cases vs. Avg_temp: 0.19
  Cases vs. Temp_range: 0.17

State: Florida
  Cases vs. Precip_avg: 0.06
  Cases vs. Avg_temp: 0.18
  Cases vs. Temp_range: -0.09

State: Minnesota
  Cases vs. Precip_avg: -0.11
  Cases vs. Avg_temp: -0.18
  Cases vs. Temp_range: -0.03
```

Figure 4. Sorted Correlation matrix by Absolute highest correlation

As demonstrated in the graphic the best starting point for analysis was Arizona, followed by Delaware, New Hampshire, and Florida. When conducting more analysis Arizona was eliminated based on a lack of cases and analysis began on Delaware, New Hampshire, Florida, and Minnesota. This selection of states provided a robust starting point as all of the selected states represented different regions of the United States where climatic factors may have had greater or lesser effects on lyme disease. For example, Delaware represents the more urbanized and moderate climate of the northeast, while Florida represents a more humid subtropical environment. The goal was to examine each state, and see if it would be possible to build a model that could accurately predict cases in that state, and then see if that model was

generalizable enough to be able to predict cases in the region, and then if possible the entire country.

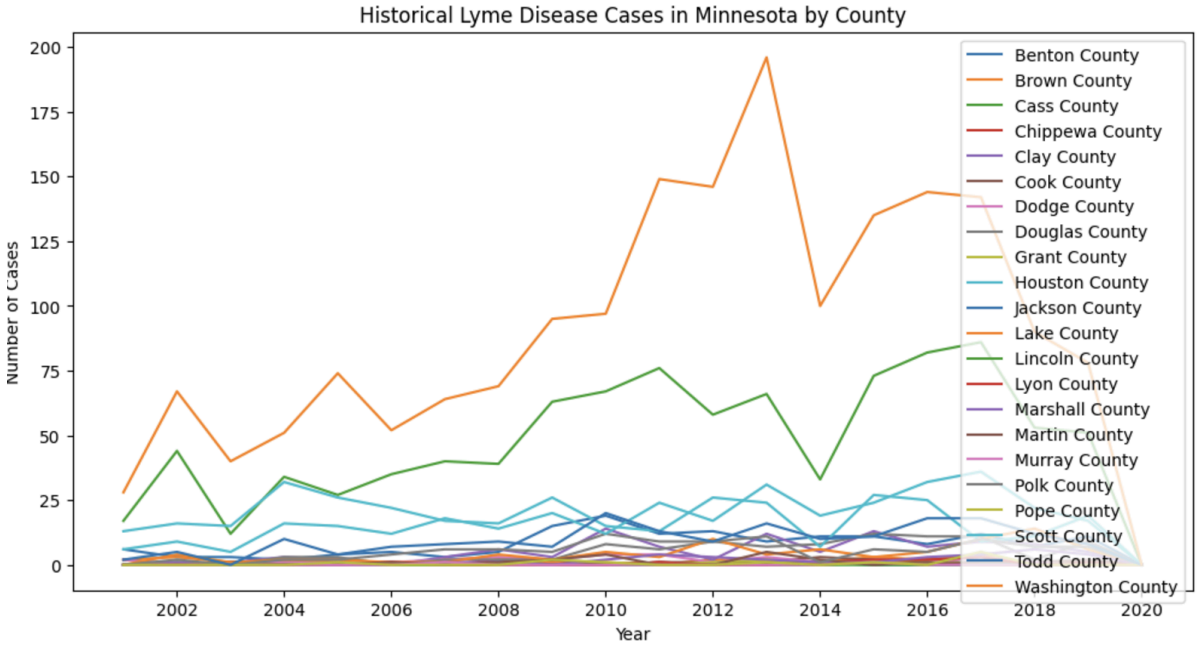

Figure 5. Lyme Disease Cases in Minnesota

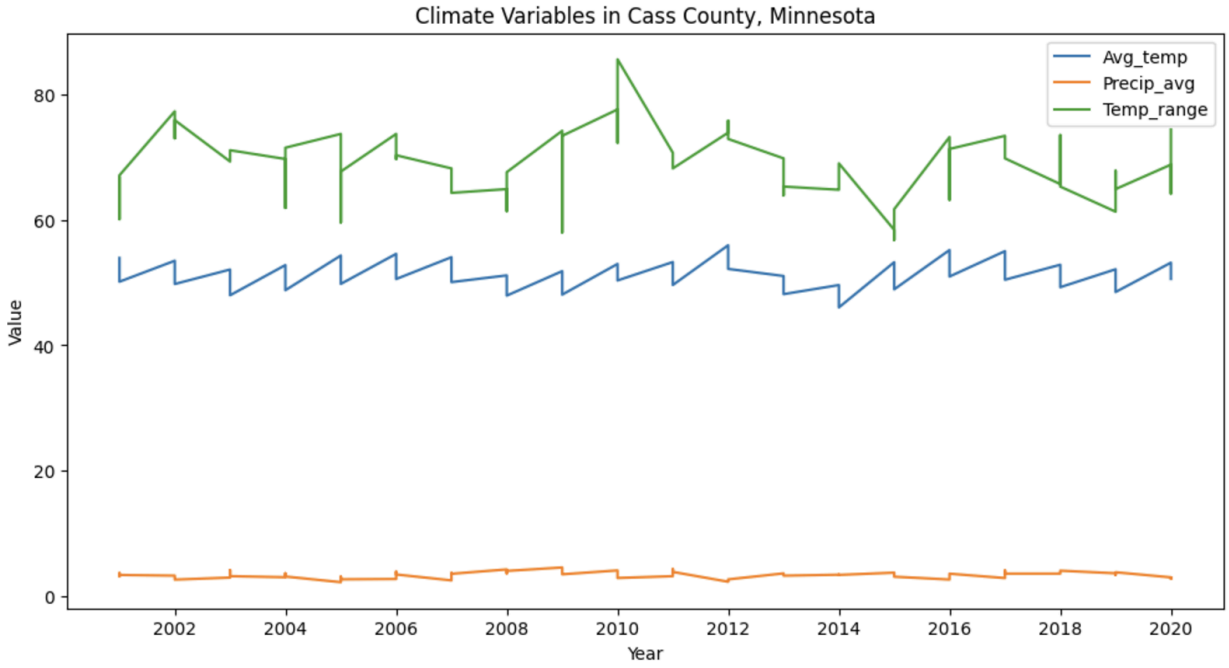

Figure 6. Climate Variables in Cass County Minnesota

To start, many of the top counties for cases in the state were evaluated, for example in figures five and six Cass County Minnesota. This approach was to start at the most granular level of data we have for both lyme disease cases and climate variables and see which counties would be best fit for the model. Having a good understanding of which counties, within states, within regions would work best, it was time to start building the models.

Methods

Several different models were tested across different parts of the dataset in hopes of creating a generalized model that could be trained and reproduced throughout other adjacent datasets. The two model types that were tried were a Long Short Term Memory Model (LSTM), and a Convolution Neural Network model (CNN) which was eventually chosen.

The first type of model tested on the data was an LSTM model. This model was chosen for its ability to handle large datasets with temporal features, particularly data that might include temporal lags, which could be significant in tick growth cycles. While the model did not perform well in predicting case counts—achieving an R^2 of approximately 0.2 on the training data and near zero on the test data—it offered valuable insights into the model-building process. One key observation from this model was that over 68% of the rows in the dataset contained zero cases, with a mean of around 10 cases. This created a large imbalance in the data with the range being well over 1,000. Thus it was necessary to implement scalers in the weather data as well as eliminating many of the states in the study that were not receiving cases, and thus a second model was chosen to better capture spatio-temporal trends that were prevalent in the data.

The second model featured a convolution neural network with K-means validation to better capture spatial heterogeneity between features. The model was run using the Keras package of TensorFlow.The model was tested using a five-fold temporal cross validation package to ensure that the test results were indicative of the training data, while also still considering temporal heterogeneity. This is demonstrated in the code in figure seven.

```
# Prepare data for the CNN
seq_length = 50
features = data[['Cases', 'Precip_avg', 'Max_temp', 'Avg_temp', 'Temp_range']].values
X, y = create_sequences(features, seq_length)
# Initialize TimeSeriesSplit with 5 splits
n_splits = 5
tscv = TimeSeriesSplit(n_splits=n_splits)
```

Figure 7. Time series with five fold validation

There were standardized epochs and batches sized at 50 and 16 respectively such that the model could capture trends without becoming overfit. The model was then trained on 80% of the training data with the last five years being saved to be tested on later. Before the model was able to be applied to the test data set it was cross validated based on the code in figure eight.

```
# Cross-validation loop
for train_index, val_index in tscv.split(X):
    X_train_cv, X_val_cv = X[train_index], X[val_index]
    y_train_cv, y_val_cv = y[train_index], y[val_index]

    # Reshape input data for CNN (samples, timesteps, features)
    X_train_cv = X_train_cv.reshape(X_train_cv.shape[0], X_train_cv.shape[1], X_tra
    X_val_cv = X_val_cv.reshape(X_val_cv.shape[0], X_val_cv.shape[1], X_val_cv.shap

    # Create and train the CNN model for each fold
    model = Sequential()
    model.add(Conv1D(filters=64, kernel_size=3, activation='relu', input_shape=(X_t
    model.add(MaxPooling1D(pool_size=2))
    model.add(Flatten())
    model.add(Dense(50, activation='relu'))
    model.add(Dense(1))  # Output layer for regression

    model.compile(optimizer='adam', loss='mse')

    # Train the model on the current fold
    model.fit(X_train_cv, y_train_cv, epochs=50, batch_size=16, verbose=0)

    # Make predictions on train and validation sets
    y_train_pred_cv = model.predict(X_train_cv).flatten()
    y_val_pred_cv = model.predict(X_val_cv).flatten()
```

Figure 8. Cross Fold Validation

This cross fold validation allowed the model to be fitted to its neighbors' spatio-temporal neighbors, such that each county was checked with other counties within the state to make sure the values being received each fold were consistent with their neighbor values. This is important because it provides the basis for the model to improve itself, not only with each epoch, but also with each fold. This adjustment improved the R^2 score of the dataset, raising it from the mid-70% range to over 90%. While this method proved to be highly effective, there is a possibility that it introduced an overfitting element to the data, which should be considered in future evaluations. Overall this CNN model was able to outperform the LSTM model significantly with a 91.46% R^2 score on the training data and a 63.35% R^2 on the test data as shown in figure nine. These results suggest that, despite being overfit, the model was able to achieve its primary goal of predicting Lyme disease cases using climatic variables, at least to some degree.

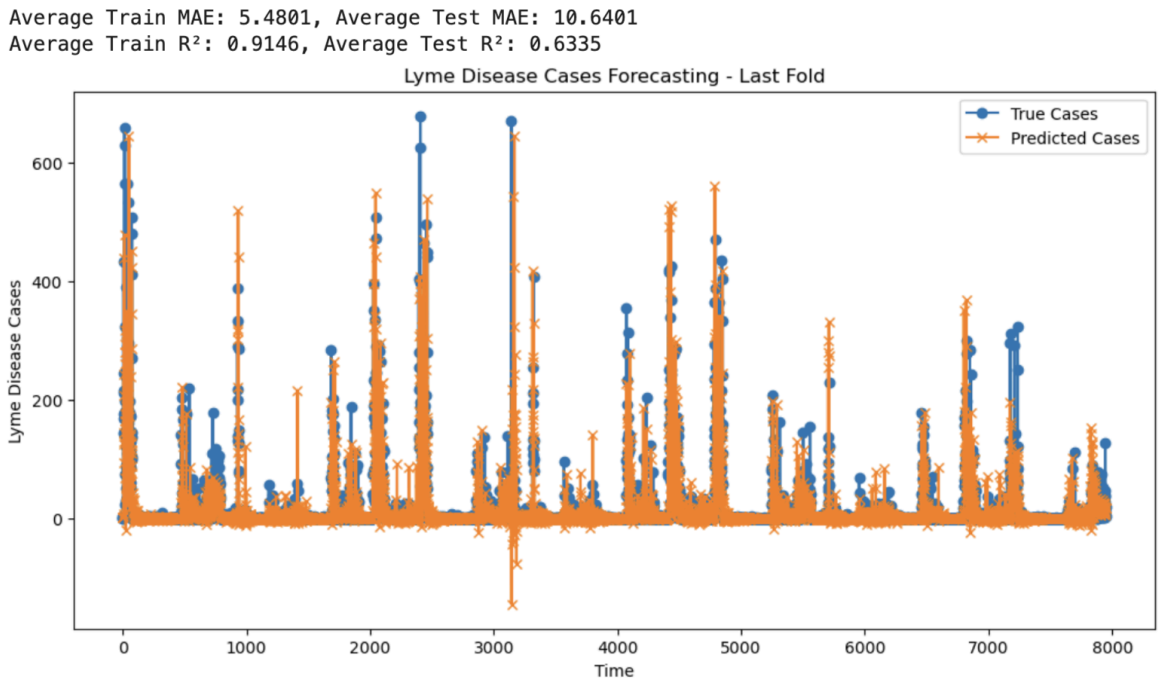

Figure 9. Final Model Output

Limitations

Despite the extreme success of the model on training data, and moderate success on testing data there are still many limitations that inhibit its usefulness. The biggest limitation of the model is that it is overfit, based on the sizable differences between accuracy on training and test data sets. The model was clearly specialized to the first four folds of data, and does not provide effective predictions on the fifth fold of test data. This suggests that the model is not useful in predicting future infection of lyme disease to the same degree that it was at predicting past lyme disease. For this reason it may be concluded that the model fails in its goal of being able to predict future lyme disease trends based on climatic variables, but it does make important strides in providing a view of lyme disease trends that can be cross validated on spatio-temporal neighbors. This means that the model interpolates and compares several counties within a state in comparable time periods to provide the model with similar outputs to cross validate and improve the models efficacy every fold. This spatio-temporal connection should provide the basis for future code and is by far the most important stride that the research team was able to make.

Discussion

After examining the limitations of the research study, it is crucial to interpret the results to extract meaningful insights that can guide future research and foster a deeper, shared understanding of the spread of Lyme disease. This research study, and culminating model are important to future research because they provide the foundation for future neural network models that explore the linkage between climate variables and the spread of lyme disease. One of the most important discoveries of the research team is there is a general lack of scientific consensus on how much each of the various factors actually impact the spread of lyme disease. This suggests that the most important piece of future research into this subject is an approach that

quantifies the importance of climatic factors within the spread of lyme disease. This is a large undertaking because there is very little research conducted on all of the different variables, their different relative importance, and even region by region variation.

Understanding the spread of lyme disease simply cannot be a nationwide approach. There are too many regional differences based on urbanization trends, climate, and cultural phenomena that impact human outdoor activity. For example, outdoors activities such as hiking are much more prevalent in the New England region as opposed to the more urbanized/suburbanized northeast. As can be seen in the research of this study even though there are regional differences in climate and thus regional differences in lyme disease there must be more research done to find a definitive link between the two, because ultimately this study fails to make concrete conclusions about that relationship and instead provides a model for future prediction, that only provides some ability to be generalized.

Despite the shortcomings of the research in regards to understanding the importance of climate in the spread of lyme disease, the convolution neural network (a black box model) was able to provide statistically significant predictions and thus provides a baseline for further research in the field. This model could be used today by local public health officials at a county level to better understand how future climate trends are going to affect their counties. This could help prevent the spread of the disease through public awareness campaigns in the case of increased tick populations, or the diversion of lyme disease resources to other areas in the event of less cases.

Conclusion and future research

In conclusion, the research team suggests that climate variables do not have as much of an influence on the spread of lyme disease as previously thought. Additionally, the research team has determined that without monthly or seasonal data it is impossible to accurately evaluate the role that climate variables play in the spread of ticks and subsequently lyme disease.In the NOAA dataset, climate data was provided as monthly values. However, the research team made an error in aggregating this data. Since Lyme disease case data from the CDC was available only as yearly totals, the team aggregated the monthly NOAA data over an entire year. This approach was flawed, as Lyme disease cases typically occur during specific months of the year. Even without precise knowledge of the active season for each region, it would have been more accurate to limit the aggregation of climate data to the likely active period for Lyme disease in each region. This adjustment would have better aligned the climate data with the temporal dynamics of disease cases. In conclusion, eliminating climate data from the winter when there was not any spread of lyme disease would likely have improved the quality of the model and helped us better understand the relationship between the climate and the spread of lyme disease.

In future research, determining the correlation between all factors that influence the spread of lyme disease may help to make a more complete picture by contextualizing the results of the model. This could include better understanding the role that wildlife hosts, deforestation, land cover changes, and lack of awareness play in the spread. A future research project that outlines the specific role that each of these variables play in the spread of climate change could

help researchers understand how changes in climate are actually going to affect the spread of lyme disease in the context of all of the variables. For example, our climate model suggests a 10% decrease in the spread of lyme disease, but a different variable suggests a 15% increase and based on both of these models and a correlation matrix on the importance of each variable a local health department could determine what they believe is going to be the aggregate effect.

Overall, there is still much work to be done when it comes to better understanding the spread of this disease and what can be done to combat it. Spatial data plays a significant role in most diseases, but plays an elevated role in the spread of lyme disease based on the black-footed tick being the only vector for disease transmission, and its mobility limited to land that can support its lifecycle development.

Works Cited:

Jung Kjær, L., Soleng, A., Edgar, K.S. et al. Predicting the spatial abundance of Ixodes ricinus ticks in southern Scandinavia using environmental and climatic data. Sci Rep 9, 18144 (2019). https://doi.org/10.1038/s41598-019-54496-1

Chumachenko D, Piletskiy P, Sukhorukova M, Chumachenko T. Predictive Model of Lyme Disease Epidemic Process Using Machine Learning Approach. Applied Sciences. 2022; 12(9):4282. https://doi.org/10.3390/app12094282

Couper, L. I., MacDonald, A. J., & Mordecai, E. A. (2021). Impact of prior and projected climate change on US lyme disease incidence. Global Change Biology, 27(4), 738–754. doi:10.1111/gcb.15435

Simon, J. A., Marrotte, R. R., Desrosiers, N., Fiset, J., Gaitan, J., Gonzalez, A., et al. (2014). Climate change and habitat fragmentation drive the occurrence of orrelia burgdorferi, the agent of lyme disease, at the northeastern limit of its distribution. Evolutionary Applications, 7(7), 750–764. doi:10.1111/eva.12165

Stevenson, M. (2019). The effects of land cover on the spatial distribution of lyme disease in northern virginia since 2005. Retrieved from https://vtechworks.lib.vt.edu/server/api/core/bitstreams/30f9bd8d-afd2-453b-b730-91736f3ab347/content