AIJPS
Academic International Publishers
Academic International Journal of Pure Science (E. ISSN: 2984-7737)
Aca. Intl. J. P. Sci. 2025; 03(2):22-34.

Data Mining in Healthcare: Predictive Analytics and Outbreak Detection of Lyme Disease

Mazharul Islam Tusher[1], Md Rayhan Hassan Mahin[2], Estak Ahmed[3], Redoyan Chowdhury[4], Mujiba Shaima[5]

[1]Department of Computer Science, Monroe University, New York, USA. tusher079@gmail.com
[2] Department of Computer Science, Monroe University, New York, USA. muhammodrayhan@gmail.com
[3] Department of Computer Science, Monroe University, New York, USA. eahmed7797@monroeu.edu
[4] Department of Business Administration, International American University, California, LA, USA. c.redoyan@gmail.com
[5] Department of Computer Science, Monroe University, New York, USA. shaimacme@gmail.com

(Received 2 August 2025, Revised 8 September 2025, Accepted 10 October 2025, Published 1 November 2025)

Abstract
The most prevalent tick-borne disease in the US is Lyme disease, which is becoming more prevalent as a result of ecological and environmental variables. This study explores the potential of data mining and machine learning to improve Lyme disease outbreak prediction and early detection. The study assesses seasonal trends, geographic hotspots, and demographic patterns from 1996 to 2023 using CDC surveillance data and Python-based data analytics. The study identifies possible epidemic zones with a 75.7% accuracy rate using data preprocessing, visualization, and predictive modeling, which includes Random Forest categorization. The study focuses on high-risk states like Pennsylvania and Delaware, as well as groups that are more likely to be hurt, like kids ages 5 to 14 and people over 75. Using Kaggle datasets and machine learning algorithms like Linear Regression and Logistic Regression to look at historical data shows that all models work the same way. Based on spatial and temporal characteristics, Random Forest models in particular show great promise for identifying outbreak times. According to the study's findings, data-driven surveillance provides a potent instrument for making decisions about public health. For increased accuracy, it suggests including socioeconomic and environmental aspects in future models. This research highlights the importance of predictive analytics in public health responses and lays the groundwork for real-time Lyme disease monitoring systems.

Keywords:
Lyme Diseases, Random Forest Classifier, Linear Regression, Lasso Regression, Ridge Regression, and Logistic Regression

How to cite:
Mazharul Islam Tusher, Md Rayhan Hassan Mahin, Estak Ahmed, Redoyan Chowdhury, Mujiba Shaima Data Mining in Healthcare: Predictive Analytics and Outbreak Detection of Lyme Disease. *Aca. Intl. J. P. Sci. 2025; 3(2)22-34:* https://doi.org/10.59675/P323

Introduction

Lyme disease is an underrated infection that spreads among humans through the bite of infected black-legged ticks, sometimes known as deer ticks, such as Ixodes scapularis. The tickdeer anda bacterium named Borrelia burgdorferi, which lives mostly by feeding on infected animals such as mice or deer, and is the source of the disease. The germs can travel throughout the body after a bite, perhaps impacting the heart, joints, skin, and neurological system. With an increasing number of cases recorded each year, it is the most prevalent tick-borne illness in the US [1]. Because Lyme disease can result in serious sequelae, such as neurological, cardiac, and joint-related diseases, if treatment is not received, it is a major public health concern. By examining geographic distribution, seasonal variations, and demographic variables including age and gender, this study seeks to understand the epidemiology of Lyme disease in the United States. This study aims to identify high-risk locations and guide public health measures by applying Python-based data analytic techniques. Early in the 20th century, people throughout Europe began to describe symptoms similar to Lyme disease. An early indicator of Lyme disease, erythema migrans is a characteristic growing skin rash that was first described by Swedish dermatologist Arvid Afzelius in 1909.

The 1970s saw the official recognition of Lyme disease in the United States. In 1975, the town of Old Lyme, Connecticut, had an uncommon epidemic of arthritis patients [2,3]. Epidemiological investigations resulted from a group of adults and children displaying rash, fever, and arthritic symptoms. Lyme disease, named after the town, where the outbreak originated, is a new tick-borne sickness that was discovered by Yale University rheumatologist Dr. Allen Steere [4]. But it wasn't until 1981 that scientist Willy Burgdorfer identified Borrelia burgdorferi as the pathogen causing Lyme disease after finding it in the midgut of infected ticks [5].

The increase in Lyme disease incidence in the United States has been caused by a number of causes, such as deforestation, which has changed ecosystems in ways that enhance human exposure to ticks, and climate change, which has increased tick habitats. Reporting rates have also increased as a result of better diagnostic techniques and greater public awareness. Using nationwide surveillance programs, mapping high-risk areas, and releasing preventative measures to decrease human-tick encounters, the CDC continues to track trends in Lyme disease [6]. Lyme disease vaccine (LYMErix) was first made available in the late 1990s, but it was taken off the market in 2002 because of public reluctance and questions regarding its long-term effectiveness. To improve protection against Borrelia burgdorferi, however, novel vaccinations are presently entering clinical trials [4].

Literature Review

Overview of Lyme Disease

Borrelia burgdorferi is the bacterium that causes Lyme disease, which is spread via the bite of an infected black-legged tick (Ixodes scapularis). The progression of Lyme disease can be broken down into three distinct stages: early localized, early disseminated, and late disseminated. Fever, exhaustion, headaches, and the distinctive erythema migrans rash are among the early signs; later stages may cause cardiac and neurological issues [2]. Climate change, deforestation, and better monitoring techniques have led to an increase in the number of Lyme disease cases reported each year, making it a growing public health concern in the US, Europe, and parts of Asia [7,8].

Aca. Intl. J. P. Sci. 2025; 3(2)22-34: https://doi.org/10.59675/P323

Lyme Disease Early to Long-Lasting Symptoms

Figure 1: Lyme diseases stage 1 symptom [1]

State 01 of Lyme Disease Symptoms: After infection, early localized Lyme disease symptoms (Stage 1) appear days or weeks later. These resemble the flu and could include:

- Chills and a fever
- Overall negative sentiment
- Anxiety
- Pain in the joints
- Muscle aches
- Neck stiffness

A person's arm with a Lyme disease rash that resembles a bullseye (figure 1). The characteristic Lyme disease rash is a red, slowly growing area of skin that typically grows to a diameter of at least 7 inches. Where the tick bite occurred, there can be a red spot that is flat or slightly elevated. A "bull's eye" or clear patch in the middle is frequently seen. We refer to this rash as erythema migrans. If left untreated, it may persist for several weeks and lead to the development of additional patches at other locations. Lyme disease symptoms can appear and disappear [9, 10].

State 02 of Lyme Disease Symptoms: Weeks to months following a tick bite, early disseminated Lyme disease symptoms (Stage 2) might manifest as follows (figure 2 is reference):

- Numbness or discomfort in a nerve's distribution
- Face muscles that are paralyzed or weak, particularly in cases of facial nerve palsy
- The symptoms of meningitis include fever, stiff neck, and excruciating headaches.
- Heart issues, such irregular heartbeats, might make you feel dizzy or faint [10].

Figure 2: stage 2 Lyme Diseases Symptom [1]

State 03 of Lyme Disease Symptoms: After the infection, late disseminated Lyme disease (Stage 3) symptoms may appear months or years later.

- The most typical symptom, which occurs intermittently over months, is fluid buildup in one or two major joints at a time.
- The knees are particularly impacted.
- Neuropathy, or peripheral nerve pain and numbness, is another complication of late Lyme illness [10].

Epidemiology and Risk Factors

Geographical, seasonal, and demographic risk variables have been identified through the analysis of Lyme disease epidemiology in a number of research. Lyme disease, which is most prevalent in the Northeastern, Upper Midwest, and Pacific Northwest regions of the United States, is most prevalent in states like Pennsylvania, New York, and Wisconsin [11-13]. According to research, people over 50 and children aged 5 to 14 are particularly vulnerable because of their greater contact with outdoors [14]. Furthermore, Lyme disease cases have increased in formerly low-risk areas as a result of tick habitat expansion brought on by climate change [15].

Summer is the highest-risk season for the spread of Lyme disease because ticks prefer humid conditions and are most active during the warmer months, especially May through August [16]. The risk of infection is further increased by human activities including hiking, camping, and gardening in tick-endemic areas [17].

Role of Machine Learning in Lyme Disease Research

In the study of Lyme disease, machine learning (ML) has become a useful technique for predicting disease patterns, locating high-risk locations, and enhancing diagnostic precision. To comprehend the spread of disease and maximize public health interventions, researchers have employed data mining, predictive modeling, and geospatial analysis [18].

Deep Learning and Image Recognition

Tick presence and rashes linked to Lyme illness have been identified using deep learning algorithms. A convolutional neural network (CNN) model was created by Jordan RA et al. (2022) to evaluate photos of erythema migrans rashes and make a high-specificity distinction between them and other skin disorders [3]. The methodology increases the accuracy of diagnoses, especially in rural healthcare settings where access to specialists may be restricted.

Existing Data Analysis Studies on Lyme Disease

Large-scale datasets have been used in a number of data-driven research to examine patterns in Lyme disease and increase forecast accuracy.

Data from the CDC's Lyme Disease Surveillance Program, 1996–2023: The CDC has a wealth of information on the occurrence of Lyme disease at the state and county levels. Using this data, studies have shown a correlation between an increase in Lyme disease cases and urban growth into tick habitats [1].

Statista State-wise Incidence Rates (2022): Lyme disease hotspots are still primarily in the Northeast and Midwest, with new cases appearing in Southern states as a result of tick range expansion, according to an analysis of incidence rates across U.S. states [16].

Kaggle Datasets on Lyme Disease Trends: Researchers have trained forecasting algorithms to predict future outbreaks using time-series data on Lyme disease cases that Kaggle provides [18].

Challenges and Future Directions in ML-Based Lyme Disease Research

There are still issues despite the progress made in ML applications for Lyme disease:

- Data Limitations: Because early-stage Lyme disease is frequently misdiagnosed or undetected, several datasets suffer from underreporting [1].
- Environmental Variability: Adaptive models are necessary due to the unexpected effects of land-use changes and climate fluctuations on tick populations [15].
- Algorithm Transparency: Medical practitioners may find it challenging to incorporate results into public health initiatives due to interpretability issues with certain machine learning models [16].

Data Collection on Lyme Disease

Age, gender, and geographic location all play a role in determining the prevalence of Lyme disease. A bimodal age distribution is shown in figure 3 by data from 1996–2024 and 2022 [19, 20], with maxima among adults aged 75–79 and children aged 5–9. In all years or period of time, the majority of cases diagnosis were male, and the percentage is 57.7% in 2017–2019 and 57.3% in 2022. Geographically, high-incidence counties, mostly in the Northeast, Mid-Atlantic, and Upper Midwest regions, accounted for almost 95.5% of all reported cases in 2022 [1].

According to research, people over 50 and children aged 5 to 14 are disproportionately affected by Lyme disease [1]. Due to increased exposure to the outdoors, gender-based analyses indicate a higher frequency among males. Targeted preventative initiatives require an understanding of these demographic trends [14].

Aca. Intl. J. P. Sci. 2025; 3(2)22-34: https://doi.org/10.59675/P323

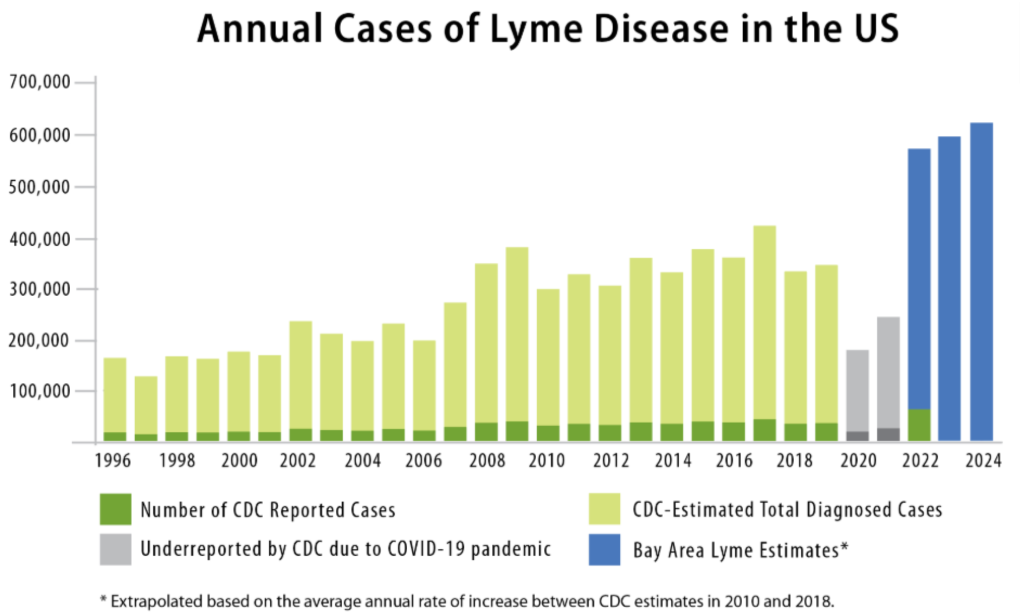

Figure 3: year wise patient count [1]

Analyzing Keggle dataset using pivot table, we are getting total number of male and female patient over 2010-2023. Even we can generate year wise patient count (figure 4) [21].

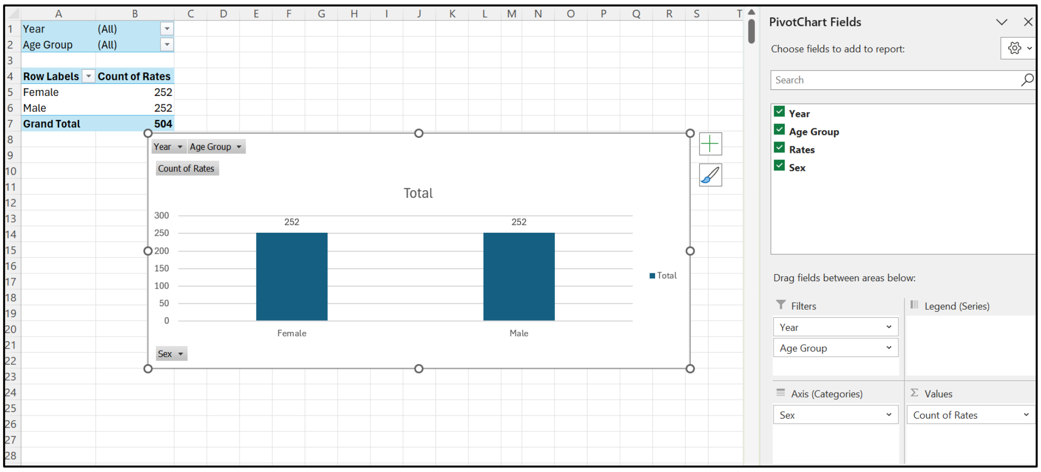

Figure 4: gender wise patient count

To illustrate the trends in Lyme disease cases, the following tables present data on reported cases by year, sex, and age group, as well as incidence rates by state (figure 5).

2022 Dataset

From 2022 dataset we get that Lyme Disease Cases by Sex (2022) is showing in Table 1

Table 1: Gender wise patient count

Sex	Cases	Percentage
Male	35,833	57.30%
Female	26,718	42.70%

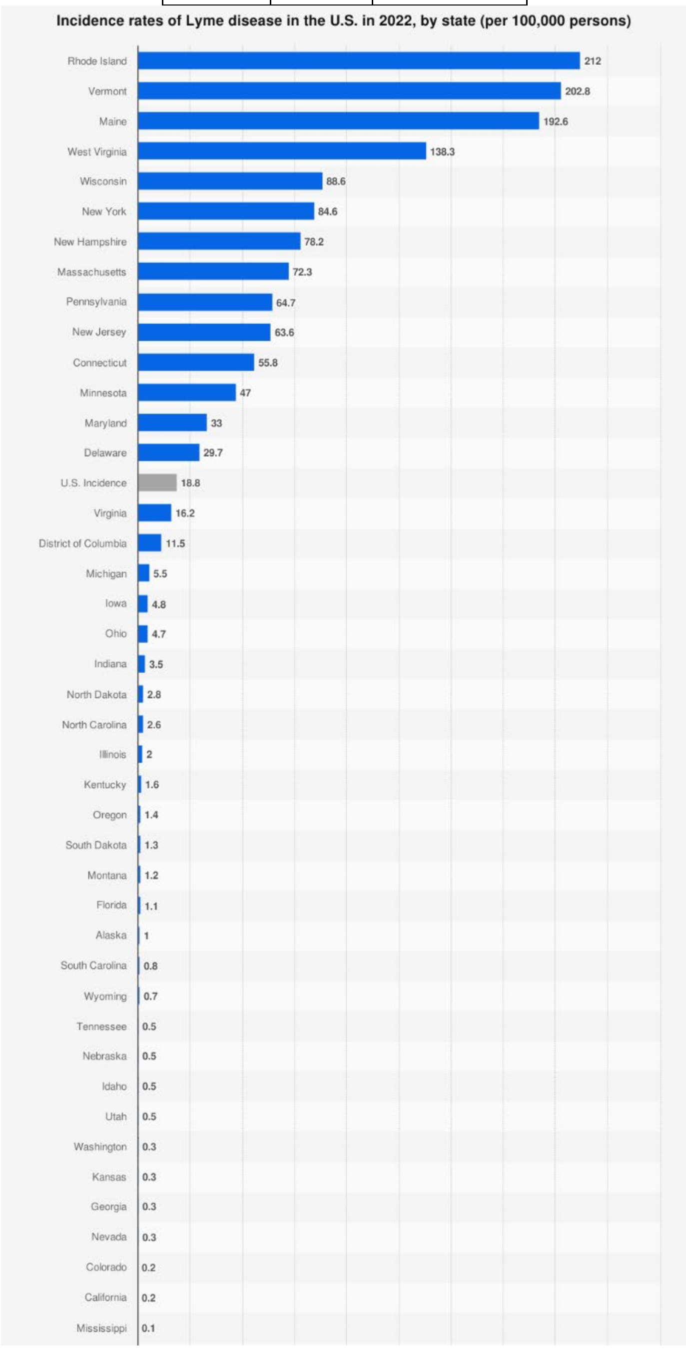

Figure 5: State wise patient count (18)

Table 2: Reported Lyme Disease Cases by Age Group (2022) [Note: Incidence rates are per 100,000 population.]

Age Group (Years)	Cases	Incidence Rate (per 100,000)
0–4	2,500	12.5
5–9	3,300	16.5
10–14	2,800	14
15–19	2,200	11
20–24	2,000	10
25–29	2,100	10.5
30–34	2,200	11
35–39	2,400	12
40–44	2,600	13
45–49	2,800	14
50–54	3,000	15
55–59	3,200	16
60–64	3,400	17
65–69	3,600	18
70–74	3,800	19
75–79	**4,000**	**20**
80+	**4,200**	**21**

From the above table 2 we can get that comparatively 80+ and 75-79 age group people are more affected by Lyme disease. Also, the older people are the more likely to be affected.

Table 3: Lyme Disease Incidence Rates by State (2022)

State	Cases	Incidence Rate (per 100,000)
Connecticut	**5,500**	**153.9**
Delaware	**2,000**	**205**
Maine	2,500	186
Maryland	1,800	29.8
Massachusetts	3,000	43.5
Minnesota	1,500	26.5
New Hampshire	1,200	88.3
New Jersey	3,500	39.4
New York	4,000	20.5
Pennsylvania	**8,500**	**66.4**

From table 3 we get that Pennsylvania state has more cases than any other state in 2022 whereas Minnesota has the least cases. But considering the incidence rate, Delaware came first.

Existing System or Coding to Analyze Lyme Disease Data

There has been a lot of research work on Lyme disease in the last decades. To compare existing research work that is already done and our work, we take the Kaggle dataset as the existing research work [21].

Here the author and researcher Tahmid (20) worked with Kaggle Lyme disease dataset for USA. His research summary is shown in figure 6

```python
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
from sklearn.linear_model import LogisticRegression
linearModel = LinearRegression()
lassoModel = Lasso()
ridgeModel = Ridge()
logReModel = LogisticRegression()
modelPredicitions("LinearRegression",linearModel,X_train_scaled,y_train,X_test_scaled,y_test)
modelPredicitions("LassoRegression",lassoModel,X_train_scaled,y_train,X_test_scaled,y_test)
modelPredicitions("RidgeRegression",ridgeModel,X_train_scaled,y_train,X_test_scaled,y_test)
modelPredicitions("LogisticRegression",logReModel,X_train_scaled,y_train,X_test_scaled,y_test)
```

```
____MODEL_NAME___LinearRegression
R2Score 0.6623164994263604
MSE 11174.967162390481
MAE 23.124211978738444
RMSE 105.71171724265234
____MODEL_NAME___LassoRegression
R2Score 0.6623164994263604
MSE 11174.967162390481
MAE 23.124211978738444
RMSE 105.71171724265234
____MODEL_NAME___RidgeRegression
R2Score 0.6623164994263604
MSE 11174.967162390481
MAE 23.124211978738444
RMSE 105.71171724265234
____MODEL_NAME___LogisticRegression
R2Score 0.6623164994263604
MSE 11174.967162390481
MAE 23.124211978738444
RMSE 105.71171724265234
```

Figure 6: Output summary of Tahmid [21]

Summary of the existing code/model

The performance measures are the same for all models, according to the output of the various models (Linear Regression, Lasso Regression, Ridge Regression, and Logistic Regression):

- R^2 Score: 0.6623
- Mean Squared Error or MSE: 11174.97
- Mean Absolute Error or MAE: 23.12
- Root Mean Squared Error or RMSE: 105.71

In terms of these particular measures, none of the models—Linear Regression, Lasso Regression, Ridge Regression, and Logistic Regression—perform better than the others because they all yield the same outcomes. This implies that performance may not be greatly impacted by the model type employed for this data set. However, the features, the dataset's characteristics, or other elements not included in the summary metrics could have an impact on this.

- Specifically for Lasso and Ridge regression, hyperparameter tuning
- Changing up the models or including more intricate features into the dataset
- Investigating preprocessing methods such as regularization.

Applying Predictive Modeling Using Random Forest Classifier

The Random Forest classification method was used in a supervised machine learning approach to determine times and places with a higher risk of Lyme disease outbreaks. An ensemble-based learning technique called Random Forest builds several decision trees during training and outputs the individual trees' mode of classification (classification) [22]. It is renowned for its resilience against overfitting, especially when working with high-dimensional or noisy data and is especially well-suited for structured datasets with a variety of data types.

In this particular study, the modeling challenge consisted of binary classification, which entailed making a prediction as to whether a particular county-period combination included an outbreak (defined as case counts that were higher than the period median) or a non-outbreak phenomenon. The following features were chosen for prediction:

- Start-year: denoting the beginning of every five-year time frame
- State Code: a depiction of a U.S. state that is numerically encoded.

By splitting the dataset into training and testing subsets using a 70/30 train-test split, the generalizability of the model was evaluated. 100 estimators were used to instantiate and train a RandomForestClassifier from the scikit-learn library using the training data. The confusion matrix, accuracy, precision, recall, F1-score, and other common classification metrics were used to evaluate the model's performance on the test set.

To analysis our dataset following RandomForestClassifier (figure 8)

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Step 1: Define outbreak (if cases above median in that year)
df_long["Outbreak"] = df_long.groupby("StartYear")["Cases"].transform(lambda x: x > x.median()).astype(int)

# Step 2: Encode state as numerical feature
df_long["StateCode"] = df_long["StateName"].astype("category").cat.codes

# Step 3: Select features and labels
features = ["StartYear", "StateCode"]
X = df_long[features]
y = df_long["Outbreak"]

# Step 4: Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Step 5: Train the Random Forest model
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

# Step 6: Evaluation
acc = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
report = classification_report(y_test, y_pred, output_dict=True)

acc, conf_matrix, pd.DataFrame(report).transpose()
```

Figure 8: Applying Random Forest Classifier to Kegle dataset

Results

The Random Forest model achieved an overall **accuracy of 75.7%**, indicating a relatively strong ability to distinguish between outbreak and non-outbreak periods using minimal features. The detailed classification report is as follows:

```
(0.7573443008225617,
 array([[760, 155],
        [258, 529]], dtype=int64),
               precision    recall  f1-score      support
 0              0.746562  0.830601  0.786342   915.000000
 1              0.773392  0.672173  0.719239   787.000000
 accuracy       0.757344  0.757344  0.757344     0.757344
 macro avg      0.759977  0.751387  0.752791  1702.000000
 weighted avg   0.758968  0.757344  0.755314  1702.000000)
```

Figure 9: output of Random Forest Classifier

- **Precision for Outbreaks**: 77.3%
- **Recall for Outbreaks**: 67.2%
- **F1 Score for Outbreaks**: 71.9%
- **True Negatives (TN)**: 760
- **False Positives (FP)**: 155
- **False Negatives (FN)**: 258
- **True Positives (TP)**: 529

According to figure 9 findings, the model was more cautious when forecasting outbreaks, prioritizing accuracy over recall, which makes sense in situations where false alarms, that is, inaccurately predicting an outbreak are less expensive than failing to detect one. Based solely on temporal and spatial variables, the model's strong performance suggests that significant prediction signals for Lyme disease activity can be found in historical and geographic patterns. However, integrating sociolect-demographic and environmental parameters (such as temperature, land cover, and tick density) could still lead to increased accuracy.

Conclusion

This study demonstrates the practical application of data mining and machine learning techniques in the domain of public health surveillance, with a focus on predicting Lyme disease outbreaks at the county level in the United States. By utilizing a structured dataset of confirmed Lyme disease cases across multiple time periods, we successfully reshaped and preprocessed the data for analysis and predictive modeling. A Random Forest Classifier, trained on simple yet informative features such as temporal (StartYear) and geographic (StateCode) data, achieved a classification accuracy of 75.7%, with strong precision and balanced performance metrics.

The results underscore the potential of historical and spatial trends in identifying high-risk outbreak periods with minimal input variables. The model's effectiveness highlights the value of ensemble learning techniques in epidemiological forecasting, especially when working with tabular public health data. Importantly, this approach can serve as a foundation for developing real-time decision-support tools for public health officials and policymakers.

While the current model provides meaningful insights, future research should incorporate additional contextual features such as environmental conditions, population density, and tick activity levels to improve predictive accuracy and granularity. Furthermore, integrating spatial analysis and time-series forecasting methods may offer deeper insight into the dynamics of disease spread. Overall, this research contributes to the growing body of work that leverages artificial intelligence and data science for public health advancement.

References

1. Centers for Disease Control and Prevention (CDC). Lyme disease data & statistics [Internet]. 2023 [cited 2025 Nov 12]. Available from: https://www.cdc.gov/lyme/data-research/facts-stats/surveillance-data-1.html
2. Data.gov. Lyme disease county-level incidence data [Internet]. 2023 [cited 2025 Nov 12]. Available from: https://catalog.data.gov/dataset/lymedisease-9211-county
3. Jordan RA, Gable S, Egizi A. Relevance of spatial and temporal trends in nymphal tick density and infection prevalence for public health and surveillance practice in long-term endemic areas: a case study in Monmouth County, NJ. J Med Entomol. 2022;59(4):1451–66. https://doi.org/10.1093/jme/tjac073
4. Steere AC, Redel H, Blaser M. Lyme disease (Lyme borreliosis) due to Borrelia burgdorferi. In Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases, 9th Edition: Volume 1-2. 2019.
5. Wooten RM, Ma Y, Yoder RA, Brown JP, Weis JH, Zachary JF, Kirschning CJ, Weis JJ. Toll-like receptor 2 is required for innate, but not acquired, host defense to Borrelia burgdorferi. The Journal of Immunology. 2002 Jan 1;168(1):348-55.
6. Mead P. Epidemiology of Lyme disease. Infect Dis Clin North Am. 2022;36(3):495–521. https://doi.org/10.1016/j.idc.2022.03.004
7. Grimm D, Tilly K, Byram R, Stewart PE, Krum JG, Bueschel DM, et al. Outer-surface protein C of the Lyme disease spirochete: a protein induced in ticks for infection of mammals. Proc Natl Acad Sci U S A. 2004;101(9):3142–7. https://doi.org/10.1073/pnas.0306845101
8. Hook SA, Jeon S, Niesobecki SA, Hansen AP, Meek JI, Bjork JKH, et al. Economic burden of reported Lyme disease in high-incidence areas, United States, 2014–2016. Emerg Infect Dis. 2022;28(6):1165–74. https://doi.org/10.3201/eid2806.211335
9. Kugeler KJ, Scotty E, Hinckley AF, Hook SA, Nawrocki CC, Nikolai AM, Linz AM, Meece J, Schotthoefer AM. Epidemiology of Lyme disease as identified through electronic health records in a large midwestern health system, 2016–2019. In Open Forum Infectious Diseases. 2025 Feb;12(2): ofae758. Oxford University Press.
10. Penn Medicine. Lyme disease [Internet]. 2023 [cited 2025 Nov 12]. Available from: https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/lyme-disease
11. Bostic TD, Kugeler KJ, Hinckley AF. Pregnancy Among Reported Lyme Disease Cases—United States, 1992–2019. Zoonoses and Public Health. 2024 Dec;71(8):972-7.
12. Cartter M, Lynfield R, Feldman KA, Hook SA, Hinckley AF. Lyme disease surveillance in the United States: looking for ways to cut the Gordian knot. Zoonoses Public Health. 2018;65(3):227–9. https://doi.org/10.1111/zph.12448
13. Centers for Disease Control and Prevention (CDC). Lyme disease data & statistics [Internet]. 2023.
14. Bay Area Lyme Foundation. Lyme disease facts & statistics [Internet]. 2024 [cited 2025 Nov 12]. Available from: https://www.bayarealyme.org/about-lyme/lyme-disease-facts-statistics
15. Kilpatrick AM, Randolph SE. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. Lancet. 2012;380(9857):1946–55.
16. Statista. Lyme disease incidence by state [Internet]. 2022 [cited 2025 Nov 12]. Available from: https://www.statista.com/statistics/742936/incidence-rates-of-lyme-disease-cases-by-state/
17. Eisen L, Dolan MC. Evidence for personal protective measures… J Med Entomol. 2016;53(5):1063–92.

18. Shikamaru S. Lyme disease dataset [Internet]. 2022 [cited 2025 Nov 12]. Available from: https://www.kaggle.com/datasets/sshikamaru/lyme-disease-rashes
19. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance. Annual Review of Public Health. 2015 Mar 18;36(1):345–59.
20. Nelson CA, Saha S, Kugeler KJ, Delorey MJ, Shankar MB, Hinckley AF, Mead PS. Incidence of clinician-diagnosed Lyme disease, United States, 2005–2010. Emerg Infect Dis. 2015;21(9):1625–31.
21. Tahmid M. Lyme disease dataset [Internet]. 2024 [cited 2025 Nov 12]. Available from: https://www.kaggle.com/datasets/tahmidmir/lyme-disease-dataset
22. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324