

Development of a Machine Learning-Based Framework for Predicting the Prevalence and Early Diagnosis of Bubonic Plague

Original Research Article | Volume 1 | Issue 5 | 2026 | Article Number: 015

Accepted: 24 July 2026 | Published: 10 October 2026 | ISSN: 2979-8582 (Online)

 Crossref : <https://doi.org/10.67693/BJCR-CK6EPD5N>



Ekuewa J.B¹, Nwaekpe O.C¹, Owonibi S.K²

¹Department of Networking and Cloud Computing, Federal Polytechnic, Ede, Osun State, Nigeria

²Department of Environmental Biology, Federal Polytechnic, Ede, Osun State, Nigeria

 **NOC** [0009-0008-7842-2066](https://orcid.org/0009-0008-7842-2066), **OSK** [0000-0001-7910-1784](https://orcid.org/0000-0001-7910-1784)

Correspondence: Ekuewa Jacob Bamidele, Allbytesconsultz@gmail.com

Abstract

Bubonic plague remains a significant zoonotic disease despite advances in public health interventions and disease control programs. The disease continues to occur in several endemic regions and presents substantial challenges to healthcare systems due to delayed diagnosis and limited predictive surveillance capabilities. This study developed and evaluated machine learning models for predicting bubonic plague prevalence and facilitating early diagnosis using epidemiological, environmental, demographic, and clinical variables. A quantitative predictive research design was adopted utilizing an integrated dataset comprising 5,000 observations obtained from epidemiological repositories, environmental databases, and healthcare records. Four supervised machine learning algorithms, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN), were implemented and evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. Results revealed that Random Forest achieved the highest predictive performance with an accuracy of 92.7%, precision of 91.8%, recall of 92.3%, F1-score of 92.0%, and ROC-AUC of 0.964. Feature importance analysis identified swollen lymph nodes, fever, rodent exposure, flea exposure, and climatic variables as the most influential predictors. The findings demonstrate the effectiveness of machine learning in supporting disease surveillance, outbreak forecasting, and clinical decision-making. The study concludes that intelligent predictive systems can significantly strengthen public health preparedness and facilitate early intervention strategies for plague control.

Keywords: Bubonic Plague, Machine Learning, Disease Prediction, Random Forest, Artificial Intelligence, Disease Surveillance, Early Diagnosis

1. INTRODUCTION

Infectious diseases continue to constitute a major threat to global health security, accounting for substantial morbidity, mortality, and economic burden worldwide. The emergence and re-emergence of zoonotic diseases, coupled with increasing environmental changes and population mobility, have intensified the need for innovative surveillance and predictive healthcare systems [1]. Recent advances in artificial intelligence (AI) and machine learning (ML) have transformed healthcare analytics by enabling the extraction of actionable insights from large-scale, heterogeneous, and real-time datasets, thereby improving disease prediction, diagnosis, and outbreak preparedness [1, 2].

Bubonic plague, caused by the bacterium *Yersinia pestis*, remains one of the most historically significant zoonotic diseases and continues to persist in natural reservoirs across Africa, Asia, and the Americas. Although modern antibiotics have substantially reduced mortality rates, plague outbreaks continue to occur in endemic regions such as Madagascar and the Democratic Republic of Congo, posing ongoing public health concerns [3]. The disease is primarily transmitted through infected flea vectors associated with rodent hosts and may progress rapidly to septicemic or pneumonic plague if not diagnosed and treated promptly [3, 4].

Conventional disease surveillance systems largely depend on laboratory confirmation, epidemiological investigations, and retrospective reporting mechanisms. While these approaches provide valuable epidemiological information, they are often limited in their ability to provide early warning signals for outbreak prevention and proactive intervention [3]. Consequently, there has been growing interest in leveraging machine learning techniques to support intelligent disease surveillance, outbreak forecasting, and clinical decision-making.

Recent studies have demonstrated the effectiveness of machine learning in predicting infectious diseases including tuberculosis, influenza, antimicrobial resistance, dengue fever, Ebola, and COVID-19 [5, 6]. Machine learning algorithms such as Random Forest, Support Vector Machine, Gradient Boosting, and Artificial Neural Networks have shown remarkable capabilities in identifying complex nonlinear relationships within healthcare datasets and generating highly accurate predictive models [1].

Despite these advancements, relatively few studies have focused specifically on integrating epidemiological, environmental, demographic, and clinical variables into a unified machine-learning framework for predicting bubonic plague prevalence and facilitating early diagnosis. This gap underscores the need for predictive intelligence systems capable of supporting public health preparedness and timely intervention strategies. Therefore, this study develops a machine learning-based framework for predicting plague prevalence and enabling early diagnosis using multidimensional datasets.

2. LITERATURE REVIEW

The epidemiology of bubonic plague is influenced by complex interactions among environmental, ecological, climatic, demographic, and socioeconomic factors. Studies have demonstrated that temperature fluctuations, rainfall patterns, humidity levels, and rodent population dynamics significantly affect flea abundance and pathogen persistence, thereby influencing plague transmission cycles [3]. The persistence of *Yersinia pestis* within natural reservoirs highlights the need for surveillance systems capable of monitoring multiple interacting risk factors simultaneously.

Recent advances in artificial intelligence have expanded the application of machine learning technologies in healthcare analytics. Machine learning enables computational systems to learn patterns from historical data and generate predictions without explicit programming, making it particularly suitable for disease surveillance and outbreak forecasting [1]. Applications of machine learning in healthcare include disease

diagnosis, clinical decision support, medical imaging, personalized medicine, and epidemic prediction (Melchane et al., 2024).

Santangelo et al. (2023) conducted a systematic review of machine learning applications in infectious disease prediction and reported that machine learning models consistently achieved superior predictive performance compared with conventional statistical approaches. Their findings indicated that Random Forest, Gradient Boosting, and Deep Learning algorithms were particularly effective for modeling infectious disease dynamics and outbreak trends.

Similarly, Al Meslamani et al. (2024) reviewed the applications of machine learning in infectious disease management and reported substantial improvements in disease detection, risk prediction, and healthcare resource optimization. The study emphasized that machine learning algorithms can effectively analyze multidimensional healthcare datasets and identify latent relationships that may not be detectable through traditional analytical methods.

Furthermore, Lv and Wang (2024) conducted a comprehensive systematic review and meta-analysis of machine learning-based predictive models in healthcare and concluded that machine learning approaches consistently achieved high predictive accuracy across various disease domains. Their findings reinforced the growing importance of predictive analytics in modern healthcare systems.

Although significant progress has been made in infectious disease prediction, plague-related research remains largely focused on epidemiological characterization, ecological modeling, and historical outbreak analysis [3]. Existing studies provide limited evidence regarding the integration of environmental, epidemiological, demographic, and clinical variables within a unified machine learning framework specifically designed for plague prediction and diagnosis. Consequently, a significant research gap remains regarding the development of intelligent predictive systems capable of supporting early diagnosis and outbreak forecasting for bubonic plague.

The present study addresses this gap by developing and evaluating an integrated machine learning framework that combines multidimensional datasets to enhance plague surveillance, prevalence prediction, and early diagnostic decision-making.

3. MATERIALS AND METHODS

This study adopted a quantitative predictive research design utilizing machine learning techniques to develop predictive and diagnostic models for bubonic plague. Quantitative predictive designs are particularly suitable for studies involving numerical data analysis, pattern recognition, and predictive modeling [7].

Data were obtained from multiple sources including the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), published epidemiological studies, environmental monitoring databases, and geographic information systems repositories. Variables incorporated into the study included demographic characteristics, clinical symptoms, environmental conditions, rodent exposure indicators, and healthcare accessibility measures.

Data preprocessing involved missing value treatment, duplicate removal, outlier correction, feature encoding, normalization, and feature selection. According to [8], data preprocessing significantly improves model performance and predictive reliability.

Four supervised machine learning algorithms were evaluated:

- a. Logistic Regression
- b. Random Forest

- c. Support Vector Machine
- d. Artificial Neural Network

The dataset was partitioned into training (80%) and testing (20%) subsets. Ten-fold cross-validation was employed to improve model robustness and reduce overfitting.

4. RESULTS

The integrated dataset consisted of 5,000 observations, with 41% positive plague cases and 59% negative cases. Data preprocessing activities included treatment of 312 missing values, removal of 84 duplicate records, correction of 57 outliers, encoding of 18 variables, and normalization of 24 features.

Feature importance analysis revealed that swollen lymph nodes, fever, rodent exposure, flea exposure, and environmental conditions were the most influential predictors of infection status.

The performance of the evaluated algorithms is summarized below:

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	84.2	83.4	82.8	83.1	0.871
Random Forest	92.7	91.8	92.3	92.0	0.964
SVM	89.4	88.6	88.9	88.7	0.931
ANN	91.3	90.4	90.8	90.6	0.951

Random Forest emerged as the best-performing model across all evaluation metrics.

5. DISCUSSION

The superior performance of Random Forest confirms the effectiveness of ensemble learning approaches in modeling complex interactions among epidemiological, environmental, and clinical variables. The prominence of swollen lymph nodes and fever as major predictors aligns with established clinical manifestations of bubonic plague.

Environmental variables also contributed significantly to predictive accuracy, supporting previous findings that climatic conditions influence plague transmission through their effects on rodent reservoirs and flea vectors. The integration of machine learning into disease surveillance systems therefore offers substantial opportunities for improving outbreak preparedness and public health response.

6. CONCLUSION

This study successfully developed and evaluated a machine learning framework for predicting bubonic plague prevalence and facilitating early diagnosis. Among the evaluated algorithms, Random Forest achieved the highest predictive performance, demonstrating its suitability for disease surveillance applications.

The findings indicate that machine learning can effectively support public health decision-making by providing timely predictive intelligence and facilitating proactive interventions. Integrating environmental, epidemiological, demographic, and clinical variables within a unified analytical framework significantly enhances disease prediction capabilities.

Recommendations

- a. Public health agencies should integrate machine learning technologies into existing disease surveillance systems.

- b. Environmental monitoring variables should be routinely incorporated into outbreak forecasting frameworks.
- c. Healthcare institutions should deploy predictive decision-support systems to facilitate early diagnosis and intervention.
- d. Future studies should incorporate real-time data streams and advanced deep-learning architectures to further improve predictive performance.

References

- [1] Al Meslamani, A. Z., Sobrino, I., & de la Fuente, J. (2024). *Machine learning in infectious diseases: Potential applications and limitations*. *Annals of Medicine*, 56(1), 2362869. <https://doi.org/10.1080/07853890.2024.2362869>
- [2] Melchane, M. A., El Bakkali, H., El Mhamdi, J., & Beni-Hssane, A. (2024). *Artificial intelligence and machine learning applications in healthcare: Recent advances and future directions*. *Healthcare Analytics*, 5, 100312. <https://doi.org/10.1016/j.health.2024.100312>
- [3] Yang, R., Atkinson, S., Chen, Z., Cui, Y., Yang, X., Song, Y., Du, Z., & Wang, X. (2023). *Yersinia pestis and plague: Some knowns and unknowns*. *Zoonoses*, 3(1), 5. <https://doi.org/10.15212/ZOONOSSES-2022-0040>
- [4] Barbieri, R., Signoli, M., Chevé, D., Costedoat, C., Tzortzis, S., Aboudharam, G., & Drancourt, M. (2021). *Yersinia pestis: The natural history of plague*. *Clinical Microbiology Reviews*, 34(1), e00044-19. <https://doi.org/10.1128/CMR.00044-19>
- [5] Lv, G., & Wang, Y. (2024). *Machine learning-based antibiotic resistance prediction models: An updated systematic review and meta-analysis*. *Technology and Health Care*, 32(5), 2865–2882. <https://doi.org/10.3233/THC-240119>
- [6] Santangelo, O. E., Gentile, V., Pizzo, S., Provenzano, S., Giordano, D., & Firenze, A. (2024). *Machine learning and prediction of infectious diseases: A systematic review*. *AI*, 5(1), 235–257. <https://doi.org/10.3390/ai5010013>
- [7] Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.
- [8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-0716-1418-1>
- [9] Al-Turaiki, I., & Alballa, N. (2023). *Machine learning approaches in infectious disease prediction: A review*. *Journal of Healthcare Informatics Research*, 7(2), 115–134. <https://doi.org/10.1007/s41666-023-00132-5>
- [10] Carlson, C. J., Albery, G. F., Merow, C., Trisos, C. H., Zipfel, C. M., Eskew, E. A., Olival, K. J., Ross, N., & Bansal, S. (2022). *Climate change increases cross-species viral transmission risk*. *Nature*, 607(7919), 555–562. <https://doi.org/10.1038/s41586-022-04788-w>
- [11] Pechous, R. D., Sivaraman, V., Stasulli, N. M., & Goldman, W. E. (2016). *The plague pathogen Yersinia pestis and the pathogenesis of plague*. *Clinical Microbiology Reviews*, 29(3), 523–561. <https://doi.org/10.1128/CMR.00048-15>
- [12] Rajkomar, A., Dean, J., & Kohane, I. (2019). *Machine learning in medicine*. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMr1814259>

- [13] Topol, E.J.(2019). *High-performance medicine: The convergence of human and artificial intelligence*. Nature Medicine, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- [14] World Health Organization. (2023). *Plague fact sheet*. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/plague>



© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>)