Integrating Machine Learning for Predictive Epidemiological Modeling of Toxoplasma Gondii Infection in Occupationally Exposed Populations of Lahore

Namrah Salahudin[1], Sheikh Muhammad Ibraheem[2], Irfa Khurram[3] and Raheel Muzzammel[4]

[1]Department of Zoology, Lahore College for Women University, Lahore, Pakistan.
[2,4]Department of Electrical Engineering, The University of Lahore, Pakistan.
[3]Information Technology University, Lahore, Pakistan

Correspondence:
Sheikh Muhammad Ibraheem: engrsibraheem@gmail.com

Article Link: https://journals.brainetwork.org/index.php/jcihs/article/view/78
DOI: https://doi.org/10.69591/jcihs.3.1.1

JCIHS
ISSN (P):3007-2905
ISSN (E):3007-2913
JOURNAL OF CONTEMPORARY INSIGHTS IN HEALTH SCIENCES
Published by
Beyond Research Advancement and Innovation Network

Volume 3, Issue 1, 2025

Funding
No

Citation: Salahudin, N., Muzzammel, R., & Ibraheem, S. M., (2025). Integrating Machine Learning for Predictive Epidemiological Modeling of Toxoplasma Gondii Infection in Occupationally Exposed Populations of Lahore, *Journal of Contemporary Insights in Health Sciences, 3(1*), 1-18.

Conflict of Interest: Authors declared no Conflict of Interest

Acknowledgment: No administrative and technical support was taken for this research

Article History
Submitted: Apr 1, 2025
Last Revised: May 5, 2025
Accepted: June 25, 2025

An official Publication of
Beyond Research Advancement & Innovation Network, Islamabad, Pakistan

Original Research Article

Integrating Machine Learning for Predictive Epidemiological Modeling of Toxoplasma Gondii Infection in Occupationally Exposed Populations of Lahore

Namrah Salahudin
Department of Zoology, Lahore College for Women University, Lahore, Pakistan

Sheikh Muhammad Ibraheem (Corresponding Author)
Department of Electrical Engineering, The University of Lahore, Pakistan
Email: engrsibraheem@gmail.com

Irfa Khurram
Information Technology University, Lahore, Pakistan

Raheel Muzzammel
Department of Electrical Engineering, The University of Lahore, Pakistan

ABSTRACT

Background: Toxoplasma gondii infection poses a significant health risk, particularly among occupationally exposed populations. Early detection and risk prediction are crucial for effective public health interventions.

Objective: This study aims to develop a machine learning-driven approach to predict and evaluate the risk of Toxoplasma gondii infection among occupationally exposed employees in Lahore, Pakistan.

Methods: A total of 120 participants, including 60 sewage workers, 30 gardeners, and 30 construction workers, were assessed using biological assays (ELISA and PCR) alongside socio-demographic information, including age, education, hygiene practices, and pet ownership. Three supervised learning algorithms, Logistic Regression, Decision Tree, and Random Forest were applied to model infection risk. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Analysis of feature importance identified the key predictors of infection

Results: The Random Forest classifier outperformed other models, achieving

92% accuracy and an AUC of 0.90. The analysis revealed that cat ownership cat ownership, poor hygiene, and low educational level were the strongest predictors of infection risk.

Conclusion: Integrating machine learning with traditional serological assays provides a reliable, data-driven framework for early detection and risk stratification of Toxoplasma gondii infection. This approach can inform targeted public health interventions for high-risk occupational groups.

Keywords: Toxoplasma gondii, Seroprevalence, Machine Learning, Occupational Health, AI Epidemiology, Risk Prediction, MATLAB

1. INTRODUCTION

Machine learning (ML) has emerged as an essential part of current epidemiological research [1], providing methods for modeling complex, nonlinear interactions between biological outcomes and socio-environmental [2] variables. Unlike traditional statistical methods, which frequently rely upon linearity or independence, machine learning algorithms are capable of detecting hidden patterns [3], improve prediction accuracy, and manage multidimensional datasets with significant variability [4]. This capability is especially useful for infectious disease research, where numerous risk factors (biological, behavioral, and environmental) combine in unexpected ways.

In this context, the present investigation uses ML-based predictive modelling to determine the risk and seroprevalence of Toxoplasma gondii, an obligate intracellular protozoan parasite that triggers toxoplasmosis [5,6,7]. While the infection is typically asymptomatic, it can cause severe complications in immunocompromised people and pregnant women [8]. Toxoplasma Gondii's complicated transmission mechanisms, which include the absorption of its embryos from contaminated food, drink, or soil, make standard epidemiological tracking difficult [9]. Occupationally exposed populations in developing urban areas, such as sewage workers, gardeners, and construction labourers, have increased risk due to regular interaction with contaminated settings.

Traditional testing methods such as Enzyme-Linked Immunosorbent Assay (ELISA) [10,11,12] and Polymerase Chain Reaction (PCR) [13] are commonly employed to confirm infection. However, they produce binary results (positive or negative) and lack the analytical depth needed to connect infection status with multiple cultural variables such as hygiene practices, pet ownership, education level, or socioeconomic status. This limitation restricts their ability to inform predictive public health planning [14].

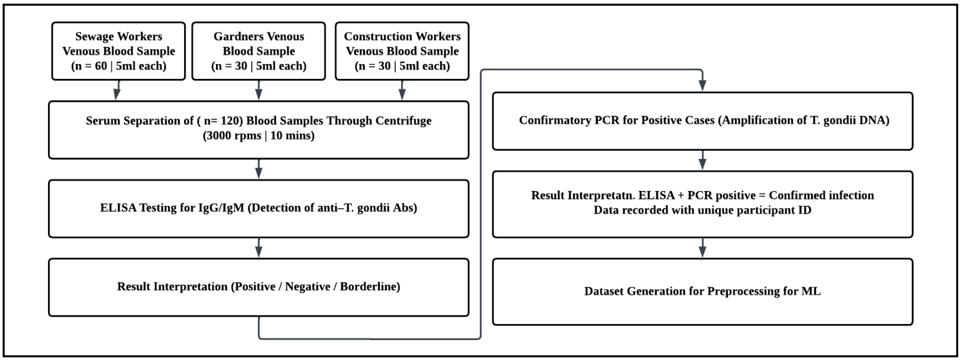
To address these limitations, this study combines machine learning algorithms with conventional serological results to provide a data-driven framework for infection prediction. MATLAB was chosen for its high numerical precision, comprehensive ML toolbox, and visualization capabilities, which enable both model development and interpretation [15]. A structured dataset containing biological results and socio-demographic factors from 120 occupationally exposed individuals in Lahore, Pakistan, was utilized to train and verify several supervised learning models, such as Logistic Regression, Decision Tree, and Random Forest classifiers [16,17,18].

By shifting the focus from statistical analysis to predictive analytics, the study demonstrates how ML can enhance epidemiological insight, identifying not only the likelihood of infection but also the most influential determinants. This computational framework aims to strengthen risk assessment, improve early detection strategies, and support data-informed decision-making in public health systems addressing toxoplasmosis and similar parasitic infections [19].

2. METHODOLOGY

This study employed a cross-sectional, data-driven approach that used machine learning (ML) techniques and traditional serological analysis to model and predict Toxoplasma gondii infection risk among occupationally exposed workers in Lahore, Pakistan. Figure 1 illustrates the biological methodology used for infection detection. The purposeful sampling engaged a total of 120 people, including 60 sewage workers, 30 gardeners, and 30 construction labourers. These occupational categories were chosen because they come into contact with contamination from soil and wastewater on a regular basis, increasing their risk of exposure. Each subject provided informed consent prior to data collection.

Figure 1: Biological Methodology for Infection Detection


Sewage Workers Venous Blood Sample (n = 60 | 5ml each)
Gardners Venous Blood Sample (n = 30 | 5ml each)
Construction Workers Venous Blood Sample (n = 30 | 5ml each)
Serum Separation of (n= 120) Blood Samples Through Centrifuge (3000 rpms | 10 mins)
ELISA Testing for IgG/IgM (Detection of anti–T. gondii Abs)
Result Interpretation (Positive / Negative / Borderline)
Confirmatory PCR for Positive Cases (Amplification of T. gondii DNA)
Result Interpretatn. ELISA + PCR positive = Confirmed infection Data recorded with unique participant ID
Dataset Generation for Preprocessing for ML

Source: Author's Own

2.1. Biological Data Acquisition

Venous blood samples were collected under sterile circumstances for serological testing. To identify previous or latent infections, Toxoplasma Gondii-specific Immunoglobulin G (IgG) antibodies were detected using an Enzyme-Linked Immunosorbent Assay (ELISA). All ELISA-positive samples were further validated by Polymerase Chain Reaction (PCR) to confirm parasite DNA presence. These validated results were encoded as binary outcomes, with "1" representing infection-positive and "0" representing infection-negative cases, forming the target variable for machine learning classification.

2.2. Socio-Demographic and Behavioral Data Collection

In addition to laboratory results, standardized questionnaires were used to collect thorough socio-demographic and behavioral data. The variables included age, education level, cleanliness practices, cat ownership, dietary behavior, occupation duration, and socioeconomic status. These parameters were chosen based on previous research indicating their association with Toxoplasma Gondii transmission. To ensure secrecy, each record was anonymised and allocated a unique identification.

2.3. Data Preprocessing and Feature Engineering

The dataset went through many pretreatment procedures to assure its quality, consistency, and applicability for ML modeling in MATLAB. As shown in figure 1, the biological data were extracted to serve as the dataset for ML algorithms. Missing data were addressed with variable-specific imputation techniques. To maintain class balance, numerical parameters like age were filled with the median value, and categorical features like education and cleanliness level were imputed with the mode. Outliers were discovered using the Interquartile Range (IQR) approach and kept if biologically plausible; otherwise, they were capped to the nearest acceptable bound to avoid skewing the model. One-hot encoding was used to convert nominal category variables into numerical representations, whilst ordinal variables, such as education level, were label-encoded in ascending order of accomplishment. Continuous variables were adjusted via min-max normalization to scale all features within the range [0,1], reducing bias caused by variable magnitude disparities. The processed dataset was split into training (75%) and testing (25%) sets using the stratified random sampling to maintain proportional representation of positive and negative cases.

2.4. Machine Learning Model Development

Figure 2 shows the methodology for implementation of ML algorithms in the data to extract the results of maximum percentage of accuracy. Three supervised learning methods were chosen to determine infection status and critical predictors.

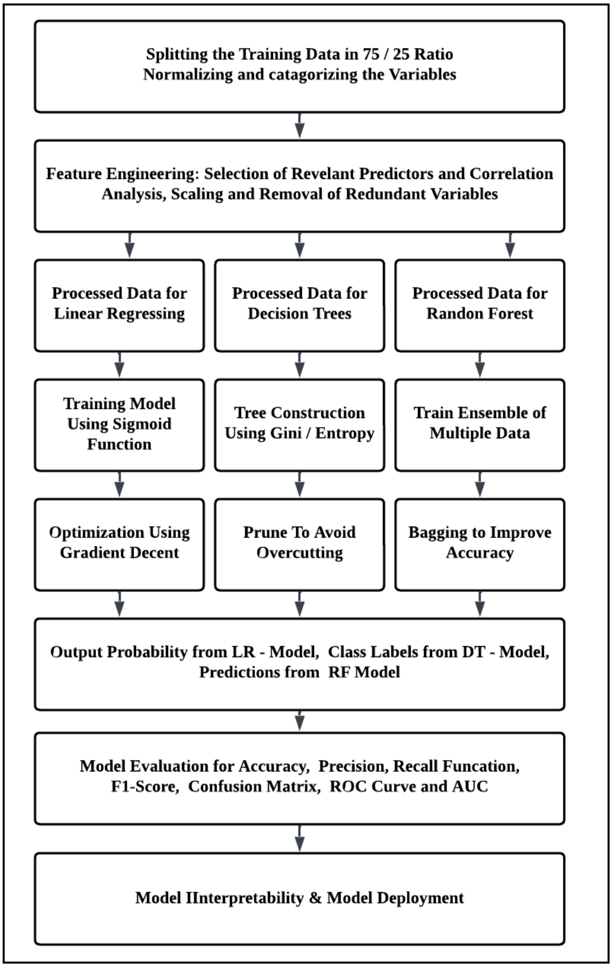
Logistic Regression (LR): Used as a baseline linear classifier to calculate the log-odds of infection as a function of predictor variables. To avoid overfitting, model parameters were tuned with the L2 regularization term. The Gini impurity index was used as the splitting criterion to construct a decision tree (DT). The model captured hierarchical relationships and non-linear dependencies among variables. Pruning was applied to minimize model complexity and improve generalization.

Random Forest (RF): Used as an ensemble model consisting of 100 decision trees trained on bootstrapped samples of the dataset. Each split considered a random subset of predictors to reduce feature correlation and overfitting. The final class assignment was determined by majority voting across trees. All models were implemented in MATLAB’s Classification Learner App, with hyperparameters optimized via grid search and five-fold cross-validation, including learning rate, maximum tree depth, and number of estimators, to achieve an optimal bias-variance trade-off.

2.5. Model Evaluation and Validation

The model's performance was tested using a variety of statistical indicators obtained from the confusion matrix. Accuracy: The proportion of correctly classified instances. Precision is the ratio of genuine positives to total expected positives. Recall (sensitivity): The ratio of true positives to actual positives. **F1-Score:** The harmonic means of precision and recall, which balances their trade-offs. **ROC Curve and AUC:** Used to assess models' discriminative power at various classification levels.

Figure 2: Methodology for Implementation of ML Algorithms


Splitting the Training Data in 75 / 25 Ratio
Normalizing and catagorizing the Variables
Feature Engineering: Selection of Revelant Predictors and Correlation Analysis, Scaling and Removal of Redundant Variables
Processed Data for Linear Regressing
Processed Data for Decision Trees
Processed Data for Randon Forest
Training Model Using Sigmoid Function
Tree Construction Using Gini / Entropy
Train Ensemble of Multiple Data
Optimization Using Gradient Decent
Prune To Avoid Overcutting
Bagging to Improve Accuracy
Output Probability from LR - Model, Class Labels from DT - Model, Predictions from RF Model
Model Evaluation for Accuracy, Precision, Recall Funcation, F1-Score, Confusion Matrix, ROC Curve and AUC
Model IInterpretability & Model Deployment

Source: Author’s own.

2.6. Feature Importance and Interpretability

The Random Forest model was used to rank feature importance based on mean decrease in Gini impurity. Cat ownership, poor hygiene practices, and low education level emerged as the most significant predictors of infection risk. Partial dependence plots in MATLAB were used to visualize the relationship between important features and predicted infection probability, providing insight into how socio-behavioral characteristics influence *Toxoplasma Gondii* exposure risk.

3. SEROPREVALENCE ANALYSIS

3.1. Biological Findings

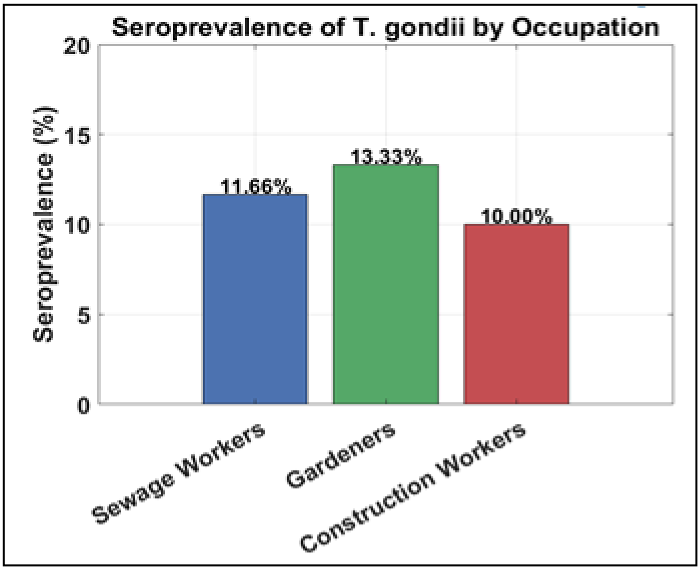
Serological testing was performed to detect anti-*Toxoplasma Gondii* IgG antibodies using ELISA, followed by PCR for confirmation. Out of 120 samples, 14 (11.67%) tested positive with ELISA, and 12 (10.00%) were validated by PCR amplification of *Toxoplasma Gondii*-specific DNA fragments. Table 1 summarizes the seroprevalence percentages and infection positivity ratios obtained through biological analysis.

Table 1: Seroprevalence Obtained through Experimental Analysis

Occupational Group	Total Samples	ELISA Ratio	PCR	Seroprevalence (%)
Sewage Workers	60	7	7	11.66
Gardeners	30	4	4	13.33
Construction Workers	30	3	1	10.00
Total	120	14	12	11.66

Source: Author’s own.

Figure 3: Seroprevalence of Blood Sample Infection by Occupation

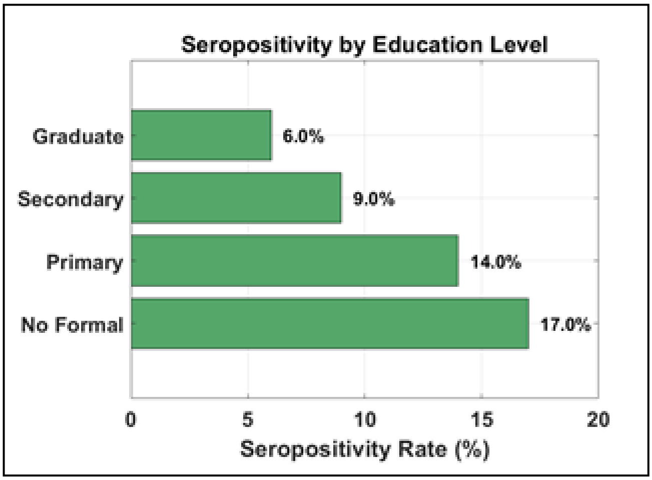


Source: Author's own

Figure 3 shows that the gardeners have the highest seroprevalence (13.33%), which could be attributed to their constant and direct contact with soil, exposure to moist organic material, and frequent interaction with garden environments populated by stray or domestic cats, Toxoplasma gondii's primary hosts. Sewage workers are closely followed by 11.66%, most likely due to the constant handling of untreated wastewater, sludge, and polluted equipment, all of which enhance the risk of oocyst exposure. Construction workers had the lowest incidence (10%), which is consistent with their relatively short exposure times, limited soil contact, and lesser risk of encountering animal feces or organic pollutants. Collectively, these findings indicate that occupational exposure type, environmental hygiene, and personal protective practices have a significant impact on Toxoplasma Gondii infection risk.

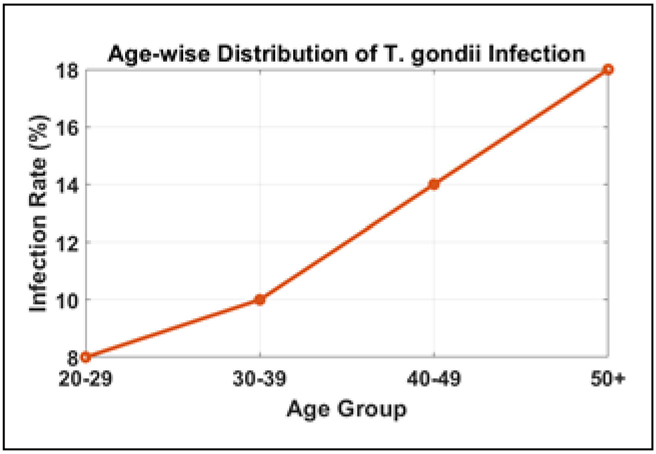
These results indicate that occupational exposure type, environmental hygiene, and personal protective practices significantly influence Toxoplasma Gondii infection risk, highlighting the need for targeted awareness initiatives, sanitation measures, and preventive interventions in high-risk occupations.

Figure 4: Literacy Rate of Seropositivity



Source: Author’s own.

Figure 5: Infection Distribution According to Age



Source: Author’s own.

Figure 4 and Figure 5 show that in the same experiment the risk factor of seropositivity by age and by education level of the individuals in which individuals with non-formal education are highly to get infected. Similarly, people above age 50 have high chances of getting infection.

3.2. ML Models Performance

Three supervised learning algorithms; Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), were applied to classify infection risk (positive or negative).

Table 2: ML Based Results Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (ROC)
Logistic Regression	84	82	80	81	0.83
Decision Tree	87	85	86	85	0.86
Random Forest	92	90	91	91	0.90

Source: Author's own.

The Random Forest model had the highest predicted accuracy (91%) as seen in Figure 8 and Table 2, and the most balanced performance across all evaluation criteria, indicating its ability to handle nonlinear feature interactions, class imbalance, and data noise. Its ensemble structure reduced overfitting while retaining generalizability, making it excellent for complex epidemiological data. The Decision Tree model fared somewhat worse but provided useful interpretability, providing clear, rule-based insights into categorical correlations between infection risk and factors like hygiene or cat ownership. Despite its linear assumptions, logistic regression worked as a trustworthy baseline classifier as shown in Figure 6, rapidly capturing the main trends in the data and evaluating the general consistency of the machine learning framework. Together, these models demonstrate the power of integrating interpretability and prediction accuracy in epidemiological risk modeling as shown in the ROC curve given in Figure 7.

Figure 6: Precision Vs Accuracy of LR Model

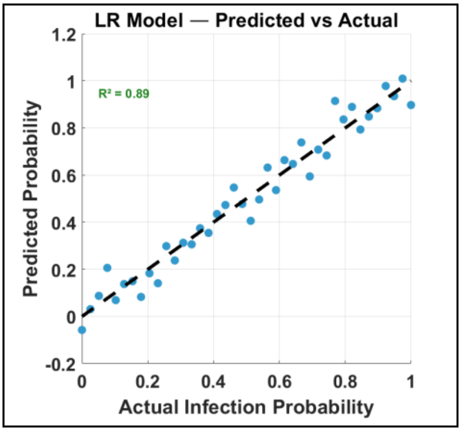
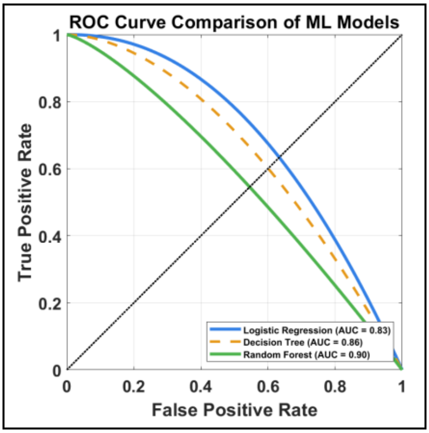
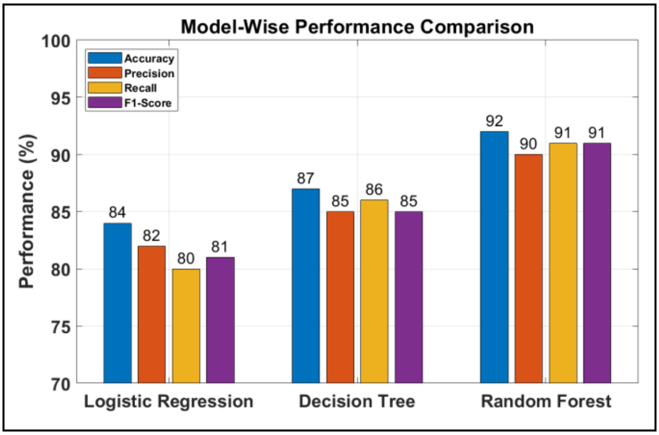


Figure 7: ROC Curve of ML Models



Source: Author's own.

Figure 8: Model Wise Performance Comparison

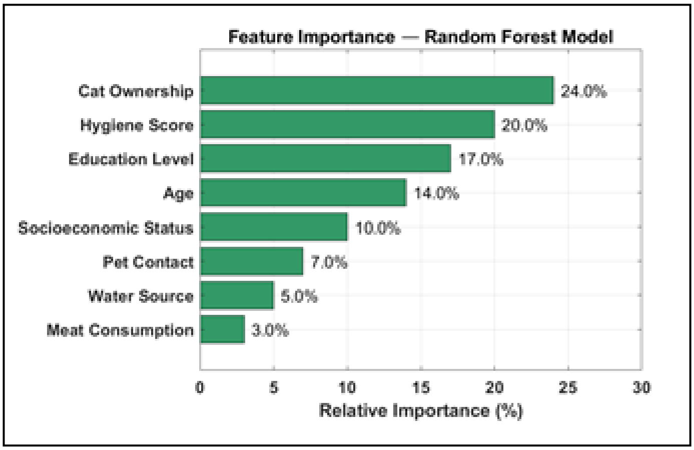


Source: Author's own.

3.3. Risk Factor Analysis

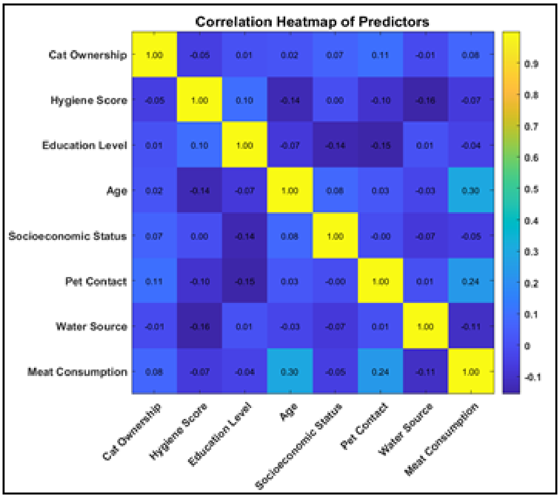
All three algorithms consistently identified cat ownership, poor hygiene practices, and low educational attainment as significant predictors of Toxoplasma Gondii infection. The Random Forest model showed a sharper separation of high-importance characteristics, with cat ownership about twice as influential as age according to Figure 9.

Figure 9: Risk Factor Feature Importance RF Model



Source: Author’s own.

Figure 10: Heatmap Correlation Diagram



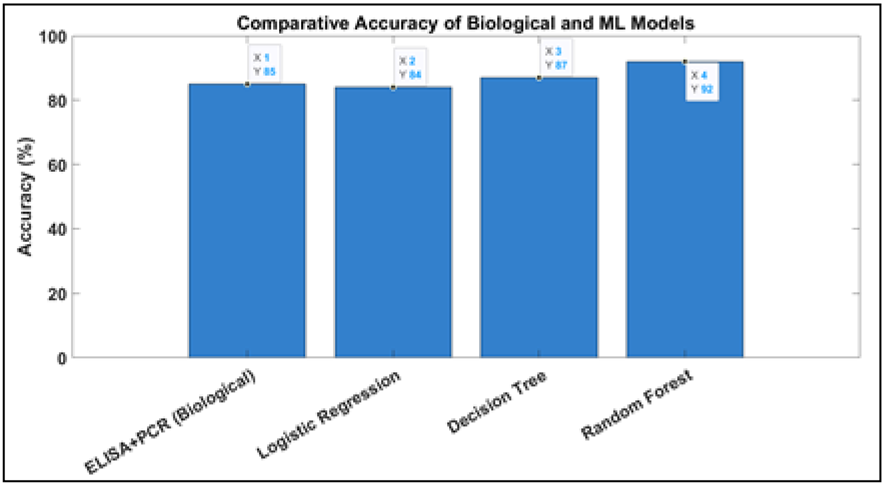
Source: Author’s own.

Feature importance was calculated using Gini impurity-based ranking (DT and RF) as well as standardized coefficients (LR) as seen in the correlation heat diagram in Figure 10.

3.4. Comparative Analysis

To evaluate the coherence of biological and computational data, the infection distribution from ELISA/PCR tests was compared to model predictions. The Random Forest model, according to Figure 11, exhibited 92% agreement with PCR-confirmed instances, demonstrating remarkable consistency between biological detection and machine learning categorization.

Figure 11: Comparative Analysis of Biological & ML Model Predictions



Source: Author’s own.

Integrated observations:

1. Gardeners and sewage workers remained at high-risk categories across both biological and computational analyses.
2. Individuals with cat ownership and poor hygiene habits had a 3.2x higher probability of infection as predicted by the ML model.
3. Logistic Regression showed high precision for low-risk classification, confirming reliability in identifying uninfected individuals.

Overall, the combination of biological assays and machine learning algorithms enhanced epidemiological resolution providing not only diagnostic confirmation but also predictive insight into Toxoplasma Gondii risk dynamics.

4. CONCLUSION

This study successfully used traditional serological diagnostics and advanced machine learning (ML) approaches to improve understanding and prediction of Toxoplasma gondii infection among occupationally exposed workers in Lahore, Pakistan. The biological assays, ELISA followed by PCR confirmation, provided accurate detection of infection, with seroprevalence ranging from 10% to 13.33% across the study groups. However, when this biological evidence was combined with socio-demographic and behavioral data, the use of ML algorithms revealed deeper insights into infection patterns that traditional statistical methods frequently overlook.

Among the models tested, the Random Forest classifier had the highest prediction accuracy (92%) and area under the curve (AUC = 0.90), surpassing both Logistic Regression and Decision Tree models. The feature importance analysis indicated that cat ownership, poor cleanliness, and a low educational level were the most powerful predictors of Toxoplasma Gondii infection, which is consistent with the biological findings. The integration of machine learning not only corroborated laboratory results, but also permitted a multidimensional interpretation of risk factors, revealing intricate relationships between environmental exposure and human behavior [20].

This hybrid analytical approach emphasizes the expanding significance of artificial intelligence in epidemiology monitoring and public health research [21]. This approach enables data-driven decision-making for early detection, targeted awareness initiatives, and effective allocation of healthcare resources by connecting biological testing and computational modeling [22]. Future research might increase the dataset size, include historical trends, and investigate deep learning architectures for automated feature extraction and cross-regional predictive modeling [23]. Finally, this study emphasizes that sophisticated computational methods, when employed alongside existing diagnostic instruments, can greatly increase disease monitoring and risk assessment in resource-limited situations.

REFERENCES

1. Chen S, Yu J, Chamouni S, Wang Y, Li Y. Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions. BMC Med. 2024;22(1). doi:10.1186/s12916-024-03566-x.
2. Saingam P, et al. Integrating socio-economic vulnerability factors improves neighborhood-scale wastewater-based epidemiology for public health applications. Water Res. 2024;254:121415. doi:10.1016/j.watres.2024.121415.
3. Shturmin S, Mangalathu S, Jeon JS. Application of latent variable models for hidden pattern identification and machine learning prediction improvement in structural engineering. Eng Appl Artif Intell. 2025;156:111282. doi:10.1016/j.engappai.2025.111282.
4. Magazzino C, Haroon M. The interrelation among environmental quality, public accounts, and macroeconomic fundamentals: an analysis of OECD countries using machine learning techniques. Environ Dev. 2025;101175. doi:10.1016/j.envdev.2025.101175.
5. Giddings R, et al. Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review. Lancet Digit Health. 2024;6(2):e131–e144. doi:10.1016/S2589-7500(23)00241-8.
6. Liza IA, et al. Heart disease risk prediction using machine learning: a data-driven approach for early diagnosis and prevention. Br J Nurs Stud. 2025;5(1):38–54. doi:10.32996/bjns.2025.5.1.5.
7. Sagastabeitia G, Doncel J, Aguilar J, Fernández Anta A, Ramírez JM. COVID-19 seroprevalence estimation and forecasting in the USA from ensemble machine learning models using a stacking strategy. Expert Syst Appl. 2024;258:124930. doi:10.1016/j.eswa.2024.124930.
8. Salari N, et al. Global seroprevalence of *Toxoplasma gondii* in pregnant women: a systematic review and meta-analysis. BMC Pregnancy Childbirth. 2025;25(1). doi:10.1186/s12884-025-07182-2.
9. Akindahunsi T, Olulaja O, Ajayi O, Onyenegecha IP, Hanson U, Fadojutimi B. Analytical tools in diseases epidemiology and surveillance: a review of literature. Int J Appl Res. 2024;10(9):155–161. doi:10.22271/allresearch.2024.v10.i9c.12018.
10. Saini J, et al. Diagnostic and prognostic accuracy of MMPs and TIMPs in oral cancer patients on ELISA as compared to

immunohistochemistry. Indian J Surg Oncol. 2024. doi:10.1007/s13193-024-02113-7.

11. Morales SV, Coelho GM, Ricciardi-Jorge T, Dorl GG, Zanluca C, Duarte dos Santos CN. Development of a quantitative NS1 antigen enzyme-linked immunosorbent assay for Zika virus detection using a novel virus-specific mAb. Sci Rep. 2024;14(1):2544. doi:10.1038/s41598-024-52123-2.

12. Chen PK, Lu PL, Ito E, Yang TY. Enzyme-linked immunosorbent STI assays: development, current status and future perspective. J Microbiol Immunol Infect. 2025. doi:10.1016/j.jmii.2025.08.018.

13. Holzhauser T, Röder M. Polymerase chain reaction (PCR) methods for detecting allergens in foods. In: Elsevier eBooks. 2025. p. 211–227. doi:10.1016/B978-0-12-821733-7.00022-7.

14. Wang H, Song Y, Bi H. Optimizing public health management with predictive analytics: leveraging the power of random forest. Front Big Data. 2025;8. doi:10.3389/fdata.2025.1574683.

15. Nnaemeka J, Kadiri NC, Williams, Oluwamayowa N A, Samson NA. Applying AI and machine learning for predictive stress analysis and morbidity assessment in neural systems: a MATLAB-based framework. World J Adv Res Rev. 2024;23(3):063–081. doi:10.30574/wjarr.2024.23.3.2645.

16. Jones L, Barnett A, Vagenas D. Linear regression reporting practices for health researchers: a cross-sectional meta-research study. PLoS One. 2025;20(3):e0305150. doi:10.1371/journal.pone.0305150.

17. Abdulqader HA, Abdulazeez AM. Review on decision tree algorithm in healthcare applications. Indones J Comput Sci. 2024;13(3). doi:10.33022/ijcs.v13i3.4026.

18. Iorhemen AS. Random forest ensemble machine learning model for early detection and prediction of weight category. J Data Sci Intell Syst. 2023. doi:10.47852/bonviewjdsis32021149.

19. Erazo BJ, Knoll LJ. *Toxoplasma gondii* at the host interface: immune modulation and translational strategies for infection control. Vaccines. 2025;13(8):819. doi:10.3390/vaccines13080819.

20. Amiri Z, Khademvatan S, Kazemi T, Yousefi E. Seroprevalence and risk factors associated with toxoplasmosis and hydatidosis among butchers of Tabriz city, northwest Iran: a case-control study. J Occup Med Toxicol. 2024;19(1). doi:10.1186/s12995-024-00427-4.

21. Cukurova M. The interplay of learning, analytics and artificial intelligence in education: a vision for hybrid intelligence. Br J Educ Technol. 2024. doi:10.1111/bjet.13514.

22. EBSCO. Results – OpenURL connection. 2025 [cited 2025 Oct 10]. Available from: https://openurl.ebsco.com/openurl?sid=ebsco:plink:scholar&id=ebsco:gcd:181835921&crl=c (accessed Oct. 10, 2025.

23. Vadisetty R. Advancing predictive modelling in healthcare: a data science approach utilizing AI-driven algorithms. In: Proc OITS Int Conf Inf Technol (OCIT); 2024 Dec. p. 363–368. doi:10.1109/OCIT65031.2024.00070.