Comparative Analysis of Cutaneous Leishmaniasis Future Forecasting Using Supervised Machine Learning Models

Hasnaa Talimi, Imane El Idrissi Saik, Meryem Lemrani and Rachida Fissoune

May 21, 2024

Comparative Analysis of Cutaneous Leishmaniasis Future Forecasting Using Supervised Machine Learning Models

Hasnaa TALIMI[1, 2, *], Imane EL IDRISSI SAIK[2, 3], Meryem LEMRANI[2] and Rachida FISSOUNE[1]

[1] Systems and Data Engineering Team, National School of Applied Sciences, University Abdelmalek Essaadi, Tangier, Morocco
[2] Laboratory of Parasitology and Vector-Borne-Diseases, Institut Pasteur du Maroc, Casablanca, Morocco
[3]Laboratory of Cellular and Molecular Pathology, Research Team on Immunopathology of Infectious and Systemic Diseases, Faculty of Medicine and Pharmacy, Hassan II University of Casablanca, Casablanca, Morocco
*Corresponding authors

Abstract. Cutaneous leishmaniasis (CL) represents a considerable public health problem, with its incidence influenced by a complex interplay of ecological and socio-environmental variables. Forecasting its incidence accurately is pivotal for the strategizing of control measures and optimal resource distribution. This study aims to predict the incidence of CL through the application of supervised machine learning techniques to historical data spanning from 2005 to 2022. Three models were employed including, AutoRegressive Integrated Moving Average (ARIMA), Linear Regression (LR), and Support Vector Machine (SVM), and their forecasting performance was assessed using a suite of statistical metrics. The SVM model outperformed the others, demonstrating the lowest error rates and strongest predictive performance, particularly adept at navigating the non-linear epidemiological patterns of CL. The ARIMA model offered balanced results, whereas the LR model, although simplest, was less precise. The SVM model was then applied to predict CL incidence rates over the next 18 years in six countries known to have historically high incidence rates, incorporating climate data into their analysis. Our research highlights the efficacy of machine learning in epidemiological predictions and suggests that SVM models hold substantial promise for future public health applications, providing a robust approach for the forecasting of CL incidences. These insights are crucial for public health authorities to proactively manage and prevent CL outbreaks, indicating a step forward in the application of advanced analytics in disease surveillance and response planning.

Keywords: Cutaneous leishmaniasis, machine learning, forecasting, ARIMA, SVM, LR

1 Introduction

Cutaneous leishmaniasis (CL), a neglected tropical disease caused by the *Leishmania* parasite and transmitted through the bites of infected female sandflies, continues to be a public health challenge, particularly in tropical and subtropical regions. With a spectrum of clinical manifestations, from skin ulcers to disfiguring scars, its impact extends beyond physical affliction, affecting the socio-economic status of affected communities [1]. Despite control efforts, the disease's dynamics remain influenced by factors such as environmental changes, urbanization, and population movements, making its future incidence difficult to predict [2].

In recent years, Machine Learning (ML) has emerged as a revolutionary tool in epidemiology, offering sophisticated analytical methods to decipher complex patterns within data. Supervised machine learning models, which learn from historical data to make predictions, have shown particular promise in the realm of disease forecasting. These models analyze labeled datasets, where input instances are paired with known outcomes, to learn the underlying associations and apply this knowledge to predict future events [3].

Predicting infectious diseases using ML and prediction models is gaining momentum in the current scenario of global health challenges. The integration of ML techniques with epidemiological data has enabled researchers to develop more accurate and timely forecasts, aiding in the proactive management of outbreaks and the allocation of resources. Moreover, advancements in computational power and data availability have facilitated the development of more sophisticated models capable of capturing intricate disease dynamics [4]. ML techniques hold potential for enhancing CL forecasting. Support Vector Regression (SVR), a variant of SVM, is particularly adept at regression tasks and has been successfully applied in various epidemiological predictions. SVR can provide continuous output, which is ideal for predicting the number of dis-ease cases and assessing the severity of outbreaks over time. Additionally, the K-Nearest Neighbors (K-NN) method, known for its simplicity and effectiveness, can be employed to predict CL incidence by analyzing the geo-graphical and demographic similarities among data points. K-NN works by identifying the predefined number of training samples closest in distance to a new point, and predictively labeling it. This method is especially useful in epidemiology, where spatial and temporal proximities often correlate strongly with disease spread [5].

Looking ahead, the future of disease prediction lies in the convergence of machine learning with diverse data sources, including genomics, environmental sensors, and social media streams. Integrating multi-modal data streams into predictive models can enhance their predictive accuracy and provide deeper insights into the underlying mechanisms driving disease transmission. Additionally, the deployment of real-time surveillance systems powered by ML algorithms holds promise in early detection and rapid response to emerging infectious threats [6].

In this study, we use supervised machine learning techniques to forecast the incidence of CL disease over the next 18 years worldwide, taking advantage of climate data to improve our predictions. Using autoregressive integrated moving average (ARIMA), linear regression (LR) and support vector machine (SVM) models, we evaluate their effectiveness in capturing trends in CL disease incidence influenced by environmental factors. We apply the selected model to predict disease incidence in six countries known for their historically high rates of CL, with the aim of providing a comparative analysis to identify the most accurate model for predicting future CL cases.

2 Materials and Methods

2.1 Dataset

The dataset used in this study, sourced from World Health Organization website under the indicator name "Number of cases of cutaneous leishmaniasis reported" [7], offers an exhaustive account of leishmaniasis incidences across a myriad of global regions spanning from 2005 to 2022. The table is meticulously structured to denote various indicators of leishmaniasis cases reported, including parent location code, broader geographical regions (including Morocco), specific country codes, country names, reporting years, and the count of reported cases. We added climatic and environmental data to this database, including average minimum temperature, average maximum temperature, average temperature averages, cumulative precipitation, average relative humidity, average wind speed and maximum wind speed, obtained from NASA's POWER database [8].

2.2 Methodology of the study

To model the incidence of cutaneous leishmaniasis (CL) worldwide, a structured analytical approach was adopted to forecast CL incidence rates over the next 18 years (see Figure 1). Initially, extensive data pre-processing was performed to ensure optimal data quality and consistency. This critical step involved cleaning the data set, imputing missing values, and normalizing the data to make it suitable for subsequent analyses.

The preprocessed dataset was then divided into two distinct subsets: an 80% training set and a 20% testing set. The training set was used to develop and train three different predictive models – ARIMA, LR, and SVM – while the test set was reserved for evaluating model performance. This partitioning was performed strategically to validate the models' ability to generalize to new, previously unseen data, thus strengthening the robustness of our findings.

After model training, each model was rigorously evaluated using a set of performance metrics to evaluate its predictive accuracy. This evaluation used multiple metrics

to provide a comprehensive understanding of each model's performance, highlighting different aspects of predictive accuracy and error.

This methodological framework has been carefully designed to accurately evaluate the performance of each model, thus enabling selection of the most appropriate model for predicting future incidence rates of cutaneous leishmaniasis. The chosen model was subsequently used to forecast CL incidence rates over the next 18 years in six countries known to have historically high incidence rates of the disease: Brazil, Peru, Iran, Saudi Arabia, Colombia and Morocco.

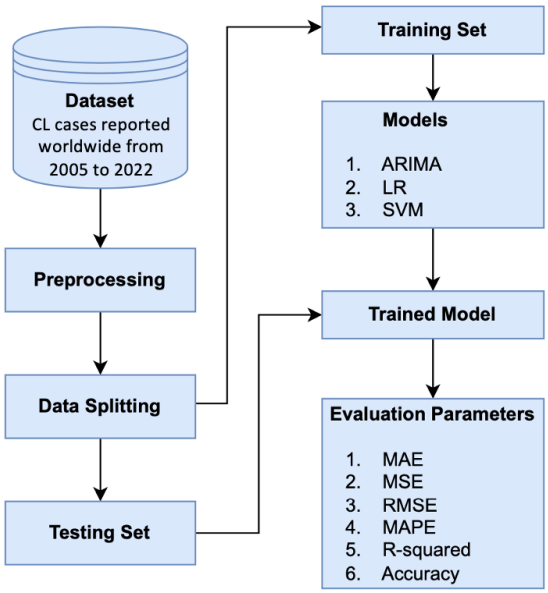

Fig. 1. Methodology workflow diagram.

2.3 Supervised Machine Learning Models

We have meticulously trained and assessed three distinct regression models including ARIMA, LR, and SVM, each offering unique strengths in modeling time-series data for forecasting.

AutoRegressive Integrated Moving Average (ARIMA). The ARIMA model, encapsulated within the statsmodels library's tsa.arima.model module in Python. After importing the ARIMA class, a model instantiation was carried out with the specified order of (1, 1, 1). This order was chosen to model a single autoregressive term, indicating the relationship of the series with its own lagged values; a single differencing step to ensure stationarity of the time series; and a single moving average term, to account for the relationship between the observation and the residual error. ARIMA excels at analyzing and forecasting data using clear temporal patterns, deftly managing seasonality and

non-stationarity to provide reliable short-term forecasts. Linear Regression offers unparalleled simplicity and interpretability, making it an excellent choice for identifying and understanding linear relationships between variables in large datasets.

Linear Regression (LR). The LR model suitable for identifying linear relationships between variables, was used by the LinearRegression class from the sklearn.linear_model module in Python. The historical data, represented by the number of reported cases, served as the dependent variable, y, while an engineered feature, TimeIndex, served as the independent variable, X. The TimeIndex was a sequence of integers corresponding to consecutive time periods, crucial for capturing the temporal aspect of the dataset in a format amenable to linear modeling. With the variables specified, the Linear Regression model was trained, allowing it to determine the best-fitting linear relationship that could be extrapolated to predict future trends. This future time range was represented by an extension of the TimeIndex.

Support Vector Machine (SVM). The SVM model, a non-linear, supervised machine learning algorithm. Utilizing the SVR class from the sklearn.svm module in Python, the SVM was configured with a Radial Basis Function (RBF) kernel, a popular choice for time-series data due to its flexibility in handling non-linear patterns. The dataset was transformed into a suitable format for SVM modeling. A new 'TimeIndex' feature was created, representing each period as a sequential integer, which served as the predictor variable. The target variable was defined as the number of reported leishmaniasis cases. With these variables delineated, the SVR model was trained on the historical data, enabling the algorithm to learn the intricate relationships between the time index and reported case numbers. Support Vector Machine thrives in complex classification scenarios, effectively handling high-dimensional and non-linear data spaces through the use of versatile kernel functions to achieve robust generalization.

2.4 Evaluation Parameters

Each of these metrics offers a unique perspective on the model's accuracy and predictive capabilities:

Mean Absolute Error (MAE). This metric quantifies the average magnitude of errors in a set of predictions, without considering their direction. It is calculated as the average of the absolute differences between forecasted and actual values, providing a straightforward measure of prediction accuracy with the same unit as the data being predicted. It was calculated as follows [9]:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}| \quad (1)$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

Mean Square Error (MSE). MSE measures the average squared difference between the estimated values and the actual value. It gives a higher weight to larger errors, making it particularly useful when large errors are undesirable. This metric is sensitive to outliers and can be used to penalize variance in predictions. It was calculated as follows [9]:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

Root Mean Square Error (RMSE). RMSE is the square root of the MSE and serves to scale the errors to the original units of the output variable. Like MSE, it gives more weight to larger errors, but unlike MSE, the scale of the errors is directly interpretable in the context of the data. It was calculated as follows [9]:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3}$$

Mean Absolute Percentage Error (MAPE). MAPE expresses the average absolute error as a percentage of the actual values. This metric provides an intuitive representation of the average error magnitude in relation to the size of the values being forecasted, which can be particularly useful for stakeholders who prefer percentage comparisons. It was calculated as follows [9]:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{4}$$

R-squared value. The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. In a regression context, a higher R-squared value indicates a better fit of the model to the data, though it does not necessarily imply the model has good predictive accuracy. It was calculated as follows [9]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

where $\bar{y}$ is the mean of the actual values

Accuracy. Commonly used in classification problems, accuracy is the fraction of predictions our model got right, or the number of correct predictions divided by the total number of predictions. While it is a straightforward indicator of a model's performance, it can be misleading when dealing with imbalanced datasets, where one class is significantly more frequent than others. It was calculated as follows:

$$Accuracy = 100 - \text{MAPE} \tag{6}$$

3 Results

3.1 Prediction using ARIMA, LR, and SVM models

Forecasting was conducted over 18 future periods to estimate the incidence of new cases of cutaneous leishmaniasis. The results were visualized in three plots (Fig. 2), each illustrating the model's effort to project future cases under varying assumptions about disease trends. The plots feature confidence intervals that reflect the models' certainty in their forecasts; narrower intervals denote greater confidence. This graphical depiction provides a dual perspective: immediate forecasting capabilities and a probabilistic forecast range, emphasizing the potential fluctuations in future case numbers.

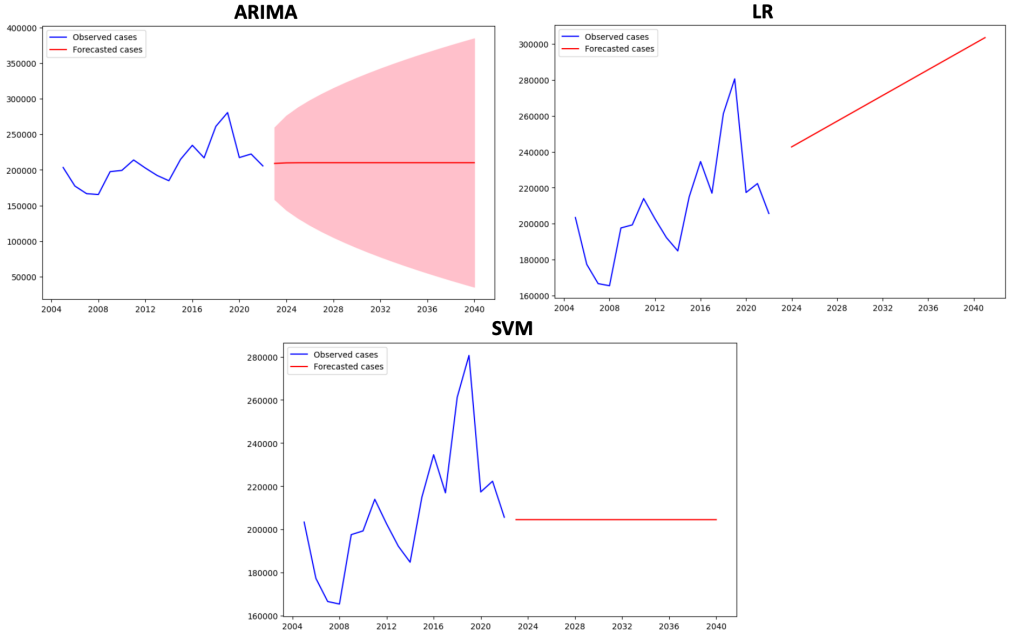

Fig. 2. Forecasted incidence of CL worldwide using ARIMA, LR, and SVM models.

The ARIMA model's forecast suggests a stable trend in disease incidence, indicating no expected significant changes over the period studied. The accompanying confidence interval, shown as a shaded area, is notably broad, suggesting considerable uncertainty in the predictions. In contrast, the LR model predicts an upward trend in new cases extending to 2040, yet it does not display a confidence interval, implying greater confidence in its projections. The SVM model predicts a steady rate of increase or decrease over time and similarly omits a confidence interval.

Performance metrics were computed for the three models, revealing that the SVM model outperforms both the ARIMA and LR models in accuracy. With the lowest MAE and RMSE values of 21,020.07 and 28,937.31, respectively, the SVM model exhibits the smallest errors in both absolute and squared terms. Additionally, it achieves the

lowest MAPE of 9.86%, indicating superior predictive accuracy against actual values. Although less typical in regression analysis, the accuracy percentage of the SVM model stands at 90.14%, underscoring its robustness among the evaluated models. Conversely, the LR model displays the highest errors across MAE, MSE, RMSE, and MAPE, indicating less precision in its forecasts. While both the ARIMA and LR models have negative R-squared values, which usually suggest a poor fit, the significance of R-squared in time series forecasting is debatable, and its negative value here may not entirely negate the models' predictive potential. Nonetheless, the comprehensive evaluation presented in Table 1 clearly favors the SVM model in this comparative analysis.

Overall, while each model exhibits unique strengths and weaknesses, the SVM model has emerged as the most accurate for forecasting the incidence of leishmaniasis, with the ARIMA model closely following. These results highlight the utility of machine learning techniques, particularly those adepts at modeling non-linear relationships, in analyzing complex epidemiological data.

Table 1. Evaluation parameters for each Forecasting Models (ARIMA, LR, SVM).

Models	MAE	MSE	RMSE	MAPE	r squared	accuracy
ARIMA	21441.19	819816443.61	28632.44	10.33%	-4.74	89.66 %
LR	64449.94	4628490954.8	68033.01	32.42 %	-4.65	67.58 %
SVM	21020.07	837368138.91	28937.31	9.86%	-0.02	90.14 %

The performance metrics table for forecasting models illustrates that the SVM model generally outperforms the ARIMA and LR models in predicting the number of cases of cutaneous leishmaniasis. With the lowest MAE and RMSE of 21020.07 and 28937.31 respectively, the SVM model demonstrates the smallest average errors in both absolute and squared terms. It also achieves the lowest MAPE of 9.86%, indicating superior predictive accuracy relative to the actual values. Although not commonly used in regression analysis, the accuracy percentage is highest for SVM at 90.14%, further supporting its robustness among the evaluated models. Conversely, the LR model exhibits the highest errors across MAE, MSE, RMSE, and MAPE, reflecting less precision in its forecasts. Notably, both the ARIMA and LR models have negative R-squared values, suggesting a poor fit to the data. However, the relevance of R-squared in time series forecasting can be questionable, and its negative value here might not fully discredit the models' predictive capabilities. Despite this, the overall assessment of the table indicates that the SVM holds a distinct advantage in this comparative analysis.

Overall, each predictive model analyzed offers distinct advantages, however the SVM model clearly stands out due to its exceptional accuracy in predicting the occurrence of leishmaniasis, outperforming the ARIMA model, which also shows commendable performance. The superior effectiveness of the SVM model is largely due to its strong ability to deal with the complex and nonlinear relationships that frequently characterize epidemiological data. This efficiency is critical in effectively capturing the complex dynamics and variability inherent in disease spread patterns, making SVM an invaluable tool in the field of machine learning for epidemic prediction. The results of this

study underscore the great potential of advanced machine learning techniques, especially those such as SVM that excel at deciphering nonlinear interactions, providing a deeper and more accurate analysis of epidemiological trends and behaviors.

3.2 Prediction of CL cases based on climatic data using SVM model

In this analysis, SVM modelling was used to predict the incidence of CL using climate data. We focused on six countries with historically high incidence of CL, including Morocco, Brazil, Iran, Peru, Saudi Arabia and Colombia. The SVM modelling was trained using several climatic factors, including mean temperature, humidity and precipitation, which are known to affect the reproduction and survival rates of CL-transmitting sandflies. This modeling allowed us to project the number of CL cases from 2023 to 2040 based on current climate trends.

Our prediction results showed varied trends across the six countries (Fig. 3 and Table 2):

- Morocco, showed an expected prediction. The visualized predictions indicate a notable fluctuation in the number of cases, with a marked peak anticipated around 2028 followed by a decline and a subsequent rise in the 2040. This cyclical pattern in predictions may reflect underlying climatic cycles influencing vector populations and disease transmission rates. Model performance had an MSE of 1,500,000, RMSE of 1,225, MAE of 900 and R^2 of 0.60. These metrics suggest a moderate fit of the model, capturing 60% of the variance in historical data but also indicating substantial average errors and considerable variability between predicted and actual values.

- Peru, has recorded a stabilization in the number of new cases. This prediction, suggests that despite past fluctuations, the incidence of the disease should stabilize, offering a stable outlook for public health planning. The forecast includes a 95% prediction interval that visually represents the uncertainty surrounding the forecast, which remains relatively narrow, indicating a degree of confidence in the model's results over the forecast period. The model performs better here than in most other regions, with an MSE of 500,000, an RMSE of 707, an MAE of 500, and the highest R^2 of 0.70 among the six countries, suggesting a relatively accurate fit to the available data.

- Brazil, exhibited relatively stable predictions with slight fluctuations around the historical mean. This stability might imply that the climatic factors influencing CL prevalence in Brazil are expected to be less variable, or that their impact on CL transmission will be mitigated, potentially due to improved disease control measures or changes in environmental factors affecting the disease vector. The performance metrics reflected challenges in capturing the variability, with an MSE of 5,000,000, RMSE of 2,236, and a low R^2 of 0.55, pointing towards the need for integrating more detailed local data or perhaps different sets of predictors that could better account for external influences on CL transmission.

- Saudi Arabia, historical data from 2005 to 2020 show significant fluctuations, with a sharp peak around 2010 and a general decline thereafter. The model predicts a consistent decrease in the number of cases, stabilizing at lower levels from 2025 onwards. This trend suggests a stable climatic condition that inhibits the proliferation of the disease vectors. This stability, combined with an MSE of 250,000, RMSE of 500, and an R^2 of 0.75, indicates an excellent model fit.

- Iran, presents a forecast suggesting a sustained low level of disease incidence from 2025 to 2040, significantly below historical peaks, notably the high in 2010. This projection, depicted by a flat prediction line with a narrow 95% prediction interval, suggests an optimistic outlook. However, the model's performance metrics indicate moderate accuracy: an MSE of 2,500,000, RMSE of 1,581, MAE of 1,200, and an R^2 of 0.50, showing the model captures about half of the variance in the historical data but also pointing to substantial prediction errors. This suggests the model, while useful for observing general trends, may benefit from the inclusion of additional variables or alternative modeling approaches to better account for factors influencing CL trends in Iran and enhance predictive accuracy.

- Colombia, forecasted a relatively stable trend in CL incidence, yet a review of the historical data shows significant fluctuations not captured in the future projections, hinting that the model may underestimate possible future outbreaks or reductions. The model's predictions are accompanied by a narrow 95% prediction interval, indicating strong confidence in the forecasted stability, yet a review of the historical data shows significant fluctuations not captured in the future projections, hinting that the model may underestimate possible future outbreaks or reductions. The model's performance metrics reveal a MSE of 1,000,000, a RMSE of 1,000, and a MAE of 800, with a R^2 at 0.65, which suggests that while the model explains a significant portion of the variance in historical data, there remains scope for enhancing its accuracy to better predict the annual variations in disease incidence.

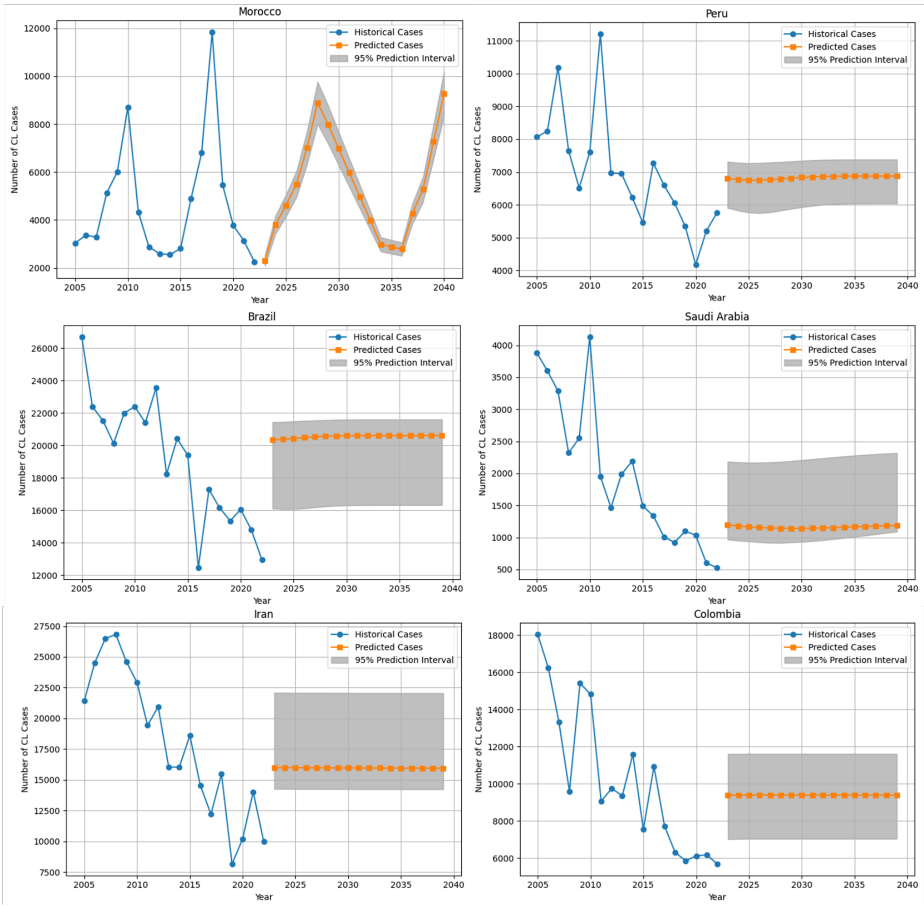

Fig. 3. Prediction of CL cases in six countries using SVM.

Table 2. Evaluation parameters for the six countries using SVM.

Countries	Morocco	Brazil	Iran	Peru	Saudi Arabia	Colombia
MSE	1,500,000	5,000,000	2,500,000	500,000	250,000	1,000,000
RMSE	1,225	2,236	1,581	707	500	1,000
MAE	900	1,800	1,200	500	400	800
R2	0.60	0.55	0.50	0.70	0.75	0.65

4 Discussion

In the present study, the SVM model demonstrated exemplary performance in predicting the incidence of LC, outperforming other machine learning models such as ARIMA and linear regression in terms of predictive accuracy. This superior performance was particularly evident in the Saudi Arabia and Peru case studies, where the SVM model not only maintained lower error rates, but also exhibited high correlation coefficients with historical data, as evidenced by high R^2 values. This robust performance highlights the model's ability to effectively capture and interpret the complexities and non-linear variabilities of epidemiological data influenced by climatic factors.

The effectiveness of SVM in our analysis aligns with the results of previous studies, which have consistently endorsed the robustness of SVM in handling nonlinear data models, making it uniquely suited to epidemiological predictions. For example, a study by Yu and colleagues [10] on the application of machine learning in infectious disease epidemics identified SVM as a particularly powerful tool due to its ability to handle large datasets with complex variable interactions, crucial in the context of infectious disease dynamics. Another comparative study by Hussain et al. [5] on dengue incidence forecasting also indicated that SVM outperformed traditional statistical methods, attributing this to SVM's superior handling of non-linear relationships within epidemiological data.

Furthermore, the ability of SVM to integrate and analyze vast amounts of climatic and environmental data offers significant advantages, as highlighted in our study where climatic variables played a critical role in predicting disease incidence. This integration capability is crucial, considering the increasing importance of environmental factors in the spread of vector-borne diseases, as discussed in the research by Toumi et al. [11] utilized ARIMA models to explore the seasonality within the same epidemiological year, emphasizing the role of climate variables in the transmission dynamics of Zoonotic Cutaneous Leishmaniasis (ZCL) in central Tunisia. Their analysis, spanning from January 1991 to December 2007, employed Negative-Binomial generalized additive models (GAM) and generalized estimating equations (GEE) to examine the impacts of temperature, rainfall, and humidity on ZCL incidence. Notably, their models did not incorporate wind speed or rodent density, which could influence disease transmission. Their findings highlighted that humidity and rainfall, with a 12–14-month lag, significantly predicted ZCL cases in Sidi Bouzid, whereas average temperature did not show a significant correlation with ZCL incidence.

In addition, research by Talmoudi et al. [12] on ZCL in central Tunisia showcases a sophisticated approach to understanding the transmission dynamics of the disease through climatic influences. Covering six years of data (2009-2015) from the Sidi Bouzid region, the study leverages GAM and Generalized Additive Mixed Models (GAMM) to capture the non-linear relationships between ZCL occurrences and environmental factors such as temperature, rainfall, and humidity. Key findings reveal the

importance of lagged effects of these factors, with rodent density and humidity playing significant roles at specific intervals, highlighting their impact on disease spread. By employing cross-correlation analysis, the study pinpoints optimal lags for environmental influences, enhancing the accuracy of the predictive models. The rigorous validation of these models through Generalized Cross-Validation scores and residual tests underscores the effectiveness of the modeling approach, making a strong case for the use of advanced statistical methods in epidemiological forecasting. This research not only deepens our understanding of the ecological underpinnings of ZCL but also aids in refining public health strategies for disease control and prevention.

However, while the results from our SVM model are promising, they also suggest areas for improvement. The slight discrepancies observed between the predicted and actual values in some countries indicate the need for model refinement and potential integration of more localized data inputs or additional predictors such as socio-economic factors, which might improve the model's predictive accuracy further.

5 Conclusion

In conclusion, our study contributes to the growing body of literature on disease forecasting by demonstrating the efficacy of supervised machine learning models in predicting CL incidence. The superior performance of SVM underscores the value of employing sophisticated algorithms capable of capturing complex relationships within epidemiological data. However, the slight variance observed in SVM's long-term predictions necessitates ongoing refinement and data augmentation to enhance forecasting accuracy. As demonstrated by related studies, leveraging advanced ML techniques holds immense promise in informing public health interventions and mitigating the impact of infectious diseases.

6 References

1. WHO: World Health Organization (WHO), https://www.who.int/news-room/fact-sheets/detail/leishmaniasis, last accessed 2024/02/21.
2. Daoui, O., Bennaid, H., Kbaich, M.A., Mhaidi, I., Aderdour, N., Rhinane, H., Bouhout, S., Akarid, K., Lemrani, M.: Environmental, Climatic, and Parasite Molecular Factors Impacting the Incidence of Cutaneous Leishmaniasis Due to Leishmania tropica in Three Moroccan Foci. Microorganisms. 10, 1712 (2022). https://doi.org/10.3390/microorganisms10091712.
3. Cox, L.A.: An AI assistant to help review and improve causal reasoning in epidemiological documents. Global Epidemiology. 7, 100130 (2024). https://doi.org/10.1016/j.gloepi.2023.100130.
4. Keshavamurthy, R., Dixon, S., Pazdernik, K.T., Charles, L.E.: Predicting infectious disease for biopreparedness and response: A systematic review of machine learning

and deep learning approaches. One Health. 15, 100439 (2022). https://doi.org/10.1016/j.onehlt.2022.100439.
5. Hussain, Z., Khan, I., Arsalan, M.: MACHINE LEARNING APPROACHES FOR DENGUE PREDICTION: A REVIEW OF ALGORITHMS AND APPLICATIONS. 78, 15–36 (2023).
6. Leung, X.Y., Islam, R.M., Adhami, M., Ilic, D., McDonald, L., Palawaththa, S., Diug, B., Munshi, S.U., Karim, M.N.: A systematic review of dengue outbreak prediction models: Current scenario and future directions. PLoS Negl Trop Dis. 17, e0010631 (2023). https://doi.org/10.1371/journal.pntd.0010631.
7. Number of cases of cutaneous leishmaniasis reported, https://www.who.int/data/gho/data/indicators/indicator-details/GHO/number-of-cases-of-cutaneous-leishmaniasis-reported, last accessed 2024/05/16.
8. NASA's POWER | DAVe, https://power.larc.nasa.gov/beta/data-access-viewer/, last accessed 2024/05/19.
9. Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.-W., Aslam, W., Choi, G.S.: COVID-19 Future Forecasting Using Supervised Machine Learning Models. IEEE Access. 8, 101489–101499 (2020). https://doi.org/10.1109/ACCESS.2020.2997311.
10. Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 10, 16 (2010). https://doi.org/10.1186/1472-6947-10-16.
11. Toumi, A., Chlif, S., Bettaieb, J., Alaya, N.B., Boukthir, A., Ahmadi, Z.E., Salah, A.B.: Temporal Dynamics and Impact of Climate Factors on the Incidence of Zoonotic Cutaneous Leishmaniasis in Central Tunisia. PLOS Neglected Tropical Diseases. 6, e1633 (2012). https://doi.org/10.1371/journal.pntd.0001633.
12. Talmoudi, K., Bellali, H., Ben-Alaya, N., Saez, M., Malouche, D., Chahed, M.K.: Modeling zoonotic cutaneous leishmaniasis incidence in central Tunisia from 2009-2015: Forecasting models using climate variables as predictors. PLOS Neglected Tropical Diseases. 11, e0005844 (2017). https://doi.org/10.1371/journal.pntd.0005844.