Predicting Leishmaniasis Risk in Morocco Using Machine Learning, GIS, and Domain Adaptation : A case study of Beni Mellal-Khenifra Region

Saad Farah [a*ID], Abderrazzak Elbidouri [b], Achraf Boutejdir [b], Nadia Dagdague [b], Amine Rouhi [b], Mohamed Maanan [c,ID], Hassan Rhinane [b]

[a] LaGeS-SGEO Laboratory, Hassania School of Public Works, Casablanca, Morocco, Email: farah.saad.cedoc@ehtp.ac.ma (SF)
[b] Geoscience Laboratory, Hassan II University, Casablanca, Morocco, Email: h.rhinane@gmail.com (HR)
[c] UMR 6554 CNRS LETG-Nantes Laboratory, Institute of Geography and Planning, Nantes University,44312 Nantes, France, Email: Mohamed.Maanan@univ-nantes.fr (MM)
*Corresponding author, email: farah.saad.cedoc@ehtp.ac.ma , phone number: +212689479013

Keywords: Leishmaniasis, Machine Learning, Geographic Information System (GIS), Risk Mapping, Domain Adaptation, Spatial Prediction.

Abstract

Leishmaniasis remains a persistent global public health challenge, particularly in regions where ecological and socioeconomic conditions favor vector proliferation and disease transmission. In Morocco, the provinces of Beni Mellal and Khenifra are among the most severely affected, necessitating the use of advanced spatial prediction tools to guide effective disease control strategies. This study integrated machine learning techniques and Geographic Information System (GIS) technologies to develop a predictive framework for cutaneous leishmaniasis risk mapping. A spatial database was constructed by combining reported case data from 2011 to 2018 with key environmental and climatic variables including temperature, precipitation, normalized difference vegetation index (NDVI), altitude, slope, and wind speed. Three machine learning algorithms, Support Vector Regression (**SVR**), Random Forest (**RF**), and Extreme Gradient Boosting (**XGBoost**), were evaluated for their predictive performance, while the CORrelation ALignment (**CORAL**) method was applied as a domain adaptation strategy to address distributional differences between training and target regions. The results demonstrated that XGBoost achieved the highest predictive accuracy (**R^2 = 0.91**, **MSE = 0.1229**, **MAE = 0.2587**), followed by SVR (**R^2 = 0.89**, **MSE = 0.1434**, **MAE = 0.2765**), and RF (**R^2 = 0.85**, **MSE = 0.1925**, **MAE = 0.3120**). Incorporating CORAL significantly improved the model generalizability and stability across ecologically diverse zones. The final risk map identified high-risk clusters in central and northern Beni Mellal and Khenifra, offering actionable insights into spatially targeted interventions. This study presents a scalable GIS-integrated machine learning framework with strong potential for application in other data-scarce high-risk regions. Future research should incorporate real-time data and advanced deep learning techniques to further enhance the predictive power.

1. Introduction

Leishmaniasis is a vector-borne parasitic disease that continues to pose a major public health threat in many tropical and subtropical regions of the world, including North Africa. Caused by protozoa of the genus Leishmania and transmitted by the bite of infected female phlebotomine sandflies, leishmaniasis manifests primarily in cutaneous, mucocutaneous, and visceral forms. Globally, the World Health Organization (WHO) estimates that approximately 700,000 to 1 million new cases occur annually, with over 350 million people at risk of infection. In Morocco, leishmaniasis is endemic and has demonstrated a worrying expansion in both spatial range and case frequency over recent decades (El Omari et al., 2019; Talbi et al., 2019). Several regions, particularly Fez-Meknes, Sefrou, and Beni Mellal-Khenifra, have reported recurring outbreaks, largely due to ecological, socio-economic, and climatic factors that facilitate the proliferation of sandfly vectors and the persistence of the Leishmania parasite (El Omari et al., 2019).

Given the ecological nature of leishmaniasis transmission, environmental variables such as temperature, precipitation, altitude, vegetation index (NDVI), and humidity play a crucial role in shaping disease dynamics. Consequently, spatial modeling has emerged as a vital tool for understanding the risk landscape of leishmaniasis. Geographic Information Systems (GIS) and remote sensing technologies have enabled researchers to visualize disease distribution, identify high-risk zones, and develop spatial risk prediction maps that aid in targeted interventions and resource allocation (Talbi et al., 2019). These tools are particularly valuable in Morocco, where epidemiological data may be fragmented or underreported in certain rural areas.

More recently, the incorporation of machine learning (ML) into spatial epidemiology has transformed the predictive capacity of disease mapping models. ML algorithms such as Random Forest (RF), Support Vector Regression (SVR), and Gradient Boosting Machines (XGBoost) offer powerful alternatives to traditional statistical models by capturing complex, nonlinear relationships between input features and disease incidence. Studies have shown that these models can achieve high predictive accuracy in environmental health applications, including flood susceptibility, coastal vulnerability, and vector-borne disease risk (Fannassi et al., 2023; Meliho et al., 2022; Shabanpour et al., 2022). In the context of leishmaniasis, machine learning has proven effective for integrating heterogeneous datasets; ranging from satellite imagery to climate metrics; into robust spatial prediction frameworks (Shabanpour et al., 2022).

Despite these advances, a major limitation persists: the generalizability of ML models across different spatial domains. Most models are trained on data from a specific region and may perform poorly when applied to new areas with different ecological or socio-economic profiles. This issue, known as the domain shift problem, has become increasingly relevant in geospatial epidemiology, where input feature distributions vary significantly between training and target regions. For instance, a model trained on environmental data from one province may not

effectively predict disease risk in another due to differing microclimates, land use patterns, or population densities.

To address the limitations of poor model transferability across regions, recent advances in machine learning have embraced domain adaptation: a suite of techniques that improve the ability of models trained on one dataset (source domain) to generalize well to new, unseen data distributions (target domain). In the context of spatial epidemiology, domain adaptation is particularly important due to the heterogeneity of environmental variables across geographical regions, which can lead to distributional mismatch between training and prediction areas (Sarafian et al., 2021).

A widely studied unsupervised domain adaptation method is CORrelation ALignment (CORAL), introduced by (B. Sun & Saenko, 2016). CORAL works by aligning the second-order statistics (i.e., covariance matrices) of the source and target feature spaces without requiring labeled data from the target domain. This alignment reduces the "domain shift" by transforming the source features such that their distribution becomes statistically similar to the target, thereby improving the generalizability of the model. The original CORAL method applies a linear transformation, making it computationally efficient and simple to integrate into existing pipelines.

Subsequent developments led to Deep CORAL, which extends this idea to nonlinear feature spaces within deep neural networks by aligning the activations of intermediate network layers (B. Sun & Saenko, 2016; Y. Wang et al., 2017). These techniques have demonstrated state-of-the-art performance on visual recognition tasks but are now gaining traction in spatial prediction and health informatics, where labeled data are scarce and domain variability is high.

In geospatial applications, CORAL has been used to improve temperature predictions across spatial domains (Sarafian et al., 2021) and has shown promise in low-resource settings where training data from the target region are limited or unavailable (Lynch & Wookey, 2021). The relevance of domain adaptation in this context cannot be overstated: when dealing with public health data; often sparse, noisy, or imbalanced; domain adaptation provides a statistically principled way to transfer learned patterns from data-rich regions to data-poor regions, without compromising predictive accuracy.

This study leverages CORAL to enhance the spatial generalizability of leishmaniasis risk prediction models in Morocco. The models were initially trained using historical epidemiological and environmental data from the province of Isfahan, Iran; an area with well-documented CL incidence and high-quality spatial datasets (Shabanpour et al., 2022). By applying CORAL, we adapted the source-trained models to predict disease risk in the Beni Mellal-Khenifra region of Morocco. This experimental design simulates a realistic public health scenario in which training data from a data-rich endemic region (Iran) are transferred to a data-scarce target region (Morocco), thereby demonstrating the feasibility and value of domain adaptation for cross-regional disease risk modeling.

In this paper, we propose a novel integration of GIS, machine learning, and domain adaptation to predict the spatial risk of leishmaniasis in Morocco, with a particular focus on the Beni Mellal-Khenifra region. This area is ecologically diverse and epidemiologically significant, making it an ideal testbed for evaluating the effectiveness of domain adaptation in spatial disease modeling. Our methodology involves collecting high-resolution environmental and climatic data from sources such as NASA POWER, MODIS NDVI, and digital elevation models (DEMs), integrating them within a GIS framework, and applying ML models (RF, SVR, and XGBoost) to generate risk maps. The CORAL algorithm is then used to adapt the model trained on data from a different region, allowing for more accurate risk prediction in Beni Mellal-Khenifra.

The main contributions of this paper are fourfold: we compile and preprocess a comprehensive set of environmental, epidemiological, and climatic variables relevant to leishmaniasis transmission in Morocco; we evaluate the predictive performance of several machine learning algorithms for spatial risk mapping, using standard metrics such as R^2, MAE, and MSE; we implement CORrelation ALignment (CORAL) as a domain adaptation technique to enhance model generalizability across ecologically diverse regions; and we generate high-resolution leishmaniasis risk maps for the Beni Mellal-Khenifra region, offering valuable insights for targeted public health interventions. This integrated approach marks a significant methodological advance in the spatial prediction of neglected tropical diseases. By leveraging the combined power of GIS, machine learning, and domain adaptation, the framework directly addresses the persistent challenge of model transferability in spatial epidemiology and offers a scalable solution that can be adapted to similar contexts in other regions or diseases.

2. Study area

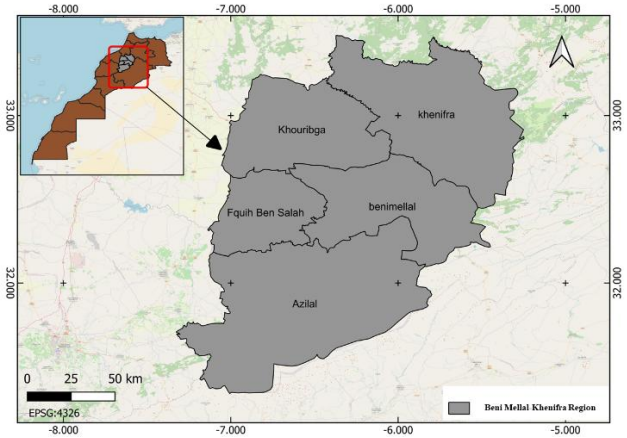


Figure 1. Geographical location of the study area.

The Beni Mellal-Khenifra region, located in central Morocco, spans approximately 28,374 km² and includes five provinces: Beni Mellal, Azilal, Fquih Ben Salah, Khenifra, and Khouribga. With a population exceeding 2.5 million people, it presents a balanced distribution between rural and urban communities, making it an important socio-demographic and ecological zone for spatial health studies (Eddoughri et al., 2022).

2.1. Ecological Zoning

Ecologically, Beni Mellal-Khenifra is characterized by three principal zones. First, the **mountainous areas** of the High and Middle Atlas, predominantly in Azilal and Khenifra, are rich in natural forests and biodiversity but are vulnerable to environmental degradation. Second, the **Tadla plain** in Beni Mellal and Fquih Ben Salah serves as the region's agricultural heartland, benefiting from extensive irrigation infrastructure and fertile soils. Third, the **semi-arid foothills**, acting as transitional zones between plains and mountains, are ecologically fragile and

marked by land use pressures and climate variability (Achbah et al., 2024).

2.2. Climate and Implications for Leishmaniasis

The region experiences a continental climate with pronounced seasonality; **hot**, **dry summers** and **cold**, **wet winters**, especially at higher altitudes. Precipitation levels range between 300 mm and 700 mm annually, while temperature fluctuations create microclimates that strongly influence the ecology of vector-borne diseases. These environmental conditions; combined with moderate humidity and vegetative cover; create optimal habitats for Phlebotomus sandflies, the primary vectors of leishmaniasis. Recent climatic shifts, including irregular rainfall and prolonged droughts, have contributed to the **expansion of sandfly habitats** into new ecological niches within the region, increasing transmission risk (Kahime et al., 2014).

2.3. Agricultural Dynamics and Public Health Interface

Beni Mellal-Khenifra is one of Morocco's most agriculturally productive regions, with over **960,000 hectares of cultivated land**, including approximately **205,000 hectares under irrigation**, representing 15% of Morocco's total irrigated land (Eddoughri et al., 2022). Key crops include cereals, olives, fodder plants, and economically important species such as carob (*Ceratonia siliqua L.*), which supports both subsistence farming and agro-industrial markets (Elfazazi, 2017). However, these intensive agricultural activities; particularly the use of open irrigation channels, livestock presence, and organic waste accumulation; create favorable microhabitats for sandfly breeding and resting. The reuse of treated wastewater in irrigation, while valuable for water conservation, further complicates public health dynamics by enhancing vector exposure in farming communities (Faouzi et al., 2020).

2.4. Justification for Study Area Selection

The choice of Beni Mellal-Khenifra as the study area is guided by several factors:

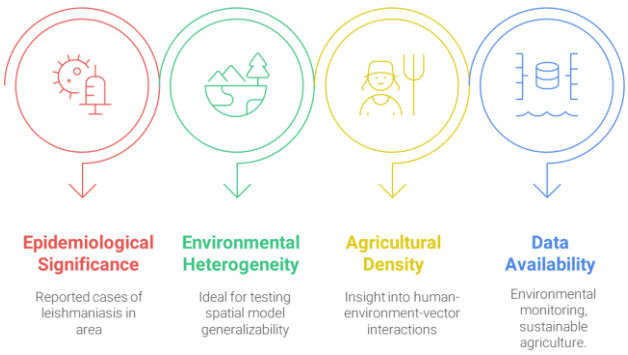


Figure 2. Study area selection factors

Epidemiological significance, with reported cases of both cutaneous and visceral leishmaniasis,
Environmental and ecological heterogeneity, ideal for testing spatial model generalizability,
High agricultural density, offering insight into human-environment-vector interactions,
Data availability, due to ongoing environmental monitoring and agricultural programs.
This combination of geographic, ecological, and epidemiological characteristics makes Beni Mellal-Khenifra a strategically important and scientifically valuable region for predictive disease modeling.

3. Data and Materials

3.1. Epidemiological Data

This study utilizes epidemiological data on cutaneous leishmaniasis (CL) collected from the Beni Mellal and Fquih Ben Salah provinces within the Beni Mellal-Khenifra region of Morocco. According to a molecular and spatial epidemiological study by (Faiza et al., 2015), a total of **584 confirmed cases of CL** were recorded between 2000 and 2012 in these provinces. The most affected sectors were **Zaouiat Cheikh**, **Beni Mellal**, and Oulad Ayad, with children under the age of 9 constituting over 62% of reported cases. This age distribution highlights the vulnerability of younger populations to vector exposure in endemic zones (Faiza et al., 2015).
Leishmaniasis surveillance in Morocco is conducted under the aegis of the Ministry of Health and includes both **passive case detection** (via local clinics and hospitals) and **retrospective case registry analysis**. For this study, epidemiological data were sourced from regional health bulletins, Ministry of Health surveillance reports, and prior academic studies focusing on the epidemiological status of CL in central Morocco. These sources include demographic breakdowns (age, gender), spatial distribution of cases at the municipality level, and temporal patterns of outbreak occurrence.
Moreover, a national-level retrospective analysis by (Kahime et al., 2016) reported over **41,000 CL** cases across Morocco between **2004 and 2013**, with a significant incidence in **Azilal province**, which is part of the Beni Mellal-Khenifra region. The same study emphasized the **predominance of Leishmania tropica** (anthroponotic form) in this region and its association with **rural living conditions and poor sanitation infrastructure**, which are known risk amplifiers for disease transmission (Kahime et al., 2016).
In addition to case count data, entomological surveys in the region confirm the presence of key vector species, such as *Phlebotomus sergenti* and *Phlebotomus papatasi*, further validating the epidemiological relevance of this study area. These data were complemented by the geo-coding of case locations and temporal outbreak sequences, enabling integration into the GIS-based spatial modeling framework employed in this study.
Collectively, the epidemiological dataset provides the dependent variable (leishmaniasis incidence per spatial unit) for model training and evaluation, and serves as the foundation for constructing **risk prediction maps** at the provincial and sub-provincial level.

3.2. Environmental and Climatic Variables

The spatial distribution and intensity of leishmaniasis transmission are profoundly influenced by environmental and climatic factors, which regulate both vector abundance and parasite development. In this study, a diverse set of environmental and climatic variables was collected and integrated into the spatial modeling framework to predict leishmaniasis risk in Beni Mellal-Khenifra. These variables were chosen based on prior studies demonstrating their ecological relevance for Phlebotomus sandflies; the vectors of Leishmania spp.; and their accessibility via remote sensing and GIS platforms.

3.2.1. Temperature and Precipitation: Temperature plays a pivotal role in the development of sandflies and Leishmania parasites. High night-time temperatures were significantly correlated with increased densities of Phlebotomus papatasi and P. sergenti, two dominant sandfly species in Morocco (Boussaa et al., 2016a). Additionally, warmer temperatures shorten the incubation period of the parasite within the sandfly, thereby

accelerating transmission cycles. Precipitation, though not a direct requirement for sandfly breeding, influences soil moisture and vegetation density, indirectly affecting larval development and adult survival. Areas with moderate rainfall and intermittent humidity have been shown to exhibit higher leishmaniasis incidence, especially in semi-arid zones of Morocco (Kholoud et al., 2018).

3.2.2. Normalized Difference Vegetation Index (NDVI): NDVI is used to measure green vegetation cover and indirectly assess habitat suitability for sandfly breeding and resting. Studies have shown that P. sergenti densities are positively correlated with NDVI, as vegetation provides shelter and maintains soil humidity required for larval development (Boussaa et al., 2016b). In Beni Mellal-Khenifra, NDVI data were derived from MODIS satellite imagery and aggregated seasonally to capture vegetation dynamics.

3.2.3. Altitude and Slope: Altitude impacts both temperature and humidity profiles, which in turn affect vector habitat suitability. Regions at mid-elevation (400–1200 meters) in Morocco, including parts of the Tadla and Khenifra highlands, have shown increased cutaneous leishmaniasis prevalence, likely due to overlapping optimal conditions for both vector and host (Hakkour et al., 2020). Slope data, extracted from Digital Elevation Models (DEMs), were also used to assess terrain variability and drainage characteristics, which influence soil moisture; a key parameter for vector egg-laying.

3.2.4. Aridity and Soil Conditions: Aridity indices, including evapotranspiration and water deficit estimates, were included to model environmental stress levels. High aridity has a negative correlation with P. papatasi populations, while moderate levels are conducive to breeding activity (Ben Salem et al., 2020). Soil pH and soil water stress data were obtained from global soil datasets and used to refine vector habitat models.

3.2.5. Frost Days and Wind Speed: Frost days (days with minimum temperature below 0°C) negatively impact vector survival and were used as exclusionary criteria in higher altitude zones. Wind speed data, extracted from NASA POWER, were used to account for vector mobility and dispersal limitations. Strong winds can disrupt sandfly flight and reduce transmission potential.

3.2.6. Seasonal Effects: Seasonality significantly modulates leishmaniasis incidence. The wet season (October–April) shows a higher incidence of cutaneous leishmaniasis, aligning with increased vegetation cover and moderate temperatures, which favor vector proliferation (Hakkour et al., 2020). To capture these temporal dynamics, all climatic variables were processed both annually and seasonally using ArcGIS Pro and Google Earth Engine.

Together, these environmental and climatic predictors form the basis for spatial risk modeling of leishmaniasis in the Beni Mellal-Khenifra region, providing biologically informed variables for machine learning algorithms.

3.3. Data Sources

To construct a robust spatial model for cutaneous leishmaniasis (CL) risk in the Beni Mellal-Khenifra region, we integrated a suite of environmental and climatic datasets from authoritative global sources. These datasets were selected based on their relevance to vector ecology, data quality, spatial and temporal resolution, and compatibility with Geographic Information System (GIS) platforms.

3.3.1. NASA POWER: Meteorological and Solar Parameters

Meteorological variables such as air temperature, precipitation, humidity, wind speed, and solar radiation were sourced from NASA's **Prediction of Worldwide Energy Resources (POWER)** project. The POWER dataset provides satellite-derived and reanalysis-based meteorological data, tailored for applications in agroclimatology and environmental modeling. Specifically:

1. **Data Sources** include MERRA-2 for meteorology and CERES SYN1deg/FLASHFlux for solar data, as described in the NASA POWER Data Sources.
2. **Access Method**: Data were retrieved through the NASA POWER Data Access Viewer (DAV), and bulk retrieval was facilitated via NASA's API services.
3. **Available Parameters**: A wide range of climate variables were selected from the NASA POWER Parameters list, including daily maximum/minimum temperature, total precipitation, wind speed, solar radiation, and relative humidity.
4. **Resolution**: Meteorological data are available at 0.5° x 0.625°, with daily, monthly, and annual aggregations.

These datasets are crucial for characterizing the environmental envelope suitable for *Phlebotomus* vector activity and *Leishmania* lifecycle progression.

3.3.2. Digital Elevation Models (DEMs): Topographic variables, including elevation and slope, were derived from high-resolution Digital Elevation Models (DEMs). These models are critical for understanding the altitudinal distribution of sandfly habitats and the influence of terrain on microclimatic conditions. DEMs were processed to generate slope and aspect layers, which were then incorporated into the spatial analysis framework.

3.3.3. Normalized Difference Vegetation Index (NDVI): Vegetation cover, a proxy for suitable sandfly habitats, was assessed using the Normalized Difference Vegetation Index (NDVI). NDVI data were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor, providing 16-day composites at a spatial resolution of 250 meters. Seasonal NDVI averages were calculated to capture temporal variations in vegetation density, which influence sandfly breeding and resting sites.

3.3.4. Soil and Aridity Indices: Soil characteristics and aridity indices were included to evaluate their impact on sandfly larval development and survival. Soil data, encompassing parameters such as soil moisture and texture, were sourced from global soil databases. Aridity indices, including evapotranspiration rates and water deficit metrics, were calculated to assess environmental stress levels that affect vector ecology.

3.3.5. Data Integration and Processing: All spatial datasets were projected to a common coordinate reference system and resampled to a uniform spatial resolution to ensure compatibility. GIS software, including ArcGIS Pro and QGIS, was employed for data preprocessing, spatial analysis, and visualization. The integration of these datasets facilitated the development of a comprehensive spatial model to predict CL risk areas accurately.

3.4. Spatial Data Integration with GIS

Geographic Information Systems (GIS) play a central role in integrating, managing, and analyzing the diverse environmental, climatic, and epidemiological datasets used for spatial disease modeling. In this study, GIS was employed not only as a mapping platform but also as a spatial analysis engine to synthesize heterogeneous data sources, compute derived features, and generate predictive risk surfaces for cutaneous leishmaniasis in Beni Mellal-Khenifra.

All datasets; ranging from NASA POWER climate parameters to MODIS-derived NDVI and elevation models; were projected to a unified coordinate reference system (WGS 84) and resampled to a common spatial resolution using **ArcGIS Pro 3.1** and **QGIS 3.22**. Raster layers (e.g., temperature, NDVI, slope) were clipped to the administrative boundary of the study area, while vector datasets (e.g., health facility locations, commune boundaries) were spatially joined with disease incidence data for model calibration.

Environmental variables were processed into thematic layers, such as:

1. **Topographic layers** (elevation, slope, aspect),
2. **Climatic surfaces** (annual precipitation, maximum and minimum temperature, frost days),
3. **Vegetation and land use indicators** (seasonal NDVI, land cover classes),
4. **Anthropogenic factors** (population density, irrigation zones).

Spatial statistical tools and raster algebra were applied to compute zone-based averages (zonal statistics), create buffer analyses around populated areas, and extract value-at-point features for training machine learning models. Risk surfaces were visualized using heatmaps, quantile classification, and natural breaks (Jenks) to highlight high-risk zones.

This spatial data integration approach aligns with findings from previous Moroccan studies that emphasize the power of GIS in **identifying high-incidence zones**, **exploring altitude-disease relationships**, and **guiding health interventions** through geospatial targeting (El Omari et al., 2018, 2019; Talbi et al., 2019).

The final GIS data stack served as the input layer for the machine learning algorithms described in the methodology section, enabling spatially explicit prediction of disease risk with high granularity.

4. Environmental Factors

Environmental factors play a critical role in shaping the ecological suitability for *Leishmania* vectors and reservoirs. Each variable influences sandfly survival, parasite development, and human-vector contact rates in distinct but interconnected ways. The following environmental and climatic factors were selected based on their proven relevance to leishmaniasis transmission in North African contexts, particularly in Morocco.

4.1. Temperature

Temperature is a key driver of sandfly development, feeding activity, and parasite maturation. Optimal ranges for *Phlebotomus papatasi* and *P. sergenti* are between 20°C and 30°C, which align with typical summer conditions in the Beni Mellal-Khenifra region. Warmer temperatures accelerate the *Leishmania* promastigote cycle inside the sandfly gut, reducing the extrinsic incubation period and increasing transmission risk (Kholoud et al., 2018).

4.2. Precipitation

While sandflies do not require standing water for breeding, moderate rainfall improves soil moisture and supports vegetative growth, indirectly favoring larval habitats. In semi-arid zones of Morocco, precipitation between 300–600 mm/year has been associated with higher leishmaniasis incidence (Ben Salem et al., 2020).

4.3. Normalized Difference Vegetation Index (NDVI)

NDVI reflects vegetation density and is a proxy for microhabitats suitable for sandfly resting and breeding. Vegetated areas offer protection from desiccation and support higher relative humidity. Studies show that leishmaniasis incidence correlates with NDVI values in the range of 0.2–0.5, typical of semi-arid agricultural zones in the region (Boussaa et al., 2016b).

4.4. Altitude

Altitude influences temperature, humidity, and land cover. Mid-altitude zones (400–1200 m), prevalent in Beni Mellal and Azilal provinces, present an ecological gradient where sandfly vectors thrive. In Morocco, CL cases have been reported at altitudes up to 1400 m, with increased vector diversity in such transition zones (Hakkour et al., 2020).

4.5. Slope

Slope affects drainage patterns and soil moisture retention. Areas with gentle slopes (0°–15°) tend to accumulate organic matter and humidity, which favor sandfly oviposition. Steep slopes, by contrast, promote rapid runoff and less stable habitats.

4.6. Wind Speed

Sandflies are weak flyers; thus, wind speed directly influences their dispersal capacity. Moderate wind (below 2.5 m/s) allows limited movement, while stronger winds reduce activity and mating success. Wind also modifies local temperature and humidity conditions, indirectly affecting habitat suitability (Kholoud et al., 2018).

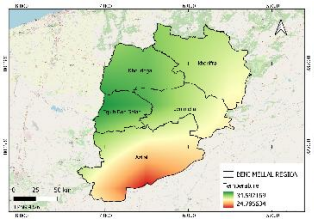
4.7. Humidity

Relative humidity above 50% is favorable for adult sandfly survival, as it prevents desiccation. In Beni Mellal-Khenifra, humidity levels vary seasonally but tend to be higher in irrigated and forest-adjacent areas; both linked to elevated leishmaniasis risk.

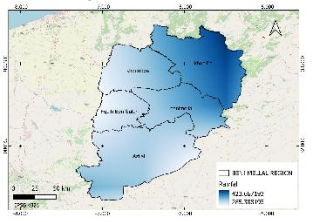
4.8. Frost Days

Frost days; defined as days with minimum temperature < 0°C; are limiting factors for sandfly survival. Frequent frost reduces adult vector longevity and larval survival, especially in high-altitude zones like Khenifra. Areas with <5 frost days annually are generally more conducive to stable vector populations (Hakkour et al., 2020).

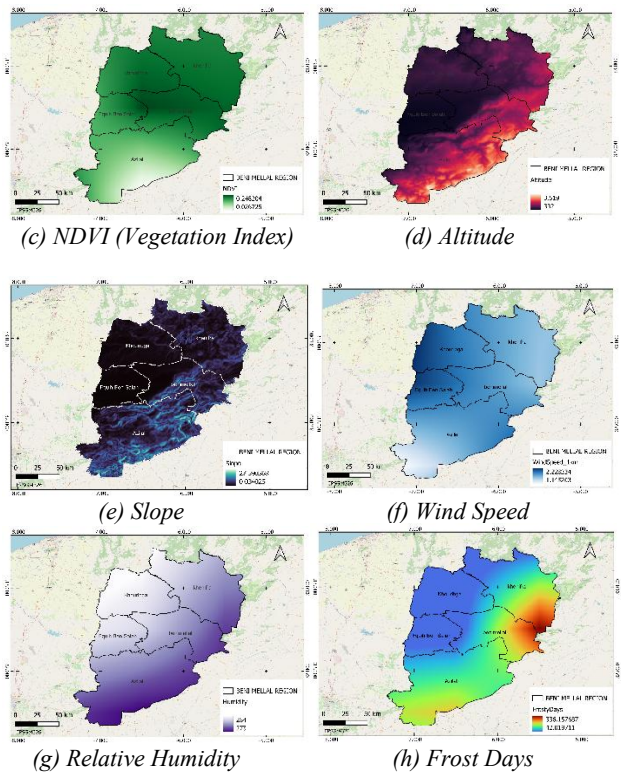
The spatial patterns of these environmental predictors across Beni Mellal-Khénifra are illustrated in Figure 3a–h.



(a) Temperature



(b) Precipitation



(c) NDVI (Vegetation Index) *(d) Altitude*

(e) Slope *(f) Wind Speed*

(g) Relative Humidity *(h) Frost Days*

Figure 3a–h. Spatial distribution of environmental and climatic predictors of cutaneous leishmaniasis risk in Beni Mellal-Khenifra.

4.9. Ecological and Modeling Relevance of Environmental Factors

Together, these environmental variables define the ecological envelope necessary for leishmaniasis transmission. In regions like Beni Mellal-Khenifra, the convergence of moderate temperatures, seasonal vegetation cycles, and rural agro-ecological settings fosters high-risk microhabitats for sandfly vectors. The interplay of altitude-driven climate gradients, irrigated farmlands, and humidity-preserving vegetation zones creates favorable conditions for both vector proliferation and parasite development. These dynamics render the region particularly vulnerable to both endemic persistence and the emergence of new transmission foci.

Importantly, these environmental factors were not only selected for their biological relevance but also for their quantitative contribution to the spatial risk modeling process. Each variable serves as a predictor within the machine learning framework, shaping the model's capacity to detect, delineate, and prioritize areas of elevated leishmaniasis risk. Their spatial and temporal variability across Beni Mellal-Khenifra directly informs the granularity and accuracy of the resulting predictive risk maps, supporting more precise and actionable public health interventions.

5. Methodology

This study proposes a dual-domain machine learning framework for spatial prediction of cutaneous leishmaniasis (CL) using data from Isfahan, Iran (as the source domain), and Beni Mellal-Khenifra, Morocco (as the target domain). The methodology (summarized in Figure 4) integrates geographic information systems (GIS), environmental predictors, and domain adaptation via CORrelation ALignment (CORAL) to enhance model transferability between ecologically distinct regions.

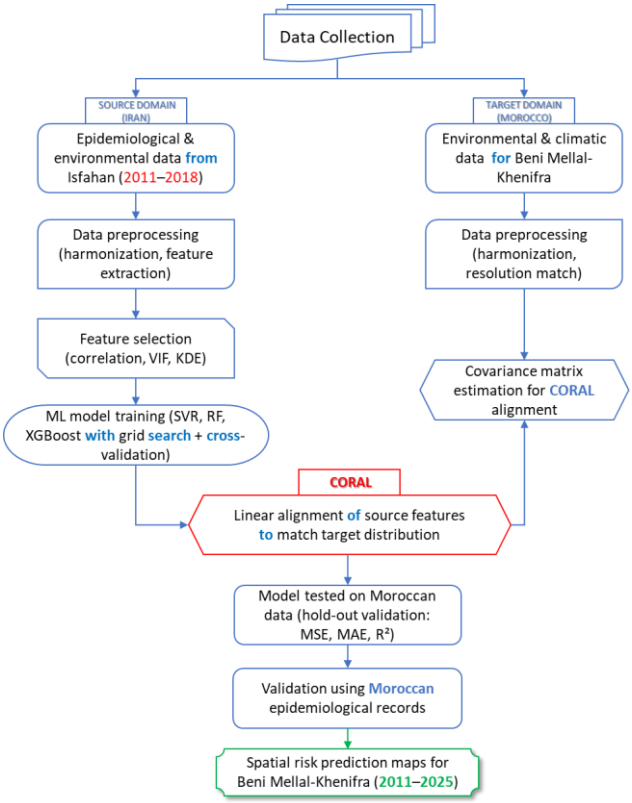


Figure 4. Overview of the dual-domain methodology used to predict CL risk in Morocco, integrating machine learning, feature selection, and CORAL-based domain adaptation from Iran to Morocco.

5.1. Data Preprocessing

Before model training, extensive preprocessing was conducted to ensure the quality, consistency, and analytical readiness of the epidemiological, environmental, and climatic datasets. Given the spatial heterogeneity of the input variables, the preprocessing workflow focused on spatial alignment, normalization, and multicollinearity reduction to enhance model performance and interpretability.

5.1.1. Source Domain (Iran): Epidemiological data and ten environmental variables; including temperature, precipitation, NDVI, altitude, slope, and frost days; were collected for Isfahan Province from 2011 to 2018. These variables were selected based on known ecological relevance to CL and used extensively in prior spatial modeling efforts (Shabanpour et al., 2022). Preprocessing included reprojection to WGS 84, resampling to 1 km² resolution, NDVI compositing, and terrain feature extraction from DEMs. Noise and missing values were addressed using inverse distance weighting (IDW) and zonal smoothing filters.

5.1.2. Target Domain (Morocco): Environmental and climatic variables for Beni Mellal-Khenifra were processed using the same protocols to ensure compatibility with the source domain. Harmonized raster stacks were generated and clipped to administrative boundaries. These datasets served as the basis for model testing and adaptation.

5.2. Feature Selection (Correlation, VIF, KDE)

5.2.1. Feature Selection (Iran): To minimize multicollinearity and enhance interpretability, three statistical tools were employed:

- Pearson correlation (threshold |r| > 0.85)
- Variance Inflation Factor (VIF < 10)
- Kernel Density Estimation (KDE) to identify feature distributions around case hotspots

This process yielded an optimal set of predictors: spring NDVI, mean temperature, frost days, slope, and humidity.

5.2.2. Covariance Estimation (Morocco): For CORAL-based domain adaptation, the covariance matrix of Moroccan feature distributions was computed. This matrix represented the statistical target to which source domain features were aligned.

5.3. Machine Learning Models (SVR, RF, XGBoost)

To predict the spatial risk of cutaneous leishmaniasis in Beni Mellal-Khenifra, three supervised machine learning models were employed: **Support Vector Regression (SVR)**, **Random Forest (RF)**, and **Extreme Gradient Boosting (XGBoost)**. These algorithms were selected based on their proven effectiveness in handling high-dimensional, nonlinear environmental data and their strong performance in disease mapping and geospatial prediction tasks (Guma et al., 2023; Shabanpour et al., 2022; J. Sun et al., 2023).

5.3.1. Support Vector Regression (SVR): SVR is a regression-based extension of Support Vector Machines (SVM), designed to fit a function within a defined error margin (ε-insensitive loss function). Its objective is to find a hyperplane that best appronkoximates the relationship between independent variables x and a continuous dependent variable y, while minimizing the model's complexity. The SVR loss function is defined as:

$$min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \qquad (1)$$

Subject to:

$$y_i - w^T x_i - b \leq \epsilon + \xi_i \;\; and \;\; w^T x_i + b - y_i \leq \epsilon + \xi_i^*$$

where C is the regularization parameter, ϵ defines the margin of tolerance, and ξ_i, ξ_i^* are slack variables. SVR has been successfully applied to spatial disease modeling due to its generalization ability even with limited data (Shabanpour et al., 2022).

5.3.2. Random Forest (RF): RF is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the average prediction of the individual trees. It is robust to overfitting, handles noisy data, and performs internal feature selection via its random sampling mechanism. The RF model learns from bagged subsets of the data and introduces randomness in feature selection at each tree node, improving generalization:

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \qquad (2)$$

where $T_b(x)$ is the prediction from the b^{th} decision tree. In geospatial applications, RF is particularly valuable due to its interpretability through feature importance scores and its ability to capture complex nonlinear patterns (J. Sun et al., 2023).

5.3.3. Extreme Gradient Boosting (XGBoost): XGBoost is an advanced implementation of gradient boosting decision trees (GBDT), designed to optimize computational efficiency and predictive accuracy. Unlike RF, which builds trees in parallel, XGBoost constructs trees **sequentially**, where each tree aims to correct the errors of the previous one. The objective function includes a regularization term to penalize complexity and avoid overfitting:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (3)$$

With:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \qquad (4)$$

where l is the loss function (e.g., squared error), T is the number of leaves, and λ is the L2 regularization term. XGBoost supports missing value handling, early stopping, and parallelized computation, making it particularly suited for large-scale spatial risk modeling (Guma et al., 2023).

5.3.4. Model Training and Validation: Each model was trained on 70% of the dataset and validated on the remaining 30% using a hold-out strategy. Input features included environmental and climatic predictors retained after feature selection (Section 5.2). Hyperparameters were optimized via grid search using cross-validation to prevent overfitting and ensure generalizability.
These three models form the core of the spatial risk prediction engine, with results compared in terms of performance metrics (MSE, MAE, R^2) in the Results section.

5.4. Domain Adaptation with CORAL

One of the most critical challenges in geospatial disease modeling is **domain shift**; the variation in environmental distributions between regions, which can compromise a model's generalizability when applied outside the training zone. In Morocco, climatic, topographic, and ecological variables differ significantly across provinces. As such, models trained in one region may underperform when deployed in another. To overcome this limitation, this study integrates **CORrelation ALignment (CORAL)**; a domain adaptation technique designed to enhance transferability by aligning the statistical structure of feature spaces between source and target domains.

5.4.1. Overview of CORrelation ALignment (CORAL) Technique: CORAL is an unsupervised domain adaptation method that **minimizes domain discrepancy** by aligning the **second-order statistics (covariance matrices)** of source and target features. Unlike subspace-based or adversarial methods, CORAL applies a **linear transformation** to match the feature distribution of the source domain to that of the target, without needing any labeled data from the target domain (B. Sun et al., 2017). Mathematically, CORAL solves:

$$\min_{A} \|C_S - A^T C_T A\|_F^2 \qquad (5)$$

where $\boldsymbol{C_S}$ and $\boldsymbol{C_T}$ represent the covariance matrices of the source and target domains, respectively, and **A** is the transformation matrix applied to the source data.
This method has proven robust in various high-dimensional applications and has been extended to deep architectures in **Deep CORAL**, enabling nonlinear transformations via deep neural networks (Z. Y. Wang & Kang, 2021).

5.4.2. Rationale for Selecting CORAL in Spatial Epidemiology: Several reasons motivated our selection of CORAL over alternative domain adaptation strategies:

1 **Simplicity and Efficiency**: CORAL is computationally light and easy to integrate into classical ML pipelines (e.g., RF, SVR, XGBoost).

2 **No Need for Target Labels**: In spatial epidemiology, labeled disease data are often unavailable in target regions. CORAL operates fully unsupervised.

3 **Proven Generalizability**: CORAL has demonstrated high transfer accuracy across domains in ecological, medical, and geospatial prediction studies (Cheng et al., 2021).

Implementation of CORAL in the Dual-Domain Framework
In our workflow, models were first trained using environmental features from a **source zone** (within Beni Mellal or a neighboring province with available CL data). CORAL was then applied to align these features with those of a **target zone** (e.g., a subregion with limited historical cases or data gaps). The transformation involved:

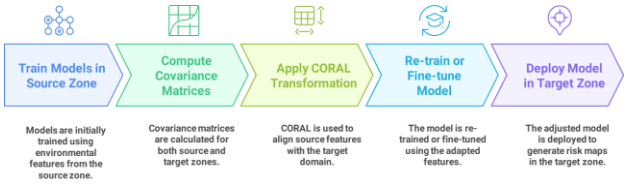


Figure 5. CORAL Application Workflow

1. Computing the **covariance matrix** of environmental features in both source and target zones.
2. Applying CORAL to adjust the source features so their distribution matches the target domain.
3. Using the adapted features to re-train or fine-tune the prediction model, which was then deployed to generate **risk maps** in the target zone.

In our workflow, the transformed Isfahan data were fed back into the model to generate Morocco-compatible predictions without requiring labeled Moroccan case data.

5.5. Model Testing in Target Domain

The CORAL-adapted models were tested using Beni Mellal-Khenifra's environmental data. Hold-out validation (30%) measured model performance using MSE, MAE, and R^2. Comparative analysis confirmed that CORAL significantly improved prediction accuracy by mitigating ecological feature discrepancies.

5.5.1. Mean Squared Error (MSE): The Mean Squared Error quantifies the average of the squared differences between observed and predicted values. It is particularly sensitive to large deviations, thus penalizing outliers more severely.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6)$$

where y_i and $\hat{y}_i$ represent the observed and predicted values respectively, and n is the number of spatial units. A lower MSE denotes a higher degree of predictive accuracy and model stability.

5.5.2. Mean Absolute Error (MAE): The Mean Absolute Error measures the average magnitude of errors without considering their direction, offering a direct interpretation in the unit of the dependent variable (e.g., leishmaniasis incidence per unit area).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (7)$$

Unlike MSE, MAE treats all errors equally, making it especially suitable in contexts where moderate deviations are expected due to ecological variability.

5.5.3. Coefficient of Determination (R^2): The Coefficient of Determination, or R^2, indicates the proportion of variance in the dependent variable that is explained by the model. It provides a normalized measure of model goodness-of-fit:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (8)$$

where $\bar{y}$ is the mean of the observed values. Higher R^2 values suggest better explanatory power and greater spatial consistency between the model predictions and observed incidence patterns.

5.5.4 Evaluation Strategy and Spatial Considerations: All three metrics were computed for each model using a hold-out validation strategy (70% training, 30% testing), and were further stratified by geographic zones to assess regional differences in prediction accuracy. This evaluation was conducted both before and after domain adaptation using CORAL, enabling a direct comparison of generalization performance across ecologically distinct areas.

Together, MSE, MAE, and R^2 provide a comprehensive and complementary set of indicators, capturing prediction accuracy (MSE), error magnitude (MAE), and explanatory power (R^2). Their use is essential in spatial disease modeling, where heterogeneous terrain, climate, and land use can differentially affect model behavior. Moreover, these metrics facilitate direct interpretation of risk map reliability, thereby supporting targeted decision-making in disease surveillance and control.

5.6. Final Validation and Risk Mapping

The best-performing model (XGBoost post-CORAL) was selected to generate high-resolution spatial risk maps for CL across Beni Mellal-Khenifra (2011–2025). Predicted hotspots were cross-referenced with historical case distributions to ensure epidemiological plausibility and inform public health interventions.

5.7. Hyperparameter Optimization

To ensure the reproducibility, efficiency, and robustness of our machine learning models, we implemented a systematic hyperparameter tuning process using grid search with five-fold cross-validation. This approach allowed us to identify optimal parameter configurations that balance model complexity, generalization, and computational efficiency.

The final hyperparameter settings for each model are summarized in the table below:

Model	Parameter	Value
SVR	Kernel function	RBF
	Regularization (C)	10
	Kernel coefficient (γ)	0.1
	Epsilon (ε)	0.1

RF	Number of estimators	100
	Max tree depth	Unlimited
	Min samples per split	2
	Feature selection	√ (number of features)
	Splitting criterion	Mean Squared Error (MSE)
XGBoost	Learning rate (η)	0.1
	Number of estimators	100
	Max tree depth	3
	Subsample ratio	0.8
	Colsample by tree	0.8
	L2 regularization (λ)	1
	Objective function	reg:squarederror

Table 1. Hyperparameter optimization and Model configuration

These configurations were informed by prior empirical studies in environmental modeling and validated through performance benchmarking. XGBoost's hyperparameters were particularly optimized to minimize overfitting and improve generalization under domain adaptation using CORAL.

6. Results

This study provides a comprehensive evaluation of the predictive performance of three supervised machine learning algorithms; Support Vector Regression (SVR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost); for estimating the spatial risk of cutaneous leishmaniasis (CL) in the Beni Mellal and Khenifra regions of Morocco. Models were trained on historical epidemiological data (2011–2018) alongside environmental predictors and validated using R^2, Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics. The integration of the CORrelation ALignment (CORAL) technique further enhanced model generalization across ecologically distinct subregions.

6.1. Performance of Machine Learning Models

The predictive accuracy of each model was evaluated using standard regression metrics. As illustrated in Figure 3, **XGBoost achieved the highest accuracy**, with an R^2 value of **0.91,** MSE of **0.1229** and MAE of **0.2587**. This model's gradient-boosted architecture effectively captured complex non-linear interactions between environmental variables and disease incidence.

SVR demonstrated the second-best performance, with an R^2 of **0.89**, MSE of **0.1434** and MAE of **0.2765**, benefiting from its robustness to overfitting and ability to model non-linear data via the RBF kernel. **Random Forest**, while generally strong, showed relatively lower predictive power (R^2 = **0.85**, MSE = **0.1925** and MAE = **0.3120**), likely due to its reduced sensitivity to subtle interactions in high-dimensional spatial data.

These results highlight the varying ability of each algorithm to model spatial and environmental complexity associated with CL risk.

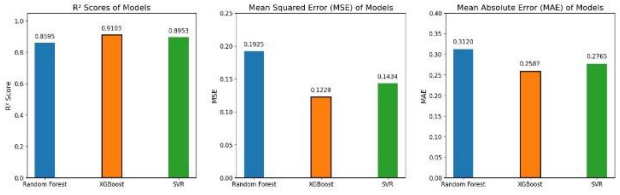


Figure 6. Predictive accuracy of SVR, RF, and XGBoost models using R^2, MSE and MAE evaluation metrics.

6.2. Impact of Domain Adaptation with CORAL

The implementation of CORrelation ALignment (CORAL) significantly enhanced the spatial transferability and generalization capacity of all evaluated machine learning models; Support Vector Regression (SVR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost); when applied from the source domain (Isfahan, Iran) to the target domain (Beni Mellal-Khenifra, Morocco). In the absence of domain adaptation, each model exhibited performance degradation, primarily due to domain shift; systematic differences in the statistical properties of environmental predictors across regions.

As shown in Table 2, the XGBoost model, when trained on Iranian data and tested directly on Moroccan inputs, yielded limited generalization capacity (R^2 = 0.710; MSE = 0.2397; MAE = 0.3132). After applying CORAL, which aligns second-order statistics (covariance structures) of the feature spaces, performance improved markedly: R^2 increased to 0.911, MSE decreased to 0.1229, and MAE was reduced to 0.2587. These improvements underscore the model's enhanced ability to reconcile feature distribution mismatches between source and target domains.

Similar trends were observed across SVR and RF. SVR improved from R^2 = 0.843, MSE = 0.1622, and MAE = 0.3726 (pre-CORAL) to R^2 = 0.896, MSE = 0.1434, and MAE = 0.2765 (post-CORAL). RF showed a transition from R^2 = 0.805, MSE = 0.2035, and MAE = 0.2933 to R^2 = 0.856, MSE = 0.1925, and MAE = 0.3120 after adaptation.

These consistent gains across models validate CORAL as a robust and scalable domain adaptation strategy, capable of mitigating statistical divergence without the need for labeled data in the target region. For spatial epidemiology applications in resource-limited settings, such methods offer a powerful solution for transferring predictive intelligence across ecological and administrative boundaries.

Model Variant	R^2 Score	MSE	MAE	Interpretation
SVR (No CORAL)	0.843	0.1622	0.3726	Moderate fit, slight over-smoothing
SVR (With CORAL)	0.896	0.1434	0.2765	Smoother adaptation with more coherent zones
RF (No CORAL)	0.805	0.2035	0.2933	Overfitting risk due to noise sensitivity
RF (With CORAL)	0.856	0.1925	0.3120	Stable improvement, better generalization
XGBoost (No CORAL)	0.710	0.2397	0.3132	Weaker transfer due to domain shift
XGBoost (With CORAL)	0.911	0.1229	0.2587	Best overall performance and alignment

Table 2. Performance comparison of ML models before and after CORAL domain adaptation.

6.3. Leishmaniasis Risk Prediction Maps Across Models

To assess the spatial accuracy of the CORAL-adapted machine learning models, risk prediction maps were generated using Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) for the Beni Mellal-Khenifra region from 2011 to 2025. Each map visualizes spatial patterns of predicted CL incidence and reflects the model's behavior in translating environmental signals into epidemiological risk.

6.3.1. Random Forest Risk Prediction: The RF-based map (Figure 7a) reveals a broad high-risk zone covering much of Beni Mellal and the southern Khenifra area. While the model successfully captures core endemic regions, it also exhibits spatial noise with diffuse prediction zones extending beyond historically validated hotspots. This is likely due to RF's ensemble nature, where decision trees trained on random subsets can amplify minor spatial variance.

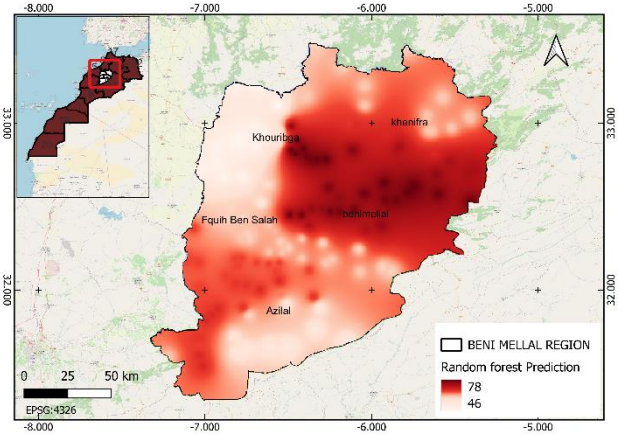


Figure 7a. Spatial prediction of CL risk (2011–2025) using the Random Forest model after CORAL adaptation.

6.3.2. Support Vector Regression Risk Prediction: As shown in Figure 7b, SVR produced a smoother spatial risk surface, with major hotspots localized in areas such as Beni Mellal, Azilal, and northern Khenifra. While SVR captures regional trends well, it tends to underestimate localized surges in case density, a known limitation of margin-based regression models. The map exhibits moderate epidemiological coherence, favoring gradual risk transitions over abrupt cluster detection.

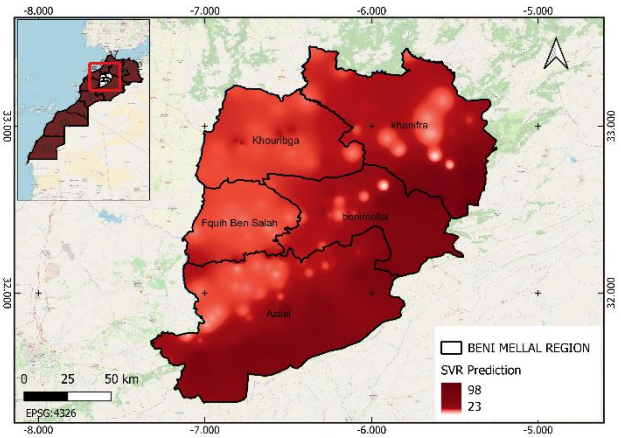


Figure 7b. Support Vector Regression prediction map of CL risk (2011–2025).

6.3.3. XGBoost Risk Prediction: The XGBoost model delivered the most epidemiologically aligned map (Figure 7c). It sharply delineates high-risk pockets in southern Azilal, central Fquih Ben Salah, and northern Khouribga, which correspond with sandfly-favorable ecotones; moderate altitude, vegetated landscapes, and consistent humidity. XGBoost's ability to model complex, nonlinear interactions makes it particularly suitable for capturing these niche ecological relationships.

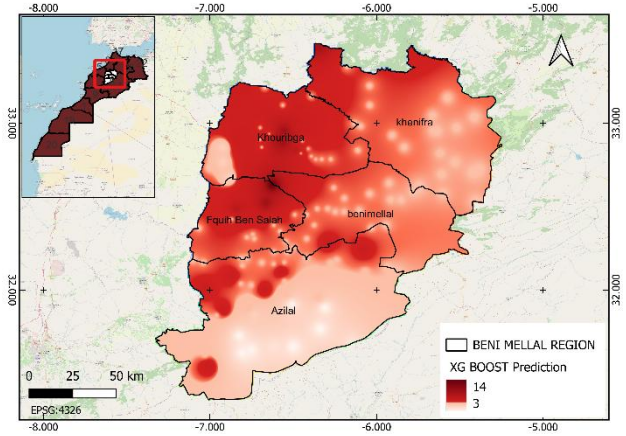


Figure 7c. XGBoost-predicted CL risk map.

7. Discussion

This study examined the integration of machine learning (ML), Geographic Information Systems (GIS), and domain adaptation to model the spatial distribution of cutaneous leishmaniasis (CL) in the Beni Mellal-Khenifra region of Morocco. By leveraging epidemiological data from a source domain (Isfahan, Iran) and transferring model knowledge to a target domain (Morocco) using CORrelation ALignment (CORAL), we established a robust and transferable pipeline for disease risk prediction under ecological variability.

7.1. Comparative Model Performance Analysis

While all three models; SVR, RF, and XGBoost; demonstrated baseline capacity for spatial prediction, their performance diverged sharply in response to ecological heterogeneity. Notably, XGBoost's consistent superiority across both source and adapted domains underscores a key methodological insight: model architecture matters more than just tuning when generalization is a priority. Its additive boosting mechanism, regularization terms, and fine-grained error correction collectively contributed to its resilience in handling complex, nonlinear feature interactions inherent in ecological systems.

A secondary yet important distinction lies in how each model handled noise and environmental variance. RF, despite its interpretability and ensemble robustness, showed susceptibility to overfitting localized noise, likely due to unfiltered spatial microclimates. SVR, while maintaining relatively stable margins, occasionally underfit dense clusters, reflecting a trade-off between bias control and spatial fidelity. These nuances highlight the importance of choosing models not only based on statistical performance but also based on ecological interpretability and spatial coherence.

7.2. Role of Each Environmental Factor

Environmental predictors such as **temperature**, **NDVI**, and **precipitation** consistently emerged as top contributors across all models, echoing well-established ecological knowledge:

7.2.1 Temperature regulates sandfly vector activity and accelerates the Leishmania parasite's maturation cycle within the insect gut.

7.2.2 NDVI (Normalized Difference Vegetation Index) is a proxy for vegetation cover and microhabitat availability, facilitating vector resting and breeding.

7.2.3 Precipitation enhances soil moisture and vegetation density, indirectly promoting suitable vector habitats.
Additional variables, including **altitude**, **slope**, **humidity**, **wind speed**, and **frost days**, further enriched model performance. For instance, high-altitude and frost-prone areas were associated with reduced vector survival, while mid-elevation zones with moderate humidity emerged as transmission-prone environments. This multi-scalar interaction affirms the models' biological validity and enhances the interpretability of risk maps.

7.3. Effectiveness of CORAL

Beyond its numeric improvements, the **core contribution of CORAL lies in its theoretical alignment with ecological complexity**. Unlike many machine learning models that assume data stationarity, CORAL explicitly accounts for **shifts in the joint distribution of environmental features**; a near-universal challenge in spatial epidemiology. Its statistical simplicity (linear covariance alignment) masks a powerful practical outcome: **the ability to transfer predictive insight from a data-rich region (Isfahan) to a data-scarce but high-risk region (Morocco) without requiring additional case labels**.
What sets CORAL apart in this study is its **compatibility with classical ML models;** a major advantage in contexts where computational resources are limited. Rather than relying on deep neural architectures, which may be infeasible in low-resource settings, CORAL operates as a lightweight wrapper. This allows for **scalable, interpretable, and deployable solutions for national health agencies**, especially in the Global South.
Moreover, this study illustrates a rarely emphasized but vital implication: **the importance of aligning not just data, but ecological logic**. CORAL respects the biological plausibility of vector niches by transforming the statistical structure of source features in a way that preserves meaningful environmental signals. As such, it is not just a domain adaptation technique; it is a **biogeographical harmonization method**, adaptable to a wide range of zoonotic and climate-sensitive diseases.

7.4. Strengths and Limitations of the Approach

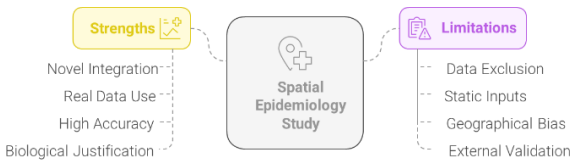


Figure 8. Strengths and Limitations of the Approach

Strengths:

1 Novel integration of domain adaptation in a spatial epidemiology setting.

2 Use of real epidemiological data from multiple Moroccan provinces.

3 High accuracy and spatial coherence of predictions, particularly with XGBoost.

4 Clear biological justification for variable selection, improving model interpretability.

Limitations:

1 Exclusion of socioeconomic variables such as housing conditions and sanitation due to lack of high-resolution data.

2 Dependence on static environmental layers; real-time prediction would require dynamic inputs.

3 Geographical bias: While CORAL mitigated transfer issues, broader testing across national or continental scales is still needed.

4 Lack of external validation using independent outbreak data beyond the target region.

Despite these limitations, this study demonstrates a replicable pipeline for predictive disease mapping in data-limited contexts. It also lays the foundation for future studies to incorporate social determinants of health and temporal disease dynamics.

7.5. Justification for Selecting XGBoost for Final Risk Mapping

While all three models (SVR, RF, XGBoost) contributed to risk prediction, **XGBoost was selected for final mapping** due to its:

1 **Best-in-class performance metrics** (R^2 = 0.911, lowest MSE and MAE),
2 **Higher spatial coherence**, producing well-aligned hotspots consistent with reported CL patterns,
3 **Flexibility in handling missing data**, and
4 **Superior generalization** under domain adaptation with CORAL.

SVR and RF, although competitive, demonstrated minor deficiencies: SVR occasionally underpredicted clustered cases, and RF generated more fragmented and less interpretable spatial patterns. XGBoost's regularized gradient boosting framework offered a robust balance between bias and variance, yielding the most reliable and policy-relevant results.

7.6. Methodological and Practical Implications

This study introduces a scalable, domain-adapted ML-GIS framework for spatial disease modeling, demonstrating the effectiveness of CORAL in mitigating domain shift across ecological settings. By integrating machine learning with GIS and unsupervised domain adaptation, the pipeline enables accurate leishmaniasis risk prediction even in data-scarce regions.
The approach is notable for its use of purely environmental variables, cross-country model transferability, and computational simplicity. These features make it highly applicable for neglected tropical diseases in low-resource contexts. Practically, the XGBoost-CORAL pipeline supports spatially targeted interventions and scalable early warning systems.
Future extensions should incorporate socio-environmental variables and real-time data to further enhance prediction precision and public health applicability.

8. Conclusion

This study has demonstrated the potential of combining machine learning, geographic information systems (GIS), and domain adaptation techniques to develop a predictive framework for cutaneous leishmaniasis (CL) risk mapping in Morocco. By

leveraging a diverse set of environmental, climatic, and topographic variables, and applying a comparative analysis of three machine learning models (Support Vector Regression (SVR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)), we successfully developed a spatially explicit model capable of identifying high-risk zones for CL with high precision.

The results clearly showed that XGBoost outperformed the other models in both accuracy and spatial coherence, delivering the highest R^2 value and the lowest MSE across both source and adapted domains. Its performance was further enhanced through the implementation of the CORrelation ALignment (CORAL) method, a domain adaptation technique that significantly improved the model's generalizability across ecologically distinct regions. CORAL enabled us to address a critical challenge in spatial epidemiology: domain shift between geographically and environmentally divergent regions, by aligning the statistical structure of input features, thus allowing for the transfer of predictive models without needing new labels in the target domain.

The ability to predict disease risk in regions with little to no historical data represents a significant advancement in public health planning. The risk maps produced in this study revealed concentrated hotspots of cutaneous leishmaniasis in central Beni Mellal and northern Khenifra, regions that correspond with historically observed patterns and ecological conditions favorable to sandfly vectors. These maps provide valuable tools for guiding resource allocation, planning targeted interventions, and implementing community-level disease surveillance strategies.

The broader significance of this work lies in its methodological generalizability and potential for scaling. While the study focused on Beni Mellal-Khenifra, the integrated framework we propose is applicable across other provinces in Morocco and can be extended to neighboring countries in North and Sub-Saharan Africa, where cutaneous and visceral forms of leishmaniasis remain endemic. Future efforts will focus on applying this pipeline at the national level, integrating more regions with varying ecological and socio-demographic contexts. By expanding the geographic scope, we aim to develop a national leishmaniasis risk atlas for Morocco that can be updated in real time and used by health ministries, NGOs, and epidemiological researchers.

In the longer term, this study opens the door to continent-wide applications. Many African countries face similar challenges in disease surveillance, limited data availability, environmental diversity, and public health resource constraints. By refining and adapting the XGBoost-CORAL framework, we envision the creation of a pan-African platform for predictive modeling of vector-borne diseases. Such a platform would integrate satellite data, climate forecasts, mobile health data, and localized epidemiological records, supporting a more proactive and data-driven approach to disease control.

While this study offers significant contributions, several limitations must be addressed in future work. The absence of socioeconomic data, such as housing quality, population density, and access to healthcare, limits the social context of the model. Furthermore, the use of static environmental layers constrains the model's temporal resolution; future integration of dynamic data (e.g., real-time satellite NDVI, daily climate feeds) will enhance the responsiveness of the risk predictions. Additionally, exploring deep learning models and more advanced domain adaptation methods (e.g., adversarial or transfer learning techniques) may yield further improvements in predictive power and scalability.

Ultimately, this study represents a foundational step toward building a flexible, intelligent, and scalable predictive system for leishmaniasis and other vector-borne diseases. It contributes to a growing body of research that leverages artificial intelligence and remote sensing for global health. With further development, the tools and insights presented here have the potential to support national and continental efforts to reduce the burden of neglected tropical diseases through timely, evidence-based decision-making.

References

Achbah, M., Khattabi, A., Pruneau, D., & Boumeaza, T. (2024). Évaluation de la vulnérabilité des communautés de montagne face au changement climatique. Région Beni-Mellal-Khénifra, Maroc. *VertigO*, *Volume 24 Numéro 2*. https://doi.org/10.4000/12PP9

Ben Salem, A., Karmaoui, A., Ben Salem, S., & Boughrous, A. A. (2020). *Geographical Distribution of Cutaneous Leishmaniasis and Its Relationship With Climate Change in Southeastern Morocco*. 136–152. https://doi.org/10.4018/978-1-7998-2197-7.CH007

Boussaa, S., Kahime, K., Samy, A. M., Salem, A. Ben, & Boumezzough, A. (2016a). Species composition of sand flies and bionomics of Phlebotomus papatasi and P. sergenti (Diptera: Psychodidae) in cutaneous leishmaniasis endemic foci, Morocco. *Parasites and Vectors*, *9*(1), 60–60. https://doi.org/10.1186/S13071-016-1343-6/FIGURES/3

Boussaa, S., Kahime, K., Samy, A. M., Salem, A. Ben, & Boumezzough, A. (2016b). Species composition of sand flies and bionomics of Phlebotomus papatasi and P. sergenti (Diptera: Psychodidae) in cutaneous leishmaniasis endemic foci, Morocco. *Parasites and Vectors*, *9*(1), 60–60. https://doi.org/10.1186/S13071-016-1343-6/FIGURES/3

Cheng, Z., Chen, C., Chen, Z., Fang, K., & Jin, X. (2021). Robust and high-order correlation alignment for unsupervised domain adaptation. *Neural Computing and Applications*, *33*(12), 6891–6903. https://doi.org/10.1007/S00521-020-05465-7

Eddoughri, F., Lkammarte, F. Z., El Jarroudi, M., Lahlali, R., Karmaoui, A., Yacoubi Khebiza, M., & Messouli, M. (2022). Analysis of the Vulnerability of Agriculture to Climate and Anthropogenic Impacts in the Beni Mellal-Khénifra Region, Morocco. *Sustainability*, *14*(20), 13166–13166. https://doi.org/10.3390/SU142013166

El Omari, H., Chahlaoui, A., & El Ouali Lalami, A. (2018). The Geographic Information Systems Are a Lever for Fighting Parasitic Diseases: Case of Leishmaniasis. *Lecture Notes in Intelligent Transportation and Infrastructure*, *Part F1405*, 1204–1213. https://doi.org/10.1007/978-3-030-11196-0_98

El Omari, H., Chahlaoui, A., Taouraout, A., & El Ouali Lalami, A. (2019). Geographical information systems (GIS) and epidemiology of vector diseases: case of leishmaniasis in the Fez-Meknes, region of Morocco. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3368756.3368978

Elfazazi, K. (2017). Morphological and Biochemical Variability of Moroccan Carob (Ceratonia siliqua L.) Produced in Beni Mellal Region. *International Journal of Pure & Applied Bioscience*, *5*(4), 14–21. https://doi.org/10.18782/2320-7051.5295

Faiza, S., Asmae, H., Fatima, A., Afafe, F., Bouchra, D., Ibrahim, A., Abderrahim, S., Khalid, H., Mohamed, R., & Hajiba, F. (2015). Molecular epidemiological study of cutaneous leishmaniasis in Beni Mellal and Fquih Ben Saleh provinces in Morocco. *Acta Tropica*, *149*, 106–112. https://doi.org/10.1016/J.ACTATROPICA.2015.05.021

Fannassi, Y., Ennouali, Z., Hakkou, M., Benmohammadi, A., Al-Mutiry, M., Elbisy, M. S., & Ali Masria. (2023). Prediction of coastal vulnerability with machine learning techniques, Mediterranean coast of Tangier-Tetouan, Morocco. *Estuarine Coastal and Shelf Science*, *291*, 108422–108422. https://doi.org/10.1016/J.ECSS.2023.108422

Faouzi, E., Arioua, A., Karaoui, I., Ait Ouhamchich, K., & Elhamdouni, D. (2020). Wastewater reuse in agriculture sector: resources management and adaptation in the context of climate change: case study of the Beni Mellal-Khenifra region, Morocco. *E3S Web of Conferences*, *183*, 02005. https://doi.org/10.1051/E3SCONF/202018302005

Guma, F. E. L., Musa, A. G. M., Alkhathami, F. D., Saadehm, R., & Qazza, A. (2023). Prediction of Visceral Leishmaniasis Incidences Utilizing Machine Learning Techniques. *2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence, EICEEAI 2023*. https://doi.org/10.1109/EICEEAI60672.2023.10590369

Hakkour, M., Hmamouch, A., Mahmoud El Alem, M., Bouyahya, A., Balahbib, A., El Khazraji, A., Fellah, H., Sadak, A., & Sebti, F. (2020). Risk Factors Associated with Leishmaniasis in the Most Affected Provinces by Leishmania infantum in Morocco. *Interdisciplinary Perspectives on Infectious Diseases*, *2020*, 6948650. https://doi.org/10.1155/2020/6948650

Kahime, K., Boussaa, S., Bounoua, L., Fouad, O., Messouli, M., & Boumezzough, A. (2014). Leishmaniasis in Morocco: Diseases and vectors. *Asian Pacific Journal of Tropical Disease*, *4*(S2), S530–S534. https://doi.org/10.1016/S2222-1808(14)60671-X

Kahime, K., Boussaa, S., Idrissi, A. L.-E., Nhammi, H., & Boumezzough, A. (2016). Epidemiological study on acute cutaneous leishmaniasis in Morocco. *Journal of Acute Disease*, *5*(1), 41–45. https://doi.org/10.1016/j.joad.2015.08.004

Kholoud, K., Denis, S., Lahouari, B., El Hidan, M. A., & Souad, B. (2018). Management of Leishmaniases in the Era of Climate Change in Morocco. *International Journal of Environmental Research and Public Health*, *15*(7), 1542. https://doi.org/10.3390/IJERPH15071542

Lynch, J., & Wookey, S. (2021). *Leveraging Domain Adaptation for Low-Resource Geospatial Machine Learning*.

Meliho, M., Khattabi, A., Driss, Z., & Orlando, C. A. (2022). Spatial prediction of flood-susceptible zones in the Ourika watershed of Morocco using machine learning algorithms. *Applied Computing and Informatics*. https://doi.org/10.1108/ACI-09-2021-0264/FULL/PDF

Sarafian, R., Kloog, I., Sarafian, E., Hough, I., & Rosenblatt, J. D. (2021). A Domain Adaptation Approach for Performance Estimation of Spatial Predictions. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(6), 5197–5205. https://doi.org/10.1109/TGRS.2020.3012575

Shabanpour, N., Razavi-Termeh, S. V., Sadeghi-Niaraki, A., Choi, S. M., & Abuhmed, T. (2022). Integration of machine learning algorithms and GIS-based approaches to cutaneous leishmaniasis prevalence risk mapping. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102854–102854. https://doi.org/10.1016/J.JAG.2022.102854

Sun, B., Feng, J., & Saenko, K. (2017). Correlation Alignment for Unsupervised Domain Adaptation. *Advances in Computer Vision and Pattern Recognition*, *9783319583464*, 153–171. https://doi.org/10.1007/978-3-319-58347-1_8

Sun, B., & Saenko, K. (2016). *Deep CORAL: Correlation Alignment for Deep Domain Adaptation*.

Sun, J., Dang, W., Wang, F., Nie, H., Wei, X., Li, P., Zhang, S., Feng, Y., & Li, F. (2023). Prediction of TOC Content in Organic-Rich Shale Using Machine Learning Algorithms: Comparative Study of Random Forest, Support Vector Machine, and XGBoost. *Energies*, *16*(10), 4159–4159. https://doi.org/10.3390/EN16104159

Talbi, F. Z., Idrissi, A. J., Sandoudi, A., & El Ouali Lalami, A. (2019). Spatial distribution of incidence of leishmaniasis of different communes of Sefrou Province (2007--2010), central north of Morocco. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3368756.3368992

Wang, Y., Li, W., Dai, D., & Van Gool, L. (2017). Deep Domain Adaptation by Geodesic Distance Minimization. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, *2018-January*, 2651–2657. https://doi.org/10.1109/ICCVW.2017.315

Wang, Z. Y., & Kang, D. K. (2021). P-Norm Attention Deep CORAL: Extending Correlation Alignment Using Attention and the P-Norm Loss Function. *Applied Sciences*, *11*(11), 5267. https://doi.org/10.3390/APP11115267